



FACULTY OF SCIENCE AND ENGINEERING

DEPARTMENT OF MATHEMATICS AND STATISTICS

MID TERM EXAMINATION PAPER 2 2016

MODULE CODE: MA4128

SEMESTER: Spring 2016

MODULE TITLE: Advanced Data Modelling DURATION OF EXAM: 1 hours

LECTURER: Kevin O'Brien

GRADING SCHEME: 50 marks
20% of total module marks

INSTRUCTIONS TO CANDIDATES

Attempt All Questions

Scientific calculators approved by the University of Limerick can be used.

Question 1: Miscellaneous

- (i) In the context of statistical modelling, describe what is meant by Overfitting. You can support your answer with sketches.
- (ii) In the context of statistical modelling, describe what is meant by the Law of Parsimony
- (iii) In the context of statistical modelling, describe what is meant by overfitting?
- (iv) (2 Marks) What is Dimensionality Reduction
- (v) (2 Marks) Compare and contrast dimensionality reduction techniques such as Variable Selection (e.g. Forward Selection and Backward Selection) with techniques such as Principal Component Analysis.
- (vi) (1 Mark) Suppose the odds of an outcome are 9. What is the probability of that outcome?
- (vii) (1 Mark) Suppose the probability of an outcome is 80%. What is the odds of that outcome occurring?
- (viii) (2 Marks) Suppose that, out of a sample of 100 women and 100 men, 80 men drank alcohol in the last week, while 20 women drank alcohol in past week. Compute the odds ratio for Women to men
- (ix) Law of Parsimony
- (x) Ordinal and Multinomial Logistic Regression
- (xi) Multicollinearity
- (xii) Tolerance and VIF

Question 2: Logistic Regression

- (i) (2 Marks) What is logistic regression? How does it differ from linear regression? Under what circumstances would you use it?
- (ii) (2 Marks) Compare and contrast Binary Logistic Regression with Multinomial and Ordinal Logistic Regression.
- (iii) (2 Marks) What is a dummy variable? Explain how it is used in Logistic Regression. Support your answer with an example.
- (iv) What is a logit? how is it computed into a probability?

Variables in the Equation								95% C.I. for EXP(B)	
		B	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper
Step 1 ^a	age	.085	.028	9.132	1	.003	1.089	1.030	1.151
	weight	.006	.022	.065	1	.799	1.006	.962	1.051
	gender(1)	1.950	.842	5.356	1	.021	7.026	1.348	36.625
	VO2max	-.099	.048	4.266	1	.039	.906	.824	.995
	Constant	-1.676	3.336	.253	1	.615	.187		

a. Variable(s) entered on step 1: age, weight, gender, VO2max.

Figure 1:

$$\pi_i = \frac{e^{\eta_i}}{1 + e^{\eta_i}}$$

Question 3: Model Metrics and Variable Selection

- (i) Given that there are N predictor variables that can be used to predict an outcome. Consider a statistical model as a subset of these predictor variable. How many distinct models can be constructed (including a constant-term only model)
- (ii) AIC
- (iii) Describe how you would use the Variance Inflation Factor to make an assessment about multicollinearity.
- (iv) Describe how to use to the Akaike Information Criterion for model selection.
- (v) State two ways of methodically diagnosing the severity of multi-collinearity. How are these techniques related? How are they used to make decisions about the data?
- (vi) Discuss a multiple regression technique could be affected by severe multicollinearity?
- (vii) Explain what variable selection procedures are used for.
- (viii) Compare and contrast three types of variable selection procedure.

Question 4: Truncated, Censored and Missing Data

- (i) (2 Marks) What is meant by missing data? Discuss the implications of Missing data in the context of a statistical analysis.
- (ii) (4 Marks) Missing Data is commonly classed into three different categories. What are these three categories? Compare and Contrast each of these three categories.
- (iii) (2 Marks) Consider the question in the figure below. This questionnaire is to be answered by parents of small children.

E12. Can I just check, are you currently pregnant? Yes ☐ ₁ No ☐ ₂

E13. Approximately how many weeks? _____ weeks

Figure 2:

- (iv) Compare and contrast the following types of missing data: Missing At Random, Missing Not At Random, Missing Completely at Random.
- (v) What is meant by missing data? Discuss the implications of Missing data in the context of a statistical analysis.
- (vi) Compare and contrast Censored Data and Truncated Data
- (vii) Compare and contrast Missing Data, Interval Data and Censored Data.
- (viii) Describe two cases of censored data. Give a Practical exam for both cases.

Question 5: Principal Component Analysis

- (i) What is PCA.
- (ii) Methods of choosing number of components to extract.
- (iii) KMO Bartlett
- (iv) Principal Component Analysis is a data reduction technique. Explain what this term means.
- (v) What is the KMO statistic? Describe how to interpret the KMO statistic.
- (vi) What is the Bartlett Test of Sphericity used for?
- (vii) varimax, quartimax and equamax are the commonly used methods in a certain procedure. What is this procedure? What is the purpose of the procedure. Which method is the most commonly used?
- (viii) Describe how to use a Scree plot in the context of dimensionality reduction techniques.
- (ix) The KMO is used to measure what characteristic of the data. Explain how the KMO measure should be interpreted.
- (x) 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.
- (xi) Discuss three techniques for determining the appropriate number of principal components.
- (xii) What is the purpose of a principal component analysis? Compare and contrast PCA and variable selection procedures such as backward selection.

Component Matrix^a

	Component			
	1	2	3	4
y1	.532	-.394	-.257	-.062
y2	.563	-.367	-.314	.188
y3	.548	-.405	-.276	.101
y4	.549	-.281	-.159	.033
y5	.475	-.022	.539	.242
y6	.524	-.140	.503	.069
y7	.471	-.195	.530	.066
y8	.484	.145	-.058	-.590
y9	.550	.284	-.005	-.464
y10	.571	.133	.062	-.424
y11	.492	.453	-.094	.356
y12	.492	.396	-.284	.295
y13	.527	.468	-.066	.253

Extraction Method: Principal Component Analysis.

a. 4 components extracted.

Figure 3:

	Component		
	1	2	3
Vehicle type	-.101	.095	.954
Price in thousands	.935	-.003	.041
Engine size	.753	.436	.292
Horsepower	.933	.242	.056
Wheelbase	.036	.884	.314
Width	.384	.759	.231
Length	.155	.943	.069

Figure 4: