Chapter 1: A detailed overview of multivariate methods

For an overview of key terms, refer to page 1-3
Multivariate analysis = All statistical techniques that simultaneously analyze multiple measurements on individual objects under investigation. To be considered multivariate, all variables must be random and interrelated in such ways that their different effects cannot meaningfully be interpreted separately. The multivariate character lies in the multiple variates and not only in the number of variables observed.

- **Factor analysis** = identifies the underlying structure of a set of variables
- **Discriminant analysis** = Differentiates among groups based on a set of variables

The basic concepts

- **The variate** = A linear combination of variables with empirically determined weights. Variables are determined by the researcher, the weights by the multivariate technique. The result is a single value representing a combination of the entire set of variables that best achieves the objective of the specific multivariate analysis.
- **Measurement scales** = A researcher cannot identify variation unless it can be measured. Data can be divided into 2 types: metric and nonmetric.
    1. **Nonmetric data** = These are qualitative measures. They describe differences in kind by indicating the presence or absence of a characteristic of a property. In this category we have
       1. **Nominal scales** = these provide the number of occurrences in each class
       2. **Ordinal scales** = Here we see an order and we can rank the classes, but the distance between the classes is unknown.
    2. **Metric Data** = These are quantitative measures and describe differences in the amount/degree in a particular attribute.
       In this category we have **interval and ratio scales**. These scales are very similar because there are constant units of measurement. The only difference is that interval data have an arbitrary zero point, where ratio scales have an absolute zero point.

The impact of choice of measurement scales
Understanding the difference in the different types of measurement scales is important for 2 reasons:
1. The researcher must identify the measurement scale of each variable used.
2. It is important for determining which multivariate techniques should be used.

Measurement error = The degree to which the observed values are not representing the true values. There are 2 important characteristics of a measure:

- **Validity** = Does the measure represent what it is supposed to?
- **Reliability** = The degree to which the observed variable measures the true value and is free of error

To reduce the measurement error there are multivariate measurements (summated scale) in which several variables are joined in a composite measure to represent a concept.

Statistical significance versus statistical power
Interpreting statistical inferences requires the researcher to specify the acceptable levels of statistical error that result from using a sample.

Type I error = The probability of rejecting the null hypothesis when it is actually true. This is the most common approach to determine the level of statistical error and is also called ‘false positive’, and alpha α. Alpha sets the acceptable limits for error
Type II error = The probability of not rejecting the null hypothesis when it is actually false.
Power = Probability of correctly rejecting the null hypothesis when it should be rejected

		Reality	
Statistical Decision		No difference	Difference
	H0; No difference	1 – α	B ; Type II error
	H1; Difference	α ; Type I error	1 – β ; Power

Power is determined by 3 factors:
1. **The effect size** = Helps determining whether or not the observed relationship is meaningful
2. Alpha (α); the more restrictive alpha, the less power
3. Sample size; the larger the sample size, the greater the power

The relationship between these variables is shown in table 1-1.

A classification of multivariate techniques, based on 3 judgments
1. Can the variables be divided into independent/ dependent classifications based on some theory?
2. If they can, how many variables are treated as dependent in a single analysis?
3. How are the variables measured?

Dependence technique = The dependent variable is explained by other variables.
Interdependence techniques = Variables cannot be classified as either dependent or independent, but all variables are analyzed simultaneously in order to find an underlying structure to the entire set of variables/ subjects.
The structure of the variables should be analyzed >> Factor analysis and confirmatory factor analysis
The variables should be grouped to represent structure >> Cluster analysis

	# of dependent variables		
Type of measurement scale		Single	Multiple
	Metric	Multiple regression, conjoint analysis	Multivariate analysis of variance (if IV is nonmetric), Canonical (if IV is metric)
	Non-metric	Multiple discriminant analysis, linear probability	Dummy variable coding -> canonical

Types of multivariate techniques:

