**ACS234 Systems Engineering Mathematics II**

**Semester 2 Test 1**

**Multiple Linear Regression**

**Briefing**

**Assignment weighting:** 5%

**Assignment released:** Monday 15 April 2024 (4 pm)

**Assignment Due:** Monday 22 April 2024 (4 pm)

**Aims and Objectives**

This assignment is designed to enhance your understanding of linear regression and develop your skills for real world problem solving.

**Submission Style**

You should submit your report online through the ACS234 BB page linked to Turnitin. Your report must be submitted as a single document which includes your registration number at the top of page one.

Solutions should be edited with a word processing software e.g. WORD or LaTeX. It is suggested to use A4 page with a margin of 2 cm on each side.

Name your report as: ACS234T1-RegistrationNo. For example, if your Registration Number is 111111111, then name your report as: ACS234T1-111111111.

**Late Submission**

This is a small test which is expected to be completed within 2 hours. So, late submissions are not accepted.

**Unfair Means**

The assignment should be completed individually. You should not share your answers with other students and should not work together in completing the assignment. The assignment must be wholly your own work. References must be provided to any other work that is used as part of this assignment. Any suspicions of the use of unfair means will be investigated and may lead to penalties. See <http://www.shef.ac.uk/ssid/exams/plagiarism> for more information.

**Special Circumstances**

If you have medical or personal circumstances which cause you to be unable to submit this assignment on time or that may have affected your performance, please complete and submit an Extenuating Circumstance Form along with documentary evidence of the circumstances to the student support team/office of your department. See [Extenuating Circumstances Form: Explanatory Notes - Forms - SSiD - The University of Sheffield](https://www.sheffield.ac.uk/ssid/forms/circsnotes).

**Background Information**

**The Dataset to be Used in Test 1**

The dataset, named, LondonAirQuality(2010\_2018).csv, contains information on air quality (monthly averaged data) in London for the period of 2010 to 2018.

The dataset consists of a total of 108 monthly averaged observations of the following 15 (columns A to O) features/variables:

* Column A, Month-Year: The date that the data were measured.
* Column B, Mean Roadside Nitric Oxide (ug/m3) of the month.
* Column C, Mean Roadside Nitrogen Dioxide(ug/m3).
* Column D, Mean Roadside Oxides of Nitrogen(ug/m3).
* Column E, Mean Roadside Ozone (ug/m3).
* Column F, Mean Roadside Sulphur Dioxide(ug/m3).
* Column G, Mean Roadside PM10 Particulate (ug/m3).
* Column H, Mean Roadside PM2.5 Particulate(ug/m3).
* Column I, Mean Background Nitric Oxide(ug/m3).
* Column J, Mean Background Nitrogen Dioxide(ug/m3).
* Column K, Mean Background Oxides of Nitrogen(ug/m3).
* Column L, Mean Background Ozone(ug/m3).
* Column M, Mean Background Sulphur Dioxide(ug/m3).
* Column N, Mean Background PM10 Particulate(ug/m3).
* Column O, Mean Background PM2.5 Particulate(ug/m3).

**Variables to Be Used in Test 1**

Test 1 is concerned with the following variables (features):

* x1 = Mean Background Nitric Oxide(ug/m3). [Column I]
* x2 = Mean Background Nitrogen Dioxide(ug/m3). [Column J]
* x3 = Mean Background Oxides of Nitrogen(ug/m3). [Column K]
* x4 = Mean Background Ozone(ug/m3). [Column L]
* x5 = Mean Background Sulphur Dioxide(ug/m3). [Column M]
* x6 = Mean Background PM10 Particulate(ug/m3). [Column N]
* y = Mean Background PM2.5 Particulate(ug/m3). [Column O]

**Test 1 – Questions (Problem Set)**

**Question 1**

1. Calculate the correlation coefficient between and (), respectively. Use the following form to report your results, and highlight (in **bold**) the variable that has the stronger (highest) correlation with .

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |
| Correlation  Coefficient |  |  |  |  |  |  |

Note: = correlation coefficient between and ; Consider a precision of TWO decimal places.

[**10 marks**]

1. If only one of the 6 variables is used to represent y through a simple linear regression model, which one should be used? Why?

[Write down your answer below]

Denote by *z* the variable that you choose to use, and let the simple linear regression be

+ error

Calculate and . Write down your answer in the following boxes.

|  |
| --- |
|  |
|  |

Notes: Consider a precision of three decimal places.

[**30 marks**]

**Question 2**

1. Fit a multiple linear regression model of the following form:

+ error

Use the following form/format to report your results (write down your answer in the corresponding empty boxes):

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

Notes:

1) Use , , …, instead of , , …, ;

2) Consider a precision of three decimal places.

[**40 marks**]

1. Calculate MSE (mean squared error) and R2 (coefficient of determination) for the fitted model.

|  |
| --- |
| MSE = |
| R2 = |

[**20 marks**]