MA4128 Advanced Data Modelling

Kevin O'Brien

August 31, 2013

Practical Examination Paper

- Date : August 31, 2013(10:00 am),
- The duration of this exam is 75 minutes.
- Answer all questions.
- **Write all your answers in this script**, and return it at the end of the examination.
- This exam is worth 30%.
- The exam is an 'open book' exam. Candidates may use web resources, with the exception of social media websites and sites where dynamic interaction is facilitated.
- There is a table of contents on the next page.

Contents

1 MANOVA and Discriminant Analysis [6 Marks]

The data give growth measurements on Tammar wallabies (Macropus eugenii). Each line is a set of measurements on an animal at a particular time. Most lengths are in tenths of millimetres. The data set for this exercise is `Wallaby3.sav`
The **Fixed Factor** (categorical variable) is:

- Loca : Location of animal (G,Ha,Hb,K,X)

For this exercise, the **Dependent Variables** are ***Ear, Arm, Leg, Weight*** :

- Head : Head length,
- Ear : Ear Length,
- Arm : Arm length,
- Weight : Weight in grams.

Questions:

a. Compute the significance value of Wilk's Lambda. Interpret this value [1 Mark]

b. For each of the four dependent variables, state the significance value for the test of between-subject effects and interpret these value [1 Marks]

c. Use Tukey's Post Hoc test to determine the difference of mean Arm measurement between locations ***Ha*** and ***Hb*** , also stating the confidence interval for this difference [1 Mark]

d. Use the LSD Post Hoc test to determine the difference of mean Leg measurement between the locations ***G*** and ***X***, and the corresponding significance value. Interpret this significance value [1 Mark]

e. Sketch all four of the profile plots. Include in your sketches the values of the mean for each location. [2 Marks]

2 Linear Models [6 Marks]

Data set : Water Usage of Production Plant ***water.sav***

A production plant cost-control engineer is responsible for cost reduction. One of the costly items in his plant is the amount of water used by the production facilities each month. He decided to investigate water usage by collecting seventeen observations on his plant's water usage and other variables. The variables are listed below(including whether or not they are independent or dependent (IV and DV respectively).

Variable	Description
Temperature (IV)	Average monthly temperate (F)
Production (IV)	Amount of production (M pounds)
Days (IV)	Number of plant operating days in the month
Persons (IV)	Number of persons on the monthly plant payroll
Water (DV)	Monthly water usage (gallons)

a. Compute the Pearson correlation coefficient for the variables Temperature and Production. [1 Mark]

b. Using all independent variables, write down the regression equation [1 Mark]

c. Compute the adjusted R-squared value for this model [0.5 Marks]

d. Write down the confidence intervals for each of the regression coefficients. State whether the coefficient is 'significant' or 'non-significant'. [1 Mark]

e. Using Forward Selection, write down in order the Independent Variables that get selected for the final regression model [0.5 Mark]

f. Using all variables selected by forward selection, write down the regression equation [1 Mark]

g. Using Backward Selection, write down in order the Independent Variables that get selected for the final regression model [1 Mark]

3 Dimensionality Reduction Techniques [6 Marks]

The data set lists estimates of the percentage of body fat determined by underwater weighing and various body circumference measurements for 252 men. For this exercise, use the data set ***Body.sav***. Use only the numeric variables in your solution. Performing a Principal Component Analysis procedure, answer the following questions.

```
Dens  Density determined from underwater weighing
Fat  Percent body fat from Siri's (1956) equation
Age  Age (years)
Wgt  Weight (lbs)
Hgt  Height (inches)
Neck  Neck circumference (cm)
Chest  Chest circumference (cm)
Abdo  Abdomen 2 circumference (cm)
Hip  Hip circumference (cm)
Thi  Thigh circumference (cm)
Knee  Knee circumference (cm)
Anl  Ankle circumference (cm)
Bic  Biceps (extended) circumference (cm)
Farm  Forearm circumference (cm)
Wrist  Wrist circumference (cm)
```

