UNIVERSITY OF DUBLIN

XST30071

TRINITY COLLEGE

FACULTY OF ENGINEERING, MATHEMATICS AND SCIENCE

School of Computer Science and Statistics

JS-SS MSISS & Maths **Trinity Term 2010**

Multivariate Linear Analysis and Applied Forecasting (ST3007)

Thursday, May 20, 2010 **Exam Hall** **9.30 — 12.30**

Dr. Rozenn Dahyot, Dr. Brett Houlding & Prof. Philip Scarf

Attempt two questions out of three in each section A and B

All questions carry equal marks

Non-programmable calculators are permitted for this examination—please indicate the make and model of your calculator on each answer book used.

You may not start this examination until you are instructed to do so by the Invigilator.

Section A - Multivariate Linear Analysis

1. The length and width of 46 jellyfish from the Hawkesbury River in Australia were recorded.

Variable	Description
Width	Width of the jellyfish (mm).
Length	Length of the jellyfish (mm).

A *k*-Means analysis of the data was performed with relevant summary output provided in Appendix A at the end of this question.

a) Describe the difference between hierarchical and iterative clustering methods.
[3 marks]

b) Provide a description of the *k*-Means clustering algorithm.
[6 marks]

c) Describe what type of cluster structure *k*-Means clustering is effective at finding.
[2 marks]

d) Explain why a *k*-Means algorithm should be re-run from differing initializations.
[2 marks]

e) Provide a description of the *k*-Means clustering output. In particular, describe what the numerical summaries tell us about the clusters for what appears to be the most appropriate value of *k*. Justify your choice of *k*.
[9 marks]

f) Give a brief account on the use of standardization within cluster analysis methods.
[3 marks]

Appendix A

Number of clusters: 1

	Number of Obs.	Within Sum of Squares	Avg. Distance From Centroid
Cluster 1	46	1538	234
Sum	46	1538	

Cluster Centroids:

	Cluster 1	Total Data
Width	13.3	13.3
Length	15.8	15.8

Distance Between Cluster Centroids:

	Cluster 1
Cluster 1	0.0

Number of clusters: 2

	Number of Obs.	Within Sum of Squares	Avg. Distance From Centroid
Cluster 1	18	157.6	2.7
Cluster 2	28	196.8	2.3
Sum	46	354.4	

Cluster Centroids:

	Cluster 1	Cluster 2	Total Data
Width	8.9	16.2	13.3
Length	11.2	18.7	15.8

Distance Between Cluster Centroids:

	Cluster 1	Cluster 2
Cluster 1	0.0	10.4
Cluster 2	10.4	0.0

Number of clusters: 3

	Number of Obs.	Within Sum of Squares	Avg. Distance From Centroid
Cluster 1	16	55.9	1.6
Cluster 2	16	111.6	2.3
Cluster 3	14	54.6	1.8
Sum	46	222.1	

Cluster Centroids:

	Cluster 1	Cluster 2	Cluster 3	Total Data
Width	14.5	8.5	17.5	13.3
Length	16.8	10.9	20.2	15.8

Distance Between Cluster Centroids:

	Cluster 1	Cluster 2	Cluster 3
Cluster 1	0.00	55.8	1.3
Cluster 2	55.8	0.00	57.0
Cluster 3	1.3	57.0	0.00

Number of clusters: 4

	Number of Obs.	Within Sum of Squares	Avg, Distance From Centroid
Cluster 1	7	7.4	0.9
Cluster 2	14	54.6	1.8
Cluster 3	11	17.7	1.2
Cluster 4	14	27.7	1.3
Sum	46	107.4	

