MATH4/68052 (GLM's and Survival Analysis) Coursework 2023-24

The marks awarded for this coursework constitute 20% of the total assessment for the module.

Your solutions to the coursework should be no more than several pages long, including text, plots, `R` output and the code used. It should take, on average, around 10 hours to finish. You are advised to complete all the computing parts first before typing your submission document.

Please read all the instructions and advice given below carefully.

The submission deadline is 11:00 am on Friday 22 March 2024.

Late Submission of Work: Any student's work that is submitted after the given deadline will be classed as late, unless an extension has already been agreed via mitigating circumstances or a DASS extension
The following rules for the application of penalties for late submission are quoted from the University guidance on late submission document, version 1.3 (dated July 2019):

"Any work submitted at any time within the first 24 hours following the published submission deadline will receive a penalty of 10% of the maximum amount of marks available. Any work submitted at any time between 24 hours and up to 48 hours late will receive a deduction of 20% of the marks available, and so on, at the rate of an additional 10% of available marks deducted per 24 hours, until the assignment is submitted or no marks remain."

Your submitted solutions should all be in one typed document. This should preferably be prepared using LaTeX or R Markdown. Word is also permissible. For each question you should provide explanations as to how you completed what is required, show your workings and also comment on computational results, where applicable.

When you include a plot, be sure to give it a title and label the axes correctly.

When you have written or used `R` code to answer any of the parts, then you should list this `R` code after the particular written answer to which it applies. This may be the `R` code for a function you have written and/or code you have used to produce numerical results, plots and tables. `R` code should also be clearly annotated.

Do not use screenshots of `R` code/output. Instead, to include `R` code use the `verbatim` environment.

Your file should be submitted through the module site on Blackboard to the Turnitin assessment under Assesment & Feedback entitled 'Coursework (March 2024)' by the above time and date. The work will be marked anonymously on Blackboard so please ensure that your filename is clear but that it does not contain your name and student id number. Similarly, do not include your name and id number in the document itself.

Turnitin will generate a similarity report for your submitted document and indicate matches to other sources, including billions of internet documents (both live and archived), a subscription repository of periodicals, journals and publications, as well as submissions from other students.

Please ensure that the document you upload represents your own work and is written in your own words. The Turnitin report will be available for you to see shortly after the due date.

This coursework should hopefully help to reinforce some of the methodology you have been studying, as well as skills in `R`.

The data we will be using for this coursework comprise $n = 712$ records of the passengers sailing on RMS Titanic that sank in the North Atlantic Ocean on 15 April 1912 after hitting an iceberg. The estimated total number of passengers and crew on board was 2224. The `titanic.df` data frame for this coursework contains the following variables:

- `Survived`: 1 = yes, 0 = no
- `Pclass`: Passenger class - 1 (1st), 2 (2nd), 3 (3rd)
- `Sex`: 'male', 'female'
- `Age`: 1 = child (under 18), 2 = adult (18 to 60), 3 = senior (over 60)
- `Parch`: Number of parents and/or children on board for a passenger
- `Embarked`: Port of embarkation - C = Cherbourg, S = Southampton, Q = Queenstown

`Survived` is a binary response variable and the exercise will be to look at logistic regression models which can be used to predict the probability that a passenger survived, given their particular set of covariate values.

`Parch` is to be regarded as a numeric variable, while each of the others are factors. The code in the `R` script file `start.r` can be run to load the data and convert the relevant variables to factors in `R` using the constraint that the first (reference) level is set to zero.

1. Write down the full additive (but no interactions) logistic regression model with a logit link and explain the notation you have used, including the terms in the linear predictor. Fit this model to the data. Present the `R` 'summary' of your fitted model. Explain and comment on the individual Z-tests of the hypotheses that the true parameter values are equal to zero.

   [5 marks]

2. Fit a reduced model which excludes the variables `Parch` and `Embarked`. Perform an analysis of deviance to show that these two variables do not make a significant contribution to the fit.

   [1 mark]

   Can this model, now just containing the variables `Pclass`, `Sex`, and `Age`, be reduced any further? Provide statistical evidence for your answer.

   [3 marks]

The following questions are all just based on using the fitted model which includes the three covariates `Pclass`, `Sex`, and `Age`.

3. Calculate the values for a new binary variable in `R` called `pred.surv` whose i'th element is equal to 1 if $\hat{p}(x_i) > 0.5$ and equal to 0 if $\hat{p}(x_i) \leq 0.5$. Here, $\hat{p}(x_i)$ denotes the estimated probability of survival for the i'th sample case who has a vector of covariates x_i.

   [1 mark]

   Tabulate the values in `Survived` against `pred.surv` and calculate the proportion of sample cases correctly classified by the model. Briefly comment on the result.

   [2 marks]

4. (i) Estimate the odds of survival for an adult female travelling 2'nd class. Briefly comment on the result.

[1 mark]

(ii) Estimate the odds ratio of survival for a adult female travelling 1'st class to a senior male travelling 3'rd class. Briefly comment on the result.

[2 marks]

5. Estimate the probability of survival for an adult female travelling 2'nd class and find an approximate 95% confidence interval for the true value.

[5 marks]

[The total marks for all the parts is 20]