MATH38172 Generalised Linear Models

Coursework 2021

Instructions

Attempt the questions below and submit your work online via Blackboard by the deadline of **Monday 26th April 2021 at 11am**. Your submission must be a single file. It may contain any sensible mix of word-processed and (scanned) handwritten parts, for example using LaTeX, RMarkdown or Microsoft Word. You should include any R code used. The coursework may take up to 10 hours to complete.

The submitted work MUST be your own. Plagiarism may result in zero marks for all parties involved and/or disciplinary action.

Background

The file `cwkdata.csv` on Blackboard contains data on 120 patients under the age of 50 admitted to hospital with a respiratory disease. For each patient the data include the patient's anorexia status (yes/no), their temperature on admission (°C), their blood lymphocyte level (in billions of cells per litre), and whether the disease became severe.

Questions

1. Import the dataset into R and inspect its structure. What do you notice? (2 marks)
2. i) Fit a GLM to predict the probability of severe disease using **only** anorexia status as an explanatory variable (plus the intercept). (2 marks)

   ii) Write down your fitted model in equation form, and interpret its parameters. Make sure to fully define any variables involved, and any assumed distributions. (2 marks)

   iii) Use this model to assess whether there is significant evidence that anorexia status is associated with the probability of severe disease. Explain your testing methodology. (2 marks)
3. Now fit two further models to predict the probability of severe disease:

   i) using **only temperature** as an explanatory variable (plus the intercept)

   ii) using **only lymphocyte level** as an explanatory variable (plus the intercept)

   Use these models to assess whether there is significant evidence that temperature and lymphocyte level are associated with the probability of severe disease. (1 mark)
4. i) Now use R to fit a model including *all three* explanatory variables in the linear predictor (anorexia, temperature, and lymphocyte level) as well as the intercept. Which variables now show significant association with the probability of severe disease? (1 mark)

   ii) Compare your answer in (i) above to the results of your previous tests in Questions 2 and 3. What do you think is the correct conclusion about which variables affect the probability of severe disease? Explain the reasoning behind your answer, and also the reasons behind any differences in the conclusions from the different tests. (4 marks)

   iii) Overall, which of the four models would you advise is used to explain the data? Why? (4 marks)
5. Using your final model, estimate the probability of severe disease for a non-anorexic patient with a temperature of 38°C and lymphocyte level of 1.3 billion cells per litre of blood. (2 marks)