Chapter 4

Multivariate Techniques in Social Sciences

Multivariate analysis is essentially the statistical process of simultaneously analysing multiple independent variables with multiple dependent (outcome or criterion) variables using matrix algebra. During the last two or three decades, multivariate statistical analysis has become increasingly popular. The theory has made great progress, and with the rapid advances in computer technology, routine applications of multivariate statistical methods are implemented in several statistical software packages. The traditional approach to the teaching of multivariate statistical analysis, as exemplified by Anderson (1958), relies heavily on advanced matrix mathematics. On the other hand, Hury and Riedwyl (1988) suggest that it is possible to understand most of the basic ideas underlying multivariate statistical analysis without a mastery of such mathematics, provided that these are conveyed with the help of real data sets. In this chapter we propose a non-mathematical data-driven approach for teaching multivariate statistical methods to students. Despite this, we are mindful of the need for students to know some basic linear algebra and univariate statistical concepts. Such basic knowledge provides students with the foundation

necessary for the application of the appropriate multivariate statistical procedures and for the interpretation of results

4.1 Basic Research Question

A critical aspect of conducting good research involves the type of design used and the type of statistical analysis. These flow from the basic research question asked.

1. Degree of relationship between the variables. Bivariate correlation and regression (simple, multiple, and multivariate/canonical) are used to 1) explain the association between variables or 2) predict the value of one or more criterion variables knowing the value of one or more predictor variables. Path analysis is used estimate direct and indirect causal relationships between variables which are observed (directly measured) [note: structural equation modelling and confirmatory factor analysis are similar but also include latent (or unobserved) variables].

2. Measure significant differences between group means. MANOVA (multivariate analysis of variance) is used when group differences are measured on two or more dependent variables that are related to one another in the real world (e.g., depression and anxiety). Like ANOVA, MANOVA controls for Type I errors if two or more ANOVAs are run, and different kinds of MANOVA analyses can be conducted: one-way MANOVA, factorial MANOVA, repeated-measures MANOVA, and MANCOVA (using one or more covariates in controlling variance).

3. Predicting membership in two or more groups from one or more variables. Outcome studies involving the classification or prediction of subjects into two groups (e.g., success/failure, yes/no, hire/don't hire, etc.)

are conducted through logistic regression. When subjects are classified into three or more groups, Discriminant analysis is used. Predictive discriminant analysis (PDA) is used to predict group membership, while descriptive discriminant analysis (DDA) is used to explain the best linear combination of dependent variables that maximizes group differences. Logistic analysis can use categorical predictors without difficulty, but in discriminant analysis, categorical predictors pose some problems but they are not insurmountable.

4. Explaining underlying structure. Often researchers infer that variables may be clustered together based on their common correlations which implies that they are correlated because they are representing some common underlying trait or factor. Principal components analysis (PCA) and common factor analysis (CFA) reduce the variables into smaller subsets based on shared variance.

4.2 MULTIVARIATE TECHNIQUES:

4.2.1 MULTIPLE REGRESSIONS:

Multiple regression is the most commonly utilized multivariate technique. It examines the relationship between a single metric dependent variable and two or more metric independent variables. The technique relies upon determining the linear relationship with the lowest sum of squared variances; therefore, assumptions of normality, linearity, and equal variance are carefully observed. The beta coefficients (weights) are the marginal impacts of each variable, and the size of the weight can be interpreted directly. Multiple regression is often used as a forecasting tool. Goal is to use the linear composite of two or more continuous and/or categorical

variables (predictors) to: 1) predict scores on a single continuous variable (criterion), or to 2) explain the nature of the single continuous criterion variable from what is known about the predictor variables. In prediction, the criterion is the main emphasis because decisions are made on its value, but often times, it is difficult to directly measure or obtain a subject's actual score on the criterion; therefore, it is important to estimate or predict one's criterion score based on the value of the predictor scores.

4.2.2. LOGISTIC REGRESSION:

Logistic regression is used to predict a categorical (usually dichotomous) variable from a set of predictor variables. With a categorical dependent variable, discriminant function analysis is usually employed if all of the predictors are continuous and nicely distributed; logit analysis is usually employed if all of the predictors are categorical; and logistic regression is often chosen if the predictor variables are a mix of continuous and categorical variables and/or if they are not nicely distributed (logistic regression makes no assumptions about the distributions of the predictor variables). Logistic regression has been especially popular with medical research in which the dependent variable is whether or not a patient has a disease. For a logistic regression, the predicted dependent variable is the estimated probability that a particular subject will be in one of the categories (for example, the probability that Suzie Cue has the disease, given her set of scores on the predictor variables).

