STAT6123 Generalised Linear Modelling

Assignment

- This assignment is worth 50% of the overall mark for STAT6123.
- The deadline for submission is **16.00 on Thursday 1 December 2022**.
- Standard University policies and procedures will be followed for late submission, extensions and academic integrity (see the Module Outline for details).
- Submission is via Blackboard.
  - You should submit a report containing your answers via TurnitinUK on Blackboard (see Module Outline for details) in a file called `report-`*ID*`.pdf`, where *ID* is your student ID number, for example `report-1234567.pdf`. In the STAT6123 Assignments folder, click on View/Complete to submit your report. Please enter this file name as the Submission title.
  - You should not include R code used in your analysis in your report, but you must submit a separate R script via Blackboard containing your code called `code-`*ID*`.R`, for example `code-1234567.R`. This code should reproduce the results contained in your report. Please rename and use the R template `code-yyy.R` provided. In the STAT6123 Assignments folder, click on Assignment code submission to submit your code.
- The page limits given below for each task are strict and is easily sufficient to receive full credit. Any pages beyond the limits will not be marked.

Task 1 [Total 65 marks, max. 9 pages]

How household expenditure varies with household income and other variables is of key interest in socio-economic studies. In order to investigate this, you are provided with data from a survey that collected high quality data on expenditure, income and other socio-economic variables. Your task is to use these data to develop a model to explain variation in household expenditure. A description of the available variables is presented below. The data (1200 observations) is included in the file `expenditure.txt` (available on Blackboard).

`id`	Household identifier
`expenditure`	Total household expenditure (GBP) [Includes food, clothing, transport, housing, education, etc.]
`income`	Gross weekly average household income (GBP)
`house.ten`	Household tenure: 1 = Public rented, 2 = Private rented, 3 = Owned
`sex.hh`	Sex of the household head: 1 = Male, 2 = Female
`lab.force`	Employment status: 1 = Full time working, 2 = Part time working, 3 = Unemployed, 4 = Economically inactive
`hh.size`	Household size: 1 = 1 person, 2 = 2 persons, 3 = 3 persons, 4 = 4 persons, 5 = 5 persons or more
`hh.adults`	Number of adults in the household: 1 = 1 adult, 2 = 2 adults, 3 = 3 adults, 4 = 4 adults or more

1. Produce and briefly discuss appropriate tables or plots to assess the distribution of `expenditure` and the relationship between `expenditure` and `income`, `house.ten`, `sex.hh`, `lab.force`, `hh.size` and `hh.adults`.

   [10 marks]

2. Regress `expenditure` on `income` and present the estimated coefficients and their standard errors. Assess the regression assumptions using appropriate plots.

   [4 marks]

3. Regress `expenditure` on `income` and `income` squared, and present the estimated coefficients and their standard errors. Assess the regression assumptions using appropriate plots.

   [4 marks]

4. Regress the natural logarithm of `expenditure` on `income`, and present the estimated coefficients and their standard errors. Assess the regression assumptions using appropriate plots.

   [4 marks]

5. Regress the natural logarithm of `expenditure` on `income` and `income` squared, and present the estimated coefficients and their standard errors. Assess the regression assumptions using appropriate plots.

   [4 marks]

6. Which of the above four models best describe the relationship between `expenditure` and `income`? Justify your answer and summarise the relationship between `expenditure` and `income` based on your preferred model.

[4 marks]

7. By considering the addition of the other variables and interactions to your preferred model from question 6, propose a suitable regression model for `expenditure`. Document your model building process and use diagnostic tools to assess the fit of your model.

[18 marks]

8. Describe the relationship between `expenditure` and the explanatory variables in your model.

[12 marks]

9. Up to 5 marks will be allocated for general presentation of the results in the report.

[5 marks]

Task 2 [Total 35 marks, max. 2 pages]

For this task you need to (a) submit R code using the R template, which will be used to replicate your answers, and (b) include the answers to the questions below in your report. **You are not allowed to use existing R functions that fit models.** However, you are allowed to use other R functions, for example, those required for matrix algebra computations.

The dataset for this task includes data on the number of days ahead travellers purchase their airline tickets (y) and the distance in kilometres they plan to travel (x). The data file (available on Blackboard) is called `airline.txt` and contains 1000 records on the two variables y and x. One way to model the number of days ahead travellers purchase their airline tickets is by using a distribution with probability density function (p.d.f.) given in (1), where θ denotes the parameter of interest:

$$f(y;\theta) = \theta e^{-\theta y}, \quad y > 0, \quad \theta > 0. \tag{1}$$

The p.d.f. in (1) is a member of the exponential family with the following components (using the same notation as in the lecture notes):

$$b'(\theta) = -1; \qquad b''(\theta) = 0; \qquad c'(\theta) = \frac{1}{\theta}; \qquad c''(\theta) = \frac{-1}{\theta^2}; \qquad \theta = \frac{1}{\mu}.$$

Let $Y_1, \ldots, Y_n$ be independent random variables from (1) and assume that the mean number of days, μ_i, can be modelled as a function of distance x_i using the following link function and systematic component, $\log \mu_i = \beta_0 + \beta_1 x_i$, with $\mu_i = E(Y_i)$, for $i = 1, \ldots, n$.

1. Use the expressions provided in the lecture notes and the information above to obtain the score $\boldsymbol{u}(\boldsymbol{\beta})$ and the information $\boldsymbol{I}(\boldsymbol{\beta})$, where $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$, under the link function and systematic component specified above. Present your derivations and report the score and information.

   [10 marks]

2. Using the score and the information, write R code that implements the Fisher scoring algorithm to fit a glm to dataset `airline.txt` under the distribution specified in (1) with the link function and systematic component specified above. Obtain the maximum likelihood estimate (m.l.e.) of $\boldsymbol{\beta} = (\beta_0, \beta_1)$. Report the point estimates of the model parameters.

   [10 marks]

3. Calculate and report the variance-covariance matrix of the m.l.e. of $\boldsymbol{\beta}$.

   [5 marks]

4. Compute the t-statistic for testing the significance of β_1. Is the variable significant? Justify your answer.

   [5 marks]

5. Present well-structured code with brief comments and a concise but informative report.

   [5 marks]