QMM: Exercise Sheet 3 - Logistic Regression

Fabien Baeriswyl, Jérôme Reboulleau, Tom Ruszkiewicz

Exercise 1. For this exercise, use the `Profiling.csv` datafile. Suppose the managers of a major mobile communication company is interested in profiling their clients. To do so, they select a sample of 180 clients ($i \in \{1, \ldots, 180\}$) for who they have the following information:

- `mobile`$_i$`:` a dummy variable indicating whether client i has subscribed to a minutes package (type of mobile subscription) or not;
- `income`$_i$`:` monthly income in USD of the ith client;
- `hours`$_i$`:` hours spent by client i using internet;
- `where`$_i$`:` a dummy variable indicating where the internet connection of client i is mostly used, 0 for mostly used at home and 1 for mostly used at the office.

a) First, explore the variables and specifically the target response (i.e. `mobile`) and determine whether a linear regression is appropriate or not.

b) Fit a logistic regression model to predict the variation in `mobile` using `income`, `hours` and `where` as explanatory variables. Which coefficients are significantly different from 0 using a confidence level of $\alpha = 5\%$? Compute $e^{\beta_{\text{where}}}$ and interpret this quantity.

c) Ensure that there is no significant multicollinearity in your model.

d) Fit a new logistic regression to predict the variation in `mobile` using only `income` and `hours`. Test whether this model is an adequate simplification of the model fitted in a) at $\alpha = 1\%$ and at $\alpha = 5\%$.

e) You want to test the predicting ability of the model in a). To do so, you use a new dataset (the `Profiling-Test.csv` datafile). Discuss it with $c = 0.5$ first. Then, raise $c = 2/3$ and discuss the changes.

f) Manually compute the probability associated to the score, i.e. $\hat{\pi}(x_i)$, for the first observation in your test set.

g) Compute the AUC for the ROC of each individual variable. Interpret it.

h) Compute the 95% confidence interval for the regression coefficients.

Exercise 2. For this exercise, use the `ship3.csv` datafile. This is the same data as for the Data Case on the shipping problem we discussed some weeks ago, except that we added a column to indicate whether the year in which stands observation i is a Crisis period or not (this variable is 1 only for 2007 and 2008).

a) Plot the selling price against the crisis variable. Discuss what you see.

b) Fit a logistic regression in trying to predict the variation in the `crisis` variable with the `selling price` as regressor. Interpret $e^{\beta_{\text{selling price}}}$.row

c) Discuss the overall fit of the model in b).

d) Plot the regression line of your model in b) on the graph obtained in a).

e) Split the dataset into a train set and a test set. Note that it is customary to take 75% of the observations in the training set and the remaining 25% in the testing set. Re-fit your model in b) using only the observations in the train set and evaluate its predicting ability on the train set.

Exercise 3. For this exercise, use the `admission.csv` datafile. Suppose a researcher is interested in knowing what are the factors that affect admission to a specific graduate school. The researcher knows the GPA (Grade Point Average) of student i (`GPA` variable), the rank of his undergraduate institution, between 1 and 4 (`rank` variable), rank 1 being the most prestigious category and finally whether the application to graduate studies is successful or not (`admit` variable).

a) Plot the relationship between the GPA and the admission status. Discuss it briefly.

b) Split the dataset between a train set and a test set.

c) Fit a logistic regression in trying to predict the admission status with the GPA and the rank of the undergraduate institution. Be careful: the nature of the undergraduate institution being such that it takes (integer) values between 1 and 4 means that you should interpret each level of it as a factor in your fit. Interpret the values of $\exp(\beta_{\text{rank}_i})$ for $i \in \{1, 2, 3, 4\}$ (there are four ranks for the undergraduate institution).

d) Plot the regression lines (bear in mind that there are multiples of them) on your graph in a).

e) Discuss the predicting ability of your model in c).

f) Draw the ROC for the full model in a) and compute the area under it. Discuss your findings.

g) Try to simplify the model. Can you do so at $\alpha = 5\%$?