



## **FACULTY OF SCIENCE AND ENGINEERING**

### **DEPARTMENT OF MATHEMATICS AND STATISTICS**

## **END OF SEMESTER EXAMINATION**

MODULE CODE: MA4128

SEMESTER: REPEAT 2018

MODULE TITLE: Advanced Data Modelling    DURATION OF EXAM: 2.5 hours

LECTURER: Kevin O'Brien

GRADING SCHEME: 100 marks

60% of total module marks

EXTERNAL EXAMINER: Prof. A Marshall

### **INSTRUCTIONS TO CANDIDATES**

This paper is comprised of five questions, each worth 25 marks. Attempt any four questions. Scientific calculators approved by the University of Limerick can be used. Statistical tables are provided at back of exam paper.

1. (a) (4 Marks) Provide a brief description for three tests from the family of Grubb's Outliers Tests. Include in your description a statement of the null and alternative hypothesis for each test, any required assumptions and the limitations of these tests.
- (b) The following statistical procedure is based on this dataset.

{6.98, 8.49, 7.97, 6.64, 8.80, 8.48, 5.94, 6.94, 6.89, 7.47, 7.32, 4.01}

```
> grubbs.test(x, two.sided=T)

Grubbs test for one outlier

data:  x
G = 2.4093, U = 0.4243, p-value = 0.05069
alternative hypothesis: lowest value 4.01 is an outlier
```

- (i) (1 Mark) Describe the purpose of this procedure. State the null and alternative hypotheses.
- (ii) (1 Mark) Write the conclusion that follows from the code output above.
- (iii) (1 Mark) State any relevant assumptions for this procedure.
- (c) Consider the following inference procedure performed on data set  $X$ .

```
> shapiro.test(X)

Shapiro-Wilk normality test

data:  X
W = 0.84987, p-value = 2.143e-13
```

- (i) (2 Marks) Describe the purpose of this procedure. Include in your answer how the outcome of the procedure is to be interpreted.
- (ii) (3 Marks) What are the null and alternative hypotheses for this test? Write the conclusion that follows from this procedure.

*This question is continued on the next page.*

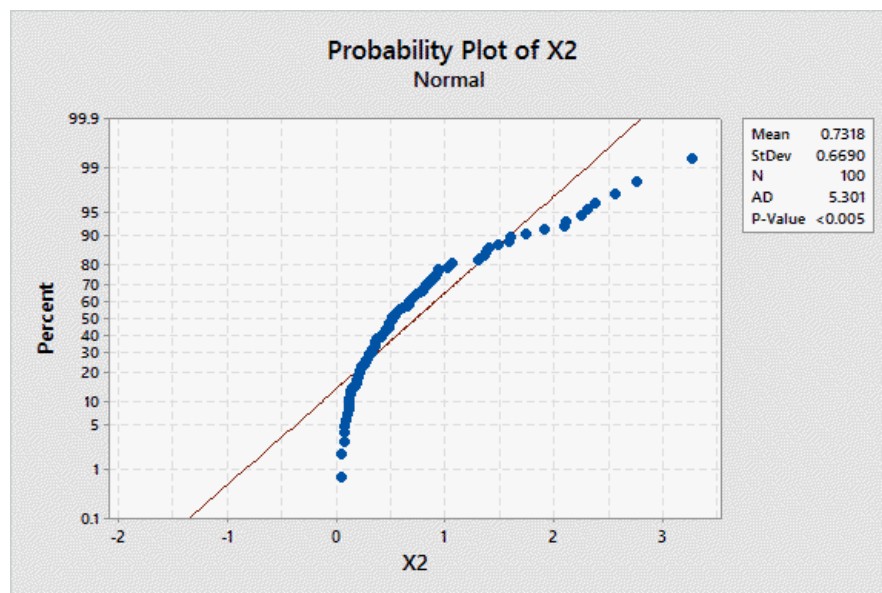
- (iii) (2 Marks) A subsequent procedure is reported below. Describe what was attempted in the procedure, and the outcome. Suggest a possible reason for this outcome.

```
> X <- log(X)
>
> shapiro.test(logX)

Shapiro-Wilk normality test

data:  X
W = NaN, p-value = NA
```

- (d) A graphical procedure was carried out to assess whether or not this assumption of normality is valid for data set Y. Consider the normal probability plot (i.e. Q-Q plot) in the figure below.



- (i) (2 Marks) Provide a brief description on how to interpret plots such as this. Support your answer with sketches.
- (ii) (1 Mark) What is your conclusion for this procedure? Justify your answer.
- (e) The following questions relate to Missing Data.
- (i) (2 Marks) What is Missing Data? Discuss the implications of Missing Data in the context of a statistical analysis.
- (ii) (3 Marks) Compare and contrast the following types of missing data: Missing At Random, Missing Not At Random, Missing Completely at Random.
- (iii) (3 Marks) Explain what is meant by Censored Data and Truncated Data. Describe two different types of Censored data.

