FCTUC FACULDADE DE CIÊNCIAS
E TECNOLOGIA
UNIVERSIDADE DE COIMBRA

Bioestatistics

Integrated Master in Biomedical Engineering

Group 10
Inês Félix 2015230932
Xavier Pinho 2014210242

Final report

about the ***Final work of Biostatistics*** [1]

Abstract

The present work proposes to determinate which is the main age group that accesses a website. To do so, one should consider if it is **younger than**/ **older than**, to be able to adapt the offer to the target audience. These data can't be obtained directly (by asking a question to the users), but through a set of characteristic data, such as the writing pace, the sequence of clicks and double clicks, etc. With the due treatment of these data, it is expected to obtain an approximate value of the age of the target public, so that the publisher of the site can adapt the offer to visitors as much as possible.
With the accomplishment of this work it is possible to apply the knowledge acquired in the subject of Biostatistics.

I. Introduction

With the purpose of optimize the number of views of a website, the publisher must know exactly the intended audience.

From a broad set of data collected from several collaborators of which the age group is known, it is now intended to implement the method that will be used on the site.

We have a set of nine quantitative variables. For each one of them, tests will be carried out to evaluate it and can extract all necessary information and its significance with respect to the **qualitative variable** that is the **age**-if the visitor has age superior to the legal one (1= yes 'older than'; 0= no 'younger than').

Afterwards, we adjusted each model (GLM and SVM).

II. Method

For the development of this work we use the software *RStudio*, integrated for the programming language *R*. We also use *Matlab* to execute the SVM model.

We start by analysing the following **quantitative variables** provided:

- **rhythm1**, **rhythm2** and **rhythm3**- writing pace on form 1, 2 and 3, respectively;

- **click1**, **click2** and **click3-** time between clicks on form 1, 2 and 3, respectively;
- **double1**, **double2** and **double3**- interval in double click on form 1, 2 and 3, respectively.

Since these are quantitative variables we will proceed the analyses through **boxplot** (which are not conclusive) and then apply the **Shapiro-Wilk test**, to evaluate the normality. Considering the results, we perform the **ANOVA test** to the normalized variables and the **Wilcoxon test** to the non-normalized ones.

After analyzing the variables, we divided the database into two distinct groups: the **training group** (allows to adjust the statistical model) and the **test group** (tests and validates the model). This division was made randomly, using 70% of the database for the training and 30% for the test.

We construct a Generalized Linear Model (**GLM** function) using the best combination of variables and then evaluate its adjustment with the **Hosmer-Lemeshow**, **R^2 Cox-Snell** and **R^2 Naguelkerke methods**.

Then, we perform 500 runs of new training and test groups to determinate the best values for the sensitivity, specificity and accuracy, which is going to be the mean values.

At last, we perform the Supported Vector Machines (**SVM**).

III. Results and Discussion

First we are going to present the results for the **analysis of the quantitative variables**.

		Younger than	Older than
Rhythm1	**W**	0.972	0.97906
	p-value	0.01638	0.007047
Click1	**W**	0.96427	0.98096
	p-value	0.003677	0.01262
Double1	**W**	0.97444	0.97218
	p-value	0.02671	0.0009459
Rhythm2	**W**	0.97705	0.97664
	p-value	0.0455	0.00341
Click2	**W**	0.98132	0.98346
	p-value	0.1092	0.02781
Double2	**W**	0.97649	0.98099
	p-value	0.04054	0.01275
Rhythm3	**W**	0.97421	0.97626
	p-value	0.02551	0.003042
Click3	**W**	0.97843	0.97689
	p-value	0.06036	0.003671
Double3	**W**	0.96578	0.97307
	p-value	0.004889	0.001215

Tab. 1. Results for the Shapiro-Wilk test.

We can verify that the variables click2 and click3 are normalized (p-value < 0.05). To them, we performed the ANOVA test.

	p-value
Click2	0.0971
Click3	0.114

Tab. 2. Results for the ANOVA test.

