National College of Ireland

Masters in Science in Data Analytics

(MSCDAD_A / MSCDAD_B / MSCDAD_C / PGDDA_SEP24)

Statistics & Optimisation

Continuous Assessment (35%)

Release Date: 8th November 2024
Submission Date: 2nd December 2024

Duration: 25 days

1 Introduction

This assessment is designed to evaluate the learning outcomes of the Statistics and Optimisation module as outlined below:

LO1 Apply appropriate statistical inference techniques to the analysis of data across a variety of domains.

LO2 Interpret the results from statistical software packages and programming languages and report on them in a comprehensive, ethical and professional manner.

LO3 Critically evaluate, select and apply regression models appropriate to a given problem domain.

LO4 Critically appraise, select and apply suitable modelling and forecasting techniques for time series data.

2 Objectives

The objective of this assessment is to perform multiple linear regression and time series analyses.

Note: This is an **individual assessment**.

2.1 Multiple Linear Regression

On Moodle you will find a number of data sets for multiple linear regression. Each one is named *mlr* followed by a single digit and the extension *.csv*.

Important!!

For your analysis, you should use the file with the digit matching the last digit in your student number. For instance, if your student number is x12345678, you should use *mlr8.csv*. Failure to use the correct data set may result in your submission being considered invalid.
Before performing your analysis, you must split the data set into train and test partitions. This must be done using a random seed equal to your student number. For example, if your student number is x12345678, the seed should be 12345678. Failure to do so may result in your submission being considered invalid.

2.2 Time Series Analysis

On Moodle you will find a number of data sets for time series analysis. Each one is named *ts* followed by a single digit and the extension *.csv*.

Important!!

For your analysis, you should use the file with the digit matching the last digit in your student number. For instance, if your student number is x87654321, you should use *ts1.csv*. Failure to use the correct data set may result in your submission being considered invalid.

Split the data into consecutive training and test partitions. The choice of split ratio is left up to you. For instance you could use the first 80% of the data to train and the remaining 20% to test.

3 Deliverables

3.1 Report

You should document your analysis in a report of 7-8 pages. The report should use the IEEE conference template and should be in PDF format. Word and LaTeX templates are available from the IEEE website[1].

The report should contain the a separate section for the multiple linear regression and time series analyses, each of which should have the following sub-sections:

- **Introduction** - here you should provide a brief description of the objectives of the analysis.
- **Exploratory Data Analysis** - this should include levels of measurement and descriptive statistics for all variables, as well as visualisations that provide an insight into the variables in the data set.
- **Data Preparation** - here you should describe and justify all steps you took to prepare the data for modelling.
- **Modelling** - give a detailed description of the steps you followed to build your final model. You should include a rationale for rejecting intermediate models and you should describe and justify any handling of missing data and outliers in addition to any transformations applied to variables.
- **Interpretation** - For multiple linear regression, briefly interpret the model coefficients, their p values and confidence intervals. Also interpret the p value for the F statistic. For time series models, provide a short interpretation of the model parameters or coefficients.
- **Diagnostics** - using suitable visualisations and appropriate statistical tests. For multiple linear regression, discuss if the final model meets the Gauss-Markov assumptions.
- **Evaluation** - evaluate the predictive capability of your model using a test data set.

[1] http://www.ieee.org/conferences_events/conferences/publishing/templates.html

3.2 Code Artefact

You should create a *zip* or *gz* archive of program code used in the analyses.

4 Submission

The assessment carries 35% of the total marks for the module.

The submission should consist of:

- A **report** that must include your name and student number (as per NCI official documents) This must be clearly visible on the front page of the report. The report should be named *x12345678.pdf* replacing *12345678* with your student number, and should be uploaded as a PDF document to the **Report** Turnitin link on Moodle.
- A **code artefact** that should be uploaded as a *zip* or *gz* archive to the **Code Artefact** link on Moodle. This should be named *x12345678.zip* or *x12345678.gz*, replacing *12345678* with your studentnumber.

Please following the file naming conventions given above.

Late submissions will not be accepted unless an extension has been requested through NCI360 and officially approved.

5 Marking

The project will be marked according to the grading rubric provided in the last two pages of this document.

6 Academic Integrity

Any written work created by others must be properly cited and should be paraphrased or summarised where possible, otherwise it should be included in quotes. Figures not created by you should include an acknowledgment detailing the name(s) of the creator(s). Code found on the internet should not be claimed as your own, but instead a comment should be included in the source code indicating where you obtained it.

Students are strongly advised to familiarise themselves with the Guide to Academic Integrity produced by the NCI Library[2].

Note: All submissions will be electronically screened for evidence of academic misconduct, e.g. plagiarism, collusion and misrepresentation. Any submission showing evidence of such misconduct will be referred to the college's academic misconduct committee for disciplinary action.

