M.A. Research Psychology

Multivariate Statistics course

Assignment 4 : Logistic regression

The Eyewitness research group at UCT has been collaborating with a French team in Toulouse on the development of a questionnaire that will predict recognition success of eyewitnesses. This was tested in an experiment in which participants viewed a video of a staged crime, and then after a brief delay (5 mins) were shown a lineup and asked to make an identification from the lineup, if they thought the perpetrator of the crime was present in the lineup. There are two data files – one collected in 2014, and one collected in 2015.

Some key variables in the 2014 data file (SAFRANCEEWdata.Rdata):

lineuptpta	(binary, coding whether the target/perpetrator was present or absent in the lineup)
lineupacc	(binary – the outcome variable – recording accuracy of decision in lineup task (1 = correct (hit in target present lineup, correct rejection in target present lineup))
videocondition	(binary, recording whether participants saw a long, or short video of the crime in question – about 3s vs 90s).
confidenceresp	(continuous, witness' self-reported confidence in the correctness of his/her decision, 0 to 100 point scale)
lineuprt	(continuous, reaction time of witness in milliseconds in lineup task)
decisionstrategyresp.1	(self-report item recording whether the witness made the decision after much deliberation, or 'automatically')
decisionstrategyresp.2	(self-report item recording whether the witness rejected the lineup because the faces were unlike that of the perpetrator)

The file Merged_15032016.xlsx is a raw output file from the software program EPrime, which was used to control the experiment. It has all the variables listed above, plus many more, but for a replication conducted in 2015. However, the variable names are different, and need to be deduced e.g. the 2014 variable lineupacc is called Lineup.ACC. The decision strategy variables DecisionStrategyQues and DecisionStrategy.RESP need to be used in conjunction to determine the equivalent of decisionstrategyresp.1 and decisionstrategyresp.2 (these are items 1 and 8 in DecisionStrategyQues). The rest of the variables you need to get from the 2015 data file need to be ascertained in this way.

Treat the 2014 data as the training data, and the 2015 data as the test data set.

1 Load the training and test data files, cleaning them up as needed, in an R script. (15)

2 Select the variables you need, and store them in appropriate dataframes. (5)

NOTE: questions 1 and 2 are difficult, especially the importation and cleaning up of the 2015 data file. To succeed in that task you may need to do the following:

i) save the Excel file in .csv format; ii) note the following correspondence between variables in the 2014 and 2015 files [lineuptpta = LineupTPTA, lineupacc = Lineup.ACC, videocondition = VideoCondition, confidenceresp = Confidence.RESP, lineuprt = Lineup.RT]; iii) note that to select the relevant data for the variables above you need to filter cases so that only trial = 1, and `Procedure[Block]` = ExpProc (note the backticks). Then you need to think about how to get the decision strategy data into the data file. You will likely need to read them into a separate dataframe, filter to select just the ones you want, create a wide data frame, and join that. Tricky!

If you prefer not to tackle the challenge in 1 and 2, then you can load the files containing the test and training data directly (they are in the folder). Your mark will then be computed out of 80, and we will prorate it to a percentage.

3 Explore the data with whatever summary statistics / graphs you think necessary (15)

a. Interpret your findings. What do you expect to find in the modelling phase on the basis of what you see thus far?

4 Build a logistic regression model that uses some or all of the variables above. You should use the logic of training and test sets, building the model on the training set, and then testing it on the test set. (45)

a. Note that the lineuptpta, and videocondition variables are aspects of the design, and should be entered before the other variables, which are predictors.

b. The lineuptpta variable can be expected to have interactions with the decision strategy variables: when the target/criminal is present, one can expect to have 'automatic', rapid recognition, and when the target is absent one can expect to have slow, deliberate feature checking.

c. Interpret the results from your model in terms of the implications for understanding the determinants of witness decisions in lineups.

5 For a logistic regression model relating confidenceresp and lineuptpta to lineup accuracy, compute the following (10)

i. the probability that someone who is > 80% confident is accurate

ii. the odds that this person is correct

iii. the partial odds ratio in the relation between lineuptpta and accuracy of decision

b. Assess the accuracy of the model using a confusion matrix. Check the amount of difference in the prediction errors (sensitivity and specificity – hint, use the confusionMatrix function from package caret). (5)

6 Show algebraically that the logit function used for logistic regression can be written in two alternate ways, representing probability, and odds, respectively - that is, that A, B, and C below are equivalent. Use the equation editor in Word to write your answer, or do it by hand, scan it, and upload it. (5)

A

$$\frac{\hat{p}_i}{1-\hat{p}_i} = e^{(B_1X_i+B_0)}.$$

B

$$\hat{p}_i = \frac{1}{1+e^{-(B_1X_i+B_0)}}$$

C

$$\ln\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = B_1X_i + B_0.$$

Note: page limit is 10 pages
Create a repository on Github and put your R project and all files you need into it. Make regular commits to the repository while you are working on the assignment. When you submit your assignment, please explicitly indicate the link for the repository. An extra 5 marks will be awarded for successful creation and usage of the repository!