Generalised Linear Models, Assessed Practical

Week 7, MT 2020

- This practical sheet contains two sections. Write a report on the Assessed Exercise in Section 2 only.
- The report has soft word limit at 2000 words and a hard limit at 2500 words. This word limit is on the main body of the report. Equations, tables, figures, captions, appendices to your report and computer code do not contribute to the word count.
- You should use your anonymous practical ID (and not your real name) for the cover page of the report, and you should name the PDF file you upload using that same ID (e.g. "P042.pdf").
- The hand-in deadline is 12 noon Thursday 10 December.

Any queries you have about the exercise in Section 1 may be directed to the Teaching Assistant or the Lecturer during the practical session. Neither would answer questions regarding the exercise in Section 2, with the sole exception of questions relating to a limited number of programming issues.

1 Exercise for practice, NOT ASSESSED

The dataset `bw.csv` gives details of 189 babies and mothers, focusing on low birth weight. The dataset contains information on:

- `low`: birth weight status, 1 = birth weight less than 2.5 kg, 0 otherwise
- `age`: mother's age in years
- `mwt`: mother's weight in pounds
- `race`: mother's race (1 = white, 2 = black, 3 = other)
- `smoke`: 1 if smoked during pregnancy, 0 otherwise
- `ptl`: number of previous premature labours
- `ht`: indicator, 1 if mother has history of hypertension, 0 otherwise

1. Load the data, using for example `read.csv()`:
   ```
   bw <- read.csv("bw.csv")
   ```
2. Produce some suitable exploratory plots of the data, examining the relationships between the variables.
   ```
   # brief hints
   bw$race <- as.factor(bw$race)
   bw$ptl <- as.factor(bw$ptl)
   # to be able to refer to a column as e.g. race rather than bw$race
   # here it is convenient to:
   attach(bw)
   # use detach(bw) to remove it when finished, can check using search()
   # For plot examples
   (tab1 <- table(low, race))
   barplot(tab1, beside = TRUE)
   # can use e.g. names.arg and col arguments of barplot() to improve plot
   boxplot(mwt ~ low, xlab='low bw', ylab='mother weight')
   ```
3. Which GLM do you specify to analyse how the incidence of low birth weight depends on the other variables? Motivate your choice. What are your priors wrt the directions of the effects?
4. Carry out model selection using appropriate tests (ignoring any interaction terms). Focus also on whether `race` can be reclassified as a binary indicator white-nonwhite and whether `ptl` can be redefined as a binary indicator $\{0, \geq 1\}$.
5. Assess the quality of the model fit using suitable methods. If you decide that your data has outliers, explain why and say what action you took in response.
6. Interpret your findings fully.
7. Compute the average marginal effect for `mwt`. How do you interpret this effect? Sketch how you would obtain a standard error.

2 ASSESSED EXERCISE

The data in `dvis.csv` relate to the number of visits to a family doctor, or GP. Each row of the file corresponds to one individual. For each individual the variables available are either at the individual level, or at the household level in which the individual resides, and are given by:

- `docvis`: the number of visits to a family doctor in month before interview
- `age`: age in years
- `hhninc`: net monthly household income in German Marks/1000
- `female`: indicator, 1 if female, 0 otherwise
- `hhkids`: indicator, 1 if children under the age of 16 present in household, 0 otherwise
- `married`: indicator, 1 if married, 0 otherwise
- `employed`: indicator, 1 if in paid work, 0 otherwise
- `educyrs`: years of schooling
- `eductype`: highest degree of schooling obtained, ordered, 0 = none, 1 = high school,..., 5 = university
- `privateins`: indicator, 1 if having private health insurance, 0 otherwise (covered by public insurance)
- `addins`: indicator, 1 if purchased additional health insurance (on top of private/public insurance), 0 otherwise

Exercise:
Investigate and write a report on how the number of visits to a doctor depends on the other variables. The main goal here is to obtain a suitable interpretable model and to give a full interpretation of that model.

1. Perform an exploratory analysis of the data and summarise your findings. As well as producing suitable plots that examine the relationship between the number of visits to a family doctor and the available explanatory variables, you may also wish to consider some numerical summaries.
2. Model the relation between the number of doctor visits and the other variables that are available using the Poisson GLM. Using appropriate tests, carry out model selection to examine the relationship between the possible explanatory variables and the number of visits.
3. Assess the quality of the model fit using suitable methods.
4. Interpret your final model carefully. Give an interpretation/explanation of the effect of each of the variables included in your final model on the number of visits to a family doctor.
5. Calculate an estimate for the dispersion parameter ϕ. What does this estimate imply for the standard errors you found for the model in 4.?