In a logistic regression, odds ratios are commonly employed to measure the strength of the partial relationship between one predictor and the dependent variable (in the context of the other predictor variables). It may

be helpful to consider a simple univariate odds ratio first. Among the male respondents, 68 approved continuing the research, 47 voted to stop it, yielding odds of 68 / 47. That is, approval was 1.45 times more likely than non approval. Among female respondents, the odds were 60 / 140. That is, approval was only .43 times as likely as was non approval. Inverting these odds (odds less than one are difficult for some people to comprehend), among female respondents non approval was 2.33 times as likely as approval.

Because the outcome is dichotomous, the relationship between Y and the Xs is non-linear (the line is S-shaped rather than straight), therefore the goal is not to predict a score on Y but to predict the probability that a subject belongs to one group or another. A probability score of 1 indicates absolute certainty that the individual belongs to the target group, and a probability score of 0 indicates absolute certainty that the person does not belong to the target group. Generally a cutting score of 0.5 is used so that those with probabilities above .5 are classified into the target group (Y=1.0), and those below .5 are classified into the other group (Y=0.0). A linear composite X score is derived from the simultaneous solution of the data matrix, but becomes the exponent for the natural log *e*: $P/(1-P) = e^{a + b1X1 + b2X2 + \ldots bnXn}$ .Logistic regression analysis does not use ordinary least squares in weighting the X variables, but uses a more complex process of maximum likelihood estimation. MLE requires a greater sample size than does linear multiple regression.

Interpretation: Interpretation follows four steps.

1) The researcher must determine if the overall model is sufficient. An F test is conducted using the likelihood ratio which is based on $\chi 2$. Essentially the ratio involves the likelihood of the full model compared to the model using only the constant as a predictor.
2) One can examine a *pseudo-R2* value, which is not a true proportion of variance accounted for by the model, but is an estimated value.
3) Look at a goodness-of-fit index to determine if the model estimated by MLE fits well with the data at hand. The value is -2 (Log Likelihood) or -2LL. A perfect model would have a value of 0. The goal is to have values approaching 0 as more important predictor variables are entered into the model. The -2LL is evaluated each time predictors are added into the model. Selection procedures are similar to those used in multiple regression.
4) A classification table is given which indicates the hit rate of the model selected. It is based on a contingency table of True +, True -, False +, and False -. SPSS gives the frequency of the four contingencies plus the overall hit rate (True + plus True -). The final hit rate should be greater than that of chance alone which is represented by the base rate (the ratio of actual +/total sample size). This is often called *incremental validity*. Further analysis of hit rates calculates positive predictive values and negative predictive values.

4.2.3 CANONICAL CORRELATION:

Canonical correlation is another extension of multiple regression where rather than using a single outcome variable Y, two or more Y variables are

ISBN: 978-93-85822-53-7

predicted by two or more predictor X variables. The purpose is usually not to predict Ys from Xs but to explain the relationship between the X and Y variable sets. The analysis can be run in both directions, Xs predicting Ys and Ys predicting Xs, but the researcher is generally concerned about Xs predicting Ys where the Y set is the new hypothesized inter-relationship and the X set are the traditional predictors seen in prior research.

In Multiple regression, several X variables (predictors) are independent and only one Y variable (criterion) is dependent. In canonical correlation, there are multiple sets of X variables and multiple sets of Y variables.

Because canonical correlation involves multiple Xs and multiple Ys, shared variance exists along two or more dimensions or geometric axes (with one outcome variable in multiple regression, there is only one dimension or axis). In fact, the number of dimensions or axes is equal to the number of variables in the smaller of the two sets. For example, if the data had three X variables and four Y variables, then the number of dimensions would be three. If the data had two X variables and two Y variables, the number of dimensions would be two. Each dimension will be represented by two linear composites, one for the X set of variables and one for the Y set. The correlation between the two composites is called the *canonical correlation*, RC. It is analogous to the multiple R in multiple regression. The RC2, also labeled as the *eigen value*, indicates the proportion of variance shared between the two composites. Like multiple regression where the predictor variable with the greatest degree of shared variance with Y is entered into the model first, the two sets of X and Y variables with the greatest degree of shared variance are entered into the canonical

ISBN: 978-93-85822-53-7

correlation model first. This canonical correlation of the X and Y composite variables is called the 1st *canonical* function. After this first canonical function extracts its proportion of shared variance then a second canonical correlation is computed from the remaining variance. Because the second function does not include any of the variance of the first function, the two sets are totally uncorrelated (orthogonal or independent). Each successive canonical function will be uncorrelated with all previous canonical functions. It is important to also note that because each successive canonical function has less and less remaining variance in which to operate, each successive RC gets smaller than the previous canonical correlations.

