Academic year 2019-2020

PROJECT 2
Further study of the correlation structure and dimension reduction

High-dimensional data analysis

Baudenne Céline s154198
Girineza Guy s144377
Payan Palomo Fatima s190047

Master civil Engineering in Data science
Master in Mathematical Sciences

Note

All the analysis of this project required the handling of missing data. The handling process that was conducted is described in detail in annex 1. Also it should be noted that the variables *Fat.Ft2*, *BMI.BM2* and *Dst.Ds2* have been removed from the data set as they are redundant with respect to the other variables.

A. Robust outlier detection

In this section, a robust detection of the potential outliers will be performed, for which we chose to use the MCD (Minimum Covariance Determinant) method with a coverage parameter corresponding to 75% of the sample size. The results will be compared with the classical distances. In the first instance, the informed group and the control group have been considered separately.

Firstly, for the data set corresponding to the "Not Informed" group, it can be seen that for the classic Mahalanobis distance there is no outlier, while with the robust distance we can observe the appearance of 7 potential outliers.

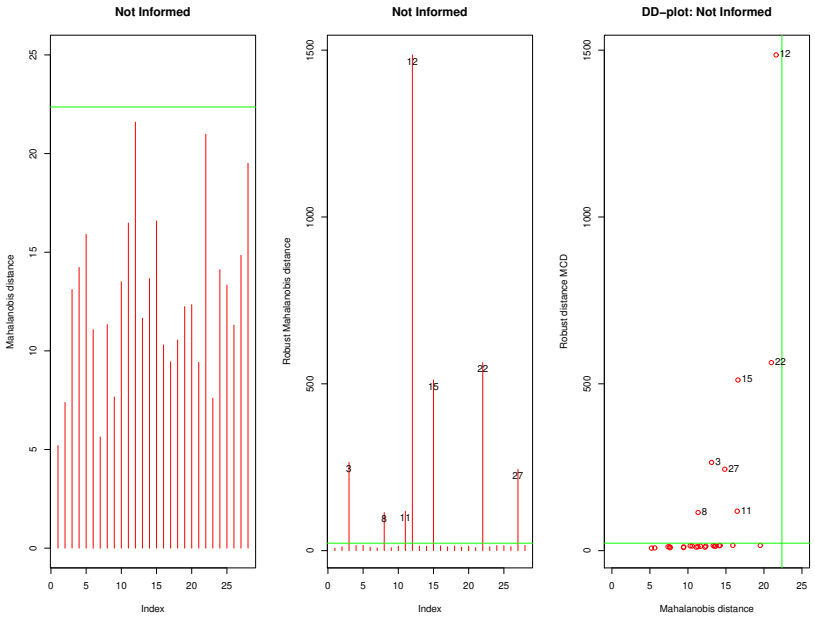

Figure 1 – Representation of the classic and robust Mahalanobis distances for the control group.

The analysis of the Mahalanobis distances for the data set corresponding to the "Informed" group gives the following results: The classic Mahalanobis distances reveal the existence of 2 outliers, while with the robust distances we observe 10 outlying observations, including those detected by the classical approach.