Both variables lead to p-values greater than 0.05, so we can't reject the null hypothesis. This means that there aren't significant statistical differences. This way, we can conclude that both variables are not discriminative.

To the rest of the variables we performed the Wilcoxon test.

Rhythm1	**W**	10508
	p-value	0.8585
Click1	**W**	10692
	p-value	0.94
Double1	**W**	7640
	p-value	3.502e-05
Rhythm2	**W**	10442
	p-value	0.7876
Double2	**W**	8230.5
	p-value	0.0008928
Rhythm3	**W**	12842
	p-value	0.002302
Double3	**W**	15826
	p-value	7.525e-13

Tab. 3. Results for the Wilcoxon test.

Since the null hypothesis of this test is that the median is equal, observing our results we conclude that only the variables **double1**, **double2**, **rhythm3** and **double3** reject the null hypothesis.

We will considerate these variables as the discriminative ones, and we are going to present the dispersion plots of each.

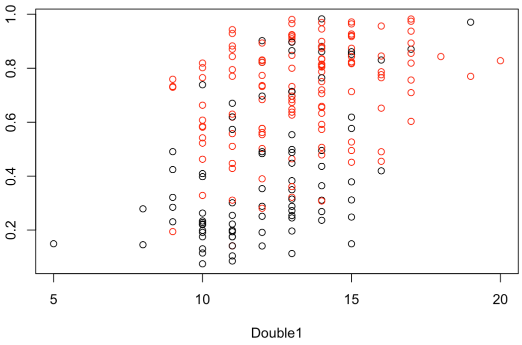


Fig. 1. Values of Double1 where the age older than is represented in red and younger than in black.

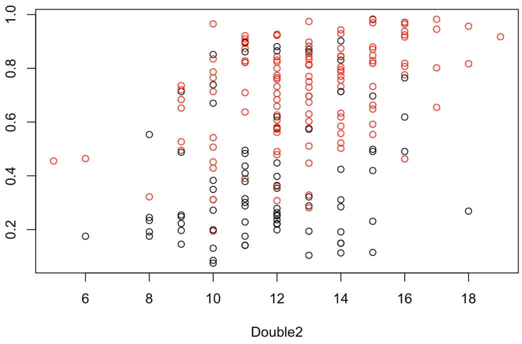


Fig. 2. Values of Double2 where the age older than is represented in red and younger than in black.

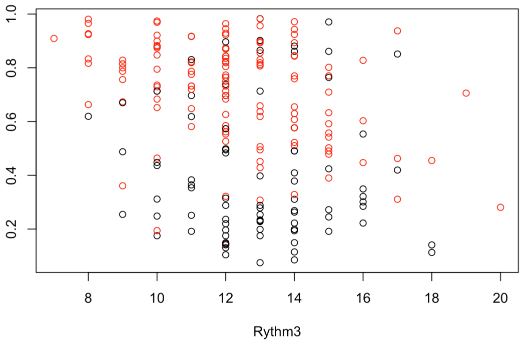


Fig. 3. Values of Rythm3 where the age older than is represented in red and younger than in black.

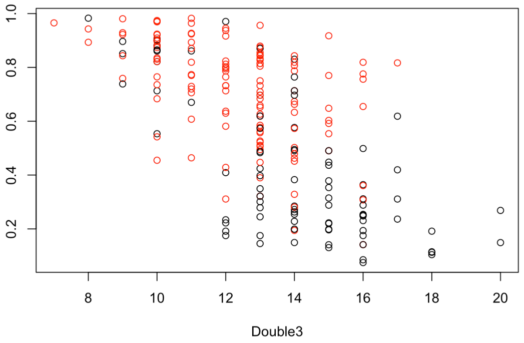
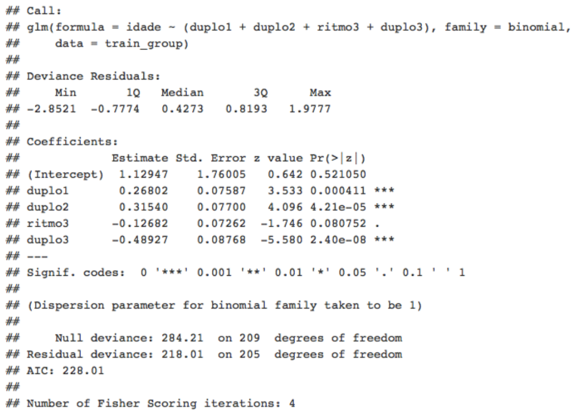


Fig. 4. Values of Double3 where the age older than is represented in red and younger than in black.

