ST346: Assessed coursework 1

Generalized Linear Models

Deadline 12 noon (GMT) Tuesday 27 October 2020

Your solutions should be submitted electronically in the form of a PDF document using the submission portal on the ST346 Moodle page. Please remember to include only your **ID number** on your submission to allow anonymous marking.

If you have any queries about the coursework please post them on the ST346 forum, but do not post any part of your solutions. This assignment counts towards **10%** of your final module mark.

The maximum score for this coursework is 20/20. Numbers in brackets indicate the points available for each question.

To access the data needed for this assignment, download the file `courseworkData1.rda` from the ST346 Moodle web page and read it into R using the function `load()`. This will create a copy of two data frames in your R workspace: `insurance` and `doctors`.

1. The `insurance` data set concerns the number of car insurance claims main by clients of an insurance company in a single year. Variables in the data set are:

   - **car** Engine size of car (1: $<$ 1 litre, 2: 1–1.5 litres, 3: 1.5–2 litres, 4: $>$ 2 litres).
   - **age** Age group: (1: < 25 years, 2: 25–29 years, 3: 30–35 years, ≥ 35 years)
   - **district** Where policy holder lived (1: urban area, *i.e.* in a city; 0: rural area, *i.e.* outside a city)
   - **y** Number of claims
   - **n** Number of insurance policies

   In this data set, individual policies have been aggregated into groups defined by the cross-classification of `car`, `age`, and `district` giving $N = 4 \times 4 \times 2 = 32$ rows.

   (a) Fit a null Poisson regression model with number of claims as the outcome, but none of the variables `car`, `age`, `district` as predictor variables. Show that the estimate for the intercept term is numerically equal to

   $$\log\left(\frac{\sum_{i=1}^{N} y_i}{\sum_{i=1}^{N} n_i}\right)$$

   i.e. the log of the rate of claims per policy across all policy-holders. [2]

   (b) Fit another Poisson model with predictor variables `car`, `age`, and `district` where `car` and `age` are factors (*i.e.* considered as categorical variables).

   If we denote the coefficient for the variable `district` by β_d then $\exp(\beta_d)$ is the ratio between the rate of claims in urban *vs.* rural areas. Give an estimate of this rate ratio. Is the rate of insurance claims higher in urban or rural areas? [2]

   (c) Use stepwise regression to determine whether the model in question 1b can be improved by removing predictor variables or adding interactions. Your minimal model should be the null model fitted in question 1a and your maximal model should be one with all predictors and all 2-way interactions. [3]

(d) Using the model chosen by stepwise regression in question 1c, test whether a linear dose-response with age is a better fit than a categorical model with the `anova()` function (If your "optimal" model does not include age then you have gone wrong. Try question 1c again).

You will need to use the 2-argument version of the `anova` function

```
anova(m1, m2, test="LRT")
```

where `m1` and `m2` are the two fitted models returned by the `glm()` function. [2]

(e) The insurance company wants to make the insurance premiums proportional to the risk of an insurance claim. A customer pays a $100 dollar premium for a car in category 1. If they change their car to one in category 4 then what should be their new insurance premium? [2]

2. The data frame `doctors` comes from the British Doctors Study (Follow the link for more information). This study, which began in 1951, was the world's first large prospective study of the effects of smoking to establish a convincing linkage between tobacco smoking and cause-specific mortality (death).

The `doctors` data set concerns deaths from coronary heart disease 10 years after the start of the study. The data on 34494 participants have been aggregated into 10 groups defined by age and smoking status. The variables in the data set are:

- **age** Age group. A factor with levels: 35–44, 45–54, 55–64, 65–74, 75–84.
- **smoking** A binary indicator of smoking habits (1=smoker, 0=non-smoker)
- **deaths** Total number of deaths that occurred in each group in 10 years of follow-up.
- **personyears** Total number of person-years of follow-up in each groups (*i.e.* if 5 doctors are followed for 10 years then the group has $5 \times 10 = 50$ person-years of follow-up)

(a) Consider the following model

$$D_i \sim \text{Poisson}(\mu_i)$$
$$\log(\mu_i) = \alpha + \beta s_i + \log(Y_i)$$

where D_i is the number of deaths in row i, Y_i is the number of person-years of follow up in group i and s_i is the smoking status.