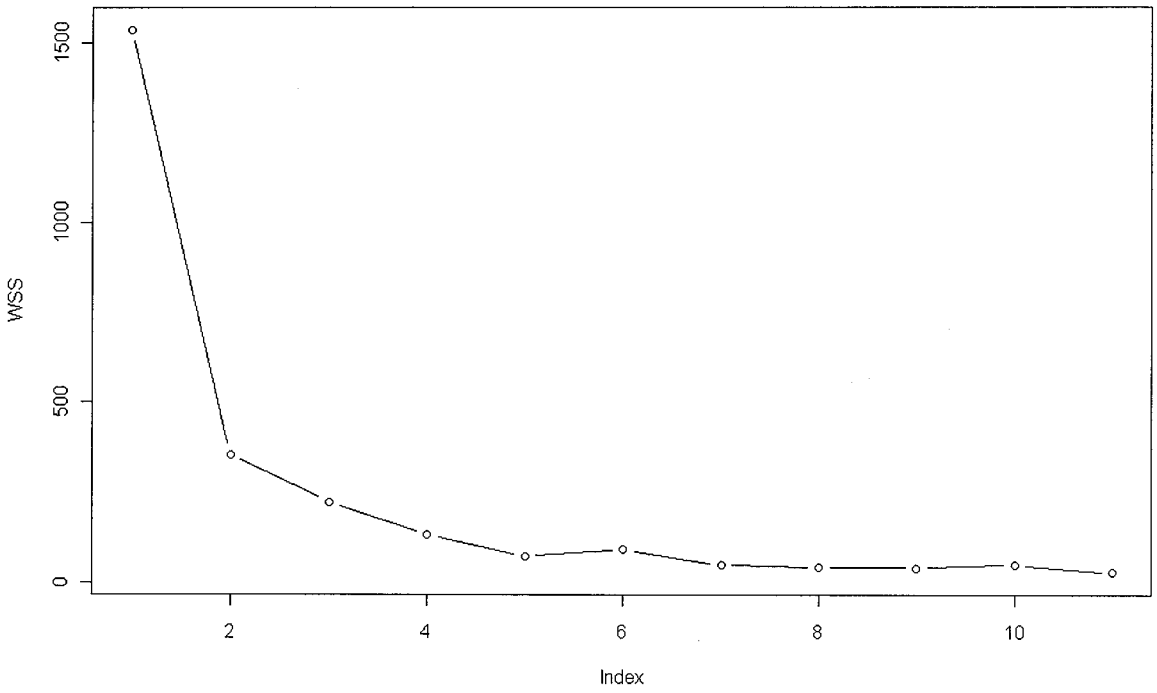
Cluster Centroids:

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Total Data
Width	11.4	17.5	7.4	14.8	13.3
Length	13.6	20.2	9.8	17.3	15.8

Distance Between Cluster Centroids:

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Cluster 1	0.00	9.0	5.6	5.0
Cluster 2	9.0	0.00	14.6	4.0
Cluster 3	5.6	14.6	0.00	10.5
Cluster 4	5.0	4.0	10.5	0.00

Total within cluster Sum of Squares plotted against number of clusters



2. Data were recorded on the survival status of 765 passengers of the Titanic.

Variable	**Description**
PClass	Binary: 1=1st Passenger Class, 0=2nd or 3rd Passenger Class.
Age	Age in years.
Sex	Binary: 0=Female, 1=Male
Survived	Binary: 1=Yes, 0=No.

Logistic regression was used to determine if *PClass*, *Age*, and *Sex* is a good predictor of survival status. Output from the logistic regression is given in Appendix B at the end of this question.

a) Provide a motivation for the use of logistic regression, rather than linear regression, for this data.

[3 marks]

b) Explain what the output from Appendix B tells us about the relationship in this model between *PClass*, *Age*, and *Sex*, and whether the passenger survived the Titanic.

[6 marks]

c) Give the formula for the probability of *Survived*=1 that results from the logistic model. Use this formula along with the output from Appendix B to predict the probability that a female from 1st passenger class who was aged 20 years would survive.

[6 marks]

d) For data with only binary entries, explain the Simple Matching (Hamming) and Jaccard dissimilarity measures (you should provide and explain the relevant formulas). Give a motivation for using the Jaccard measure instead of the Hamming measure.

[4 marks]

e) For data $\mathbf{x}^T=(x_1, x_2, \ldots, x_m)$, $\mathbf{y}^T=(y_1, y_2, \ldots, y_m)$, and $\mathbf{z}^T=(z_1, z_2, \ldots, z_m)$, give the definition of the Maximum dissimilarity measure and show how this measure satisfies the common dissimilarity properties of non-negativity, symmetry, and the triangle inequality.