From all the figures above, we can notice that there is a good association between the variables and the qualitative variable age, where the values for *older than* are superior than for *younger than*.

Considering the discriminative variables defined above, we create the Generalized Linear Model through the *glm function* from *R*.



```
## Call:
## glm(formula = idade ~ (duplo1 + duplo2 + ritmo3 + duplo3), family = binomial,
##     data = train_group)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8521  -0.7774   0.4273   0.8193   1.9777
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.12947    1.76005   0.642 0.521050
## duplo1       0.26802    0.07587   3.533 0.000411 ***
## duplo2       0.31540    0.07700   4.096 4.21e-05 ***
## ritmo3      -0.12682    0.07262  -1.746 0.080752 .
## duplo3      -0.48927    0.08768  -5.580 2.40e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 284.21  on 209  degrees of freedom
## Residual deviance: 218.01  on 205  degrees of freedom
## AIC: 228.01
##
## Number of Fisher Scoring iterations: 4
```

Fig. 5. Results of the summary obtained from the GLM model.

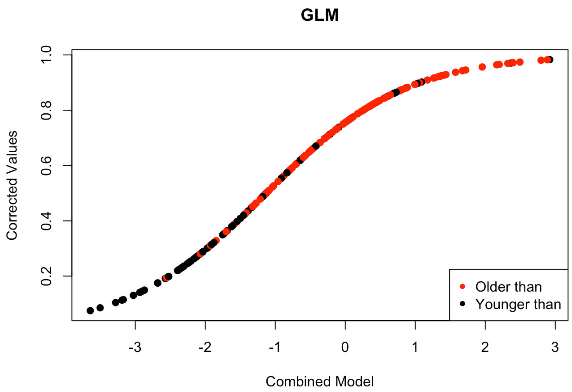


Fig. 6. GLM logistic model.

To evaluate the quality of the fit we executed the methods already mentioned.

Hosmer-Lemeshow (p-value)	5.52e-05
R^2 Cox-Snell	0.470919898
R^2 Naguelkerke	0.639942749

Tab. 4. Evaluators of the fit.

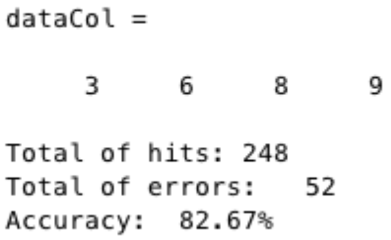
Considering the results obtained in *tab. 4* we conclude that is not a good fit.

Following the procedure presented in methods, we considered the approximated mean values (for the 500 runs) of accuracy, sensibility and specificity for the group test.

	Mean value
Accuracy	0.7520
Sensibility	0.7678
Specificity	0.7441

Tab. 5. Values of accuracy, sensibility and specificity for the group test.

Now we switch to *Matlab* programming to perform the SVM model through backward elimination.



```
dataCol =

     3     6     8     9

Total of hits: 248
Total of errors:   52
Accuracy:  82.67%
```

Fig. 7. Best results from the SVM model.

The variables to which the results lead to a greater accuracy are: **double1**, **double2**, **click3** and **double3**. We can realize that almost all of these are the same as our considered variables for the GLM model.

IV. Conclusion

Comparing both models, the one with greater accuracy is the **SVM model**.

In both models, we can confirm the presence of 3 variables: **double1**, **doube2** and **double 3**. This way, these are the most discriminative variables.