Academic year 2019-2020

PROJECT 3
Supervised classification

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

For this third project, the variables depicting the difference between two measurements have again been taken into account. Indeed, as will be shown in the next section and as it had already been noticed in the exploratory analysis of the first project, the difference induced by the informed and control group is reflected in these variables. They are thus relevant for the following analyses.

A. Preliminaries for the supervised classification.

In the exploratory analysis performed for the first project, some variables had been identified as having a discrepancy with respect to the informed or not informed group. In particular for the variable *Age*, we had observed that women in the informed group seem to have in average a younger age than women in the control group. This difference clearly originates from the data collection process but still induces a discriminating power to distinguish the two groups. The variables depicting the difference between the measures of body-mass index (*BMI.BM2*) and the difference between the measures of percentage of body fat (*Fat.Ft2*) had also been identified as containing some information about the classification.

In the meantime, some missing data processing have been performed which have slightly weakened the difference between the two groups displayed in some variables. For the variable *Age* and *BMI.BM2*, the t-test of significant difference in the mean still produces a p-value leading to a rejection of the null hypothesis of no difference in the mean (p-value = 0.011 and 0.0126 respectively). For the variable *Fat.Ft2*, the t-test produces a p-value of 0.089 which does not lead anymore to the rejection of the null hypothesis (p-value before missing data processing = 0.019). However, we still decided to consider this variable as having some discriminating power for the classification given the p-value is only slightly greater than 0.05 compared to all the other variables. Indeed, the t-test performed for all the other variables lead to no rejection of the null hypothesis of no difference in the mean[1]. Fig. 1 shows the boxplots representing the variables with some discriminating power with respect to the two groups.

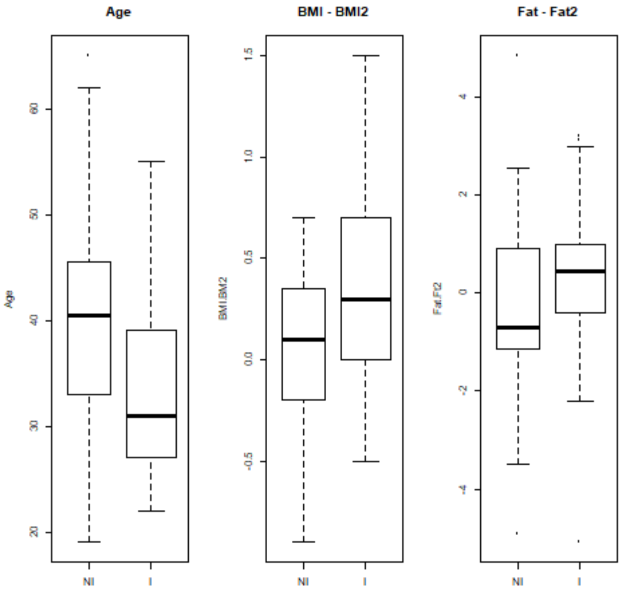

Figure 1 – Boxplot.

In parallel to this we also performed the Wilks MANOVA test to make sure the classification was relevant. This resulted in a significant difference between the mean vectors and thus a meaningful discrimination (*significance* : 0.001356).

[1] The p-values for each variables are displayed in annex 1

B. Classification using the logistic regression model.

1. Logistic regression model To get started, two first experiments to try to derive a good logistic model explaining the variable *Informed* have been performed. The first one uses as explanatory variables all the variables except the 'difference' variables (Model 1). The reason behind this is that the 'difference' variables in the end contain the same information than the other variables, therefore, taking them into account while considering all the other variables may interfere in the determination of the model. Nevertheless, as shown previously, some of these 'difference' variables still contain some information about the variable of interest justifying the second test which considers as explanatory variables the variables identified as containing some information about the binary indicator, namely *Age*, *BMI.BM2* and *Fat.Ft2* (Model 2). The following table shows the results of the two experiments only for the most significant variables.