- Principal components & Common factor analysis
  To analyze interrelationships among a large number of variables and to explain these variables in terms of their common underlying dimensions. The objective is to find a way of condensing the information contained in a number of original variables into a smaller set of variates with minimum information loss.
- Multiple regression & Multiple correlation
  The objective is to predict changes in the dependent variable in response to changes of the independent variable.
- Multiple discriminant analysis & Logistic regression
  Applicable in situations in which the total sample can be divided into groups based on a nonmetric dependent variable characterizing several known classes. The main objective is to understand group differences and predict the likelihood that an entity belongs to a particular group. It might be used to distinguish innovators from non-innovators according to demographic and psychographic profiles. Logistic regression models are a combination of multiple regression and multiple discriminant analysis. These are similar to MRA, but here the dependent variable is non-metric.
- Canonical correlation analysis
  The objective is to correlate simultaneously a single metric dependent variable with several metric independent variables.
- Multivariate analysis of variance & Covariance
  This method can be used to explore the relationship between several categorical independent variables and 2 or more dependent, metric dependent variables. MANCOVA can be used in conjunction with MANOVA to remove the effect of any uncontrolled metric independent variables.
- Conjoint analysis
  This method is allowing for the evaluation of complex products while maintaining a realistic decision context for the respondent.
- Cluster analysis
  This method is developing meaningful subgroups of individuals or objects the objective is to classify a sample of entities into a smaller number of mutually exclusive groups based on similarities among entities. The groups are not predefined. There are 3 steps:

1. Measurement of some form of similarity/ association among entities to determine the number of groups
2. The actual clustering process
3. Profile variables to determine their composition

- Perceptual mapping, or multidimensional scaling

This method is used to transform consumer judgments of similarity/ preference into distances represented in a multidimensional space.

- Correspondence analysis
  This method facilitates perceptual mapping

1. Contingency tables
2. Nonmetric data transformed to metric
3. Dimensional reduction
4. Perceptual mapping

- Structural equation modelling & Confirmatory factor analysis
  SEM allows separate relationships for each set of dependent variables, there are 2 components:
  1. **Structural model** = The path model, which relates to the independent and the dependent variables.
  2. **Measurement model** = This enables the researcher to use several variables for a single independent/dependent variable
     In a confirmatory analysis the researcher can assess the contribution of each scale item as well as to incorporate how well the scale measures the concept.

Guidelines for multivariate analysis and interpretation.

- Establish practical significance as well as statistical significance. There should be a focus on the practical side: what are the implications?
- Recognize that sample size affects all results. For a small sample size, multivariate analysis may lead to too little statistical power to identify statistical results or too easily over fitting of the data. A similar impact occurs with too large sample sizes.
- Know your data. There is a tendency to accept the results without typical examination one undertakes with univariate analysis.
- Striving for model parsimony. The researcher must avoid inserting variables indiscriminately and letting the multivariate technique sort out relevant variables with for 2 reasons:
  1. Irrelevant variables usually increase the ability to fit sample data, but at the expense of over fitting the sample data and making results less generalizable.
  2. Irrelevant variables mask true effects due to multicollinearity. This is the degree to which any variable’s effect can be predicted by the other variables
- Look at your errors
- Validate your results
  1. Splitting the sample and using one subsample
  2. Gathering a separate subsample
  3. Employing a bootstrapping technique

A structured approach to multivariate model building

A six-step approach is presented

1. Define the research problem, objectives and multivariate technique
2. Develop an analysis plan
3. Evaluate the assumptions underlying the multivariate technique
4. Estimate the multivariate model and assess overall model fit, meet specific characteristics of the data or maximize the fit to the data
5. Interpret the variates
6. Validate the multivariate model

These steps should be displayed in a flowchart. It consists of 2 sections:
Stage 1-3 = Section 1 >> Issues addressed when preparing for actual model estimation
Stage 4-6 = Section 2 >> Model estimation, interpretation and validation