[2] https://libguides.ncirl.ie/academicintegrity

Grading Rubric - Statistics & Optimisation Continuous Assessment

Semester 1 - 2024/25

MULTIPLE LINEAR REGRESSION

Criterion	**H1** $\geq$ 70%	**H2.1** $\geq$ 60% < 70%	**H2.2** $\geq$ 50% < 60%	**Pass** $\geq$ 40% < 50%	**Fail** < 40%
Descriptive Statistics & Visualisation **(10%)**	All relevant descriptive statistics are presented. Highly appropriate visualisation methods are used. Figures are very clearly presented.	All relevant descriptive statistics are presented. Appropriate visualisation methods are used. Figures are clearly presented.	All relevant descriptive statistics are presented. Visualisation choices are adequate but may be sub-optimal. Not all figures are clearly presented.	Some relevant descriptive statistics are presented. Visualisation choices are largely sub-optimal. Not all figures are clearly presented.	No relevant descriptive statistics are presented. Visualisation choices (if any) are sub-optimale and poorly presented..
Modelling Process **(15%)**	An exhaustive, technically sound and exceptionally thorough description of the modelling process used.	A technically sound and thorough description of the modelling process used.	An adequate and mostly technically sound description of the modelling process used.	A limited and somewhat flawed description of the modelling process used, lacking important details.	A poor (or missing) description of the modelling process used.
Interpretation **(15%)**	An exhaustive, technically sound and exceptionally thorough description of the modelling process used.	A technically sound and thorough description of the modelling process used.	An adequate and mostly technically sound description of the modelling process used.	A limited and somewhat flawed description of the modelling process used, lacking important details.	A poor (or missing) description of the modelling process used.
Code Artefact & Reproducibility **(5%)**	A complete code artefact is provided, along with full details to permit reproduction of the analysis.	A complete code artefact is provided, reproducibility is somewhat limited.	A complete code artefact is provided but the work is not reproducible.	A somewhat incomplete code artefact is provided. The work is not reproducible.	No code artefact is provided or the submitted artefact is largely incomplete.

TIME SERIES ANALYSIS

Criterion	**H1** $\geq$ 70%	**H2.1** $\geq$ 60% < 70%	**H2.2** $\geq$ 50% < 60%	**Pass** $\geq$ 40% < 50%	**Fail** < 40%
Assessment of raw time series **(10%)**	The time series data have been thoroughly and exhaustively explored and assessed.	The time series data have been reasonably thoroughly explored and assessed.	The time series data have been adequately explored and assessed.	The time series data have been subject to a very basic exploration and assessment.	Poor or no exploration and assessment of the time series data.
Modelling Process **(15%)**	An exhaustive, technically sound and exceptionally thorough description of the modelling process used.	A technically sound and thorough description of the modelling process used.	An adequate and mostly technically sound description of the modelling process used.	A limited and somewhat flawed description of the modelling process used, lacking important details.	A poor (or missing) description of the modelling process used.
Forecasting **(15%)**	Suitable forecasts have been produced and have been exceptionally thoroughly evaluated and interpreted.	Suitable forecasts have been produced and have been thoroughly evaluated and interpreted.	Suitable forecasts have been produced and have been reasonably well evaluated and interpreted.	Suitable forecasts have been produced and have been subject to a basic evaluation. Little or no interpretation is offered.	Forecasts produced (if any) have been poorly evaluated and interpreted.
Code Artefact & Reproducibility **(5%)**	A complete code artefact is provided, along with full details to permit reproduction of the analysis.	A complete code artefact is provided, reproducibility is somewhat limited.	A complete code artefact is provided but the work is not reproducible.	A somewhat incomplete code artefact is provided. The work is not reproducible.	No code artefact is provided or the submitted artefact is largely incomplete.

REPORT

Criterion	**H1** $\geq$ 70%	**H2.1** $\geq$ 60% < 70%	**H2.2** $\geq$ 50% < 60%	**Pass** $\geq$ 40% < 50%	**Fail** < 40%
Quality of Writing **(10%)**	Exceptionally well written, with no language errors. All figures are well conceived, readable and correctly captioned. The IEEE template is strictly adhered to. The report does not exceed the length limits. All references are appropriately and correctly used.	Well written, with no significant language errors. All figures are well conceived, readable and appropriately captioned. The IEEE template is adhered to. The report does not exceed the length limits. References are appropriately and correctly used.	Well written, but has a few significant language or style errors. Figures are well presented. The IEEE template and length limit are adhered to. References are complete and correctly used.	Adequately written, with some significant language and/or style errors. Figures may be hard to read or presented in a suboptimal manner. The IEEE template may not have been followed. References are mostly complete and correctly used.	Poorly written and littered with typographical errors and/or poor use of English. The IEEE template was not used. Figures may be hard to read. References (if any) are largely incomplete.