	Results		
Models	Most significant variables	Estimate	Pr(>|z|)
Model 1 *AIC = 78.126*	Age BMI BMI2 Fat Fat2 Syst	-0.13263 6.28173 -6.69460 0.60168 -0.91065 0.09023	0.00783 ** 0.03542 * 0.02717 * 0.02622 * 0.01565 * 0.05456 .
Model 2 *AIC = 76.849*	Age BMI.BM2 Fat.Ft2	-0.10092 1.96334 0.38321	0.00207 ** 0.01074 * 0.04245 *

Table 1 – Results Model 1 (all variables except the 'difference' ones)
- Model 2 (variables identified in the preliminaries)

Those results are indeed consistent with what has been observed in the preliminaries. For the first test, the variables identified in the model as the most significant are unsurprisingly the age, the measures of the body-mass index and the measures of the percentage of body fat. Moreover, the estimates are such that the model actually considers the difference between the first and the second measurement. Also, as the regression parameter of the age is negative, the probability of success, i.e. of being informed, will decrease as the age of the person increase. This is what we were expecting in view of the boxplot presented before. The second model has a lower AIC than the first model. This illustrates the trade-off between the inclusion of parameters and the information captured by the model. The goal is not to include too many variables while achieving good results. Here, the second model achieves a good result as the explanatory variables used are the one containing the most information.

After these preliminary analyses, we used the forward selection strategy to select in an efficient way the explanatory variables worth including in the final logistic model. The results are presented in table 2. Once again, we can find mainly the same variables. The fact that the model considers both the variables *BMI.BM2* and *BMI* is not odd because this way it has all the information. Also, using this set of variables we have no problem of multicolinearity between the variables.

Models	**AIC**
Informed $\sim$ 1	91.97
Informed $\sim$ Age	86.88
Informed $\sim$ Age + BMI.BM2	79.47
Informed $\sim$ Age + BMI.BM2 + Syst	76.35
Informed $\sim$ Age + BMI.BM2 + Syst + Fat.Ft2	73.73
Informed $\sim$ Age + BMI.BM2 + Syst + Fat.Ft2 + BMI	73.38
Informed $\sim$ Age + BMI.BM2 + Syst + Fat.Ft2 + BMI + Wt2	72.35

Table 2 – Results of the forward variable selection.

Also, again the sign and magnitude of the regression parameters (eq.1) correspond to what was expected. In particular, for both 'difference' variables, the probability to have been informed will increase as the value of those variables increases which corresponds to the difference between the two groups observed on the boxplots. This is also what is expected given the context of the dataset.

$$\begin{aligned} logit(\pi) = 2.05 - 0.125 \text{ Age} + 3.25 \text{ BMI.BM2} + 0.035 \text{ Syst} \\ + 0.622 \text{ Fat.Ft2} - 0.374 \text{ BMI} + 0.052 \text{ Wt2} \end{aligned} \tag{1}$$

Classification rule From this model can be derived a classification rule that might be used to classify the observations. As the relative importance of each group is quite well balanced, we can use 0.5 as threshold. The classification rule is thus the following :

- Observation classified as not informed if $\hat{\pi}_i < 0.5$.
- Observation classified as informed if $\hat{\pi}_i \geqslant 0.5$.

Residuals and fitted values The residuals are represented in Fig. 2 and the fitted values in Fig. 3. When looking at the residuals, two observations especially stand out. Observations 24 corresponds to an uninformed person with an estimated probability equal to 0.916 while observations 35 corresponds to an informed person with an estimated probability equal to 0.103. The examination of the data for those two particular observations shows that in fact they both seem to have values closer to the mean values of their opposite group. Concerning the fitted values, apart from some misclassified observations, the membership of each observations is quite well determined.

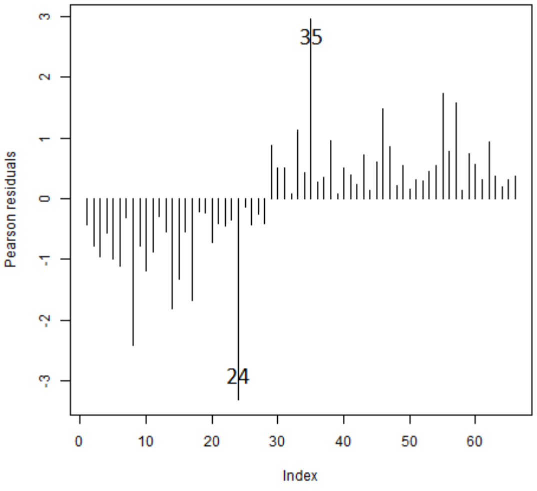

Figure 2 – Residuals.

