Probability and Statistical Inference
Continuous Assessment Supplemental
Semester I 2019/2020

Due Date: Sunday 2nd August 2020 @ 23.59

OVERVIEW

For the continuous assessment you are required to conduct and report on a statistical analysis to investigate a question for a given dataset. The dataset is available for download from the UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/student+performance) where you will find a description. It is also used in the following paper which also provides a dataset descriptor:

P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7. (https://repositorium.sdum.uminho.pt/bitstream/1822/8024/1/student.pdf)

Please ensure that you include this citation in the report you submit.

Y are required to conduct and present appropriate statistical tests including at least one multivariate inferential statistical technique covered in this module. You can choose one of the following options:

Option A:

- Using either Multiple Linear Regression or Logistic Regression:
- Develop hypotheses testable by building a regression model. These should be related to your overall research question.
- Derive and present appropriate statistical evidence.
- Build a baseline regression model.
  - Assess its fit and usefulness.
  - Illustrate your findings using appropriate examples from your data.
- Build at least one additional model which extends/amends this baseline either adding or removing predictors relevant to your hypotheses.
  - Assess its fit and usefulness.
  - Illustrate your findings using appropriate examples from your data.
  - Your regression model should include at least one nominal predictor.
- Compare the fit and usefulness of your successive models.

Option B:

- 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.

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 to achieve either option a or option b above;
- 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 submitted using the **Assignment II** in Brightspace.

- 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.

You have choices for your submission:

- Option A: R notebook which includes the commands and creates html with the nb.html created from this.
- Option B: A 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. Email 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 Sunday December 15th 2019 @ 23:59 unless an extension has been agreed.
   NB: Anything submitted later than this date without agreement will be ignored.
6. Assignments which do not adhere to the requirements or which are submitted incorrectly will attract a penalty of up to 10%.
7. No resubmission of assignments after feedback is given is allowed.

BASIC MARKING SCHEME

	Option a	Option b
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);	5	6
Option A	00	00
Assessing fit and usefulness of model(s) created using appropriate statistical evidence;	7	
Assessing how well dataset/models meet assumptions of regression using appropriate statistics;	3	

Illustration of model using example data;	3	
Comparison of successive models using appropriate statistics.	4	
Option B	0	
Assessing the suitability of the dataset for the purposes of the dimension reduction technique chosen;		6
Description and assessment of the effectiveness of the outcomes of the dimension reduction using appropriate statistics;		6
Assessing fit and usefulness of regression model created using appropriate statistics;		2
Illustration of model using example data.		2
The ability to interpret the findings from your data within the context of your question and draw conclusions from this.	3	3
Total	25	25