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

Practical 4. Poisson regression

Hand-in date: 10/11/2022

Resolve the following exercises in groups of two students. Write your solution in a Word, Latex or MarkDown document and generate a pdf file with your solution. Upload your solution to the Moodle environment of the course, no later than the hand-in date. Please take care to write your names and surnames on the first page of your report.

Poisson regression (20p). We consider a data set reporting the daily total of bike counts conducted on the Brooklyn Bridge from 01 April 2017 to 31 October 2017 (Source: NYC Open Data: Bicycle Counts for East River Bridges). The variables included in the dataset are HIGH_T; LOW_T; PRECIP and LABOR_YESNO. The number of bikes (BB_COUNT) is used as response variable in a Poisson regression. Predictors (4 variables) we consider as potentially bearing a relationship with the number of bikes are the high and low temperature (HIGH_T; LOW_T;) the precipitation (PRECIP) and the if day of the week is a working day or not (LABOR_YESNO with values 1 and 0 to indicate yes or no, respectively).

a) (0p) Read the dataset, you can use instruction attach the dataset for convenient access to the variables in the dataset.
b) (1p) Make a barplot of the table of the possible outcomes of the response variable BB_COUNT. Calculate descriptive statistics of response BB_COUNT. Is there, at the exploratory level, evidence that the response does not follow a Poisson distribution?
c) (1p) Perform Poisson regression of the number of BB_COUNT on HIGH_T, our first model. Report the regression equation. Is there evidence for association, and if so, what kind of association?
d) (1p) Make a scatter plot of BB_COUNT against HIGH_T. Add the fitted regression equation to the scatter plot.
e) (1p) Estimate the null model without predictors (BB_COUNT~1), and also plot the equation according to this model to the plot. What do you observe?
f) (2p) Interpret the first model by quantifying the effect of the predictor on the average of the response. Give a 95% confidence interval for the parameter representing that effect.
g) (1p) Is the value 0 inside the interval you obtained? Is the value 1 inside the interval? What is the relevance of this?
h) (1p) Is there any indication that overdispersion is a problem for you model? Justify your answer.
i) (1p) Formally test for overdispersion using the function dispersion test of the AER package. What is your conclusion?

j) (1p) Calculate deviance residuals according to the first model and plot these as a function of the predicted values, using a different color for each category of LABOR_YESNO. What do you observe?

k) (1p) Do a Poisson regression of BB_COUNT on HIGH_T and LABOR_YESNO. Report the fitted equation. Is there evidence for any effect of the variable LABOR_YESNO? Justify your answer.

l) (1p) Make a graphic by representing the newly fitted model in a scatterplot of BB_COUNT against HIGH_T.

m) (1p) Is there evidence for interaction between the variables LABOR_YESNO and HIGH_T? Justify your answer. Try to make a graphical representation of the fitted model with interaction in a scatterplot of BB_COUNT against HIGH_T.

n) (2p) Add the variable PRECIP to the model. Is it a significant predictor? Justify your answer.

o) (2p) Add the variable LOW_T to the model. Is it a significant predictor? Justify your answer.

p) (2p) What would be your final model for the data? Justify your answer.

q) (1p) Give examples of outcomes that can be modelled using a Poisson regression, such as the number of goals in a handball match.