LIÈGE université

FACULTÉ DES SCIENCES APPLIQUÉES
Ecole d'ingénieurs et d'informaticiens

MATH2021-1
HIGH DIMENSIONAL ANALYSIS DATA
Gentiane Haesbroeck

Project 3
-
Supervised classification

Hubar Julien - s152485
Delvoye Benjamin - s154317

Preliminary anal: $1^5/3$
Log regression: $3^{25}/8$ [this part is really not clear! It is impossible to derive which final model has been used]
LDA: $6^5/8$
Overall: $0^5/1$

Master 1 Data science Engineering
Faculty of Applied Sciences

$11^{75}/20$

Academic years 2019-2020

Contents

this is not a classic way to proceed: boxplots are much more adequate because we can compare distributions!

1 Preliminaries for the supervised classification

A very visual way to find out if other variables can explain our categorisation is to plot them individually and colour the graph using the binary variable.
So if we look at the variables "radius_mean"[10], "concave.points_mean"[10] and "concavity"[10] we can see that the variables tends to separate into two groups.
By continuing the analysis on the variables "radius_se"[11], "concave.points_se"[11], "concavity_se"[11], "radius_worst"[20], "concave.points_worst"[20] and "concavity_worst"[20]. We come to more or less the same conclusion. Let us note all the same for, "concave.points_se"[11] and "concavity_se"[11]. The two groups are very confused.

for which variables?

no conclusion on these variables!!

would be clearer if that was presented in 2 lists!

Also, as we discussed in project 2, this dataset seems to be very suitable for classification. Indeed it has been showed outliers strongly depended on the class of individuals: outstanding values of explanatory variables were most of the time related to nuclei cell with malign cancer.

discri

non discrim.

2 Classification using the logistic regression model

2.1 Logistic model

why? Is it due to a lack of overlap? : multicol.

In order to find the best model of logistic regressions explaining our binary variable. It was necessary to proceed with several iterations of the basic model. The basic model was the one that included all the variables. As can be seen in Figure 1, the model does not converge. To overcome this problem, the same regressions were applied to the 3 main dataset groups (" mean ", " se " and " worst "). This made it possible to identify the variables (perimeter_mean, perimeter_se, perimeter_worst, compactness_mean, compactness_se, fractal_dimension_worst). The identification of this one was done through the study of their p-values. This made it possible to develop model 3. Although there are still many variables with large p-values, the graphical representation shows that the model is very good. In order to improve the model, the same process was carried out to manufacture model 4 but it did not show any significant improvement. An evolution of each model can be found in the appendix.

how more precisely??

2.2 Residual deviance

I do not understand how the 3 ≠ models could be compared.

An analysis of the residual deviance of each model allows us to see that it is becoming more and more balanced and correct.
Considering the values obtained, it is not surprising to see that the model2 has the best behaviour in terms of pearson residues. Even if the analysis of the figures of the logistic regressions would have led us to think that the model 3 would have the best behaviour.