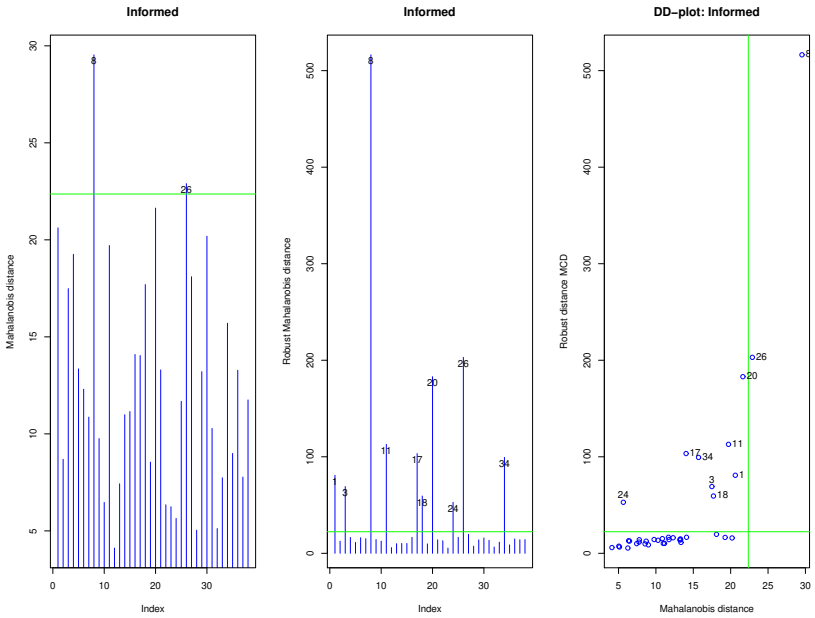

Figure 2 – Representation of the classic and robust Mahalanobis distances for the informed group.

Next, we looked at the Mahalanobis distances for the global data set. The classical approach shows the possible existence of 8 outliers, although using the robust distance technique we obtain twice as many outliers. In the DD-plot graph it can be seen that the set of data belonging to the "Informed" group is the one that provides the most possible outliers.

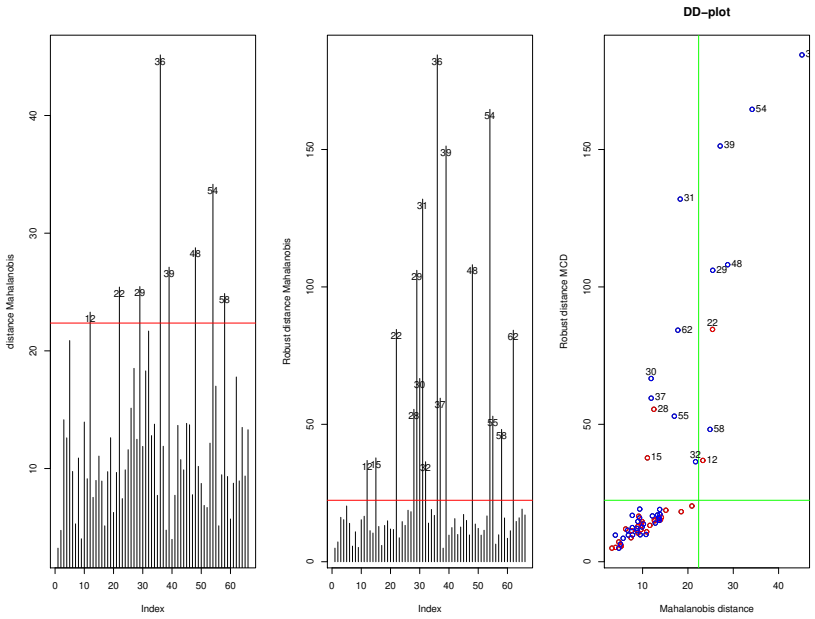

Figure 3 – Representation of the classic and robust Mahalanobis distances for the whole data set.

In all cases, considering the robust distance induces an augmentation of the number of potential outliers. This is explained by the fact that the robust estimation focuses on the 75% of observations which are the most concentrated and thus considers more observations as outlying. In the classical approach, some extreme outliers might make the intermediate outlying observations look less suspicious and thus potential outliers are masked by the classical approach.

B. Further investigation of the correlation structure of the quantitative variables

1. Robust estimation of the correlation matrix

As before, the robust estimation of the correlation matrix has been computed using the MCD estimator with 75% of the sample size for the optimal subset of observations.
For the most part, the same pattern is observable for the robust estimation (Fig. 5) as for the classic correlation matrix (Fig.4). In particular the highly positive correlation between the variables *Wt*, *BMI* and *Fat* as their measurements are based on the weight. It also includes the correlation between each variable and its second measurement. Regarding the variables about the blood pressure, their correlation are slightly reduced in the case of the robust estimation. This is due to the fact that the robust estimation focuses only on the 49 most concentrated observations. For example, Figure 6 represent the variables *Syst2* and *Diast2*, the observations represented in red are the observations selected in the optimal subset used in the robust estimation. It can be seen, in fact, that those observations have a less important positive correlation and that the outlying observations intensify the positive correlation.

