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) Use the Dixon Q-test to determine if there is an outlier present in this sample data. You may assume a significance level of 5%.

$$131, 139, 107, 117, 123, 127, 122, 132, 135$$

(i) (1 Mark) State the null and alternative hypotheses for this test.

(ii) (2 Marks) Compute the test statistic?

(iii) (1 Mark) State the appropriate critical value.

(iv) (1 Mark) What is your conclusion to this procedure?

(d) 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 is the null and alternative hypotheses for this test? Write the conclusion that follows from this procedure.

This question is continued on the next page.

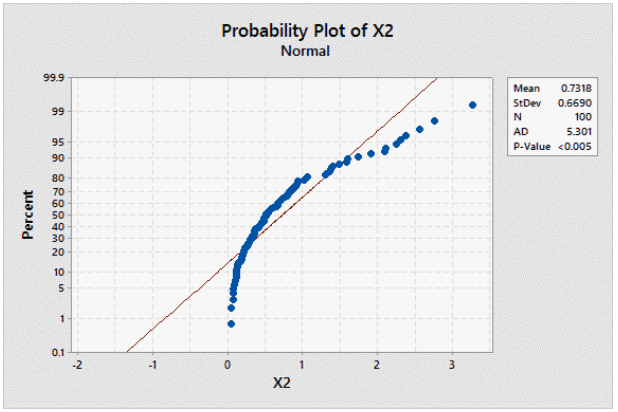
(iii) (2 Marks) A subsequent procedure is reported below. Describe what was attempted in the procedure, and the outcome. Suggest a possible reason for this outcome.

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

Shapiro-Wilk normality test

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

(e) 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.

(f) 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) (5 Marks) Describe the technique of Multiple Imputation.

(iv) (4 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.

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

(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) 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

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.

$$\{1 = \text{Inner City}, 2 = \text{Inner Suburbs}, 3 = \text{Outer Suburbs}\}$$

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

This question is continued on the next page.

(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.

(vii) (2 Marks) What is a dummy variable? Explain how it is used in Logistic Regression. Support your answer with an example.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	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[a]	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[a]	Age	1.018	1.058
	Socioeconomic Status	.264	.577
	Sector within city	1.034	4.177
	Disease Status	.413	1.904
	Constant		
Step 2[a]	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) (4 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
   (iv) Chebyshev Distance

$$A = \{5, 9, 2, 11, 4\}$$
$$B = \{3, 6, 9, 4, 7\}$$

   (v) (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 description with a sketch.
   (i) Cosine Similarity.
   (ii) Mahalanobis Distance.

   (c) The following questions relate to Hierarchical Clustering.
   (i) (1 Mark) Distinguish between agglomerative and divisive hierarchical clustering techniques.
   (ii) (3 Marks) Why do you standardize variables before carrying out a cluster analysis. Support your answer with an example.
   (iii) (3 Marks) Describe the process of Ward's Linkage in the context of cluster analysis.
   This question is continued on the next page.

(iv) (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

(v) (2 Marks) In the context of cluster analysis, What is the chaining effect? Give a brief description, supporting your answer with sketches.

4. (a)

(b)

(c) 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
```

This question is continued on the next page.

(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.

(d) 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) (4 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 multicollinearity in the context of multiple regression analysis.

(i) (1 Mark) Define multicollinearity.

(ii) (2 Marks) State two ways in which a multiple regression analysis could be affected by severe multicollinearity.

(iii) (2 Marks) State two ways of formally diagnosing the severity of multicollinearity, making reference to how both should be used to make decisions about the data.

(b) (i) (3 Marks) In the context of regression models, explain what is meant by Heteroscedascity and Ho- moscedascity. Support your answers with sketches.

(ii) (3 Marks) Explain the term Influence in the context of linear regression models. Support your answer with sketches.

(iii) (3 Marks) Explain the term Cooks Distance in the context of linear regression models.

(iv) (3 Marks) The Durbin Watson Test was carried out to test for Autocorrelation. Briefly describe autocorrelation. You may support your answer with sketches.

(v) 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
```

(vi) (3 Marks) In certain circumstances, Robust Regression may be used in preference to Ordinary Least Squares Regression. Describe what these circumstances might be.

(vii) (3 Marks) State one difference between Ordinary Least Squares and Robust regression techniques, in terms of computing regression equations.

(viii) (3 Marks) Explain the process of Huber Weighting, stating the algorithm used to compute weight- ings.

(ix) (3 Marks) Suppose that Huber Weighting, with a tuning constant of $k = 13.45$ was applied to the observations tabulated below. What would be the outcome of the procedure for each case.

Observation (i)	Residual (e_i)
11	-9.07
14	14.54
18	22.91
21	33.23

(c) 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.

(iv) (1 Mark) Explain why the adjusted R^2 value may differ in value from the corresponding multiple R^2 value for the same fitted model.

(v) (3 Marks) Describe model validation in the model-building process, with particular emphasis on the standard data partition.

(d) (8 Marks) Suppose we have 5 predictor variables: x_1, x_2,x_3, x_4 and x_5 to model a response variale 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.

Variables	AIC	Variables	AIC
$\emptyset$	200	x1, x2, x3	74
		x1, x2, x4	75
x1	150	x1, x2, x5	79
x2	145	x1, x3, x4	72
x3	135	x1, x3, x5	85
x4	136	x1, x4, x5	95
x5	139	x2, x3, x4	83
		x2, x3, x5	82
x1, x2	97	x2, x4, x5	78
x1, x3	81	x3, x4, x5	85
x1, x4	94		
x1, x5	88	x1, x2, x3, x4	93
x2, x3	87	x1, x2, x3, x5	120
x2, x4	108	x1, x2, x4, x5	104
x2, x5	87	x1, x3, x4, x5	101
x3, x4	105	x2, x3, x4, x5	89
x3, x5	82		
x4, x5	86	x1, x2, x3, x4, x5	100

Tables

Critical Values for Dixon Q Test

N	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
3	0.941	0.97	0.994
4	0.765	0.829	0.926
5	0.642	0.71	0.821
6	0.56	0.625	0.74
7	0.507	0.568	0.68
8	0.468	0.526	0.634
9	0.437	0.493	0.598
10	0.412	0.466	0.568
11	0.392	0.444	0.542
12	0.376	0.426	0.522
13	0.361	0.41	0.503
14	0.349	0.396	0.488
15	0.338	0.384	0.475
16	0.329	0.374	0.463