Queen Mary
University of London

Main Examination period 2023 – May/June – Semester B

MTH791U / MTH791P: Computational Statistics with R

Duration: 3 hours

The exam is intended to be completed within **3 hours**. However, you will have a period of **4 hours** to complete the exam and submit your solutions.

You should attempt ALL questions. Marks available are shown next to the questions.

All work should be **handwritten** and should **include your student number**. Only one attempt is allowed – **once you have submitted your work, it is final**.

In completing this assessment:

- You may use books and notes.
- You may use calculators and computers, but you must show your working for any calculations you do.
- You may use the Internet as a resource, but not to ask for the solution to an exam question or to copy any solution you find.
- You must not seek or obtain help from anyone else.

When you have finished:

- scan your work, convert it to a **single PDF file**, and submit this file using the tool below the link to the exam;
- e-mail a copy to **maths@qmul.ac.uk** with your student number and the module code in the subject line;

Examiners: M. Iacopini, A. Shestopaloff

Question 1 [22 marks].

(a) Suppose you are given two samples, $(x_1, \ldots, x_m)$ and $(y_1, \ldots, y_n)$ and you are interested in testing if the mean in the population associated with the first sample is different from the population mean for the second sample.
In R, the function `dwilcox` calculates the probability mass function of the null distribution for the Mann-Whitney statistic, U_X. Suppose that this function is run and outputs the following:

```
> dwilcox(0:3, m=2, n=3)
[1] 0.1 0.1 0.2 0.2
```

First, state what is the set of possible values that U_X can take. Then use the R output above to recover the full null distribution of U_X, then calculate the one-sided probability $P(U_X \geq 5)$. **[8]**

(b) Consider performing a Mann-Whitney rank test using a normal approximation. Discuss the advantages and problems of this choice. **[14]**

Question 2 [14 marks]. Suppose that you have observed the following data

$$(x_1, y_1), (x_2, y_2), (x_3, y_3) = (10.1, 2.3), (5.7, 6.3), (8.2, 9.0)$$

where you only know that (x_h, y_h) is independent of (x_j, y_j) for each $h \neq j$. You are interested in finding out if the distribution of the difference, $X - Y$, is symmetric about 0.

(a) Without making computations, describe an appropriate permutation test to test this hypothesis. Mention the type of dataset at hand, the null hypothesis, and define the test statistic. **[6]**

(b) Describe **all** the possible methods you can use to carry out the test, discussing the advantages of each one of them. **[8]**

Question 3 [11 marks]. Consider the problem of estimating a univariate probability density function using a kernel density estimator (KDE) or a histogram.

(a) What are the components of a KDE? Which of them, in general, has the most influence on the appearance of the estimated probability density function? **[4]**

(b) State one advantage of using a KDE over a histogram for estimating a probability density function. **[3]**

(c) For a given sample size $n > 0$, how do the bias and variance of a histogram estimator at a single point change as the bin width is made smaller? **[4]**

Question 4 [10 marks]. Consider the following function:

$$K(x) = \begin{cases} x+1 & -1 < x < 1 \\ 0 & \text{elsewhere} \end{cases}$$

It this a valid kernel? Motivate your answer.

Question 5 [17 marks]. You are given a dataset consisting of pairs (x_i, y_i), with $i = 1, \ldots, n$, where each pair (x_i, y_i) is independent from (x_j, y_j) for any $i \neq j$. You are interested in computing the standard error of the estimator

$$\hat{\theta} = \sum_{i=1}^{n} (x_i - y_i)$$

Then:

(a) Explain in your own words how to generate a nonparametric bootstrap replication for $\hat{\theta}$. **[11]**

(b) Suppose a dataset of size $n = 5$ is collected and some researchers (say, R1, and R2) design a bootstrap procedure to compute the standard error of $\hat{\theta}$, as follows:

- R1 obtains the standard error $\widehat{se}_B(\hat{\theta}) = 0.23$ using a non-parametric bootstrap method with $B = 1000$ bootstrap replications;
- R2 based on a parametric bootstrap with $B = 599$ replications, claims that $\widehat{se}_B(\hat{\theta}) = 0.12$.

Which one of them do you believe is following the correct approach (if any)? Motivate your answer. **[6]**

Question 6 [26 marks]. Suppose a dataset $(\mathbf{x}, \mathbf{y}) = ((x_1, y_1), \ldots, (x_n, y_n))$ is available and two different linear models, say Model A and Model B, may be used to analyse them. The models are then fitted to the dataset and a leave-one-out cross-validation procedure is performed.

(a) Define the $PRESS$ statistic and explain what is it useful for. **[8]**

(b) Suppose the sum of squared errors, SSE, and the $PRESS$ statistic computed for each model are such that

$$SSE_A < SSE_B, \qquad PRESS_A > PRESS_B$$

where the SSE_A represents the SSE of model A (analogous notation for the other cases). According to this result, which is the best fitting model? Motivate your answer. **[9]**

(c) Provide an explanation of a possible reason why the relationships $SEE_A < SSE_B$ and $PRESS_A > PRESS_B$ are not contradicting each other. **[9]**

End of Paper.