THE UNIVERSITY OF MELBOURNE

Student number

Semester 2 Assessment, 2024 (MOCK EXAM)

School of Mathematics and Statistics

MAST90104 A First Course in Statistical Learning

Reading time: 5 minutes — Writing time: 45 minutes

This test consists of 6 pages (including this page)

Permitted Materials

- Printed course material (lecture notes, lab/workshop sheets, solutions) are permitted.
- A Casio FX-82 calculator is permitted.
- Any mobile phones or internet-enabled devices brought into the test room must be **turned off** and placed on the floor under your table.

Instructions to Students

- You may use an off-line PDF reader on your laptop or tablet for your notes, but it must be set in flight mode or have both internet and Bluetooth disabled. Copying or saving materials to the lab's computer to use during the test are not allowed.
- Give answers to 4 decimal places.
- At the end of the test, please remain in your seat while your paper is being collected. Save your `R` code as a single script (.R or .Rmd) and submit through the LMS in the Computer Laboratory Test item.
- Write your answers in the boxes provided on the test. There is extra space you can use for answers to any question commencing on page 6. If you still do not have enough space, tick the box near the bottom of page 6 and request extra pages from an invigilator—include the question number at the top of each additional page.
- You must NOT remove this question paper, or any extra pages provided to you, at the conclusion of the examination.

Instructions to Invigilators

- Students are to write their answers on the paper. They may request extra pages if they run out of space on the test paper.
- This test paper contains examinable material and must be collected, together with extra pages if any, at the conclusion of the examination.
- Students are not allowed to leave during the test or submit early.

Blank page

Question 1 (18 marks)
In this question, you are required to generate synthetic datasets based on a logistic regression model.
Include each part's R code in the script submitted through the LMS

If correct answer then full mark. If the answer is wrong, we will check your codes. No answer - 0pt

(a) Write a function named `GenerateLogit` in R with argument n that generates independent draws $\{y_i\}_{i=1}^n$, where each y_i is drawn as:

$$y_i \sim \text{Bin}(1, g^{-1}(-0.15 + 0.005 \times i))$$

and g is the logit link function. By setting seed equals to 1, run your function once with $n = 20$ and report the numerical value of the statistic $\sum_{i=1}^n y_i$ and sample variance of $\{y_1, \ldots, y_n\}$.

$\sum_{i=1}^n y_i = 11$
sample var$\{y_1, \ldots, y_n\} = 0.2605$

(b) Using your generated data from (a), fit the following Binomial logistic regression model

$$y_i \sim \text{Bin}(1, g^{-1}(\theta_0 + \theta_1 \times i)).$$

Report your Wald's 90% confidence interval for θ_1.

Wald's estimate for θ_1 is $\widehat{\theta}_1 = 0.0274$
Wald's SE estimate for θ_1 is $SE(\widehat{\theta}_1) = 0.0784$
A 90% CI for θ_1 is $\widehat{\theta}_1 \pm z_{0.05} SE(\widehat{\theta}_1) = (-0.1015, 0.1564)$

(c) Write a function named `RepeatedCI` in R with two arguments n and `M.reps` that executes a for-loop with `M.reps` iterations. At iteration $i = 1, \ldots$ `M.reps`, `RepeatedCI` sets the seed at i, and then calls `GenerateLogit` with n as argument to generate data $\mathbf{y}_i$. The output data $\mathbf{y}_i$ is then used to construct a 90% confidence interval for θ_1. Run your function with $n = 35$ and `M.reps` $= 5$. Report the proportion of times that the confidence interval contains the number 0.005.

Proportion of times $= 0.8000$

(d) Refer to part (c). Re-run your function `RepeatedCI` with $n = 35$ and `M.reps` $= 20, 50, 500, 5000$. Report the proportion of times that the confidence interval contains the number 0.005 for each value of `M.reps`. What is the limiting proportion of confidence intervals that contain 0.005 as `M.reps` $\to \infty$?

Proportion of times for `M.reps` = 20 is 0.8500
Proportion of times for `M.reps` = 50 is 0.8200
Proportion of times for `M.reps` = 500 is 0.8900
Proportion of times for `M.reps` = 5000 is 0.9016
Limiting proportion is 0.9000 (By Strong law of Large Numbers)

(e) Refer to part (d). Answer this part without re-running your function `RepeatedCI`. Fix `M.reps` $= 300$. What is the limiting proportion of confidence intervals that contain -0.1 as $n \to \infty$?

Limiting proportion is 0 (By Strong law of Large Numbers)

Question 2 (16 marks)
The dataset, swiss, is available in R and records standardized fertility measure and socio-economic indicators for each of 47 French-speaking provinces of Switzerland at about 1888. The variables are

- Fertility: a common standardized fertility measure
- Agriculture: % of males involved in agriculture as occupation
- Examination: % draftees receiving highest mark on army examination
- Education: % education beyond primary school for draftees
- Catholic: % 'catholic' (as opposed to 'protestant')
- Infant.Mortality: live births who live less than 1 year.

You can load the data in R by running: `data(swiss)`
Include each part's R code in the script submitted through the LMS.

If correct answer then full mark. If the answer is wrong, we will check your codes. No answer - 0pt

(a) Fit a linear model to predict fertility using all of the other variables, store it and find the coefficients. What is the estimated coefficient of `Education`? What is the regression sum of squares of this model?

The estimated coefficient is -0.8709
Regression sum of squares is 236311.9

(b) Starting with the full model, perform backward selection using the F-statistic criterion and store the selected model. Write down the sequence in which the variables were excluded from your final model. Also, write down your final model.

Variables were excluded in this sequence: `Examination`
The final model is

$$\texttt{Fertility} = \beta_0 + \beta_1 \texttt{Education} + \beta_2 \texttt{Catholic} + \beta_3 \texttt{Infant.Mortality} + \beta_4 \texttt{Agriculture} + \epsilon.$$

(c) Compare the model in part(a) and part(b) using F test. Report the test statistic and p-value. What can you conclude from the test?

Test statistic is 1.0328 , p-value is 0.3155.
There is no evidence that the full model fits the data better than the model selected in (b).

Additional writing space for any question commences on the next page

Additional answer space for any question—submit this page even if blank

End of Test

You must tick this box if you have used extra pages ☐