al

model1	model2	model3	model4
32006.76421	24.43389	41.68973	42.59127

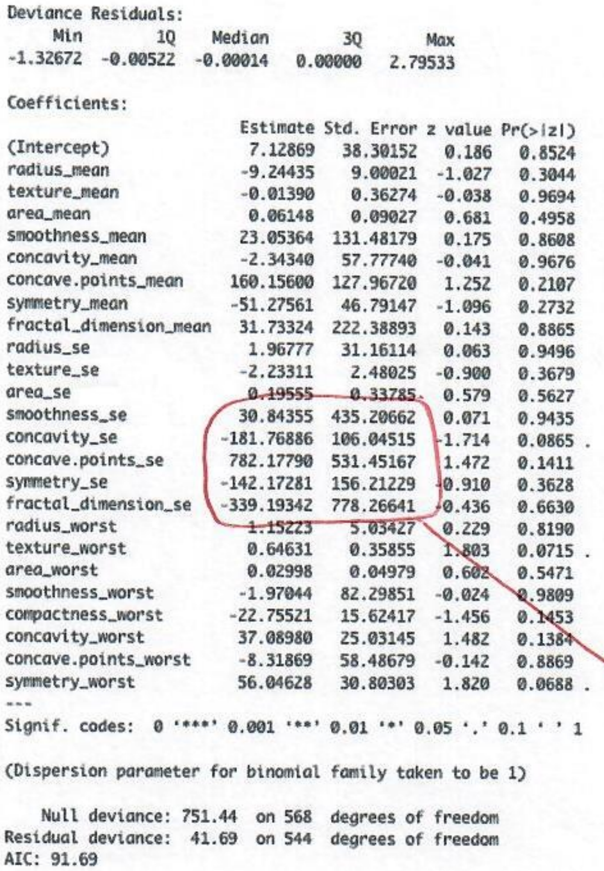
Tab. 1: Residual deviance

what do you mean by "balanced"?

but you do not take into account the number of parameters!

this indeed is interesting to look at!!

in order to compare models with ≠ numbers of parameters, an approach based on the AIC or BIC was recommanded.



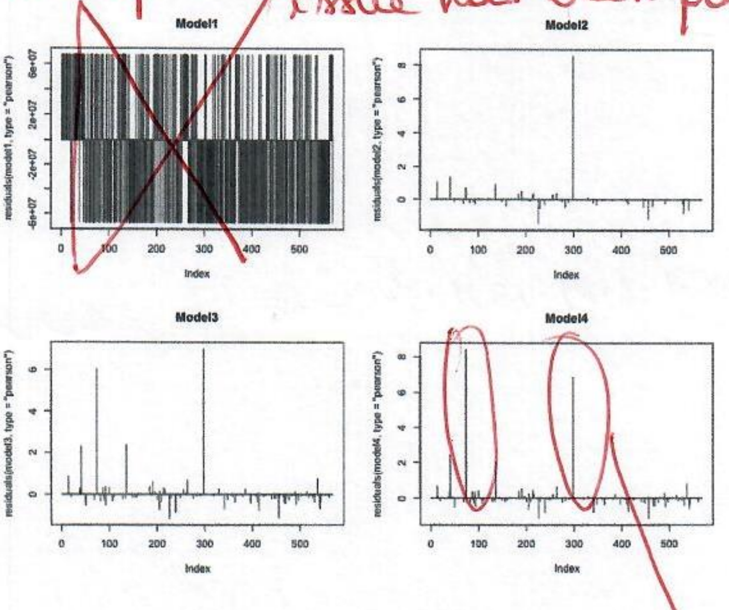
```
Deviance Residuals:
    Min        1Q    Median        3Q       Max
-1.32672  -0.00522  -0.00014   0.00000   2.79533

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                7.12869   38.30152   0.186   0.8524
radius_mean               -9.24435    9.00021  -1.027   0.3044
texture_mean              -0.01390    0.36274  -0.038   0.9694
area_mean                  0.06148    0.09027   0.681   0.4958
smoothness_mean           23.05364  131.48179   0.175   0.8608
concavity_mean            -2.34340   57.77740  -0.041   0.9676
concave.points_mean      160.15600  127.96720   1.252   0.2107
symmetry_mean            -51.27561   46.79147  -1.096   0.2732
fractal_dimension_mean    31.73324  222.38893   0.143   0.8865
radius_se                  1.96777   31.16114   0.063   0.9496
texture_se                -2.23311    2.48025  -0.900   0.3679
area_se                    0.19555    0.33785   0.579   0.5627
smoothness_se             30.84355  435.20662   0.071   0.9435
concavity_se            -181.76886  106.04515  -1.714   0.0865 .
concave.points_se        782.17790  531.45167   1.472   0.1411
symmetry_se             -142.17281  156.21229  -0.910   0.3628
fractal_dimension_se    -339.19342  778.26641  -0.436   0.6630
radius_worst               1.15223    5.03427   0.229   0.8190
texture_worst              0.64631    0.35855   1.803   0.0715 .
area_worst                 0.02998    0.04979   0.602   0.5471
smoothness_worst          -1.97044   82.29851  -0.024   0.9809
compactness_worst        -22.75521   15.62417  -1.456   0.1453
concavity_worst           37.08980   25.03145   1.482   0.1384
concave.points_worst      -8.31869   58.48679  -0.142   0.8869
symmetry_worst            56.04628   30.80303   1.820   0.0688 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 751.44  on 568  degrees of freedom
Residual deviance:  41.69  on 544  degrees of freedom
AIC: 91.69
```

FIGURE 1

— is this the so-called "basic model"?

we can indeed see that there is some convergence problems.

irrelevant if a non-convergence issue has been pointed out.



so, if this is the final model, there are still problems!