Interpretation: Interpretation involves five steps:

1) The overall significance of the model (null: explained variance = 0) is tested by Wilk's lambda. If the model is significant (explained variance > 0), then proceed.
2) Determine which canonical functions are important enough to keep. Two indices are considered simultaneously. A $\chi 2$ based on Wilks lambda is calculated for the total set of canonical functions with the null being that all of the RC = 0. If the chi-square is significant, then one can conclude that at least the first canonical correlation is significantly greater than 0. This is because the first set has the greatest RC value. The second chi-square tests the null that all of the remaining RC = 0. At the point the chi-square becomes non-significant, the remaining chi-squares will also be non-significant. As mentioned in previous sections, statistical significance is heavily influenced by sample size, so one must also look at the RC2 value of

each canonical function. A general rule given by Pedhazur (1997) is to keep functions with RC2 > 10%.

3) As in other regression analyses, regression coefficients will be calculated for each variable in the X and Y set. These are named *canonical weights or coefficients.* These are standardized and their values will be recalculated for each successive canonical function. These can be interpreted as to their respective contribution but have the same interpretation confound seen in other regression coefficients.

4) Structure coefficients are more widely used for determining the importance of each variable. The structure coefficient is the correlation between a variable and its respective linear composite. For example, the structure coefficient for X1 is the correlation of X1 with the X linear composite. The squared structure coefficient is the amount of the linear composite variance that is explained by X1. Most researchers construct a table of structure coefficients much like the table used for factor analysis. The columns represent the canonical functions retained and the rows represent each variable. The cell data represent the structure coefficients. Like factor analysis, the researcher then underscores the structure coefficients with values greater than .45 indicating that these have significant contribution to the variance of the canonical function. Furthermore, the variables which are retained provide a description of the dimension represented in that particular canonical function. These dimensions can even be named as factors are in factor analysis.

ISBN: 978-93-85822-53-7

5) As discussed, the RC2 represents only the shared variance within a specific canonical function and does not represent the proportion explained in the total X and Y variance. A *redundancy index* is calculated for each canonical function to provide an estimate of the variance explained for the full model. A separate redundancy is calculated for the X set and the Y set. It is the average structure coefficient multiplied by the canonical correlation.

4.2.4 FACTOR ANALYSIS:

The term factor analysis refers to a set of analytical techniques designed to reduce data into smaller, meaningful groups based upon their inter-correlations or shared variance. The assumption is that those items or variables that are correlated must be measuring a similar factor or trait or construct. In the case where only a few variables are used, the researcher may be able to determine groupings by simply observing the content of each variable; however, for large data sets and/or more ambiguous items, this task would be formidable. Factor Analysis reduces the information in a model by reducing the dimensions of the observations. This procedure has multiple purposes. It can be used to simplify the data, for example reducing the number of variables in predictive regression models. If factor analysis is used for these purposes, most often factors are rotated after extraction. Factor analysis has several different rotation methods—some of them ensure that the factors are orthogonal. Then the correlation coefficient between two factors is zero, which eliminates problems of multicollinearity in regression analysis.

Factor analysis is also used in theory testing to verify scale construction and operationalizations. In such a case, the scale is specified upfront and we know that a certain subset of the scale represents an independent dimension within this scale. This form of factor analysis is most often used in structural equation modeling and is referred to as Confirmatory Factor Analysis. For example, we know that the questions pertaining to the big five personality traits cover all five dimensions N, A, O, and I. If we want to build a regression model that predicts the influence of the personality dimensions on an outcome variable, for example anxiety in public places, we would start to model a confirmatory factor analysis of the twenty questionnaire items that load onto five factors and then regress onto an outcome variable.

Factor analysis can also be used to construct indices. The most common way to construct an index is to simply sum up the items in an index. In some contexts, however, some variables might have a greater explanatory power than others. Also sometimes similar questions correlate so much that we can justify dropping one of the questions completely to shorten questionnaires. In such a case, we can use factor analysis to identify the weight each variable should have in the index.