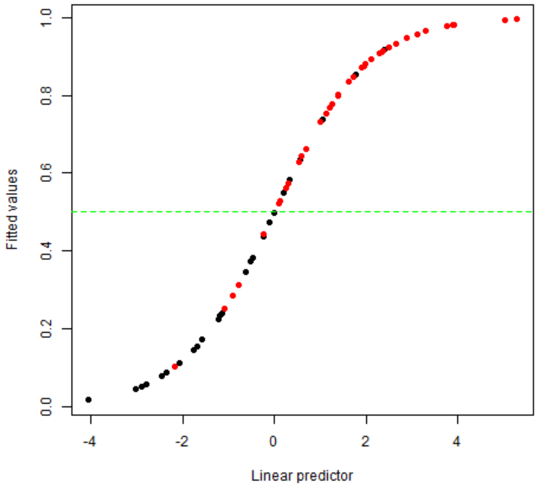

Figure 3 – Fitted values.

2. Leave-one out cross-validation A leave-one out cross-validation has been performed on the data in order to test the effectiveness of the model on the classification of unseen data. The results are presented in the confusion matrix in table 3.

	Classification	
Truth	Not informed (failure)	Informed (success)
0	18	10
1	8	30

Table 3 – Confusion matrix.

The results shown in the confusion matrix are not surprising. Indeed, here the classification is about whether the person has been informed or not about the possible benefits of her work on her health. At the end the goal is to study the placebo effect. The false positive (not informed women classified as informed) and false negative (informed women classified as not informed) could be explained by the fact that even after being informed it might be that no visible evolution in the measures is observable and in the other way, some evolution in the measures can be observed even if the women has not been informed.

The error rate is computed as follow : $\frac{FN+FP}{N} = \frac{8+10}{66} = 27\%$. As just stated, this error rate can be explained by the fact that some observations might not be fundamentally different while belonging to different groups. It is then difficult based on the model to correctly classify some unseen observations.

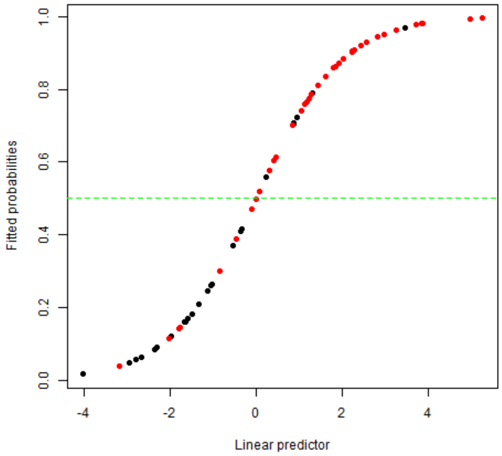

Figure 4 – Fitted probabilities from the leave-one-out cross-validation.

C. Classification based on LDA scores.

1. Canonical variable Equation 2 represents the unique canonical variable Z_1 produced with the Linear Discriminant analysis using as explanatory variables all the variables except the 'difference' variables. Indeed, those variables could not be used with the others because they clearly are collinear. A test has been performed replacing the variables by their difference but lead to a lower discriminant power. We have thus decided to keep working without the 'difference' variables as it let the model more possibilities to find an appropriate discriminating direction.

We can see on the canonical variable that, again, the coefficient of the variable *Age* is negative having thus the same effect as explained in the case of the regression model. The signs of the other coefficients are such that the model in fact takes into account the difference between the two measures of the different variables but with slight differences in coefficients which was not possible when working with the

’difference’ variables.

$$\begin{aligned}Z_1 = &- 0.076 \text{ Age}\\ &- 0.185 \text{ Wt} + 0.232 \text{ Wt2}\\ &+ 2.447 \text{ BMI} - 2.663 \text{ BMI2}\\ &+ 0.321 \text{ Fat} - 0.381 \text{ Fat2}\\ &+ 21.474 \text{ WHR} - 19.355 \text{ WHR2}\\ &+ 0.027 \text{ Syst} - 0.006 \text{ Syst2}\\ &- 0.015 \text{ Diast} + 0.014 \text{ Diast2}\end{aligned} \tag{2}$$

1D-scores. Figure 5 represents the 1D-scores obtained with the model. It can be seen that there remains some overlap between the scores of the two classes. Again, this could be explained by the fact that the two groups are not clearly separate.