[6 marks]

Appendix B

Call:

glm(formula = Survived ~ ., family = binomial(logit), data = Titanic)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.6406	-0.6052	-0.4648	0.7792	2.2699

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	1.628748	0.248278	6.560	5.37e-11	***
Age	-0.034331	0.007433	-4.619	3.86e-06	***
PClass1	1.895242	0.236271	8.021	1.04e-15	***
Sex1	-2.580914	0.194033	-13.301	< 2e-16	***

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

(Dispersion parameter for binomial family taken to be 1)

Null deviance:	1025.57 on 755 degrees of freedom
Residual deviance:	723.15 on 752 degrees of freedom
AIC:	731.15

Number of Fisher Scoring iterations: 4

3. A Motivational State Questionnaire was completed by 197 respondents. Each respondent was provided with ten words that could be used to describe a personality trait. For each such trait the respondent scored themselves on a 0-3 scale so as to indicate how strongly they agreed the attribute described their personality.

   The attributes in question were: {"sociable", "lively", "irritable", "attentive", "tense", "interested", "frustrated", "inspired", "fearful", "proud"}

   Output from an application of Factor Analysis for this data is provided in Appendix C at the end of this question.

   a) Provide a brief description on the benefits of Dimension Reduction as a statistical technique.

   [2 marks]

   b) State and explain the Factor model for multivariate data $\mathbf{x}^T=(x_1, x_2, \ldots, x_m)$. That is to say, show how the multivariate data point relates to the *common factors* and the *specific factors*, and indicate any assumptions over the expectation and covariance of these factor terms.

   [6 marks]

   c) In relation to the variance of a multivariate random variable, explain what is meant by the *communality* and *uniqueness* within a Factor Analysis, and show how these three terms are related.

   [4 marks]

   d) Explain what is meant by a *factor rotation* and why it is relevant within Factor Analysis.

   [4 marks]

   e) Explain what the output of Appendix C tells us about the data arising from the Motivational State Questionnaire.

   [6 marks]

   f) Contrast the similarities and differences between Factor Analysis and Principal Components Analysis.

   [3 marks]

Appendix C

factanal(x = personality, factors = 2, rotation = "none")

Uniquenesses:

sociable	0.545
lively	0.324
irritable	0.627
attentive	0.565
tense	0.284
interested	0.462
frustrated	0.396
inspired	0.569
fearful	0.569
proud	0.536

Loadings:

	Factor1	Factor2
sociable	0.674	
lively	0.817	
irritable	-0.419	0.444
attentive	0.648	0.122
tense		0.842
interested	0.696	0.232
frustrated	-0.302	0.716
inspired	0.567	0.331
fearful		0.654
proud	0.674	

	Factor1	Factor2
SS loadings	3.080	2.042
Proportion Var	0.308	0.204
Cumulative Var	0.308	0.512

Section B - Applied Forecasting

4. **Definitions.**

   Write short notes (approx. 150 to 200 words) on FIVE of the following topics (5 marks each):

   (a) ACF and PACF plots.

   (b) Regression methods in forecasting.

   (c) The use of the backshift operator in representing ARMA time series models.

   (d) MAPE and RMSE

   (e) AIC and BIC.

   (f) Transformations to induce stationarity.

   (g) What are the R commands used to

   - fit a linear model?
   - fit an ARIMA model?
   - fit an exponential smoothing model?
   - plot the times series and the corresponding ACF and PACF?

   (25 marks)

5. **Times series Analysis.**

(a) Discuss the 3 components that can be found in times series.

[4 marks]

(b) Is the time series in figure 1 stationary? Explain your answer.

[4 marks]

(c) Explain why the expert decides to compute the first differences of the data (cf. figure 2).

[4 marks]

(d) The expert suggests an $ARIMA(3, 1, 0)$ model. Explain this choice.

[4 marks]

(e) Write the $ARIMA(3, 1, 0)$ model using the backshift operator.

[4 marks]

(f) A few other models of the form $ARIMA(p, 1, q)$ were also tested by the expert, and their AICs were computed (cf. Table 1). Which model do you think is the best? Explain your answer.