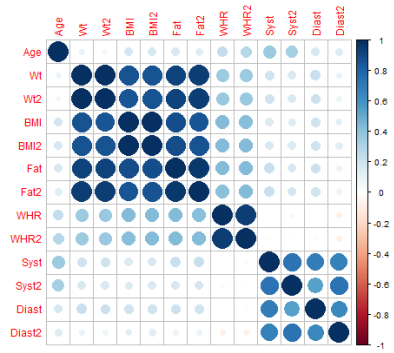

Figure 4 – Classic correlation matrix.

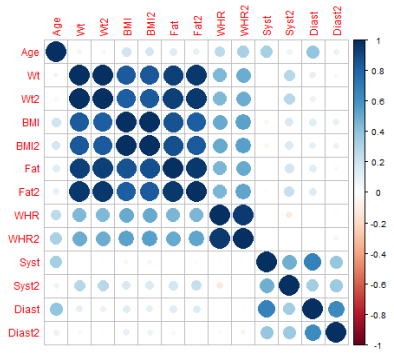

Figure 5 – Robust estimation of the correlation matrix.

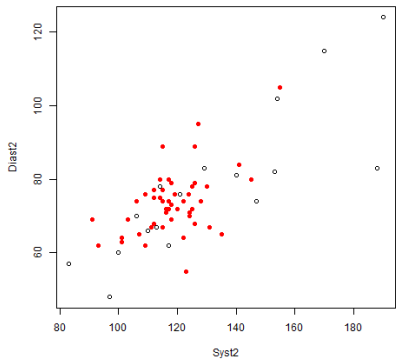

Figure 6 – Representation of the variables Syst2 and Diast2;
Red = observations of the optimal subset for the robust estimation.

2. Graphical models

The graphical models have been drawn based on a standardize version of the data as all the variables do not have the same units of measurement and the same order of magnitude.

For the L1-regularized estimation of the covariance matrix, the appropriate value of the regularization parameter has been chosen, firstly, by studying the evolution of the BIC for different values of lambda. However, this resulted in an appropriate value of 0.0 for lambda which could be explained by the fact that some dependencies between the variables being fairly high it might not be appropriate to force some elements to be equal to zero. An appropriate value for lambda has thus been determined by observing the evolution of the number of estimated parameters different from zero depending on the value of lambda. A value of 0.4 yields for a reasonable decrease in the number of non-zero estimated parameters[1].

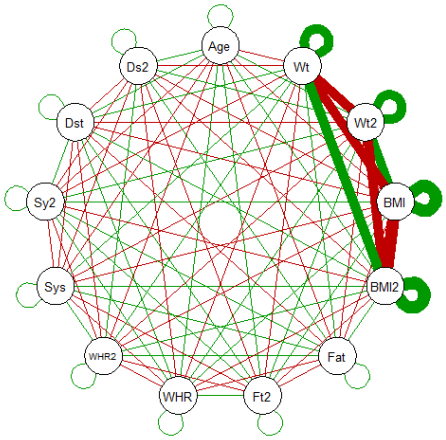

Figure 7 – Graphical model based on the classic covariance matrix.

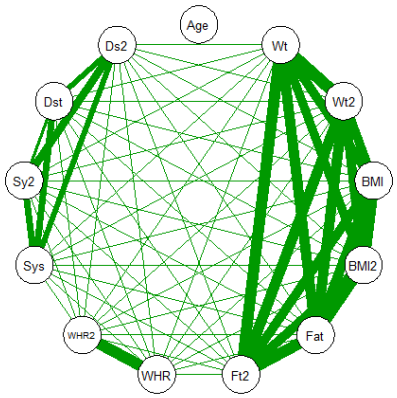

Figure 8 – Graphical model based on the L_1-regularized covariance matrix.

Figure 7 has been drawn using the inverse of the classical covariance matrix while Figure 8 has been constructed based on the inverse of the L1-regularized covariance matrix. The former is difficult to interpret as all the variables are linked to each other with different intensities and some links being represented as positive and others as negative. The graphical model based on the regularized estimation can be easily interpreted as what it represents is closer to what has already been observed in the previous analysis. Focusing on the thickest links, we can found back the expected links between each variable and its counterpart measured 4 weeks later, the links between the variables in relation with the blood pressure as well as between all variables for which the measurement is based on the weight. Those relations had previously been identified in the study of the correlation structure.

Relevance of the multivariate normality assumption.

