Exercises - B part

Prof. Paolo Girardi

November 22, 2021

This file provided a short list of exercises that can be used for educational purposes (2 for binary regression and 3 for Poisson regression). The datasets are inside the moodle folder. They can be imported with the command *db <-read.csv(file.choose(),header=T)*

The solution to this analysis will be provided on the basis of week assignment.

Additional exercises can be found on the suggested books (in particular in the textbook Faraway 2016 - Binary and binomial (pages 46-49 and 64-66) and Poisson regression (pages 95-97)).

1 Binary regression

Exercise 1
The data set **medpar.csv** is an excerpt from US national Medicare inpatient hospital database. It contains 1495 observations on the following variables:

- **los**: length of hospital stay (in days);
- **hmo**: patient belongs to a Health Maintenance Organization (1), or private pay (0);
- **white**: patient identifies themselves as primarily Caucasian (1) in comparison to non-white (0);
- **age80**: patient age 80 and over (1), or age >80 (0);
- **type**: a three-level explanatory variable related to the type of admission (1 = elective, 2 = urgent, and 3 = emergency).

We would like to investigate whether there is an association between the length of hospital stay and the other variables.

(1) Import the data. Based on the descriptive statistics, do you expect that there is a significant relation between `los` and `type`? Justify your answer.

(2) Estimate a Poisson regression model with *los* as dependent variable and type as an explanatory variable. Name this model Model 1. Interpret the parameters of the model (including the intercept).

(3) Add to the model in (2) the explanatory variables *age80*, *hmo* and *white*. Name the resulting model Model 2. Test whether Model 2 has a better fit than Model 1.

(4) Interpret the parameter related to the variable age80.

(5) Test whether the equi-dispersion assumption is matched by the data. If that is not the case:

1. Estimate an adequate model including all the explanatory variables. Name this model Model 3.
2. Compare the results of Model 2 and Model 3. Are they the same?

Exercise 2

The website blablabla.com sells magazine subscriptions. The related company is going to plan a large e-mail marketing campaign. All of the e-mails that will be sent will go to customers that have previously bought a magazine subscription at blablabla.com and who have not opted out of receiving e-mails.

The magazines advertised in each e-mail will be automatically selected for each customer when the e-mail is generated in order to maximize the probability that the customer will buy. The website blablabla.com will only include ads for three magazines in each e-mail in a row at the top of the message so that it is likely that the ads will appear in the e-mail preview (and therefore actually be viewed without the receiver having to open the e-mail). Moreover, the managers believe that including more ads is ineffective.

To evaluate the efficacy of the campaign, the company run an experiment. They sent 673 e-mails to customers containing the ad for the “Art with you” magazine and recorded whether or not the customer purchased this magazine.

The company has also collected data on the customers by matching the information provided by third-party data (which can be purchased from data sources such as the credit scoring agencies) and the recipient of the e-mails when he/she made a purchase at blablabla.com.
The data set **magazine.csv** contains the following information:

- Purchased “Art with you” magazine (Buy = 1 if purchased “Art with you”, 0 otherwise)
- Household Income (Income; rounded to the nearest 1, 000.00)
- Gender (IsFemale = 1 if the person is female, 0 otherwise)
- Marital Status (IsMarried = 1 if married, 0 otherwise)
- College Educated (HasCollege = 1 if has one or more years of college education, 0 otherwise)
- Employed in a Profession (IsProfessional = 1 if employed in a profession, 0 otherwise)
- Retired (IsRetired = 1 if retired, 0 otherwise)
- Not employed (Unemployed = 1 if not employed, 0 otherwise)
- Length of Residency in Current City (ResLength; in years)
- Dual Income if Married (Dual = 1 if dual-income, 0 otherwise)
- Children (Minors = 1 if children under 18 are in the household, 0 otherwise)

- Home ownership (Own = 1 if own residence, 0 otherwise)
- Resident type (House = 1 if the residence is a single-family house, 0 otherwise)
- Race (White = 1 if the race is white, 0 otherwise)
- Language (English = 1 if language is English, 0 otherwise)
- Previously purchased an art magazine (PrevArt = 1 if previously purchased an art magazine, 0 otherwise).
- Previously purchased a cinema magazine (PrevCin = 1 if previously purchased a cinema magazine)

(1). Estimate a logistic regression model, with Buy as the dependent variable and Gender, Not Employed, Income, and Own as explanatory variables. Is there any explanatory variable you would remove from the model? Based on which test? Remove the variable(s) and name the resulting model as Model 1.