[5 marks]

(25 marks)

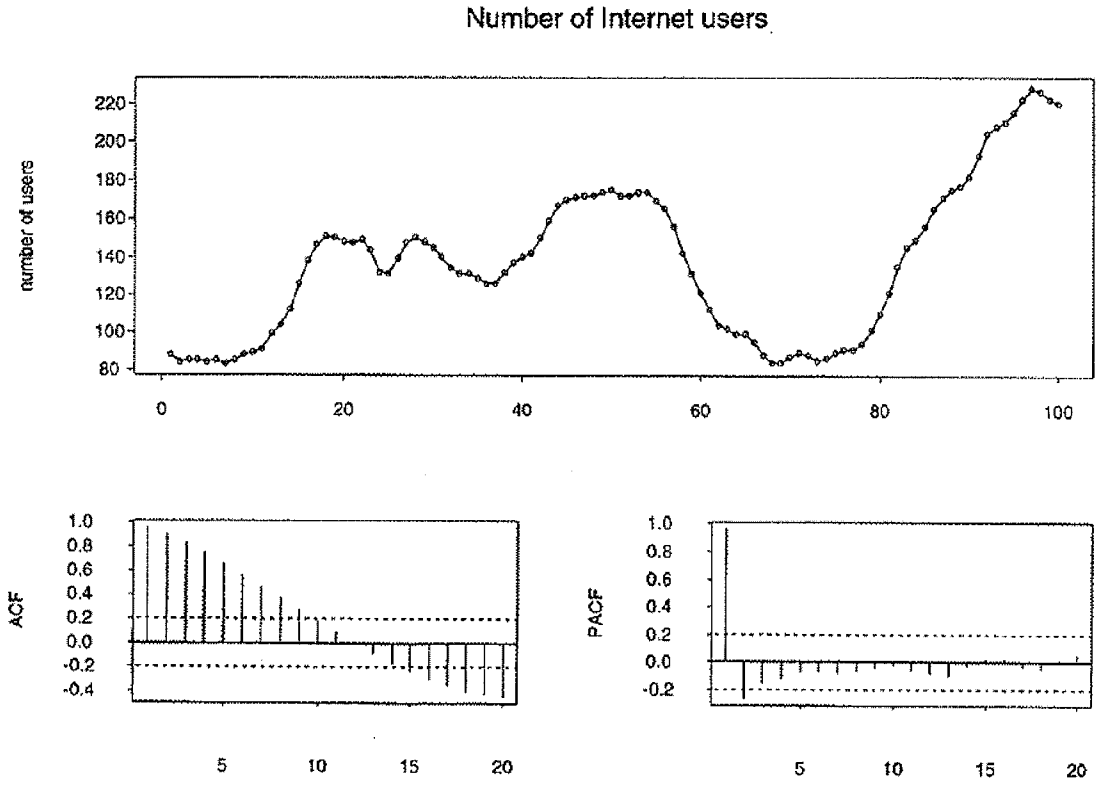


Figure 1: Time series: Number of internet users logged onto the internet server each minute over a 100-minute period.