A representation of the distribution of each variable can be found in the annex 3[2]. Overall, as can be seen on the graphs, the distributions of the different variables are not diverging too much from their corresponding univariate normal distributions. This thus supports the relevance of the multivariate normality assumption although it is not a guarantee.

C. Visualisation of the quantitative data in 2D

PCA

PCA has been performed on the classical correlation matrix. Indeed, as the variables have different units of measurement and different variability, the correlation matrix is more appropriate. Also, as shown previously when looking at the correlation plot, the non robust correlation matrix and the robust

[1]The graph representing the evolution of the number of non-zero parameters in the estimation depending on the value of lambda can be found in annex 2.

[2]We did not reuse the graphs presented in the first project as they took into account the difference between the informed and not informed groups while we do not make this difference here.

correlation matrix with a coverage parameter of 75% of the data set are fairly similar. Figure 9 shows the representation of the data in 2D using the two first principal components with a differentiation according to the groups.

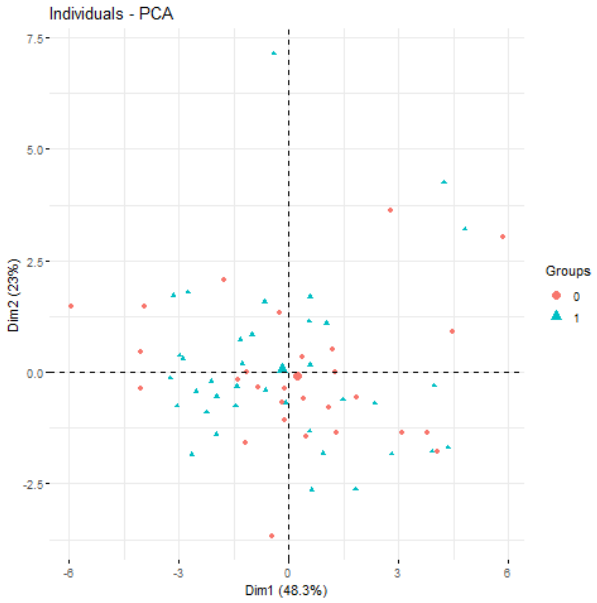

Figure 9 – 2D visualisation of the data using PCA.

Variance explained by the 2D projection

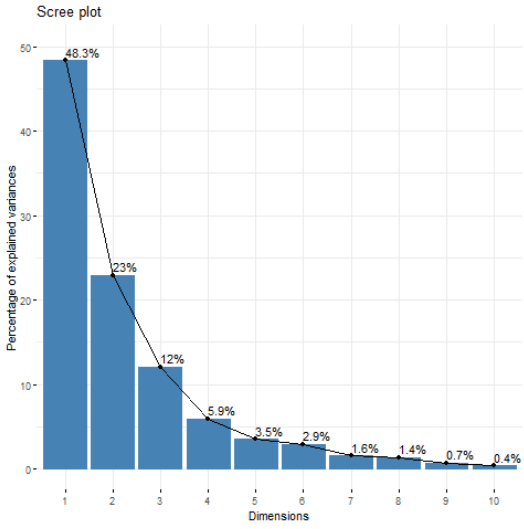

Figure 10 – Scree plot.

It can be seen on the scree plot that the two first principal components already explain $71,3\%$ of the total variance. It would be reasonable to work with the three first PCs as together they explain more than 80% of the total variance.

Interpretation of the two first PCs

It can be seen on Figure 11 and 12 that the first principal component mainly explains most of the variables related to the weight while the second principal component explains the blood pressure related variables. In other words, high values for the variables related to the weight will translate in a high score

for the first PC while high values for the variables related to the blood pressure will translate in a high score for the second PC. This has been verified by looking for particular observations in the data set. It can also be seen that in regards to the second PC, the weight related variables and the blood related variables behave in an opposite way.

