SB1: 2nd Assessed Practical

Week 8, MT 2019

- This practical sheet contains two sections. **Only the exercise in Section 2 is assessed** – it contributes 8.5% to your raw SB1 total mark.
- Write your answer to the exercise in Section 2 as a report.
- The deadline for submission, which is officially published in the Part B synopses, is:

  12 noon Monday week 2, Hilary Term 2020,
  at the Statistics Department reception, 24–29 St Giles'.

- On the cover page of your report:

  please use your candidate number, not your name (nor your student number).

Your report should be clearly written. There are no marks awarded for presentation but there are marks awarded for clarity. When you make a statistical test do not just report the p-value but also report your conclusion using plain language. The same holds for interpreting models; reporting the coefficients and their standard errors is required but try to link your results to the research question as well. You should use captions for your tables and figures and include your commented R-code, preferably in an appendix.

HINTS

- You must hand-in a paper copy of your report by the deadline. Note that hand-in is at the Department of Statistics.
- Your report must be accompanied by a completed declaration of authorship form (on paper) saying that the report you are submitting is your own work. The declaration form will be available on the course material page:
  `http://www.stats.ox.ac.uk/current_students/bammath/course_material`
- There are some tips about writing reports, an example practical report and accompanying R code on my SB1 practicals page:
  `http://www.stats.ox.ac.uk/~laws/SB1/SB1_practicals.html`
- Your report should probably be between 5 and 10 pages including any figures/tables. The best way to include your R code is as an appendix (not as part of the report). If you need, an extra page or two for R code is ok, but you should aim for the main report to be no more than 10 pages – part of writing clearly is writing concisely.
- Your report doesn't need to be in LaTeX, but you can of course use LaTeX if you wish.
- There is plenty of time until the deadline. The report should only take a fraction of this time. But please don't leave writing your report to the last minute.

1 Exercise for practice, NOT ASSESSED

The dataset `awards.csv` gives the number of awards earned by students at one high school.

The outcome variable is `num_awards`, the number of awards. Predictors of the number of awards earned are:

- `prog`: the type of program in which the student was enrolled (1 = vocational, 2 = general, 3 = academic)
- `math`: the score in their final maths exam
- `gender`: gender (0 = males, 1 = females).

1. Read the data in `awards.csv` into R using `read.csv()`.
2. Produce some suitable tables and/or explanatory plots of the data and comment on these.
3. Fit a Poisson GLM with canonical link function and begin with the model that includes all possible interactions of the explanatory variables. Now carry out model selection using likelihood ratio tests.
4. Assess the quality of the model fit using suitable goodness of fit methods. If you decide that your data has outliers, explain why and say what action you took in response.
5. Give an interpretation of your fitted model.

Use the accompanying R-file `awards.R` for various hints on how to approach this exercise. The R examples from lectures should also be helpful.

`http://www.stats.ox.ac.uk/~laws/SB1/awards.R`

2 ASSESSED EXERCISE

The dataset `workf.csv` is concerned with the participation of women aged 20–35 in the workforce in Canada.

The response variable `employed` indicates whether a woman was in employment (`employed = 1`) or not (`employed = 0`). The explanatory variables are:

- `region`: the region of the country in which the woman lived – one of `Atlantic`, `BC` (= British Columbia), `Ontario`, `Prairies`, `Quebec`
- `ch04`: the presence, or not, of children aged 0 to 4 in the household (`Yes` or `No`)
- `ch59`: the presence, or not, of children aged 5 to 9
- `ch10`: the presence, or not, of children aged 10 to 14
- `income`: family income after tax in $1000s, not including the woman's own income if any
- `educ`: the number of years of education.

Write a report on the exercise below.

Exercise:

You are asked to investigate how the employement status '`employed`' depends on the available explanatory variables. The aim is to: (i) perform some exploratory analysis; (ii) obtain a model that explains how employement status depends on the other variables; (iii) interpret the model you obtain.

1. Produce some suitable tables and/or explanatory plots of the data and comment on these.
2. Carry out a univariate logistic regression for each of the explanatory variables with canonical link function. Interpret your findings.
3. Considering all of the available explanatory variables, carry out model selection using likelihood ratio tests and write down the GLM for your selected model (including the linear predictor and link function).
4. 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.
5. Give a full interpretation of your fitted model.
6. Is there any evidence of an interaction between any of the explanatory variables, and if so what is the effect?

```
# to load the data from the web
workf <- read.csv("http://www.stats.ox.ac.uk/~laws/SB1/data/workf.csv")
# and look at first few rows
head(workf)

# maybe better is to save a copy of workf.csv and load it:
```

```
#
# if you have saved a copy of workf.csv you could use:
# workf <- read.csv(file.choose())
# and then choose the workf.csv file
#
# or you could use something like:
# workf <- read.csv("../data/workf.csv")
# where you need to specify where to find the workf.csv file
```