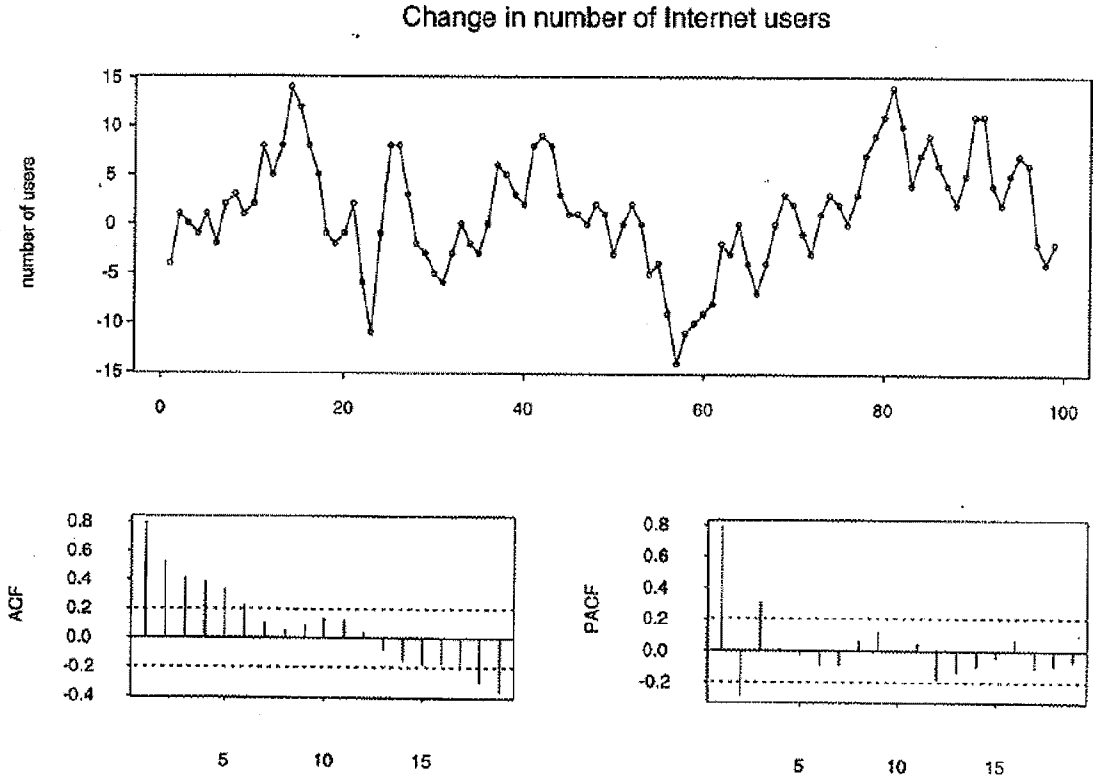


Figure 2: Time series: First differences of Number of internet users.

	$q=0$	$q=1$	$q=2$	$q=3$	$q=4$	$q=5$
$p=0$	629	548	518	518	517	516
$p=1$	527	512	514	512	513	514
$p=2$	520	514	-	-	-	-
$p=3$	509	512	514	515	-	-
$p=4$	511	510	-	-	-	-
$p=5$	513	-	-	-	-	-

Table 1: AIC values for models of the form $ARIMA(p,1,q)$ fitted on the time series *Number of internet users.*

6. **Forecasting and model selection.**

(a) We fit the model $ARIMA(1,1,1)(1,1,1)_4$ to a time series X_t. Using the backshift operator, this can be written as:

$$\underbrace{(1-\phi_1 B)}_{(1)}\ \underbrace{(1-\psi_1 B^4)}_{(2)}\ \underbrace{(1-B)}_{(3)}\ \underbrace{(1-B^4)}_{(4)}\ X_t = \underbrace{(1-\theta_1 B)}_{(5)}\ \underbrace{(1-\Theta_1 B^4)}_{(6)}\ \epsilon_t$$

Explain all the terms (1) to (6) in the model.

[6 marks]

(b) Redefine the model $ARIMA(1,1,1)(1,1,1)_4$ algebraically (without the backshift operator).

[6 marks]

(c) Holt's Linear method (i.e. Double Exponential Smoothing method) is defined as:

Init: $L_1 = X_1, \quad b_1 = X_2 - X_1, \quad F_1 = X_1$ and choose $0 < \alpha < 1$ and $0 < \beta < 1$

Compute and Forecast:

$$\begin{aligned} L_t &= \alpha\ X_t + (1-\alpha)\ (L_{t-1} + b_{t-1}) \\ b_t &= \beta\ (L_t - L_{t-1}) + (1-\beta)\ b_{t-1} \\ F_{t+1} &= L_t + b_t \end{aligned}$$

Until no more observation X_t are available

(i) What types of time series can be well modelled by the Holt's Linear method?

(ii) Calculate the level and slope of the series in table 2 using the Holt's linear method, with $\alpha = 0.5$ and $\beta = 0.1$. Compute, at each point, the 1-step ahead forecast $F_{t+1} = L_t + b_t$.

[7 marks]

t	X_t	L_t	b_t	$F_t = L_{t-1} + b_{t-1}$	$X_t - F_t$
1	3	3	1	3	0
2	4			4	0
3	2				
4					

Table 2: Fill in the blanks with the Holt's linear method in your answer book.

(d) Explain what criteria can be used to select the best Holt's Linear method.

[6 marks]

(25 marks)