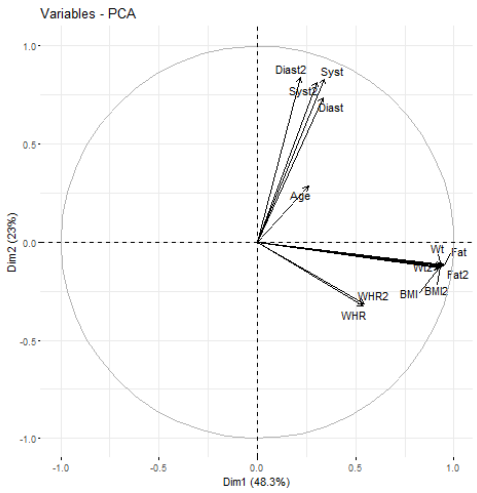

Figure 11 – Correlation circle.

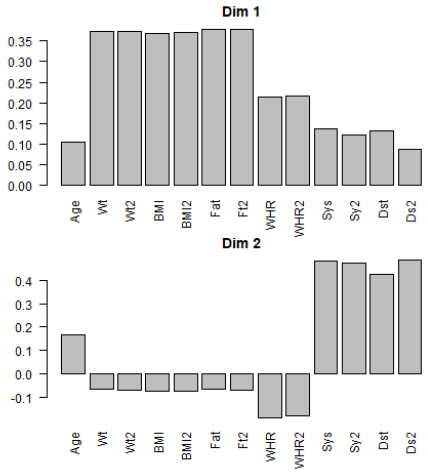

Figure 12 – Representation of the contributions of the variables for the two first PCs.

tSNE

Figure 13 is the 2D visualisation of the data produced using tSNE with a differentiation according to the groups. Tests have been performed with several values for the perplexity $(5, 10, 15, 20)$[3] but they did not result in huge differences. It turned out that the variables when considered all together, do not have discriminating power. Neither some particular structure nor some clusters or separations between the variables can be observed on the 2D representation of the data.

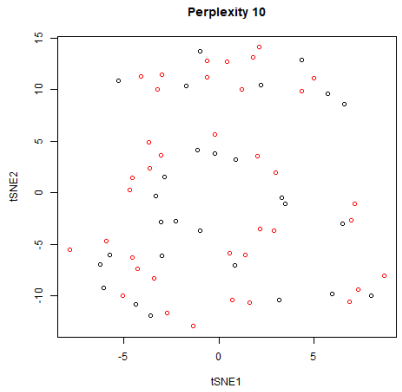

Figure 13 – 2D visualisation of the data using tSNE, *Red* = *Informed*.

In view of the two 2D visualisations of the data, the most informative projection seems to be the one produced by PCA. Indeed, this representation can easily be interpreted in terms of the original variables while the visualisation produced by tSNE does not give much information.

[3]Higher values have not been tested as, according to the documentation, the perplexity should not be greater than 3*perplexity < nrow(data) - 1.

ANNEX

1 *Processing of missing data:*

Firstly, we decided to discard the observations for which we had no information at all for the variables related to the blood pressure as well as the observation for which the *Age* was missing. The remaining observations still represent 88% of the total observations. This being done, the variables *WHR*, *Syst*, *Syst2* and *Diast* each had only one missing value left. Those missing values have been imputed by the mean of the respective variable according to the variable *Informed*. This way we hope not to interfere to much with the difference induced by the two groups. Finally, for the remaining variables (*WHR2*, *Fat* and *Fat2*) we looked for linear relations between the variables in order to find some variables that could explain those three. Taking into account the time of measurement and separating the informed and control groups, we found some linear relationships between the variables *Wt* and *Fat*; *Wt2* and *Fat2*; *WHR* and *WHR2*. The missing values have then been imputed using a linear regression. The graphs showing the linear relations between the variables as well as the comparison between the correlation matrix of the original data and the one of the imputed data are presented in Fig. 14 and 15 respectively.