FIGURE 2: residual peasron

What can be said about these extreme values?

2.3 Model

Even using the R-script, it is not possible to underst and how the models were built.

By using the coefficients table our model can be written as follows.

$$\begin{aligned}\text{Model} = {} & 7.12869 * \text{Intercept} - 9.24435 * \text{radius\_mean} \\ & -0.01390 * \text{texture\_mean} + 0.06148 * \text{area\_mean} \\ & +23.05364 * \text{smoothness\_mean} - 2.34340 * \text{concavity\_mean} \\ & +160.15600 * \text{concave.points\_mean} - 51.27561 * \text{symmetry\_mean} \\ & +31.73324 * \text{fractal\_dimension\_mean} + 1.96777 * \text{radius\_se} \\ & -2.23311 * \text{texture\_se} + 0.19555 * \text{area\_se} \\ & +30.84355 * \text{smoothness\_se} - 181.76886 * \text{concavity\_se} \\ & +782.17790 * \text{concave.points\_se} - 142.17281 * \text{symmetry\_se} \\ & -339.19342 * \text{fractal\_dimension\_sev} + 1.15223 * \text{radius\_worst} \\ & +0.64631 * \text{texture\_worst} + 0.02998 * \text{area\_worst} \\ & -1.97044 * \text{smoothness\_worst} - 22.75521 * \text{compactness\_worst} \\ & +37.08980 * \text{concavity\_worst} - 8.31869 * \text{concave.points\_worst} \\ & +56.04628 * \text{symmetry\_worst}\end{aligned} \quad (1)$$

And the classif rule??

and?? What can be said? Are these results coherent/logical?

2.4 Cross-Validation

To ensure the accuracy of the model. ~~The function cv.glm()$delta allows to analyze values showing the behavior of the model. By~~ making the difference for each delta we obtain the figure below.

No need to mention that.

What do you mean by that??

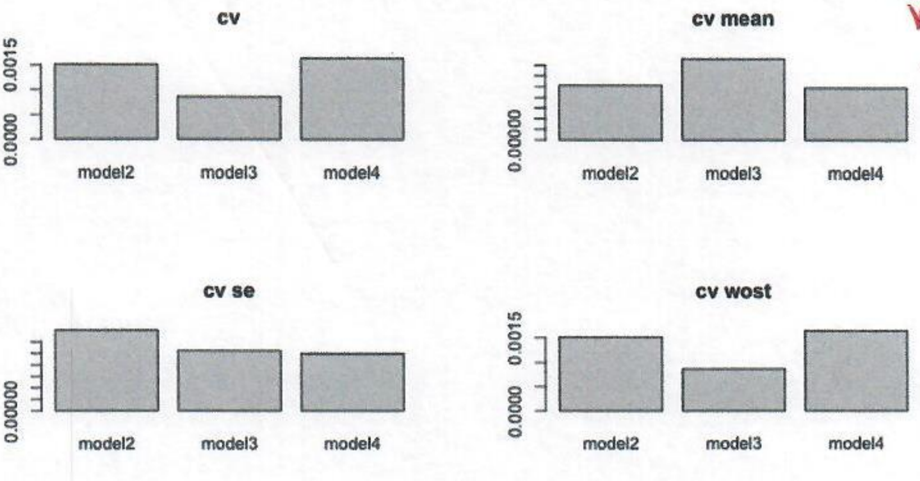


FIGURE 3

It is now model 3 which is the best??

The value for model 3 (in cv) being almost equal to zero shows us once again that our model is almost perfect

It was asked explicitely to use the leave-one-out technique while the script is based on 10-fold cross validation.

