Version: May 10th 2013

Principal Component Analysis and Factor Analysis

1.a What is Dimensionality Reduction

1.b What is the KMO statistic? Describe how to interpret the KMO statistic.

1.c What is the Bartlett Test of Sphericity used for?

1.d 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?

1.e Describe how to use a Scree plot in the context of dimensionality reduction techniques.

1.f What problems occur if a principal component analysis is done on a data matrix where the columns contain measurements on very difference scales? What can be done to overcome this problem?

1.g Principal Component Analysis is a data reduction technique. Explain what this term means.

1.h The KMO is used to measure what characteristic of the data. Explain how the KMO measure should be interpreted.

1.i 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.

1.j Discuss three techniques for determining the appropriate number of principal components.

1.k In the context of principal components what is meant by orthogonality.

1.l What is the purpose of a principal component analysis?

1.m Explain the difference between PCA and factor analysis

1.n Explain what is meant by the "true" dimension of the data? How does an analyst determine the appropriate number of factors to retain. Make reference to three difference techniques

Clustering Analysis

2.a What is the difference between agglomerative and divisive clustering?

2.b What is the difference between hierarchical and non-hierarchical clustering?

2.c How do we determine the appropriate number of clusters? Give two different visualization methods that are used to display the outcome of a cluster analysis.

2.d In the context of hierarchical cluster analyses, distinguish between agglomerative clustering and divisive clustering.

2.e What is a vertical icicle plot used for? Give a brief dscription, supporting your answer with sketches.

2.f Compute the squared Euclidean distance and the Manhattan (a.k.a. city block) distance between the following points A and B.

$$\mathrm{A} = \{2, 5, 1, 7\}$$

$$\mathrm{B} = \{1, 6, 3, 6\}$$

2.g What is a dendrogram? What does does a dendrogram depict?

2.h What is the purpose of a cluster analysis?

2.i What is two-step clustering analysis?

2.j Explain why squared Euclidean Distance may be used in preference to Euclidean Distance.

2.k Compare and contrast any two linkage methods.

2.l Explain the difference between Ward's method and K-means clustering.

2.m Discriminant Analysis is very similar to Clustering Analysis, but one key difference. Explain this difference.

2.n What is the difference between a linkage method and a similarity measure? Discuss the purpose of both.

2.o Compare and contrast any two linkage methods.

2.p Discus how to determine the appropriate number of clusters.

Linear Models

3.a What is Multicollinearity? Describe the implications of Multicollinearity?

3.b Contrast the uses of Training Data, Validation Data and Testing Data.

3.c What is meant by overfitting?

3.d Describe how you would use the Variance Inflation Factor to make an assessment about multicollinearity.

3.e Describe how to use to the Akaike Information Criterion for model selection.

3.f Explain what variable selection procedures are used for.

3.g Describe the process of model validation, with reference to training, validation and testing phases.

3.h 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?

3.i State two ways in which a multiple regression technique could be affected by severe multicollinearity?

3.j Explain what variable selection procedures are used for.

3.k Compare and contrast three types of variable selection procedure.

Logistic Regression

4.a What is a logit? How can you transform it into a probability?

4.b Suppose that the probability of a success is 0.70. Compute the corresponding odds.

4.c The usual assumptions placed on the error terms in ordinary least squares regression are:

- independently distributed
- identically distributed (equal variance)
- normally distributed

Which of these assumptions are violated when dealing with binary response data? Explain briefly how each is violated.

4.d What is a dummy variable? Explain how it is used in Logistic Regression. Support your answer with an example.

4.e There are three variants for the forward selection procedures used by SPSS. Name these three.

4.f What is the Likelihood Ration Test? Describe how it used in Logistic regression.

4.g Suppose the odds of an outcome are 4. What is the probability of that outcome?

4.h Suppose the probability of an outcome is 70%. What is the odds of that outcome occuring?

4.i What is a logit? how is it computed into a probability?

4.j Suppose that, out of a samle 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

4.k What is logistic regression? How does it differ from linear regression? Under what circumstances would you use it?

Missing Data

5.a Describe three types of missing data.

5.b what is meant by multiple imputation?

5.c Compare and contrast the following types of missing data: Missing At Random, Missing Not At Random, Missing Completely at Random.

5.d Briefly describe the technique of Multiple Imputation.

5.e Discuss some of the traditional techniques for dealing with Missing Data. For each technique discuss the limitations of that technique.

5.f What is meant by missing data? Discuss the implications of Missing data in the context of a statistical analysis.

MANOVA and Discriminant Analysis

6.a The MANOVA procedure is sensitive to Multivariate Outliers. What is a multivariate outlier? Describe a method for detecting multivariate outliers.

6.b Pillai's Trace and Wilk's Lambda are two test procedures used in MANOVA, each fulfiling the same purpose. Describe the purpose of these tests.

6.c What is the purpose of a discriminant analysis? How does discriminant analysis differ from MANOVA?

6.d Explain the following terms: confusion matrix, prior probabilities cost of misclassification, and apparent error rate.

6.e Distinguish between the True Error Rate and the Apparent Error Rate

6.f What is the confusion matrix? Explain how it is interpreted.

6.g Explain why multinomial logistic regression may be used in preference to Discriminant Analysis.

6.h The apparent error rate calculated when all observations are used to construct the discriminant rules is known to underestimate the true error rate. What can be done to overcome this problem?

6.i Compare and contrast univariate and bivariate outliers. Describe how Mahalanobis Distance is used to detect bivariate outliers. Support your answer with an illustration.

1 Logistic Regression

What is the Logit Function
The logit function that you use in logistic regression is also known as the link function because it connects, or links, the values of the independent variables to the probability of occurrence of the event defined by the dependent variable.

$$\text{logit}[E(Y)] = \log(\frac{p_i}{1-p_i}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n$$

Give a brief description of the purpose of the Cox-Snell R-square statistics

- In standard regression, R (or R squared in particular) gives you an idea of how powerful your equation is at predicting the variable of interest. An R close to 1 is a very strong prediction, whereas a small R, closer to zero, indicates a weak relationship.
- There is no direct equivalent of R for logistic regression.
- However, to keep people happy who insist on an R value, statisticians have come up with several R-like measures for logistic regression. They are not R itself, R has no meaning in logistic regression.
- Some of the better known pseudo-R measures are:
  - Cox and Snell's R-Square
  - Pseudo-R-Square
  - Hagle and Mitchell's Pseudo-R-Square

2 Supervised Learning

What is the difference between supervised and unsupervised learning?

- The difference is that in supervised learning the 'categories' are known.
- In unsupervised learning, they are not, and the learning process attempts to find appropriate 'categories'. In both kinds of learning all parameters are considered to determine which are most appropriate to perform the classification.
- Whether you chose supervised or unsupervised should be based on whether or not you know what the 'categories' of your data are.
- If you know, use supervised learning. If you do not know, then use unsupervised.

Give an example of a supervised learning methodology and an unsupervised learning methodology

- Explain how the Akaike Information Criterion would be used in the context of model selection.
- Describe the type of modelling problems that you would use Logistic Regression for.
- Discuss some of the traditional technique for dealing with Missing Data, making references to the limitations of each.