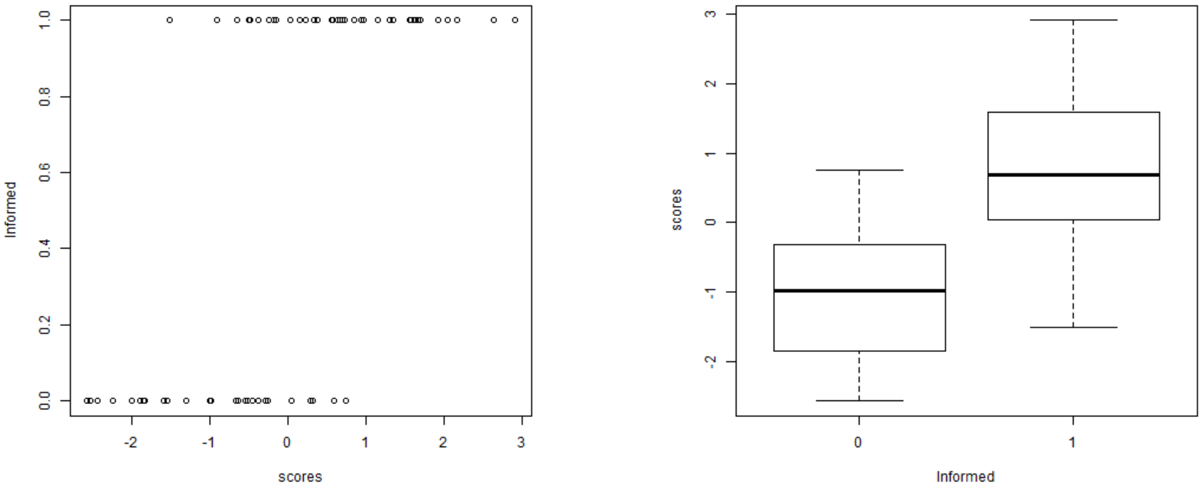

Figure 5 – 1D-scores.

Discriminant power The discriminant power of the model containing all the variables (except the ’difference’ variables) is computed in equation 3. The discriminant power γ_1 is in $[0, 1]$ by definition, thus the value obtained tells us that the variables are still able to discriminate the two groups but not perfectly.

$$\gamma_1 = \frac{\lambda_1}{1+\lambda_1} = \frac{0.778}{1+0.778} = 0.438 \tag{3}$$

2. Simplification The first step for the simplification of the canonical factor has been to try to remove one variable at a time and see which variables had the smallest impact on the discriminant power. The results are shown in table 4. It can be seen that again the variables yielding for the highest change in discriminant power are the variable *Age* and the measures of BMI and Fat.

Variable removed	Discriminant power	Variable removed	Discriminant power
LDA full	0.4378	Age	0.3364
Wt	**0.4312**	Wt2	**0.4272**
BMI	0.4067	BMI2	0.4018
Fat	0.3822	Fat2	0.3849
WHR	0.4109	WHR2	0.4151
Syst	0.4138	Syst2	**0.4367**
Diast	**0.4316**	Diast2	**0.4351**

Table 4 – First step of the simplification trial.

We have then tested to delete each variable with the smallest impact successively and see the decrease in discriminant power. The results are presented in table 5. We can see that deleting those variables does not imply drastic changes of discriminant power which means that there are not really important to differentiate the two groups. We decided not to remove other variables because doing so lead to a discriminant power dropping under 0.4 what seemed low to us. Also, we have tested to keep as explanatory variables only the one identified in the preliminaries as containing some information about the binary indicator. This model results in a discriminant power of 0.2739. We can conclude from those results that in this case the variables *Age*, *BMI.BM2* and *Fat.Ft2* are not sufficient to find a direction on which to project the data that would separate the two groups.

Model with removed variables	Discriminant power
LDA full - Syst2	0.4367
LDA full - Syst2 - Diast2	0.4351
LDA full - Syst2 - Diast2 - Diast	0.4307
LDA full - Syst2 - Diast2 - Diast - Wt	0.4228
LDA full - Syst2 - Diast2 - Diast - Wt - Wt2	0.4002

Table 5 – Second step of the simplification trial.