2. (a) (3 Marks) With reference to the table below, define each of the following appraisal metrics in the context of a binary classification procedure (*1 Mark for each*).

- (i) Accuracy
- (ii) Precision
- (iii) Recall

	Predicted Negative	Predicted Positive
Observed Negative	True Negative	False Positive
Observed Positive	False Negative	True Positive

- (b) (2 Marks) What is the F-score? Explain why the F-score is considered a more informative measure of performance than the Accuracy score.

$$\text{Hint: } \text{F-score} = \frac{2 \times P \times R}{P + R}$$

- (c) (3 Marks) Calculate the following appraisal metrics using the below table of outcomes for binary classification (*1 Mark for each*).

- (i) Recall,
- (ii) Precision,
- (iii) F-score.

	Predict Negative	Predict Positive
Observed Negative	9790	85
Observed Positive	30	95

- (d) (3 Marks) What is a Receiver Operator Character (ROC) curve? Explain its function in the context of a binary classification procedure, how it is determined, and the means of interpreting the curve. Support your answer with sketches.

- (e) (i) (1 Mark) Suppose the odds of an outcome are 6. What is the probability of that outcome?
- (ii) (1 Mark) Suppose the probability of an outcome is 75%. What is the odds of that outcome occurring?

- (f) Answer the following questions relating to the SPSS output on the next page. In this analysis, we wish to predict whether or not a person has a saving's account, based on the following demographic variables.

- Age
- Socio-economic Status
- Sector within city
- Disease Status

*This question is continued on the next page.*

There are three possible outcomes for socio-economic status.

$$\{1 = \text{Upper}, 2 = \text{Middle}, 3 = \text{Lower}\}$$

There are three possible outcomes for socio-economic status.

{1 = Inner City Districts, 2 = Outer Suburbs}

The Disease status variable is a binary variable, with 1 indicating the presence of some sort of illness.

- (i) (1 Mark) Describe how Wald's Test was used to refine the initial model. Make reference to relevant figures in the output.
- (ii) (2 Marks) What is a logit? How can you transform a logit into a probability?
- (iii) (1 Mark) State the regression equation for the final logistic regression model.
- (iv) (4 Marks) What information is contained in the column labeled **Exp(B)**? For the initial model, interpret the figures from this column for both ***Socioeconomic status*** and ***Sector within city***. As part of your answer, comment on the 95% confidence intervals for both.
- (v) (2 Marks) Predict the outcome for the following case: a 35 year old person from the upper socio-economic category residing in the outer suburbs.
- (vi) (2 Marks) Predict the outcome for the following case: a 75 year old person from the middle socio-economic category residing in the inner suburbs.

#### Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1 <sup>a</sup>	Age	.037	.010	13.802	1	.000	1.038
	Socioeconomic Status	-.941	.200	22.242	1	.000	.390
	Sector within city	.732	.356	4.221	1	.040	2.078
	Disease Status	-.120	.390	.095	1	.758	.887
	Constant	.157	.699	.051	1	.822	1.170
Step 2 <sup>a</sup>	Age	.036	.010	14.030	1	.000	1.037
	Socioeconomic Status	-.944	.199	22.377	1	.000	.389
	Sector within city	.703	.343	4.189	1	.041	2.020
	Constant	.183	.694	.070	1	.792	1.201

*This question is continued on the next page.*

### Variables in the Equation

		95% C.I. for EXP(B)	
		Lower	Upper
Step 1 <sup>a</sup>	Age	1.018	1.058
	Socioeconomic Status	.264	.577
	Sector within city	1.034	4.177
	Disease Status	.413	1.904
	Constant		
Step 2 <sup>a</sup>	Age	1.018	1.057
	Socioeconomic Status	.263	.575
	Sector within city	1.030	3.959
	Constant		

a. Variable(s) entered on step 1: Age, Socioeconomic Status, Sector within city, Disease Status.

3. (a) (3 Marks) Compute the following distance metrics between the cases A, and B, described below. (1 Mark for each).

- (i) Euclidean Distance
- (ii) Squared Euclidean Distance
- (iii) Manhattan Distance

$$A = \{1, 5, 9, 11, 4\}$$

$$B = \{3, 3, 6, 4, 7\}$$

- (iv) (1 Mark) Explain why the squared Euclidean distance may be used in preferences in to the Euclidean Distance.

- (b) Write a brief note to describe the following terms. You may support your descriptions with sketches.

- (i) (3 Marks) Cosine Similarity.
- (ii) (3 Marks) Mahalanobis Distance.

- (c) The following questions relate to Hierarchical Clustering.

- (i) (3 Marks) Why do you standardize variables before carrying out a cluster analysis. Support your answer with an example.
- (ii) (3 Marks) Describe the process of Ward's Linkage in the context of cluster analysis.

*This question is continued on the next page.*

(iii) (9 Marks) Describe any three of the following linkage methods. Support your answer with sketches (*3 Marks for each*).

- Nearest Neighbour Linkage
- Furthest Neighbour Linkage
- Centroid Linkage
- Average Linkage

4. (a) The following topics relate to techniques for predictive modeling count techniques.