2.5 Confusion matix

The study of the confusion matrix of model 3 gives us very small value for our false positive and negative and very large value for the true positive and negative. This shows once again

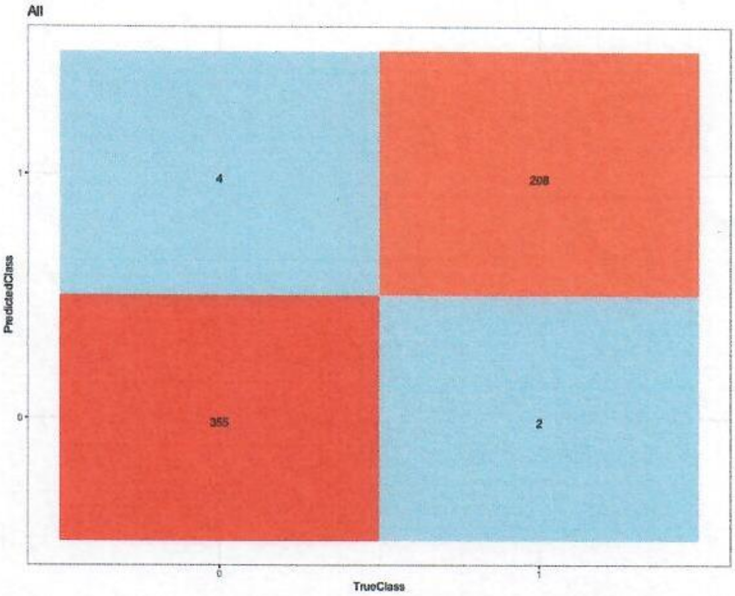


FIGURE 4

how well the classification of model 3 works. We can therefore conclude the sensitivity and specific y which is 0.9119497 and 0.9626866 respectively. Which once again shows us the good behaviour of our model

3 Classification based on LDA scores

Linear discriminant analysis idea is to find a vector (if two classes) **u** which will discriminate the best a class from the other on this vector. This means getting values on this vector as different as possible from a class to the other.
In order to find the vector **u**, the algorithm maximise variability between projected classes and minimise variance within each projected class. The lda function on R software allows to get coefficients of explanatory variables to construct vector **u**.

this is not necessary! the techniques are known.

Once the **u** vector is determined, it is then possible to obtain the projected value (score) of each individual on this new canonical variable. Distributions of classes scores define a measure of the discriminant potential of **u** (power of the statistical test).

3.1 Model 0: full.

and? interpretation?

In this part all columns of the dataset, except for the qualitative one, are used as explanatory variables. We will use this model as a baseline. Coefficients of $\mathbf{u}_0$ are displayed in 2.
This canonical vector leads to a discriminant power of 0.774 which is very satisfactory to establish a decision rule based on scores.
An approach we could have intent is to let down explanatory variables having very low coefficients. Those are poorly correlated to the new canonical variable and thus probably do not influence much power discriminant. However this approach won't be discussed as two models are already developed below.

On this plot, one can observe a difference in distributions dispersion's of classes. Let's compare their overlap evolving as we tweak the model in the section below.

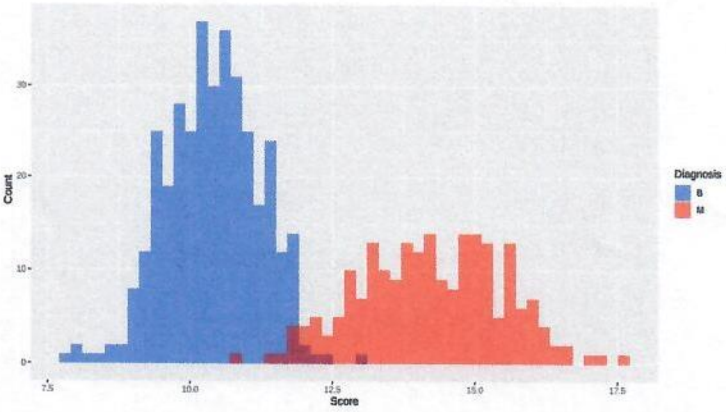


FIGURE 5: Distribution of each class on $\mathbf{u}_0$.