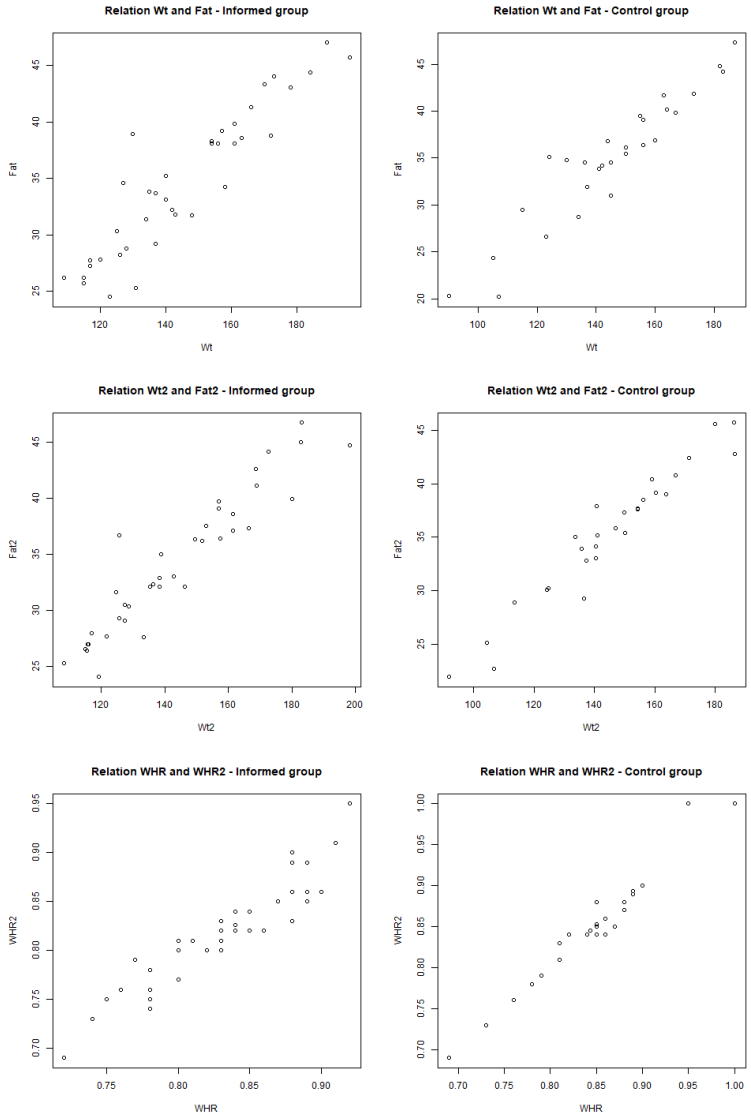

Figure 14 – Linear relations.

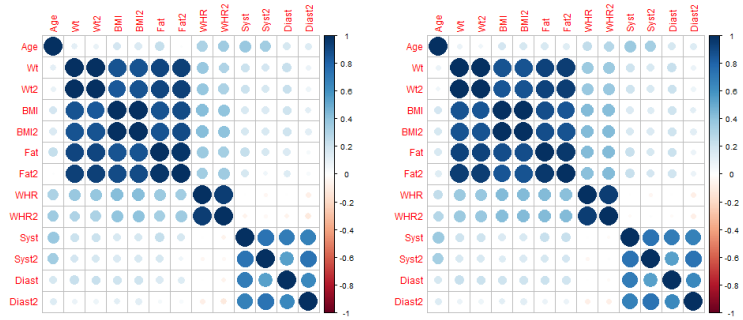

Figure 15 – Left : Original data set, right: Imputed data set.

2

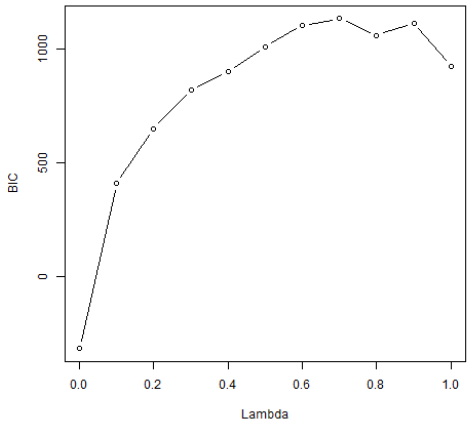

Figure 16 – Evolution of BIC as a function of lambda.

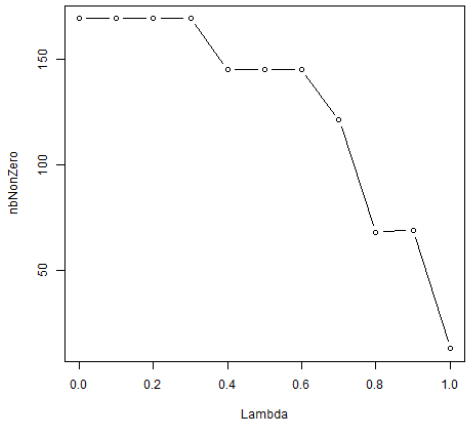

Figure 17 – Evolution of the number of non-zero parameters as a function of lambda.