We compared the classification results produced by the last simplified model with the results of the full model. The error rate of the full model on the same data is 0.1969 while the one of the simplified model is 0.2424 which shows that considering all the variables together still produces better results.

3. Classification rule The classification rule has been derived using the following equation (4) as the dataset is not exactly balanced. $\hat{\pi_0}$ and $\hat{\pi_1}$ correspond to the prior probabilities of class membership, i.e. the proportion of observations in both groups. The parameters $\hat{\mu_0}$ and $\hat{\mu_1}$ are estimated based on the scores produced by the full model and the prior knowledge about the class membership of the observations. An observation will then be classified in the success group (being informed) if the score produced by the model is greater or equal than the threshold. Figure 6 represents the 1D-scores with the indication of the threshold determined by the classification rule.

$$x \geqslant \frac{\hat{\mu_1} + \hat{\mu_0}}{2} + \frac{\ln(\frac{\hat{\pi_1}}{\hat{\pi_0}})}{\hat{\mu_0} - \hat{\mu_1}} \geqslant -0.306 \qquad (4)$$

Where $\hat{\pi_0} = 0.4242$ and $\hat{\pi_1} = 0.5758$ and $\hat{\mu_0} = -1.028$ and $\hat{\mu_1} = 0.757$.

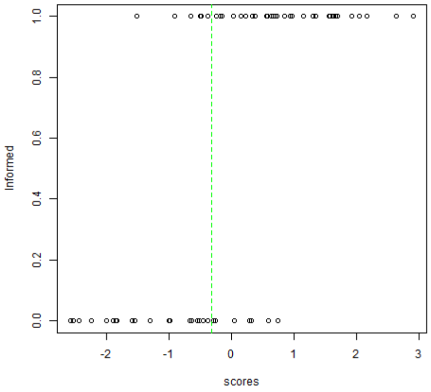

Figure 6 – 1D-scores - Green line represents the threshold of the classification rule.

Testing the classification rule on the data gives the results presented in the following confusion matrix (Table 6).[2]

	Classification	
Truth	Not informed (failure)	Informed (success)
0	21	7
1	6	32

Table 6 – Confusion matrix.

We have also tested to set the prior probabilities to equal probabilities as the prior proportions of the two groups are quite close to a balanced dataset. However, this only implies small changes.

4. Homoscedasticity The assumption of homogeneity of variances, also known as the assumption of homoscedasticity, considers that the variance is constant between different groups.

First, we can check the equality of the variances by visual inspection of the boxplots. The height of the boxes give us an idea of the variance of the data for each variable.

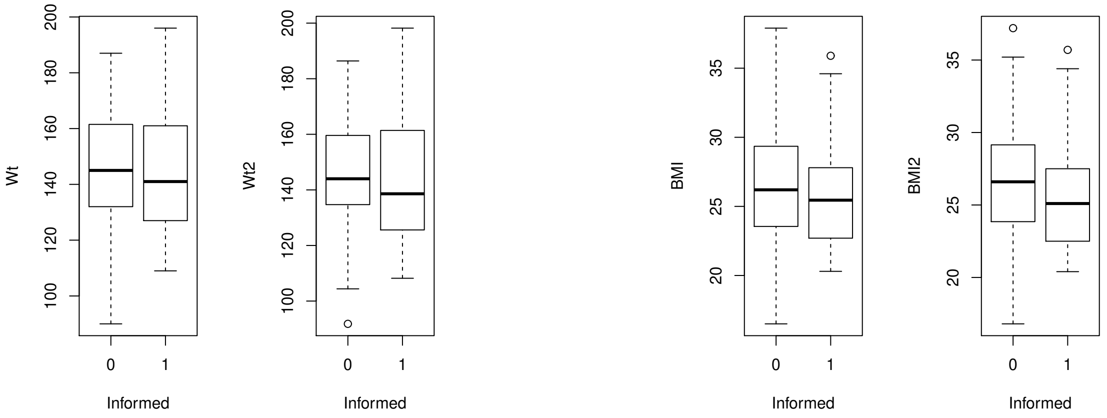

Figure 7 – Boxplots.

[2] The results are of course the same as the results provides by the predict function of R.

