Bachelor's Degree in Bioinformatics
Statistical Models & Stochastic processes
Academic year 2023-2024 1st Quarter

Practical 2. Logistic regression

Hand-in date: 29/10/2023

Resolve the following exercise in groups of two students. Write your solution in a Word, Latex or Markdown document and generate a pdf file with your solution. Upload the pdf file with your solution to the corresponding task at the Moodle environment of the course, no later than the hand-in date.

A researcher is interested in how variables, such as GRE (Graduate Record Exam scores), GPA (grade point average), prestige of the undergraduate institution, gender and month of birth affect admission into graduate school. The response variable, admitted or non-admitted is a binary variable (1: admitted, 0: Not).

The dataset has a binary response (outcome, dependent) variable called admit. There are five predictor variables: gre, gpa, rank, gender and month. We will treat the variables gre, gpa and month as continuous. The variable rank takes on the values 1 through 4; Institutions with a rank of 1 have the highest prestige, while those with a rank of 4 have the lowest. The gender takes values 0 and 1 to indicate male and female, respectively.

We will apply logistic regression to a data set of predictor variables on admission. For 10 individuals in the database the admission information is missing. Here we study the relationship between “admit” and the other variables by means of logistic regression, to see if we can predict the admission on the basis of the predictor variables. “Order” is not a predictor variable, is just used as identificator of the row position.

1. (0p) Read the file binary_v2_missing.xlsx into the R environment.
2. (1p) Make a scatterplot of gre against admit. Do you think there is any relationship between the two variables? What is the sample size? Can you think of alternative ways to better visualize the relationship between the variables?
3. (1p) Do a logistic regression of admit on gre. Write down the estimated logit.
4. (2p) How does gre affects the response variable? Interpret the regression coefficient for gre. Apply transformations as you consider adequate.
5. (2p) Is gre a significant predictor (use $\alpha = 0.05$)? Perform the corresponding hypothesis test and report the test-statistic, its reference distribution and the p-value. Is gre a significant predictor if we use $\alpha = 0.1$?
6. (1p) Make a scatter plot of gre versus admit and plot the logistic curve in the scatter plot.
7. (2p) Give a 95% confidence interval for gre using the estimated logit. Exponentiate the limits of this interval and give your interpretation of the result.

8. (2p) The admission (variable admit) of ten individuals is unknown. Use the predict function to predict the logit for these ten individuals, using the model with gre as the sole predictor. Also predict the probability of being admitted or not for each individual. How would you classify these ten individuals?
9. (1p) Calculate McFadden's pseudo R^2 for the logistic regression of admit on gre. What does a zero value for this pseudo R^2 mean?
10. (1p) Calculate the median of the gpa and create an indicator variable medgpa that registers if the gpa is less than or equal to, or above the median. Report how many individuals you have in each category.
11. (2p) Perform the logistic regression of admit on medgpa. Report the estimated logit. What is the interpretation of the exponentiated slope of this regression?
12. (2p) Perform the logistic regression of admit on gpa. Report the estimated logit. Are the results consistent with the previous analysis? Would you prefer the analysis with medgpa over the analysis with gpa? Justify your answer.
13. (2p) Make a logistic regression model with all available predictors, except the indicator medgpa you created. Determine, by doing a hypothesis test, whether the model is globally significant. Report the test statistic, its reference distribution and the p-value.
14. (1p) Try to simplify the model be eliminating non-significant predictors (one by one, use $\alpha = 0.05$). What is your final model?
15. (1p) Test by a likelihood ratio test if we can consider the joint nullity of the coefficients of all variables that you have eliminated. Report the p-value of this test.
16. (1p) Make a scatter plot matrix with the pairs instruction of using the variables in your final model, distinguishing the two varieties with a different symbol or color. Do you see a good graphical separation of the two groups?
17. (1p) Predict the admission of the individuals in the database with your final model, and make a cross table of predicted admit against observed admit.
18. (1p) How often does the model correctly predict the observed admit? Calculate the classification rate, defined as the number of correct classifications divided by the total of classifications made.
19. (1p) Also calculate the classification rate for predictions made with the full model, which contains all predictors. Does this work better? Justify your answer.