Fit this model in R and show numerically that:

$$\widehat{\beta} = \log\left(\frac{\widehat{\lambda}_1}{\widehat{\lambda}_0}\right)$$

where $\widehat{\lambda}_1$ is the estimated mortality rate in smokers and $\widehat{\lambda}_0$ is the estimated mortality rate in non-smokers.

$$\widehat{\lambda}_1 = \frac{\sum_{i\in\mathcal{S}} D_i}{\sum_{i\in\mathcal{S}} Y_i}$$
$$\widehat{\lambda}_0 = \frac{\sum_{i\in\mathcal{N}} D_i}{\sum_{i\in\mathcal{N}} Y_i}$$

where $\mathcal{S}$ is the set of rows containing smokers and $\mathcal{N}$ is the set of rows containing non-smokers. [3]

(b) Now consider this model:

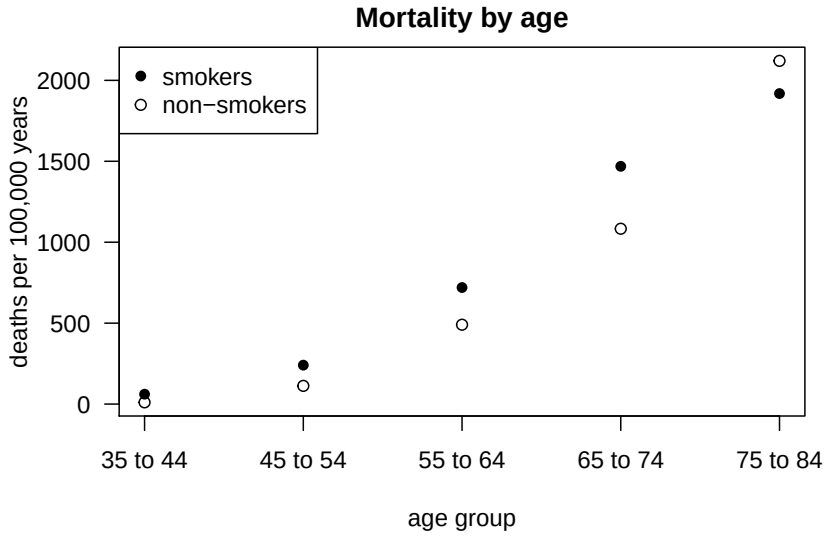
$$\log(\mu_i) = \alpha + \beta s_i + \sum_{g=2}^{G} I_{\{a_i=g\}}\gamma_g + \log(Y_i)$$

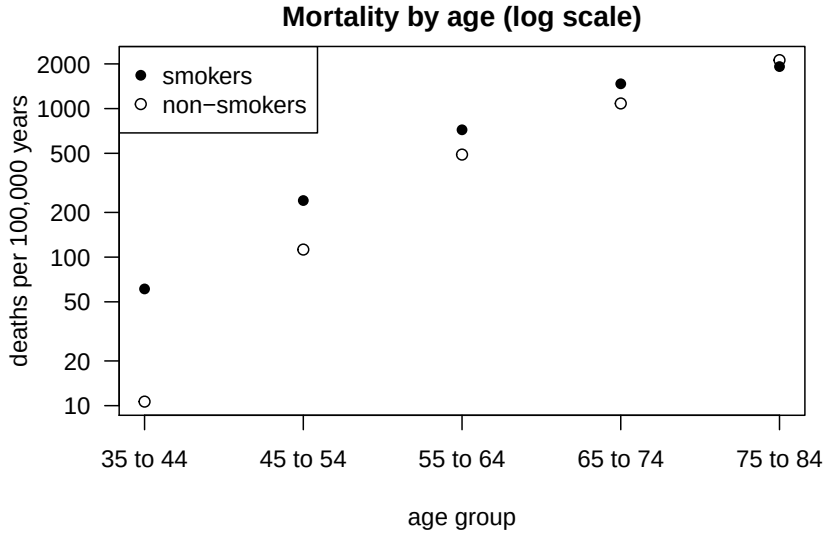
where $a_i \in \{1, 2, \ldots G\}$ is the age group in row i, and $G = 5$ is the number of age groups.

Fit this model in R. What happens to the estimate of β compared with model 2a? [3]

(c) Under the model in question 2b, the ratio of the mortality rates for smokers versus non smokers is assumed constant across age groups.

The figure below shows the estimated rates for smokers and non-smokers The top panel shows the rates on an arithmetic scale and the bottom row shows the rates on a logarithmic scale.





Is the model in question 2b appropriate? Propose an alternative model that allows the effect of smoking to depend on age. Give an estimate of the mortality rate ratio for smokers vs non-smokers among individuals aged 65–74. What is the p-value for the test that this rate ratio is equal to 1 (Hint: use the stratified parameterization and look at the output of the `summary()` function for the p-value). [3]