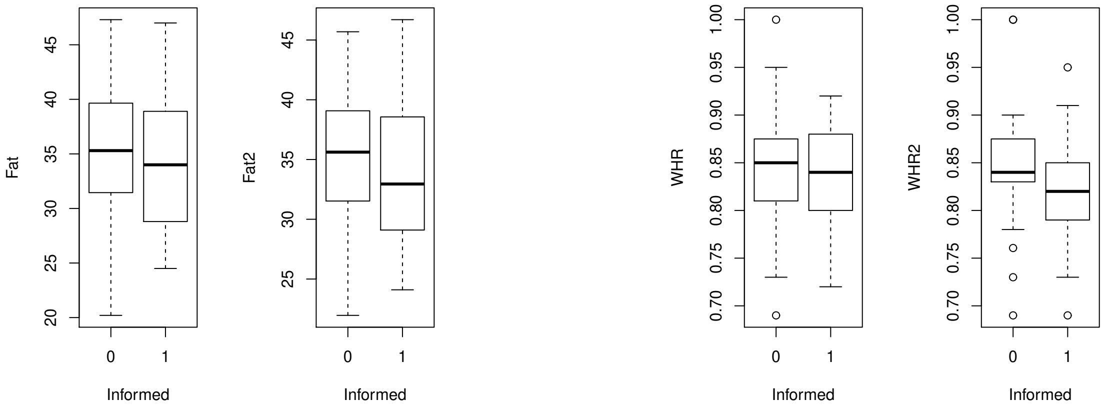

Figure 8 – Boxplots.

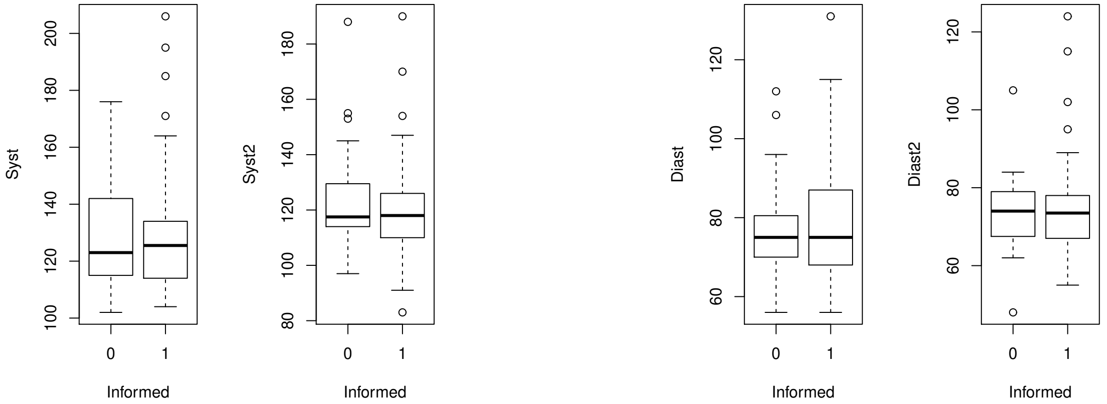

Figure 9 – Boxplots.

The variable Age is presented in Figure 1.

We can see that there are no huge differences between the boxes of the different groups. This indicates that there are no large differences of variance for the two groups.

If we do an F-test with the objective of verifying if the hypothesis of homoscedasticity holds or not, we obtain a p-value equal to 0.2934. The test does not find significant differences between the variances of the two groups as we cannot reject the null hypothesis of equal variances. This test is very sensitive to the non-normality of the data, which is not verified for all variables. However, as discussed in project 2, the distributions of the different variables are not diverging too much from their corresponding univariate normal distributions.

We can thus conclude that even if the assumption of homoscedasticity does not perfectly holds, it seems quite valid for our data.

ANNEX

1

Variables	p-value	Variables	p-value
Age	*0.011*	Dst.Ds2	0.763
Wt	0.944	Wt2	0.856
BMI	0.523	BMI2	0.373
Fat	0.822	Fat2	0.467
WHR	0.569	WHR2	0.109
Syst	0.614	Syst2	0.449
Diast	0.525	Diast2	0.617
BMI.BM2	*0.0126*	Fat.Ft2	*0.089*

Table 7 – P-value for the test of significant difference in the mean.