MSc, Generalised Linear Models, Assessed Practical

Week 8, MT 2022

- This practical sheet contains two sections. **Write a report on the Exercise in Section 2 only.**
- **The report has a word limit of 2000 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 (of the form P123 and not your name)** for the cover page of the report, and you should name the PDF file you upload using that same ID (e.g. `P123.pdf`).
- **You should submit your report via the Inspera system**. There will be instructions about how to do this on the Canvas Practicals page next to this practical handout.
- **The hand-in deadline is 12 noon Wednesday 7 December**.

Any queries you have about the exercise in Section 1 may be directed to the lecturer during the practical session. The lecturer will not 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 pre-pregnancy weight in pounds
- `race`: mother's race (1 = white, 2 = black, 3 = other)
- `smoke`: 1 if smoked during pregnancy, 0 otherwise
- `ptlp1`: 0 if no previous premature labours, 1 otherwise
- `ht`: 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)
# 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 of the other variables on the incidence of low birth weight?
4. Carry out model selection using likelihood ratio tests (ignoring any interaction terms).
5. Assess the quality of the model fit using suitable methods.
6. Interpret your findings fully.
7. Compute an estimate of the average marginal effect for `mwt`.

2 ASSESSED EXERCISE

The data in `docvis.csv` relate to the number of visits to a family doctor/GP by an individual in the two weeks before they were surveyed, in Australia, 1977-1978. Each row of the file corresponds to one individual. For each individual the variables available are given by:

- `visits`: the number of doctor visits in past 2 weeks
- `age`: age in years
- `income`: annual income in tens of thousands of dollars
- `female`: 1 if female, 0 otherwise
- `private`: 1 if individual has private health insurance, 0 otherwise
- `freepoor`: 1 if individual has free government health insurance due to low income, 0 otherwise
- `freerepat`: 1 if individual has free government health insurance due to old age, disability or veteran status, 0 otherwise
- `lchronic`: 1 if individual has a chronic condition limiting activity, 0 otherwise

Exercise:
Investigate and write a report on how the number of doctor visits 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 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 with canonical link function. Do not consider all possible interactions, but only interactions of the female indicator with the other variables. 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 and present the estimated effects and their 95% confidence intervals 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.?