STAT2004 Analytics for Observational Data

PROJECT (50% of total mark)

Overview

This assessment task is designed to provide an opportunity to actively learn the **concepts** and **skills** of *Analytics for Observational Data* through the conception and conduct of a project in a relevant area.

The project has the following conditions:

(i) You are to undertake the project **pairing with another student (only for proposal and final report).**

(ii) Your project will be assessed in **three phases**.

1. The 1st phase is a **group proposal (5%, paired)**,
2. the 2nd phase is a **group final report (30%, paired)** and
3. the final phase is **presentation (10%, individual) and peer review assessment (5%, individual)**.

The proposal must be submitted in order for the project to be marked and presented.

(iii) The project is to be conceived, conducted, analysed and reported by your group **fully complying with all aspects of Academic Integrity.**

(iv) **The proposal.** This phase is aimed to ensure that students **do not propose the same dataset AND the same R package**. Please read Phase 1 below thoroughly.

(v) The **final project** should include the analyses of the approved dataset and R package from the proposal, ie

1. **one set of publicly available multivariate data** and
2. **one R package.** Please read Phase 2 below thoroughly.

About multivariate data.

- The proposed multivariate data should consist of **at least 50 rows (subjects/cases) and at least 7 columns of variables excluding the subjects/cases column**;
- Among the columns of variables, **at least four variables/columns are numerical, with no missing data;**
- The proposed multivariate data should have more rows (*n* be the number of rows/cases/subjects) than columns (*p* be the number of columns/variables), that is **n >>> p**.

- Please discuss with the lecturer about your proposed data resources (eg from **Project Data Resources available on Blackboard** or from your workplace or else) in order to confirm that they are reasonable in satisfying the categories of being multivariate data.

About R package.

- For a final report, you have to review a statistical tool implemented via a package in R that is related to multivariate modelling (eg advanced visualising, hypothesis testing, inference, modelling\techniques tools, etc).
- **The review should include demonstration of the application of the tool in R via one example.**
- You cannot choose to review an R package that have been used or will be used in this unit or has been used in any other Statistics units. The list of R packages used in this unit will be provided on Blackboard.
- **Resources for R packages under *CRAN Task Views* for multivariate analysis is now available on Blackboard.**

The final report should **provide a link to an internet address where the dataset can be downloaded from or being uploaded on Blackboard for the final project.**

The whole project should be sufficiently substantial to warrant 50% assessment for a semester unit.

An explanation of requirements for the three phases of the project is given below.

Phase 1: Project Proposal (5% of total marks, due by Friday 26th August 8pm of Week 5, Turnitin, group proposal). This proposal must be submitted in order to receive marks for the final project report.

In this phase 1, you are asked to provide a project proposal. Much of the detail of this design can be refined as the project proceeds. Include as much detail as you can in order to allow for feedback. *You will not be marked down for "wrong" parts; you will lose marks if the plan is too brief without attempts at addressing the points below.* **Phase 1 of the project is to be no more than 3 pages.**

The proposal. This is aimed to ensure that groups **DO NOT propose the same multivariate dataset and the same R package**. The proposal should include the following, clearly **indicating your group first preference for each of the following:**

1. **two multivariate datasets AND**
2. **two R packages** that **have not been used or will not be used** in this unit or has not been used in any other Statistics units.

If two groups propose **the same data and/or R package in their first preference**, then I will look at **the submission date for a decision (ie first come first serve basis).** The other group will be advised to use their second option.

Given that a group has 2 students, we recommend to initiate your collaboration that each student contributes to search 1 set of multivariate data and 1 R package for a proposal.

The project (10 marks) plan should entail the following.

- ***Objectives/Introduction (2 marks)***
  The overall project goals for both multivariate data analysis and R package. What is the general problem or situation you wish to investigate? What are the goals of your intended data analysis (eg classification, clustering, dimensional reduction, forecasting, etc)? **This may be refined when you write your final report.**
- ***Methods/ Data Collection/R Packages (5 marks)***
  You need to propose **two multivariate datasets and two R packages clearly indicating your first preference**, in which your group will be analysing the first ones only if no overlaps with other groups. What multivariate data do you intend to analyse and from where? What R package that you will be proposing?
- ***Proposed Analysis (3 marks)***
  - Briefly state multivariate techniques that your group propose to undertake. This may be refined in Phase 2 of the final project. What techniques will you entertain? How will you perform inference? How will you validate any models you use, if applicable?
  - What is the function (eg exploratory, inference, classification, clustering, etc) of R package that you are proposing in terms of analysing multivariate data?

