MSc, Generalised Linear Models, Assessed Practical

Week 8, MT 2023

- 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.
- **You should submit your report via the Inspera system**.
- **The hand-in deadline is 12 noon Wednesday 6 December 2023** .

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. Produce summary statistics and some exploratory plots of the data, also examining the relationships between low birth weight status and the other variables.
2. Which GLM do you specify to analyse how the incidence of low birth weight depends on the other variables? Motivate your choice.
3. Carry out model selection (ignoring any interaction terms).
4. Assess the quality of the model fit using suitable methods.
5. Interpret your findings fully.
6. Compute an estimate of the average marginal effect for `mwt`.

2 ASSESSED EXERCISE

The data in `mortg.csv` relate to mortgage approvals. The data are a sample of mortgage applications in a US city in 1990. Each row of the file corresponds to one application/applicant. The variables available are given by:

- `approved:` 1 if the application was approved, 0 otherwise
- `hir:` ratio of monthly housing expenses to monthly income (if mortgage were to be approved)
- `odir:` ratio of other monthly debt payments to monthly income
- `lvr:` ratio of size of loan to assessed value of property
- `mcs:` mortgage credit score from 1 to 4 (a low value being a good score)
- `self:` 1 if applicant is self employed, 0 otherwise
- `single:` 1 if applicant is single, 0 otherwise
- `white:` 1 if applicant is white ethnicity, 0 if black ethnicity
- `uria:` 1989 state-wide unemployment rate in the applicant's industry

Exercise:
Investigate and write a report on how the probability of mortgage approval 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 bivariate plots that examine the relationship between the mortgage approval and the available explanatory variables, you should present some numerical summaries.
2. Model the relation between the mortgage approvals and the other variables that are available using the appropriate GLM with canonical link function. Carry out model selection to examine the relationship between the possible explanatory variables and mortgage approval. Do not consider all possible interactions, but only interactions of the self-employed indicator with the other variables. Investigate whether the mortgage credit score should be included as is (numeric), or as different indicators (as.factor).
3. Assess the quality of the model fit using suitable methods.
4. Interpret your final model carefully. In particular, present and interpret the estimated multiplicative effects on the odds ratio scale of the variables included in your final model. Include 95% confidence intervals for these effects.
5. Calculate, and comment on, an estimate for the dispersion parameter ϕ.