a. Implement an un-rotated solution and write down all eigenvalues greater than 1. Perform varimax rotation and write down all eigenvalues greater than 1. [2 Mark]

b. State the Kaplan Meyer Olkin (KMO) Measure of Sampling Adequacy. Interpret this value. [1 Mark]

c. Draw the Scree-plot. Indicate the relevant locations of the plot with labelled arrows [0.5 Marks]

d. For both the unrotated and rotated solutions, what percentage of variance explained for the first three principal components? [0.5 Marks]

e. For the unrotated solution, which observed variables are most associated with principal component 4. Explain your answer [1 Mark]

f. For the rotated solution, which observed variables are most associated with Principal component 3. Explain your answer [1 Mark]

4 Logistic Regression [6 Marks]

4.1 Part 1

For this exercise : The Data Set is `PostNatal.sav`. These are data based on a 5% sample of all births occurring in Philadelphia in 1990.

- Minority = Mother is from an ethnic minority (1=yes, 0=no),
- educ = Mother's years of education (0,17),
- smoke = Whether mother smoked during pregnancy (1=yes, 0=no),
- gestate = Gestational age in weeks, and
- grams = Birth weight in grams.
- Care = The baby will need Post Natal Care.

For this exercise ***Care*** is the dependent variable.
The independent variables are `Minor,educ,smoke,gestate`.
Do not use `grams`.

a. Write down the classification table. What is the overall percentage of correct predictions? [1 Mark]

b. What is the Odds ratio for the variable ***smoke*** ? [1 Mark]

c. Using forward selection based on the likelihood ratio test, state the independent variable(s) used in the model (and the order in which they are selected, if appropriate). [1 Mark]

4.2 Part 2

In this exercise use the data set `StepLog.sav`

Disregard the ID variables. Use the variables `STA` as the depedent variable, and select all other variables for construction of logistic regression model.

d. Using forward selection based on the likelihood ratio test, state the independent variables used in the model, and the order in which they are selected. [1 Mark]

e. For the intermediate and final models, write down the Cox and Snell, and the Nagelkerke R square values. [1 Mark]

f. Using backward selection based on the likelihood ratio test, state the first five independent variables to be removed from the model, and the order in which they are removed [1 Mark]

5 Cluster analysis [6 Marks]

5.1 Hierarchical Cluster analysis

This exercise concerns acid rain (use `AcidRain.sav`.) The columns in the data set represent the precipitation weighted mean concentrations of ions for the year 1986, for 47 sites in the United Kingdom. The data set is structured as follows:

- Site : Site number
- Rain : Rain (measured in mm)
- H : H+
- SO4 : SO4-2
- NO3 : NO3-
- NH4 : NH4+
- x : x-coordinate (measured in cm)
- y : y-coordinate (cm)

Perform a hierarchical clustering solution, based on the Euclidean Distance with Ward's Linkage, using the variables ***Rain, H, SO4 NO3, NH4*** only. Standardize all values using 'Z-standardization'. Label your cases using the ***Site*** variable.

a. What is the euclidean distance between sites 4 and 5, and sites 1 and 6, from the proximity matrix.[1 Mark]

b. Which two sites are the first pairing to be linked? Simply states the site numbers. [0.5 Mark]

c. Determine the cluster membership of site 12 based on a 5 cluster solution. [0.5 Marks]

d. Sketch the dendrogram (you may aggregate groups of sites to simplify the sketch i.e. treat several sites as one observation) [2 Marks]

5.2 K means Clustering

Thsi exercise concerns geographic and economic characteristics of asian countries. For this exercise, use **asia2.sav**. Use the following variables only for this exercise: `LogArea LogPop, LogGDP`. Label your cases by country.

e. For a 4 cluster solution, write down the initial cluster centres for the `LogGDP` variable.[1 Mark]

f. For a 4 cluster solution, write down the numbers assigned to each cluster. [0.5 Mark]

g. To which cluster does Australia belong, based on a 5 cluster solution [0.5 Mark]