MA318-7-AU

Department of Mathematical Sciences

Class test 7 December 2021

STATISTICAL METHODS

Time allowed: 1 hour

You have been given an additional 30 minutes in which to download the test paper and upload your answers to FASER. You must upload your answers before the deadline shown in FASER. Answers uploaded after that time will receive a mark of 0.

If you are normally allowed additional writing time your deadline has been adjusted accordingly.

The deadline times are shown in Greenwich Mean Time (GMT). Please check online for a conversion to your local time if you will be undertaking your assessment outside the United Kingdom

Number of questions: 2

Candidates must attempt ALL questions.

Contacts
If you believe there is an error in this paper or you experience any other difficulties please email your module lecturer Shenggang Hu (sh19509@essex.ac.uk.)

Extenuating circumstances
We know that students are working under different conditions at the moment. If you encounter any difficulties with your test that you believe has affected your performance you can inform the Exam Board of this by submitting an Extenuating Circumstances form https://www1.essex.ac.uk/students/exams-and-coursework/ext-circ.aspx

Academic Offences
You are reminded that copying from textbooks, articles or any online or offline sources is plagiarism, which is an academic offence. Your work must be entirely your own and collaborating with other students is collusion, which is also an academic offence. Any student suspected of committing an academic offence will be interviewed in person to determine whether the work they have submitted is entirely their own. More information about academic offences can be found at https://www.essex.ac.uk/student/exams-and-coursework/about-academic-offences

Additional instructions on completing and submitting your test answers continued on page 2.

Format of your answers

- Write your registration number clearly at the top of the first page. Do not add your name as these will be marked anonymously.
- Answers should be hand written
- Write in pen – preferably black
- Number your answers clearly
- We recommend you start a new page for each question.
- Number each page to help you check you have scanned every page
- Do not write on the reverse of the page as this can show through when you scan your answers and can affect legibility.

Scanning and uploading

- If you don't have a scanner there are also lots of **apps** you can use on your phone to scan your work into a PDF format, for example Microsoft Office Lens or Scannable – but you can use whichever one suits you. Later versions of the iPhone also have a scanning facility.
- **You should combine your pages into one continuous PDF document. Do NOT submit photographs of each individual page.**
- Pages should be scanned in portrait orientation.
- Save your document as your registration number.
- Check your final document before uploading to FASER and ensure you have captured all pages and it's legible. If the marker cannot read your answer they will be unable to mark it.
- Do not leave it until close to the deadline to upload your answers – late submissions will not be accepted.
- We recommend you check your submission after you have uploaded to ensure you have uploaded the correct document, you have scanned all your pages and it hasn't been corrupted, and retain your email receipt from FASER.
- We do not anticipate any problems with FASER but if you do experience any difficulties uploading your answers, you can email them to maths@essex.ac.uk with [module code] in the subject. You will still need to upload to FASER but you can do this when you can access the system. This is just a back-up, please do not email your answers if you are able to submit to FASER.
- Only your latest submission (within the deadline) will be marked.

MA318 R end term test

(for Postgraduate students)

Please read the following instructions carefully before start answering the questions:

- Please answer ALL TWO questions.
- Please write your code in the R code template or the R markdown template. After you have done all your questions, you should run your R code and save R outputs (in cosole or in R markdown) in the order they are produced by copy and pasting them into a separate Word file.
- Plots and Diagrams should be saved in PNG formate and uploaded as separate files. **Make sure you name your saved figures using the question number!!**
- Upload your file (or files) to FASER.
- Please do not include your name anywhere in your answers.
- There are 100 marks in total

Before you start the questions, make sure you have installed and loaded the following packages in R/RStudio:

```
library(boot)
```

Question 1:

In "binary.csv", we have an admission data of a university where 400 individuals were grouped as 0 "not admit" or 1 "admit". And the goal is to model and predict if a given individual is "admit" or "not admit", based on 3 other features, "gre", "gpa" and "rank". Rank means the prestige of the institutions that individual attended and takes values 1 through 4. Institutions with a rank of 1 have the highest prestige, while those with a rank of 4 have the lowest. Treat "rank" as a factor variable and do the following questions:

1) Load the `binary` dataset into R. Set "rank" as a factor variable.

[**5 marks**]

2) Load `boot` libarary to allow cross-validation. Conduct a cross-validation on the data set by doing the following steps:
   (i) Using your last three digits of your registration number as the random seed, split the data set into training and test set. Your training set should contain 280 randomly selected observations from the 'binary' data set.

   [**5 marks**]
   (ii) Taking 'admit' as the response variable and all other columns as features, fit a logistic model on the training set.

   [**5 marks**]
   (iii) Define a cost function to compute the classification error rate for the cross-validation pipeline. Use 0.3 as the prediction threshold, i.e. classify as 'admit' if the predicted probability is > 0.3.

   [**5 marks**]
   (iv) Compute the 10-fold coss-validation error of your model on the training set, with threshold being 0.3.

   [**5 marks**]
   (v) Define another cost function that uses 0.5 as the threshold instead. Compute the 10-fold cv-error using the new cost function.

   [**5 marks**]
3) Using the fitted model in 2), make predictions using 0.3 and 0.5 as threshold respectively on the test set and compute the test classification error rate. Note that the `admit` columns are directly 0,1 values.

[**10 marks**]

4) Using the predictions in 3) and `table()` function, create the confusion matrices on the test set. You should create two matrices, one corresponds to 0.3 as threshold and another using 0.5 as threshold.

[**10 marks**]

Question 2

The `Awards.csv` dataset contains data about the number of awards received by individuals with their level of education received and their final exam result. The number of awards received by an individual is believed to be linked to their education history. The dataset contains three columns, `num_awards` stores the number of awards received by an individual, `prog` denotes the type of education program with levels 'General', 'Academic' and 'Vocational' and `final` stores the level of their final score in the program which are either 'Pass' or 'Distinction'. Use `R` to do the following questions:

1) Download the `Awards.csv` dataset and load it into R. Report the number of observations.

[5 marks]

2) Fit the Poisson regression model to investigate the associations of number of awards with the type of education programs and final scores. Print the summary of your model and identify which covariate(s) is(are) statistically significantly associated with number of accidents under 5% significance level.

[15 marks]

3) Calculate the expected average number of awards received by someone who
   (i) got a **Pass** in the **General** program;
   (ii) got a **Distinction** in the **General** program;
   (iii) got a **Pass** in the **Academic** program;
   (iv) got a **Distinction** in the **Academic** program;
   (v) got a **Pass** in the **Vocational** program;
   (vi) got a **Distinction** in the **Vocational** program;
   On average which type of people are expected to receive the most number of awards? which gives the lowest number of awrads?

[20 marks]

4) Comment on if the fitted model in part 2) is significantly better than the NULL model under 5% significance level, based on the deviance value of the models.

[10 marks]

END OF PAPER