Factor analysis is a data reduction technique that can reduce the number of items by grouping them and by examining the content of the items in each group one can determine the structure or composition of each group thereby giving a better explanation of the data. It is important to note that factor analysis is not used in prediction or explaining the relationship between different sets of variables, nor is it used to determine group

differences. The goal is to explain the underlying structure or composition of the data; therefore we are dealing only with one set of variables.

Two types of factor analysis exist. The first, *exploratory factor analysis* (EFA) is used to explore or derive the underlying factor structure of a data matrix often without regard to theory. The purpose of EFA is to determine if underlying factors exist within a data set, and if so, what those factors are. The researcher does not necessarily need to have any expectations or theory beforehand—it can simply be "exploratory." Two types of EFA are commonly used. *Principle components analysis* (PCA) tries to account for all variance among the variables/items and so includes both shared variance and unique/error variance of each variable/item. *Principle factor analysis* (PFA) accounts for only shared variance of each variable/item. Currently, most journals prefer PFA.

The second, *confirmatory factor analysis* (CFA), is used to test a priori theory. The researcher specifies what factors exist and what variables/items constitute each factor and then "orders" these parameters into a data set to determine if indeed the factors and variables/items describe or fit the data. CFA is most commonly conducted through Structural Equation Modeling (SEM). The following discussion is on factor analysis pertaining to EFA.

The main goal of factor analysis is to explain as much variance as possible in a data set by using the smallest number of factors (groupings of variables based on high inter-correlations) and the smallest amount of items or variables within each factor. Inherently one balances explained variance with simplicity. Critical to this technique is that one wants to ensure that the variance left out of the solution is primarily error variance. In EFA, the first

factor derived is a linear composite of all variables/items such that it maximizes the amount of total variance explained (or extracted)—no other linear combination will extract as much variance. The proportion of variance extracted is called the *eigen value.* A second factor which is orthogonal to the first (uncorrelated) is then derived from the remaining variance, and its eigen value will be derived. Like the first factor, the second factor will be a linear composite of all variables/items. This process continues until all variance has been extracted, and the number of extractions is equal to the number of original variables/items.

Interpretation: Interpretation involves three steps:

1) The researcher determines the number of factors to keep. Several decision rules can be employed but generally keep those factors with eigen values greater than 1.0. The reason is that any given variable has a variance equal to 1.0 (since variables are standardized, the std. dev. = 1.0, and variance is the std. dev. Squared) which means that eigen values should explain more variance than at least one variable/item. In a satisfactory EFA, the total variance of the retained eigen values should be greater than 70%.
2) Once the number of factors has been determined, the researcher then determines which variables/items "load" on each factor. This is determined by the coefficient of the variable/item. Because each factor is a linear composite of all variables/items, each variable/item will have a different coefficient for each factor. This coefficient is actually a structure coefficient since it is the correlation between the variable/item and the factor; however, it is called the factor loading

or the factor structure coefficient. The higher the coefficient, the greater the variable/item's contribution to the factor, and the square of the coefficient is the amount of variance of the factor that is explained by the variable/item. In general, the desired outcome is that each variable/item will have a large loading on only one factor and small loadings on the remaining factors.

Before actually determining the composition of each factor, the axes which represent factor dimensions can be rotated geometrically so that the new set of axes are positioned closer to their respective factor variables/items. These axes are similar to those in a Cartesian coordinate system. For example, in a two-factor model, every variable/item will have a factor loading value for Factor I and for Factor II. Most will have a relatively high loading value (>.5) on one factor and relatively small loading on the other factor (< .5). Each factor is an axis in space and the factor loading of a variable/item is the coordinate value on that axis. If variable X1 has a factor loading of .63 on FI, and .22 on FII, then its coordinate point is (.63, .22). Also, the distance of X1 from the origin can be calculated by the Pythagorean theorem (add .63 squared and .22 squared and take the square root). By rotating the axes, one can increase the value along the FI dimension and decrease the FII dimension value as long as the distance of X1 to the origin remains unchanged. This strengthens the association of X1 with FI while decreasing its association to FII. This makes interpretation of the factor structure (composition) simpler. Usually when rotating the axes, the researcher assumes that the factors are truly orthogonal and so an orthogonal rotation is used, that is the axes remain at

90 degrees to one another. This is easy to visualize in a two-factor model. If F1 is rotated 35 degrees clockwise, FII must also be rotated 35 degrees clockwise. However, sometimes the factors are somewhat correlated and so the rotation that occurs is based upon the correlation between the factors. The cosine of the angle between the axes is equal to the correlation. By the way, this rule holds true when orthogonal rotation is used because the angles are at 90 degrees and the cosine of 90 degrees is 0.