3.2 Model 1: reducing redundancy.

Based on discussion we had in project 2, explanatory variables related to *area* and *perimeter* were removed from the dataset as a strong correlation existed between those and the *radius* measures. The model is thus reduced to 24 explanatory variables.

The discriminant power of this model is 0.750. Compared to the baseline, the loss in discriminant power is quite small (0.024). However the model size hasn't been reduced much.

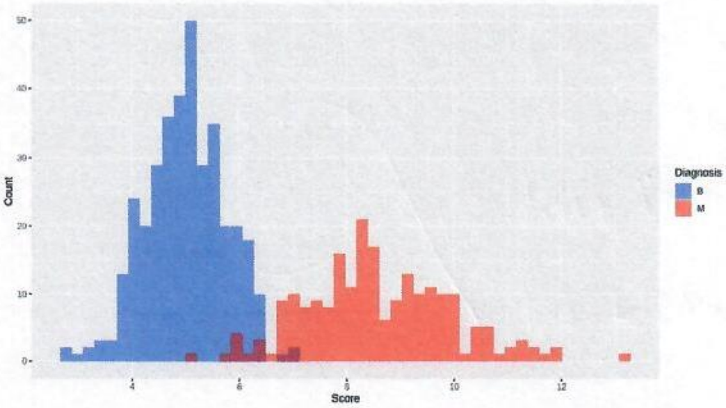


FIGURE 6: Distribution of each class on $\mathbf{u}_1$.

The loss of discriminant power can be explained here as we get a larger overlap of distributions.

3.3 Model 2: visual way.

As discussed in section 1, it is possible to visually identify which explanatory variables should provide the most discriminant power. One should look for variables having well separated centroïds and clearly defined fringe between classes on the 1D projection of individuals.

Based on the discussion above, the following explanatory variables were sliced from the dataset:

- "smoothness_mean",
- "texture_mean",
- "symmetry_mean",
- "fractal_dimension_mean",

- "texture_se",
- "smoothness_se",
- "compactness_se",
- "symmetry_se",
- "fractal_dimension_se",
- "symmetry_worst",
- "fractal_dimension_worst"

That reduces the model to 19 explanatory variables. The discriminant power decreases by only 0.01 compared to the baseline giving 0.763. This simplification seems to be better than the previous one.

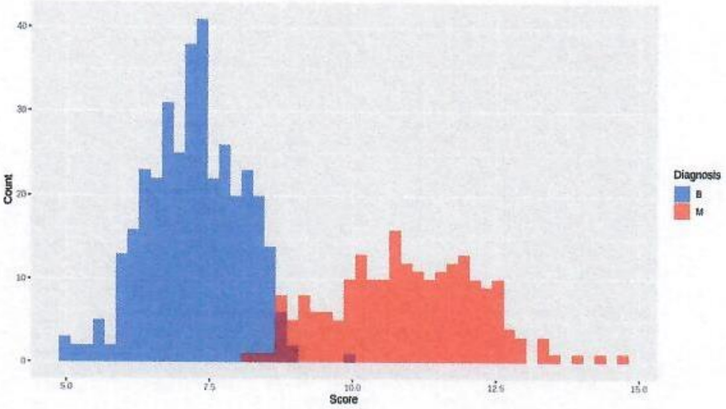


FIGURE 7: Distribution of each class on $\mathbf{u}_2$.

This plot is much more similar to the baseline we defined, with small overlap but higher count where the overlap occurs.

