QQI

HIGHER DIPLOMA IN SCIENCE IN DATA ANALYTICS

WINTER 2019 EXAMINATIONS

Module Code: **B8IT109**

Module Description: **Advanced Data Analytics**

Examiner: **Paul Laird**

Internal Moderator: **Dr Shahram Azizi Sazi**

External Examiner: **Dr Ralf Bierig**

Date: Thursday, 21st November 2019

Time: 09:30-11:30

INSTRUCTIONS TO CANDIDATES

I. *Solve all questions with R. Use a Notebook or Markdown*
II. *Answer Question 1*
III. *Answer two other questions*

Question 1

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

(a) Model sensors $S_j, j = 1, \ldots, 4$ as $N(j, 25)$, and S_5 as $\sum i$, and generate 40 samples for each sensor. Frame all samples into one dataset.

(10 Marks)

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

(5 Marks)

(c) Make a decision whether the population variance of the first sensor (σ_1^2, is significantly different from the variance of the fifth sensor (σ_5^2) at the level $\alpha = 0.05$. To do so,

I. List the assumptions, and state the null and alternative hypotheses.

(5 Marks)

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

(5 Marks)

III. Provide the 95% confidence interval for the ratio of the variance.

(5 Marks)

IV. Determine whether $\mu_i \neq \mu_j$ for any pair of sensors i,j; if so, provide the 95% confidence interval for those pairs which differ.

(10 Marks)

(TOTAL: 40 Marks)

Question 2

Use the ToothGrowth dataset

(a) Perform an ANOVA to determine whether supp or dose have a significant effect on len

(5 Marks)

(b) Use an interaction plot to determine the existence and nature (if relevant) of any interaction between the independent variables.

(5 Marks)

(c) Comment on the interaction, and the interaction which would have been observed if only doses 0.5 and 1 were analysed

(5 Marks)

(d) Perform a final ANOVA and provide all relevant coefficients.

(5 Marks)

(e) Conduct PCA on the data from **'http://users.stat.ufl.edu/~winner/data/steroid_doping.csv'** How many principal components would you use to summarise the data? Justify your answer.

(10 marks)

(Total: 30 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) Investigate the null or saturated model, and iteratively specify the significant variables on **powerp** at the level of α=0.05, and estimate the parameters of your model.

(15 Marks)

(c) Predict the value of **powerp** for:

run_id	airflux	wheelspd	regtemp	humid	drybulb	moistrem	thermalp	
1	1	550	6	100	0.6	30	1.981	0.645
2	2	550	8	110	0.7	34	3.681	1.002

(5 Marks)

(d) Provide predictions with their confidence interval.

(5 Marks)

(Total: 30 Marks)

Question 4

Using the dataset available on http://www.stat.ufl.edu/~winner/data/wage_cpi.csv, apply time series analysis, considering ‘wage’ as your time series variable:

(a) Validate the assumptions using graphical visualization.

(5 Marks)

(b) Fit the optimized model for ‘wage’ and provide the coefficient estimates for the fitted model.

(5 Marks)

(c) What is the estimated order for AR and MA?

(5 Marks)

(d) Forecast a h=10 steps ahead prediction of *wage* on the plot of the original time series.

(5 Marks)

(e) Validate your forecast by constructing a model from the data excluding the last 12 months

(10 Marks)

(Total: 30 Marks)