May 18, 2013

1 MANOVA and Discriminant Analysis (6 Marks)

The data set for this exercise is ***cicada.sav***

- The Fixed Factor (categorical variable) is **Species**
- The dependent variables are **BW**, **WL** and **BL**. (Do not use WW).

Questions:

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

b. For each of the three dependent variables, state the significance value for the test of between-subject effects and interpret these value [3 Marks]

c. Use Tukey's Post Hoc test to determine the difference of mean BW values between Tredecim and Tredicula, also stating the confidence interval for this difference [1 Mark]

d. Use LSD Post Hoc test to determine the difference of mean WL values between Tredecim and Tredicula, and the corresponding significance value. Interpret this significance value [1.5 Marks]

e. Sketch all three of the profile plots. Include in your sketchs the values of the mean for each category. [1.5 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 (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. What is the mean and standard deviation of observed values for the Production variables? [1 Mark]

b. Compute the Pearson correlation coefficient for the variables Days and Production. [1 Mark]

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

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

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

f. Determine the Variance Inflation Factor and Tolerance for each Independent Variable [0.5 Marks]

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

h. Write down the adjusted R-squared for all intermediate and final regression models [1 Mark]

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

j. 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]

For this exercise, use the data set ***Police.sav***. Use only the numeric variables in your solution. Performing a Principal Component Analysis procedure, answer the following questions. (N.B. Some of the sample questions are not tractable for this particular data set, but are included so as to give a sense of what might come up).

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. (Possible Question for other data sets) Write down the KMO statistic. Interpret this statistic. [1 Marks]

c. Draw the Scree-plot. Indicate the relevant locations of the plot with labelled arrows [1 Mark]

d. What percentage of variance explained for the first four principal components? [1 Mark]

e. For the unrotated solution, which observed variables are most associated with Principal components 4 and 5. Explain your answer [1 Mark] (Hint: suppress values lower than 0.3 in Component Matrix)

f. For the rotated solution, which observed variables are most associated with Principal component 3. Explain your answer [1 Mark] (Hint: suppress values lower than 0.3 in Component Matrix)

g. (Possible Question for other data sets) Implement a solution using ***Principal Axis Factoring***. Write down the eigenvalues, restricting to values greater than one. [1 Mark]

h. (Possible Question for other data sets) Interpret the output for the Bartlett Test for sphericity

4 Logistic Regression [6 Marks]

The Data Set is `UScrime.sav`.

Criminologists are interested in the effect of punishment regimes on crime rates. This has been studied using aggregate data on 47 states of the USA for 1960. The data set contains the following columns:

Variable	Description
M	percentage of males aged 1424 in total state population
So	indicator variable for a southern state
Ed	mean years of schooling of the population aged 25 years or over
Po1	per capita expenditure on police protection in 1960
Po2	per capita expenditure on police protection in 1959
LF	labour force participation rate of civilian urban males in the age-group 14-24
M.F	number of males per 100 females
Pop	state population in 1960 in hundred thousands
NW	percentage of nonwhites in the population
U1	unemployment rate of urban males 1424
U2	unemployment rate of urban males 3539
Wealth	wealth: median value of transferable assets or family income
Ineq	income inequality: percentage of families earning below half the median income
Prob	probability of imprisonment: ratio of number of commitments to number of offenses
Time	average time in months served by offenders in state prisons before their first release
Crime	crime rate: number of offenses per 100,000 population in 1960

For this exercise ***So*** is the dependent variable. For parts a to c, use the first six Independent variables (i.e. M to M.F) to construct a logistic regression model.

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

b. Show how to compute the precision and recall for this table. You may write your answers as fractions only. [2 Marks]

c. What is the Odds ratio for the variable ***Po.2*** ? [1 Mark]

For the remaining parts, use the all Independent variables to construct your 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 models only, write down the odds ratio and the confidence interval for these odds ratios. [1 Mark]

f. Using backward selection based on the likelihood ratio test, state the first three 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

For this exercise, use **asia.sav**. Use Squared Euclidean Distance, Wards Linkage and Z score standardization for this exercise.

- What is the squared euclidean distance between Burma and Bangladesh, from the proximity matrix.[1 Mark]
- Which two countries are the first pairing to be linked? Name the countries and the case numbers. [1 Mark]
- Determine the cluster membership of China based on a 5 cluster solution. Name the only other country in this cluster [1 Marks]
- Sketch the Dendrogram (you may aggregate groups of countries to simplify the sketch i.e. treat several countries as one observation) [2 Marks]

5.2 K means Clustering

For this exercise, use **places.sav**. Do not use standardization for this exercise.

- For a 4 cluster solution, write down the initial cluster centres for the variable recreation.[1 Mark]
- For a 4 cluster solution, write down the Final cluster centres for the variable Arts.[1 Mark]
- For a 5 cluster solution, write down the numbers assigned to each cluster. [1 Mark]
- Determine the cluster membership of case 26 based on a 4 cluster solution [0.5 Marks]
- Determine the cluster membership of case 36 based on a 5 cluster solution [0.5 Marks]