MATH 9102 - Probability and Statistical Inference

Final Assignment

Due 6pm, Sunday 26th May, 2024

40 marks

Submission guidelines:

- You will need to upload only **single** R markdown (.Rmd) file.
- File name of your Rmd file must be **YourName_StudentID_CA4**
- Do not upload given datasets (or use zip).
- Make use of R built in datasets (if mentioned in the question). If you have considered external dataset instead of R built in, upload the dataset without zipping it.
- Use the following statement if installing any package: if(!require(packageName))install.packages("packageName")

General Instructions:

- Read the questions carefully and answer all parts to secure full marks.
- Do not ask for direct solutions. This is part of your assessment.
- Assignment will be penalized if you miss any of the submission guidelines.
- Please complete assignment individually and avoid plagiarism as it will lead to penalties and negatively affect your overall grade.
- Including comments/markup in your code to explain what you did and provide answers to all questions.

OVERVIEW

For this final assignment you are required to conduct and present appropriate statistical tests including at least one multivariate inferential statistical technique covered in this module on a dataset of your choice. You need to select appropriate variables and a research question.

Using a Dimension Reduction technique followed by either Multiple Linear Regression or Logistic Regression:

- Develop hypotheses testable by dimension reduction followed by regression. These should be related to your overall research question.
- Derive and present appropriate statistical evidence.
- Assess the suitability of the dataset for dimension reduction.

- Conduct your dimension reduction.
    - Assess the effectiveness of the dimension reduction.
- If the dimension reduction succeeds, use the outcome of the dimension reduction as part of multiple linear regression or logistic regression model (must include multiple predictors).
- If the dimension reduction does not succeed, identify an alternate mechanism to derive a measure for concept for which you conducted dimension reduction. Use this as part of a multiple linear regression or logistic regression model (must include multiple predictors).
- Your regression model should include at least one nominal predictor.
- Assess the fit and usefulness of the regression model.
    - Illustrate your findings using appropriate examples from your data.

This part of the assignment is worth 40% of the CA for this module as marked out of 100%.

DESCRIPTION

You are expected to:

- Present a summary of the variables used, critically discussing relevant issues which impact statistical analysis;
    - Include statistical summaries of the variables of interest and evidence to support relationships or difference to justify their inclusion in a dimension reduction or regression model.
- Use appropriate statistical techniques;
- Present and interpret the findings;
- Briefly draw conclusions discussing your findings in terms of other related work and any implications for future work;
- Adopt the APA guidelines for reporting statistical analysis using APA citation and referencing. You must use R to conduct your analysis;
- You should cite appropriate sources (which are accessible) in order to support the guidelines you adopt in your decision making and interpretation of findings.

You will need to demonstrate:

- An ability to generate and correctly state a hypothesis or hypotheses that is/are theoretically informed;
- The ability to correctly prepare, present, analyse and critically assess the dataset used from the perspective of statistical analysis;

- The ability to correctly execute, present and interpret appropriate statistical tests using statistical software;
- The ability to analyse and present the findings gained from your statistical analysis in a clear and accurate way to a standard expected of masters/PhD level academic work;
- The ability to construct a report on a statistical inquiry.

DELIVERABLES

- You need to submit an R markdown file plus the HTML/PDF created from this.
- You must include the following information at the start of your RMD file:
    - Student Number: <<your student number>>
    - Student Name: <<your name>>
    - ProgrammeCode: <<programmecode>>
    - OptionChosen: <<optiona/optionb>>
    - The version of R used.
    - The R packages needed for your code to execute successfully.
- State clearly the hypotheses you intend to test.
- You must describe your variables.
    - In terms of their statistical measurement types and describe them with appropriate descriptive statistics and graphs.
    - You must address all issues which could impact on the choices when building a model.
    - You must present statistical evidence to support inclusion of variables as predictors in any model/use in a dimension reduction.
- You must build, present and illustrate your model as outlined in the overview.
    - Justify your choices based on your assessment of the dataset.
    - Illustrate how your model works using appropriate data.
- You must present and interpret your findings in paragraphs using APA style for reporting statistical results.
- Interpret your findings appropriately relevant to your hypotheses.
- A useful guide to creating a report of a statistical inquiry using APA guidelines is available at http://www.discoveringstatistics.com/docs/writinglabreports.pdf.

SUBMISSION

All required documents should be emailed using the subject line **PSI Final Assignment**.

- You must include the following information at the start of all files submitted:
  - Student Number: <<your student number>>
  - Student Name: <<your name>>
  - Programme Code: <<programme code>>
  - The version of R used.
  - The R packages needed for your code to execute successfully.
- All files must include your student number at the start of the file name e.g. D123456.rmd, D123455.nb.html.

Your submission should comprise pdf file including all required reporting plus an R script well commented to indicate which sections of the report commands relate to plus an output file (html, pdf, word) that includes the output from these statistical tests well commented so that the commands that generated the commands can be found.

NOTES

1. Unfair practice is a very serious offence in the TU Dublin and you must acknowledge any material used by including a referenced bibliography in your report. Any issues will be investigated and those considered serious will be handled via the TU Dublin Plagiarism policy (details are available in the General Assessment Regulations).
2. Assignments must be submitted via Brightspace through the assignment section. Emailed submissions will be ignored.
3. Extensions due to acceptable personal circumstances must be requested by email in advance of the deadline.
4. For late submissions (i.e. without an agreed extension), a penalty of 5% will be applied for every day a submission is late.
5. No submissions will be accepted after the deadline unless an extension has been agreed.

BASIC MARKING SCHEME

The ability to correctly prepare, present, analyse and critically assess the dataset used from the perspective of the proposed statistical analysis to justify use of chosen technique(s);	10
Assessing the suitability of the dataset for the purposes of the dimension reduction technique chosen;	8
Description and assessment of the effectiveness of the outcomes of the dimension reduction using appropriate statistics;	7
Assessing fit and usefulness of regression model created using appropriate statistics;	5

Illustration of model using example data.	5
The ability to interpret the findings from your data within the context of your question and draw conclusions from this.	5
Total	40

1. **Linear regression with dummy variable [5 marks]**
   Consider the dataset “weatherhistory.csv”.
   a) Describe dependent variable pressure and independent variable temperature.
   b) Explore the relationship between pressure and temperature.
   c) Build a linear model considering temperature and pressure.
   d) Identify a dummy variable and build extended model considering dummy variable.
   e) Report your findings.

2. **Multiple Linear Regression [6 marks]**
   Consider the dataset “weatherhistory.csv”.

   a) Explore the relationship between pressure and windspeed.
   b) Build a linear model considering (windspeed, humidity, temperature) and pressure.
   c) Assess how model meets key assumptions of linear regression.
   d) Investigate a differential effect by adding dummy variable.
   e) Investigate an interaction effect for windspeed and dummyvariable.
   f) Report your findings.

3. **Logistic regression [4 marks]**
   Consider the dataset “heartfailure.csv”

   a) Build a model considering diabetes as predictor.
   b) Calculate and analyze odds ratio of the model.

c) Extend the model by considering variable age. (Convert the age into categorical data, if age < 55 , category 1; age is between 55 and 68 category 2 ; otherwise category 3)
d) Report your finding.