HERIOT-WATT UNIVERSITY

F79MB Statistical Models B **Assessment Project 2 – 2020**

Your report for this project should be submitted through Turnitin by **submission date on Vision**. A link to the submission page is available through the Assessment section of the course Vision page. Please use the submission link appropriate for the campus where you are studying (Edinburgh or Malaysia).

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

The total length of your report should be no more than 12 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 12-page limit.

You may consult with me or your colleagues but your report must be your own work. **Plagiarism** is a serious academic offence 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.**

Tasks:

1. The data set in the `OECD_Pensions.txt` file, available on Vision under `Learning Materials > Data Sets`, consists of data relating to autonomous pension schemes in each of the 36 countries of the Organisation for Economic Co-operation and Development (OECD) for the year 2018. The recorded variables are as follows.

   - `Country`: name of country
   - `Benefits`: total amount paid out in benefits (millions of US dollars)
   - `Assets`: total amount held in assets (millions of US dollars)

   (a) It is proposed to model the relationship between `Benefits` and `Assets` by fitting a linear regression model with `Benefits` as the response variable and `Assets` as the explanatory variable. Fit the proposed linear regression model. Your analysis should include appropriate `R` output, plots, comments and model checking. Produce a plot showing the relationship between `Benefits` and `Assets`, with the line of best fit shown, and highlight any unusual observations. [8 marks]

   (b) As an alternative, it is proposed to fit a linear regression model with `ln(Benefits)` as the response variable and `ln(Assets)` as the explanatory variable. Fit this new proposed model. Your analysis should include appropriate `R` output, plots, comments and model checking. Produce a plot showing the relationship between `ln(Benefits)` and `ln(Assets)`, with the line of best fit shown, and highlight any unusual observations.

   [8 marks]

   (c) Of the two models in parts (a) and (b) above, which would you prefer? Explain your answer. [2 marks]

   (d) Perform a chi-squared goodness-of-fit test to assess whether the log-transformed `Assets` values can reasonably be modelled by a normal distribution, using 6 equal-probability cells for your testing procedure. Repeat the procedure, but now using 5 equal-probability cells. Comment on your results. [6 marks]

   (e) Discuss any relationship between your conclusion from part (d) and your conclusions from parts (a-c). [2 marks]

2. The data set in the `USA_Counties.txt` file, available on Vision under `Learning Materials > Data Sets`, consists of economic data for each of 3139 counties in the USA. The recorded variables are as follows.

   - `Medinc`: median household income (US dollars)
   - `Unemp`: unemployment rate (%)
   - `Metro`: an indicator variable, with 0 = non-metropolitan, 1 = metropolitan

   Note that `Metro` is a categorical variable, and should be treated in `R` as a factor.

   It is proposed to fit a multiple linear regression model, with `ln(Medinc)` as the response variable, and `ln(Unemp)`, `Metro` as explanatory variables. Fit the proposed multiple linear regression model, including both explanatory variables together with an interaction term. Carry out appropriate analyses to decide which terms to retain in the model. Your analysis should include appropriate `R` output, plots, comments, model checking and a conclusion. [12 marks]

3. The data set in the `volcanoes.txt` file, available on Vision under `Learning Materials > Data Sets`, records numbers of eruptions during the 20th century of a sample of 30 volcanoes. The recorded variables are as follows.

   - `Name`: name of volcano
   - `Elevation`: height above sea level (metres)
   - `Type`: type of volcano, either `Stratovolcano` or `Caldera`
   - `Eruptions`: number of eruptions

   Note that `Type` is a categorical variable, and should be treated in `R` as a factor.

   It is proposed to model the data in a generalized linear model framework, modelling `Eruptions` as a Poisson response variable with `Elevation` and `Type` as explanatory variables.

   Fit the proposed model, incorporating both of the explanatory variables but no interaction term. Carry out appropriate analyses to determine which terms should be retained in the model. Your analysis should include appropriate R output, plots, comments, model checking and conclusions.

   [10 marks]

[Overall presentation: 12 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.

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

The 12 marks available for the overall presentation of your report will be awarded according to the scale below.

3 marks	• Answers sometimes clearly and logically structured • Analyses mostly relevant • Statements of conclusions present and reasonably clear • Statistical calculations and methodology set out for the reader • Tables and figures relevant and referenced
6 marks	• Answers generally clearly and logically structured • Focus on key points, largely avoiding superfluous/irrelevant analyses • Statements of conclusions generally suitable for a non-statistician • Statistical calculations and methodology generally set out clearly for the reader • Tables and figures well chosen, generally clear, and correctly referenced • R code included with some comments • Sources used clearly and correctly referenced
9 marks	• Answers clearly 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 • R code included with comments • Sources used clearly and correctly referenced
12 marks	• As previous category, but in addition showing mathematical sophistication and insight, with evidence of originality of thought and reasoning