HERIOT-WATT UNIVERSITY

F79MB Statistical Models B **Assessment Project 1 – 2025**

Your report for this project should be submitted through Turnitin by **3:30pm (local time at your campus) on Monday 24th February 2025**. All required data set and a link to the submission page are available through "Assessed Project 1" in Modules on the course Canvas page. Please use the submission link provided to submit your work.

This project will account for **40%** of your grade for the course.

The total length of your report should be no more than 7 pages (11-point font, A4 size), including graphs and tables but excluding any appendices. Project title and student identification information should be in page 1, not a separate title page, and counts towards the 7-page limit. **You must complete the Standard Declaration of Student Authorship quiz on Canvas before you can access the Project 1 questions.**

Questions for the Lecturers: If you have any questions on the assignment, please ask in the discussion board rather than e-mailing lecturers directly. That way all students will see the same responses. The discussion board will be closed at 10:00 am UK time = 1:00 pm UAE time = 5:00 pm Malaysia time on Sunday, 23rd February 2025.

Students are allowed to discuss the methods used with other students, but your submitted project must be all your own work. **Plagiarism** is a serious academic offense and carries a range of penalties, some very serious – see Appendix A.

- Answer all tasks in this project, making sure that you add appropriate context and that your answers have a clear and logical structure and are well presented.
- You should explain carefully your work for each task so that your work demonstrates understanding of the methodology and computations that you use.
- You should include clearly labelled and correctly referenced graphs and add appropriate comments.
- You should use `R` to perform the required analyses and produce suitable graphs. Unless otherwise stated you do not need to explain the `R` commands that you use. However you must put the `R` code in appropriate appendices at the end of your report, and should include appropriate commenting within your `R` code.
- See Appendix B for the marks allocated to overall exposition/presentation.
- **Late project submissions will be penalised according to the University Policy on Submission of Coursework. That is, work submitted after the deadline but within 5 working days will be subject to a 30% deduction from the mark awarded; work submitted more than 5 working days after the deadline will be awarded a mark of zero. No individual extensions are permitted. In the case where a student submits coursework up to five working days late and has valid mitigating circumstances, the mitigating circumstances policy will apply.**

  You can expect to receive feedback on your work by 17th March 2025.

Tasks:

The `charges.txt` file available on Canvas under `Modules > Assessed Project 1`, contains data on a sample of 149 individual medical costs (in USD) billed by private hospitals across various regions of the United States for the year 2024.

1. Present appropriate numerical and graphical summaries, and comment on the distribution of the data.

   [2 marks]

2. Use QQ plots to explore whether (i) the original data can be adequately modelled by an Exponential distribution with parameter λ estimated from MLE, or (ii) the (natural) logarithm of the data can be reasonably modelled by a Normal distribution with mean $\mu = 9$ and variance $\sigma^2 = 1$.

   [4 marks]

3. Perform a chi-squared goodness-of-fit test to formally assess whether (i) the original data can be reasonably modelled using an Exponential distribution, and (ii) the (natural) logarithm of the data can be reasonably modelled by a Normal distribution with mean $\mu = 9$ and variance $\sigma^2 = 1$.

   You should use 9 cells for your testing procedure, beginning from zero, ensuring minimal arbitrariness in the process.

   [6 marks]

   Overall, what would your recommendation regarding of an appropriate model for the observed data?

   [2 marks]

4. For both (i) and (ii) in question (3), determine whether the Kolmogorov-Smirnov (KS) test can be applied, and justify your reasoning. If the KS test is applicable, compare its results with those from the chi-squared goodness-of-fit test conducted in part (3).

   [2 marks]

   Assume the sample size is reduced from 149 to 30 data points. Discuss the impact of this reduced sample size on both the KS test and the chi-squared goodness-of-fit test.

   [2 marks]

5. Use a non-parametric bootstrap methodology to obtain the empirical sampling distribution of the sample median, $\hat{m}$, and present the distribution graphically with comments. Compute a non-parametric bootstrap 95% confidence interval for the population median medical costs across the United States for the year 2024.

   [3 marks]

6. Under the assumption that the (natural) logarithm of the medical costs data can be reasonably modelled by a Normal distribution with mean $\mu = 9$ and variance $\sigma^2 = 1$, use asymptotic theory to compute 95% confidence interval for the population median medical costs.

Comment on the validity of the two confidence intervals obtained using asymptotic theory and non-parametric bootstrap in part (5).

[4 marks]

7. Under the assumption that the (original) medical costs data can be reasonably modelled by an Exponential distribution, perform a parametric bootstrap hypothesis test to assess whether there is evidence that the median medical costs is less than \$8,500.

[3 marks]

8. Summarize your overall conclusions.

[4 marks]

[Overall exposition/presentation: 8 marks]

[Project total: 40 marks]

[END OF PROJECT]

Appendix A: Plagiarism

- Coursework reports must be written in a student's own words and any code in their coursework must be their own code. If some text or code in the coursework has been taken from other sources, these sources must be properly referenced.
- Failure to reference work that has been obtained from other sources or to copy the words and/or code of another student is plagiarism and if detected, this will be reported to the School's Discipline Committee. If a student is found guilty of plagiarism, the penalty could involve voiding the course.
- Students must never give hard or soft (electronic) copies of their coursework reports or code to another student. Students must always refuse any request from another student for a copy of their report and/or code.
- Sharing a coursework report and/or code with another student is collusion, and if detected, this will be reported to the School's Discipline Committee. If found guilty of collusion, the penalty could involve voiding the course.
- See `https://www.hw.ac.uk/uk/students/studies/examinations/plagiarism.htm` for more details.

Appendix B: Rubric for marks allocated to overall exposition/presentation

The 8 marks available for the exposition/overall presentation of your report will be awarded according to the scale below.

2 marks	• Some answers lack of clarity and logical structure • Some analyses may not relevant • Statements of conclusions are present but may not be very clear • Some statistical calculations and methodology set out for the reader • Some tables and figures relevant and referenced
4 marks	• Answers sometimes clearly and logically structured • Focus on key points but may include some superfluous/irrelevant analyses • Statements of conclusions generally suitable for a non-statistician • Statistical calculations and methodology are generally presented okay for the reader • Tables and figures mostly well chosen, clear, and correctly referenced • Some sources are clearly and correctly referenced, with R code included
6 marks	• Answers are mostly clear and logically structured • Clear focus on key points, avoiding superfluous/irrelevant analyses • Statements of conclusions suitable (wherever possible) for a non-statistician • Statistical calculations and methodology set out clearly for the reader • Tables and figures well chosen, clear, correctly referenced and easy to interpret • Sources used clearly and correctly referenced • R code included with comments
8 marks	• As in the previous category, but the answers are very clear, strong, and logically structured. Additionally, the work demonstrates mathematical sophistication and insight, with evidence of original thought and reasoning. It also maintains a highly focused approach on key points, without superfluous or irrelevant analyses. R code is included with comments and without superfluous coding.