DBS
Dublin Business School
excellence through learning

QQI

Higher Diploma in Science in Data Analytics

SUMMER 2019 EXAMINATIONS

Module Code:	**B8IT109**
Module Description:	**Advanced Data Analytics**
Examiner:	**Paul Laird**
Internal Moderator:	**Dr Shahram Azizi Sazi**
External Examiner:	**Dr Ralf Bierig**

INSTRUCTIONS TO CANDIDATES

1. **Answer all questions using R. R code must be saved in .R format.**
2. **Submit R code together with outputs, including graphs/plots/curves as a notebook in Word, HTML or PDF format.**
3. **Attempt question 1 and 3 other questions**

Question 1

In a wireless network, six sensors sense and analyze their own datasets.

(a) Model each sensor e.g. $S_j, j = 1, \dots, 6$ with $N(j, 25)$, and $S_6 = \sum_{j=1}^{5} S_j$ generate 100 samples for each sensor. Frame all samples into one dataset.

(10 Marks)

(b) Provide descriptive analyses for your dataset (e.g. summary, boxplot, …). Give some quick insights about the simulated dataset.

(10 Marks)

(c) Make a decision whether the population variance of the first sensor (σ_1^2, is significantly different from the population variance of the sixth sensor (σ_6^2) at the level $\alpha = 0.05$. To do so,

I. List the assumptions.

(5 Marks)

II. State the null and alternative hypotheses.

(5 Marks)

III. What is your decision rule and explain your decision?

(5 Marks)

IV. Provide the 95% confidence interval for the ratio of the variances.

(5 Marks)

(TOTAL: 40 Marks)

Question 2

Use the dataset available at:

http://users.stat.ufl.edu/~winner/data/ingots.dat , with description at: http://users.stat.ufl.edu/~winner/data/ingots.txt .

Note that there are two dependent variables.

(a) Test whether there is a significant relationship between each independent variable and the breaking strain at $\alpha=0.05$.

(7 marks)

(b) Construct a model including all variables, and refine iteratively, noting your justification for each refinement.

(8 marks)

(c) Investigate whether there are any interactions between the predictors, and produce a final model.

(5 marks)

(Total: 20 Marks)

Question 3

Use dataset available on:
http://www.stat.ufl.edu/~winner/data/HVAC_perform.csv,

(a) Suggest an appropriate GLM to model **powerp** to other numerical variables.

(5 Marks)

(b) Propose and justify the optimal model for **powerp,** showing model iterations, and noting significant variables at the level of $\alpha=0.05$, and estimate the parameters of your model.

(10 Marks)

(c) Predict the value of **powerp** for a combination of predictor variables not included in the data set

(5 Marks)

(Total: 20 Marks)

Question 4

Use the dataset available at:
https://raw.githubusercontent.com/jbrownlee/Datasets/master/daily-min-temperatures.csv

(a) Decompose the data into its components, and comment on these
(5 marks)

(b) Validate the assumption of stationarity in respect of the data or any appropriate transformations of the data.
(5 marks)

(c) Investigate and fit the optimal ARIMA model
(5 marks)

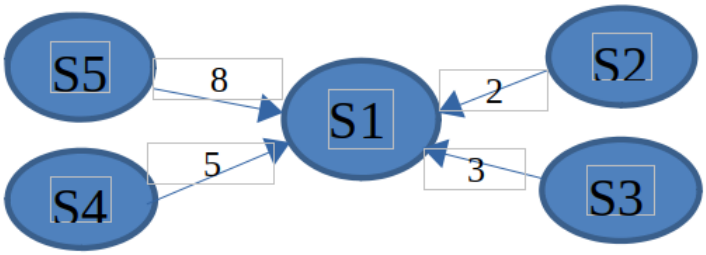
(d) Use the data excluding the last month to predict the last month's values with a 95% confidence interval.
(5 marks)

(Total: 20 marks)

Question 5

Use the simulated dataset in Question 1, sensors 1-5 in order to

(a) Adopt a centralized scheme to sensor 1, taking the weight of its own input as 1, the inputs of the other sensors are as marked on the edges: Compute the normalized weights.



(10 Marks)

(b) Compute the global arithmetic weighted mean. Please compute the global solution using R.

(10 Marks)

(Total: 20 Marks)

END OF EXAMINATION