HERIOT-WATT UNIVERSITY

F79MB Statistical Models B **Assessment Project 2 – 2025**

Your report for this project should be submitted through Turnitin by **3:30pm (local time at your campus) on Thursday 10th April 2025**. All required data set and a link to the submission page are available through "Assessed Project 2" in Modules on the course Canvas page. Please use the submission link provided to submit your work.

This project will account for **60%** of your grade for the course.

The total length of your report should be no more than 14 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 14-page limit. **You must complete the Standard Declaration of Student Authorship quiz on Canvas before you can access the Project 2 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 Tuesday, 8th April 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 7th May 2025.

Tasks:

1. The dataset `vehicle.csv`, available in the Assessed Project 2 folder on Canvas, contains a sample of 200 observations on various vehicle models collected in 2022. You can load the data into `R` using a command like `read.csv("C:/R/vehicle.csv", header=TRUE)`. Note that you will need to specify your own path to the file. Several variables were recorded for each vehicle, but the dataset retains the following:

   - `brand`: Vehicle brand (Toyota, Honda, Ford, Chevrolet, Nissan).
   - `age`: Vehicle age (as of 2022).
   - `fueltype`: Fuel type (diesel, gas, other).
   - `price`: Vehicle price in USD.
   - `type`: Vehicle type (pickup, sedan, SUV).

   (a) Present numerical and graphical summaries to explore the distributions of each variable: `brand`, `age`, `fueltype`, `price`, and `type`. Discuss any interesting findings or trends observed from these summaries.

   [4 marks]

   (b) Produce visualizations to show the relationships between vehicle price (`price`) and each explanatory variable: `brand`, `age`, `fueltype`, and `type`. Comment on any trends or associations you observe.

   [3 marks]

   (c) Compute a 99% confidence interval for the Pearson correlation between `price` and `age` using Fisher's transformation. Comment on your results, including any concerns regarding the validity or reliability of the confidence interval.

   [3 marks]

   Suggest an alternative approach to compute the confidence interval and justify your choice.

   [1 mark]

   (d) It is proposed to fit a multiple linear regression model with `price` as the response variable and `age`, `type`, `fueltype`, and `brand` as explanatory variables, including the interaction between `age` and `type`, but excluding interactions with `fueltype` and `brand`. The proposed model is specified as follows:

   $$\texttt{price} \sim \texttt{age} * \texttt{type} + \texttt{fueltype} + \texttt{brand}$$

   Carry out the appropriate analyses for this model to determine which terms should be retained. Identify any influential data points and thoroughly assess the impact of removing them from the analysis. Your report should include relevant R outputs, plots, comments, model diagnostics, justification for removing the influential points, comprehensive conclusions, and suggestions for further refinement of the fitted model.

   [14 marks]

2. The dataset in the `Airline.csv` file, available in the Assessed Project 2 folder on Canvas, contains a sample of 95 airline passenger satisfaction responses from 2024 for a UK-based airline company. This study aims to forecast the likelihood of passenger satisfaction using an appropriate logistic regression model.

   You can load the data into `R` using a command like `read.table("C:/R/Airline.csv", header=T, sep=",")`. Note that you will need to specify your own path to the file. The variables given are:

   `Gender` = Passenger's gender (Female, Male).

   `Class` = Travel class (Business, Economy, Economy Plus).

   `Distance` = Travel distance in kilometers (km).

   `Delay` = Flight departure delay in minutes.

   `Satisfaction` = Passenger satisfaction (0 = Neutral or Dissatisfied, 1 = Satisfied).

   Note that `Gender`, `Class` and `Satisfaction` are categorical variables.

   It is proposed to model the data in a generalized linear model framework, modelling `Satisfaction` as the response variable with `Gender`, `Class`, `Distance` and `Delay` as explanatory variables.

   (a) Produce appropriate plots to explore the relationship between the response variable, `Satisfaction`, and the explanatory variables `Class` and `Distance`. Comment on your results for each plot.

   [3 marks]

   (b) Carry out appropriate tests of independence to determine whether the satisfaction status (`Satisfaction`) for the flight is associated with gender (`Gender`) and travel class (`Class`). Your report should include contingency tables, justification for the tests, p-values and clear conclusions.

   [6 marks]

   (c) Analyse the data by fitting a generalized linear model with `Satisfaction` as the (Binomial) response variable and `Gender`, `Class`, `Distance`, and `Delay` as explanatory variables (with no interaction term). Carry out appropriate analyses to determine which terms should be retained in the model. Your analysis should include relevant R output, plots, comments, model checking and comprehensive conclusions.

   [12 marks]

   (d) Based on the first 10 predicted probabilities, comment on the accuracy of the preferred fitted model's predictions as discussed in part 2(c). Additionally, discuss potential improvements or enhancements that could be made to the model.

   [4 marks]

[Overall presentation: 10 marks]

[Project total: 60 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 10 marks available for the overall exposition/presentation of your report will be awarded according to the scale below.

3 marks	• Some answers lack clarity and logical structure • Some analyses may not be 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
5 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
8 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 clear comments
10 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 clear comments and without superfluous coding.