3

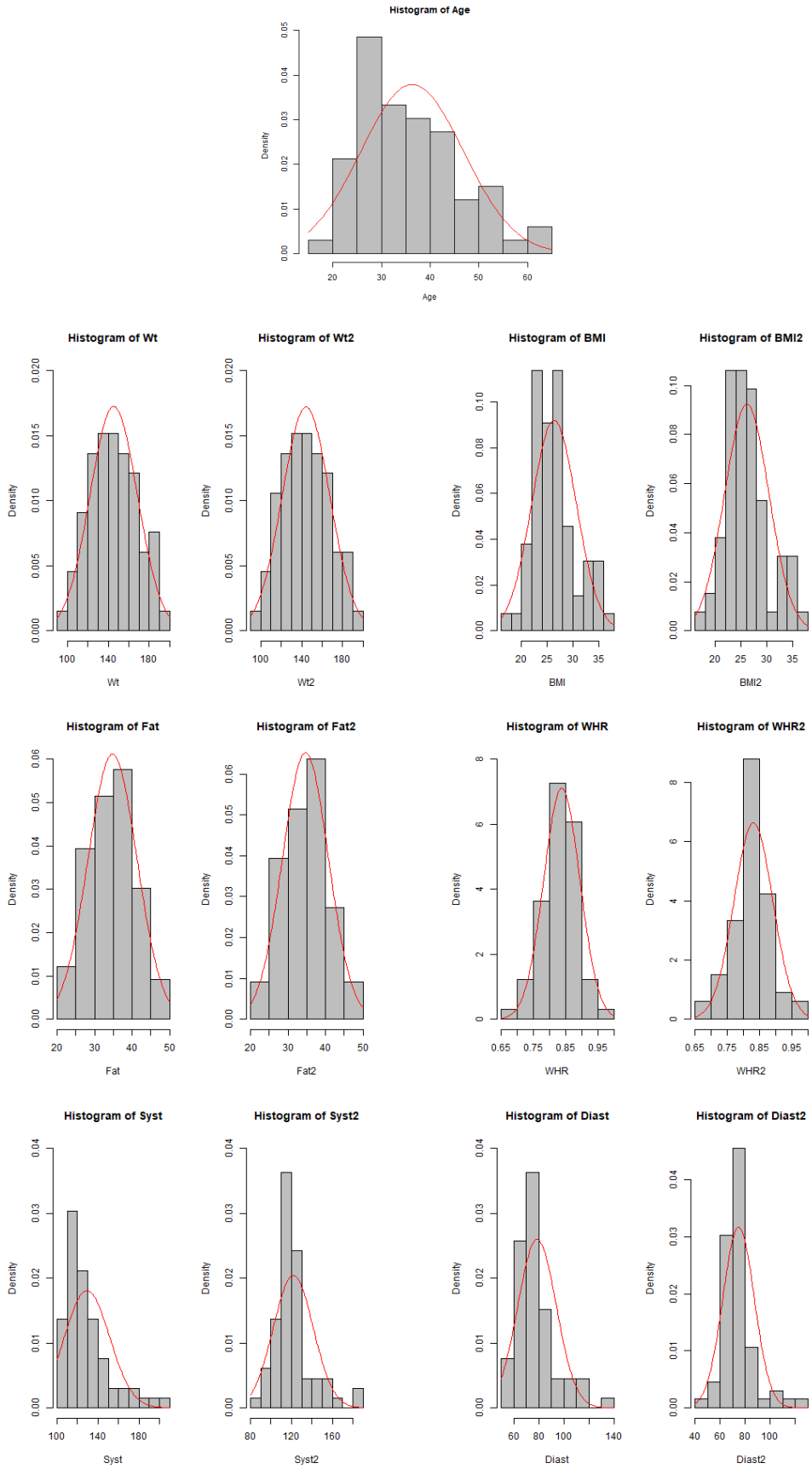

Figure 18 – Demonstration of the relevance of the normality assumption.