ST346: Assessed coursework 1

Generalized Linear Models

Deadline 12 noon (GMT) Tuesday 01 December 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 `courseworkData2.rda` from the ST346 Moodle web page and read it into R using the function `load()`. This will create a copy of 2 data frames in your R workspace: `ships` and `esoph`.

1. The data frame `ships` contains data on damage incidents in ships. The ships have been aggregated into classes defined by `type` (5 categories), `year` of construction (4 categories), and `period` of operation (2 categories), giving $5 \times 4 \times 2 = 40$ rows. For each combination of `type`, `year` and `period`, the total months of `service` of all ships in that class has been aggregated in the variable `service`, and the total number of damage incidents has been aggregated in the variable `incidents`.

   Note that some classes are empty. For these classes, the variable `service` is set to zero and the variable `incidents` is set to the missing value `NA`. Rows with missing values are automatically removed by R when you fit a statistical model.

   (a) Fit an appropriate GLM to model the risk of damage incidents in terms of the other variables. [2]

   (b) Which combination of ship type, year of construction and period of operation is at highest risk of damage? [2]

   (c) Calculate the expected number of damage incidents in twelve months for this class of shipping. [2]

2. Consider the following scenarios. Determine the appropriate model to use in each case. You should give the exponential dispersion model (or family in R terminology), the link function, the outcome variable, and the predictor variables.

   (a) A cohort of 18 year-old school leavers were surveyed regarding their future plans, with specific interest in higher education. Several variables were measured, including age, gender, smoking status, as well as whether or not they had a place at university or other higher education institute. We wish to investigate the variables which are associated with higher education attendance.[1]

   (b) Sixty rats of the same age were divided into 2 groups of 30. At the start of the experiment, all animals were weighed. Then one group was fed a control diet, whilst the other was fed a diet supplemented with vitamin D. After 6 weeks, all rats were weighed again. We wish to understand how vitamin D supplementation affects weight gain. [1]

   (c) A car manufacturer conducted a study to investigate the reliability of their cars. They measured the number of times that each of 1000 vehicles had broken down in the 10 years since it was

made. They also measured the number of miles that each car had been driven as well as the model of each vehicle and the number of times that it had been serviced. [1]

3. The data set `esoph` concerns a study of esophageal cancer conducted in the Ille-et-Vilaine department of France in the 1970s. Participants have been classified by age group (`agegp`), alcohol consumption in g/day (`alcgp`) and tobacco consumption in g/day (`tobgp`). In each class defined by these variables, `ncases` gives the number of participants in the study with esophageal cancer and `ncontrols` gives the number of healthy controls without cancer.

   (a) Fit an appropriate GLM to these data to model the risk of esophageal cancer. [2]

   (b) The variables `agegp`, `alcgp`, and `tobgp` are factors. They can be converted to numeric variables using the function `as.numeric()`. Determine whether each variable should be a factor or a numeric variable using an appropriate methodology for model choice. [2]

   (c) Using your final choice of model, estimate the odds ratio of esophageal cancer for a participant who drinks more than 120g of alcohol/day compared with one who drinks less than 40g of alcohol per day. [2]

4. Consider an experiment in which it is known that the variance of the errors of the first two observations $(i = 1, i = 2)$ is 3 times the variance of the errors for the next three observations $(i = 3, i = 4, i = 5)$. The errors are assumed to be uncorrelated. You are given the following model:

$$\mathrm{E}\left(y_i \mid \boldsymbol{\beta}\right) = \mu_i = \begin{cases} \beta_0 & i = 1 \text{ or } i = 4, \\ \beta_0 + \beta_1 & i = 2 \text{ or } i = 5, \\ \beta_0 + 2\beta_1 & i = 3. \end{cases}$$

   and you may assume that $\mathrm{Var}\left(y_i \mid \boldsymbol{\beta}\right) = \sigma^2$

   Find the design matrix for this model and then calculate the maximum likelihood estimate of β_1 to 2 decimal places when $\boldsymbol{y}^T = c(115.0, 72.5, 22.5, 110.0, 60.0)$. You should do this calculation twice: once by calling the `lm()` function in R and once by solving the weighted normal equations (or score equations) for this model. [5]