(2). Add to Model 1 the variables PrevArt and PrevCinema. Name the resulting model as Model2. Compare Model 1 and Model 2 using an appropriate test and comment on the results.

(3). The data set includes many potential explanatory variables. Automatic procedures to select the "best" model are implemented in any software. One of these procedures is called "backward elimination" and performs the following steps:

- Step 0 The binary logistic regression model including all the potential explanatory variables is fitted and the likelihood L_0 of the full model is computed.
- Step 1) All the possible p models obtained by excluding one of the possible p variables are estimated and the model with the lowest AIC (or BIC) is considered. ii) Let X_j be the variable that is excluded and $L(1)_j$ be the likelihood of the model without X_j. The significance of X_j is tested using the G statistic $G(1)_j = -2log\frac{L(1)_j}{L(0)}$ if X_j is not significant i.e., $G(1)_j<\chi^2_{1,1-\alpha}$, the considered model is selected, otherwise, the full model is selected and the selection procedure ends.
- Step 1 is repeated until the variable that is dropped at the generic step k is significant. Use the commands:
  ```
  fullmodel <− glm(Buy  .,family=’binomial’, data=advertisement)
  mod.fin <− step(fullmodel, direction = ’backward’)
  summary(mod.fin)
  ```
  to select the "best" model. Interpret the parameters describing the relation between Buy and all the selected explanatory variables

2 Poisson regression

Exercise 3
The current study assesses the relationship between cardiac vagal tone and compassionate behaviors also considering whether the vagal activity could moderate the stress-compassion link. Seventy-nine primary school children were assigned to 2 compassion-eliciting conditions (pain vs. mild distress) x 2 stress conditions (stress vs. no stress). During the test, children's reaction to the scene was video recorded and the number of compassionated behavior was registered. The children's heart rate was registered for the entire session. The data set **compassion.csv** contains the following information:

- Stress: presence of a stress condition before or after the task;
- Condition: Pain or mild-distress;
- Gender: (Gender = 1 if the person is female, 0 otherwise);
- rMSSD_base= rMSSD calculated before the test;
- rMSSD_compassion: difference of the rMSSD during and before the test;
- Compassion: occourences of a compassionated behaviour during the test.

(1). Estimate a Poisson regression model, with Compassion as the dependent variable and the other variables as explanatory. Is there any explanatory variable you would remove from the model? Based on which test? Remove the variable(s) and name the resulting model as Model 1.

(2). Check the presence of interactions in Model 1.

(3). Perform a diagnostic test and evaluate the presence of outliers or overdispersion/underdispersion.

Exercise 4
The data contained in the data frame **ants.csv** was obtained through an experiment conducted by students of a degree course in Applied Sciences at a University in Australia (Mackisack, 2017). The purpose of the experiment was to evaluate the preference of ants of the species Iridomyrmex purpureus, or meat ants, compared to several types of sandwiches. For each of

- 4 types of bread, Bread (1, rye; 2, wholemeal; 3, multi-grain; 4, white);
- 3 types of filling, Filling (1, Vegemite; 2, butter of peanuts; 3 ham and pickles);
- presence or not of butter, Butter (1, present; -1, absent).

Two pieces of the sandwich were placed near the entrance to an anthill. All 48 pieces were the same size. After 5 minutes, a glass was placed on each piece of sandwich and the number of ants captured has been counted, *Ant_count* is the variable that stores this information.

In one case, the value of the answer is equal to 67.5, since the positioning of the glass has halved one ant. It is interesting to evaluate how the number of ants depends on the characteristic of the sandwich.

(1). Estimate a Poisson regression model, with *Ant_count* as the dependent variable and the other variables, except for the *order* variable, as explanatory. Test the presence of interactions.

(2). Do you find outliers? Complete a diagnostic on the model and evaluate to fit again the model against the presence of potential outliers. Evaluate as the estimates change.

Exercise 5

The data contained in the data frame **awards.csv** is related to the number of awards earned by students at one high school. Predictors of the number of awards earned include the type of program in which the student was enrolled (e.g., vocational, general or academic) and the score on their final exam in math.

In this dataset, *num_awards* is the outcome variable and indicates the number of awards earned by students at a high school in a year, *math* is a continuous predictor variable and represents students' scores on their math final exam, and *prog* is a categorical predictor variable with three levels indicating the type of program in which the students were enrolled.

(1). Estimate a Poisson regression model, with *num_awards* as the dependent variable and the other variables, except for the *ID* variable, as explanatory.

(2). Test the presence of interactions and estimate the best predictive model.