3.4 Decision rule.

As we found a subset of explanatory variables which has a decent discriminant power in regard to the full model, a decision rule can be derived from it. To find the optimal trade-off between false positive and true positive one can inspect the ROC plot. This leads to a

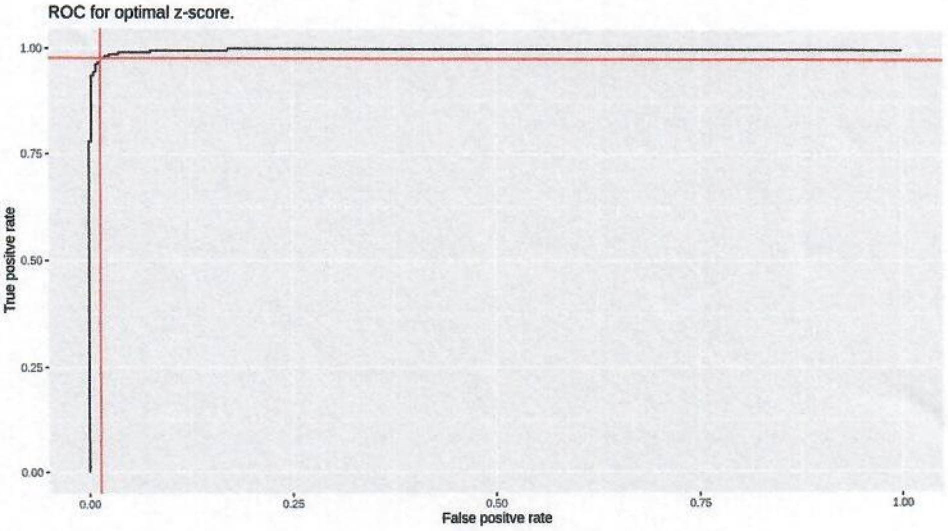


FIGURE 8

normalised z score of 0.128, a false positive rate of 0.014 and a true positive rate of 0.976. The confusion plot below shows good performance with the classification rule. ✓

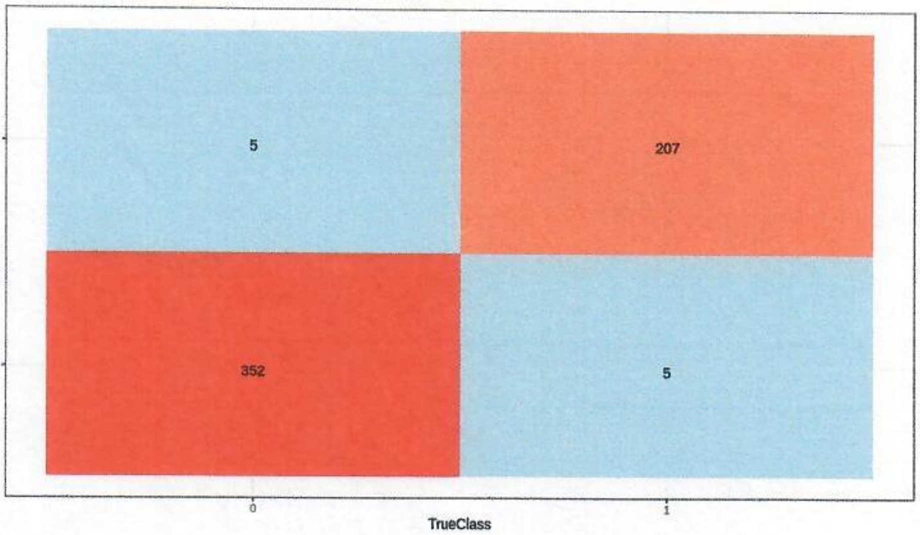


FIGURE 9: Confusion matrix with optimal cutoff of model 2.

3.5 Discuss the homoscedasticity assumption.

?? The homoscedasticity assumption concerns the original data, not the scores!

First, by looking at plots displaying distribution on **u** of each model it seems classes do not share the same variance/covariance matrix and thus do not verify homoscedasticity assumption.

Some boxplots were displayed in appendices. It is clearly observable that for those explanatory variables the behaviour of the variance is not the same from a class to the other. It is the same for most explanatory variables in the dataset.

Also we tried to perform a statistical test to verify the intuition above. Sadly results we got from `bartlett.test` function in R that were incoherent. The p-value was way to low compared to what we could expect. We probably misused this tool.(see `LDAclass.r` script)
If the p-value was not suspicious we could have compared it to a significance level of 0.05 and determine whether or not to accept the null hypothesis: distribution respect homoscedasticity.

you were expecting low p-values as you were convinced that the variances are not similar !!

no! you need to take into account the "multiple testing" effect and control the type I error!