Phase 2: Final report (30% of total marks, due Friday 21st October by 8pm of Week 12, group report) – Refer to Marking Criteria to maximise your marks of this part.

The **final project** should include two parts following the approval of proposal:

(i) **one set of publicly available multivariate data** which is to be analysed using **appropriate multivariate techniques/analyses** (or extensions of such) that you have learnt in this unit;

(ii) **one R package** which is approved in your proposal which is to be reviewed including demonstration of the application of the tool in R via one example;

(iii) **the two parts (i) and (ii) can be made dependent or independent to each other.** Independent to each other means you don't have to use the R package in (ii) to analyse your proposed dataset in (i).

In this phase of the project, your group are required to provide a group final report on the project. **Phase 2 of the project to be no more than 15 pages excluding references and appendices.** The project template will be provided in Blackboard along with the marking criteria. **Please refer to the marking criteria to maximise your marks.**

You should include at least the following:

- ***Introduction***
  An introduction to the project, reviewing the project objectives. A brief background should also be included.
- ***Methods***
  A brief discussion of the statistical methods to be used in the results (for both multivariate techniques and R package).
- ***Results***
  - Exploratory data analysis, summary statistics, discussion of possible outliers, etc with sufficient discussion/explanation.
  - Results of analyses described in the methods section.
- ***Discussion***
  -Discussion of the results and how they answer the objectives posed in the Introduction.
  -What issues arise from the analyses? Are the analyses satisfactory? Are there other research problems that spring from the analyses? Are there any limitations of the project?
- ***Conclusion***

Phase 3: Presentation (10%, individual) and Peer Review (5%, individual) (due in the Exam Period, more detail to be announced on Blackboard)

- **Recorded Presentation (Individual, 10%)**

You will be required to record a short presentation (no more than 10 minutes) to present your project and its results. By submitting your presentation, you are consenting that your presentation will be peer-reviewed by other students (see under Peer Review). A marking rubric will be provided.

Presentation is an important way to tell others about a project. There are many good advices about giving research seminars being available from the web. If you are looking for a place to start, try this link:

https://student.unsw.edu.au/support-oral-presentations

Guidelines

The focus of the presentation should be on **what you have done in your project**. You'll need to give an overview of your project's goals and methods, but mostly you should talk about how you went about the work, the results you've obtained, what the results mean, and what contribution your work has made.

- **You should plan to speak for no more than 10 minutes.**
- You should prepare PowerPoint slides (or similar, eg Beamer/LaTex or slidy) to use during your presentation.

- We recommend you should practice your presentation to make sure you know what you are trying to say and how long it will take you to say it.

There are many ways to record your presentation. We provide some information that belongs to Curtin University. More detail will be provided on Blackboard.

1. Echo Universal Capture (strongly recommended to use this): submitted using a link;

2. Collaborate Ultra or Webex or Microsoft Teams;

3. Microsoft Powerpoint (mp4 file): Create your slides and record using Microsoft PowerPoint. Then you can export it as a video to mp4 to be attached in Assignment Submission instead of as a link.

- **Peer Review (Individual, 5%)**

The main aim of peer review is to develop some skills in assessing recorded presentation. A marking rubric will be provided.

- You will be randomly allocated to peer review a presentation for a period of time, say in Week 1 of Exam.
- The peer review will be provided as an eTest on Blackboard that consists of a few short answer type of questions based on marking rubric. The test will be opened in Week 2 of Exam.
- One dummy question is about the dynamics in your group (0 mark). This question is designed to provide students with the opportunities to comment about their groups confidentially, whether everyone contributes equally or not (eg communication (in person or online); finding the data and R packages; writing the proposal; modelling; R coding; writing the final report). **Please do your best to be objective in providing these information.**