(i) (2 Marks) What does a Poisson regression model? State any assumptions that must be checked before it can be used.

(ii) (1 Mark) The R Code output given below is used to predict the number of awards won by students.

- Information is provided on which of the three school programs the student takes part in (*General, Vocational or Academic*).
- Also we are given the mathematics test score.

State the mathematical formula used to predict the number of awards won.

You can denote **progAcademic**, **progVocational** and **math** as  $x_1, x_2$  and  $x_3$  respectively.

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )	
(Intercept)	-5.2471	0.6585	-7.97	1.6e-15	***
progAcademic	1.0839	0.3583	3.03	0.0025	**
progVocational	0.3698	0.4411	0.84	0.4018	
math	0.0702	0.0106	6.62	3.6e-11	***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

(iii) (2 Marks) Use the model in Part (ii) to predict the number of awards won by a vocational program student, with a maths score of 50.

(iv) (2 Marks) Use the model in Part (ii) to predict the number of awards won by an academic program student, with a maths score of 75.

(v) (1 Mark) Describe the circumstances whereby Negative Binomial Regression Models would be used instead of Poisson Models.

(vi) (3 Marks) What is Zero Inflation? Explain the modeling process for a Zero Inflated Model. Give an example of Zero-Inflated Count Process. *Support your answer with a sketch, if necessary.*

(vii) (2 Marks) What is Zero Truncation? Give an example of a Zero Truncated Count Process.

*This question is continued on the next page.*

- (b) The following questions relate to Principal Component Analysis.
- (i) (2 Marks) Principal Component Analysis is a Dimensionality Reduction technique. Explain what this term means.
  - (ii) (5 Marks) What is meant by the “true” dimension of the data? How does an analyst determine the appropriate number of principal components to retain, making reference to three different approaches.
  - (iii) (3 Marks) The Kaiser-Meyer-Olkin (KMO) statistic is used to measure a certain characteristic of the data. What is this characteristic? Explain how the KMO statistic should be interpreted.
  - (iv) (2 Marks) Briefly describe the Bartlett Test for Sphericity, with reference to the null and alternative hypotheses, and how those statements relate to the purpose of the test.

5. (a) The following questions relate to model appraisal in the context of regression models.

- (i) (3 Marks) In the context of regression models, explain what is meant by Heteroscedasticity and Homoscedasticity. Support your answers with sketches.
- (ii) (2 Marks) Explain the term Influence in the context of linear regression models. Support your answer with sketches.
- (iii) (2 Marks) Explain the term Cooks Distance in the context of linear regression models.
- (iv) (2 Marks) The Durbin Watson Test was carried out to test for Autocorrelation. Briefly describe autocorrelation. You may support your answer with sketches.
- (v) (3 Marks) State your conclusion to the following procedure.

```
> durbinWatsonTest(myModel)
```

lag	Autocorrelation	D-W Statistic	p-value
1	-0.08428163	2.143578	0.806

Alternative hypothesis: rho != 0

(b) The following questions relate to model selection and validation in the context of multiple regression analysis.

- (i) (1 Mark) Explain the purpose of variable selection procedures.
- (ii) (3 Marks) Compare and contrast the following variable selection procedures (1 Mark for each).
  - Forward Selection
  - Backward Elimination
  - Stepwise Regression
- (iii) (1 Mark) Explain how the *Akaike information criterion* is used to compare two models fitted for the same data.

*This question is continued on the next page.*



- (c) (8 Marks) Suppose we have 5 predictor variables:  $x_1, x_2, x_3, x_4$  and  $x_5$  to model a response variable  $y$ , and that we have the Akaike Information Criterion (AIC) for models based on each possible combination of predictor variables. Use **Forward Selection** and **Backward Selection** to choose the optimal set of predictor variables, based on the AIC measure. Support your answer with sketches.

Variables	AIC	Variables	AIC
$\emptyset$	200	$x_1, x_2, x_3$	74
		$x_1, x_2, x_4$	75
$x_1$	150	$x_1, x_2, x_5$	79
$x_2$	145	$x_1, x_3, x_4$	72
$x_3$	135	$x_1, x_3, x_5$	85
$x_4$	136	$x_1, x_4, x_5$	95
$x_5$	139	$x_2, x_3, x_4$	83
		$x_2, x_3, x_5$	82
$x_1, x_2$	97	$x_2, x_4, x_5$	78
$x_1, x_3$	81	$x_3, x_4, x_5$	85
$x_1, x_4$	94		
$x_1, x_5$	88	$x_1, x_2, x_3, x_4$	93
$x_2, x_3$	87	$x_1, x_2, x_3, x_5$	120
$x_2, x_4$	108	$x_1, x_2, x_4, x_5$	104
$x_2, x_5$	87	$x_1, x_3, x_4, x_5$	101
$x_3, x_4$	105	$x_2, x_3, x_4, x_5$	89
$x_3, x_5$	82		
$x_4, x_5$	86	$x_1, x_2, x_3, x_4, x_5$	100