3) With the output arranged in a table where the columns are the factors and the rows are the variables/items, one then highlights or underscores the loadings under each factor that are greater than .50. Once all contributing variables on all factors have been identified, then one must determine the content of each factor and assign an appropriate name for each. This is done by analyzing the content or general theme of the variable/items that are highlighted.

4.2.5 PRINCIPAL COMPONENT ANALYSIS:

Principal component analysis (PCA) involves a mathematical procedure that transforms a number of (possibly) correlated variables into a (smaller) number of uncorrelated variables called *principal components*. The first principal component accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible.

Objectives of principal component analysis:

- To discover or to reduce the dimensionality of the data set.
- To identify new meaningful underlying variables.

Traditionally, principal component analysis is performed on the Covariance matrix or on the Correlation matrix. These matrices can be calculated from the data matrix. The covariance matrix contains scaled sums of squares and cross products. A correlation matrix is like a covariance matrix but first the variables, i.e. the columns, have been standardized. We will have to standardize the data first if the variances of variables differ much, or if the units of measurement of the variables differ. The mathematical technique used in PCA is called eigen analysis: we solve for the eigenvalues and eigenvectors of a square symmetric matrix with sums of squares and cross products. The eigenvector associated with the largest eigen value has the same direction as the first principal component. The eigenvector associated with the second largest eigen value determines the direction of the second principal component. The sum of the eigen values equals the trace of the square matrix and the maximum number of eigenvectors equals the number of rows (or columns) of this matrix.

There are two methods to help you to choose the number of components to keep. Both methods are based on relations between the eigen values.

- Plot the eigen values, If the points on the graph tend to level out (show an "elbow"), these eigen values are usually close enough to zero that they can be ignored.
- Limit the number of components to that number that accounts for a certain fraction of the total variance.

4.2.6 MULTIVARIATE ANALYSIS OF VARIANCE – MANOVA:

This technique examines the relationship between several categorical independent variables and two or more metric dependent variables. Whereas analysis of variance (ANOVA) assesses the differences between groups (by using T tests for two means and F tests between three or more means), MANOVA examines the dependence relationship between a set of dependent measures across a set of groups. Typically this analysis is used in experimental design, and usually a hypothesized relationship between dependent measures is used. This technique is slightly different in that the independent variables are categorical and the dependent variable is metric. Sample size is an issue, with 15-20 observations needed per cell. However, too many observations per cell (over 30) and the technique lose its practical significance. Cell sizes should be roughly equal, with the largest cell having less than 1.5 times the observations of the smallest cell. That is because, in this technique, normality of the dependent variables is important. The model fit is determined by examining mean vector equivalents across groups. If there is a significant difference in the means, the null hypothesis can be rejected and treatment differences can be determined. A multivariate analysis of variance (MANOVA) could be used to test this hypothesis. Instead of a univariate *F* value, we would obtain a multivariate *F* value (Wilks' Lambda) based on a comparison of the error variance/covariance matrix and the effect variance/covariance matrix. Although we only mention Wilks' Lambda*?*Here, there are other statistics that may be used, including Hotelling's trace and Pillai's criterion. MANOVA is useful in

experimental situations where at least some of the independent variables are manipulated. It has several advantages over ANOVA.

(a) By measuring several dependent variables in a single experiment, there is a better chance of discovering which factor is truly important.

(b) It can protect against Type I errors that might occur if multiple ANOVA's were conducted independently. Additionally, it can reveal differences not discovered by ANOVA tests.

However, there are several cautions as well.

(a) It is a substantially more complicated design than ANOVA, and therefore there can be some ambiguity about which independent variable affects each dependent variable. Thus, the observer must make many potentially subjective assumptions.

(b) Moreover, one degree of freedom is lost for each dependent variable that is added. The gain of power obtained from decreased SS error may be offset by the loss in these degrees of freedom.

(c) Finally, the dependent variables should be largely uncorrelated. If the dependent variables are highly correlated, there is little advantage in including more than one in the test given the resultant loss in degrees of freedom. Under these circumstances, use of a single ANOVA test would be preferable.

Assumptions:

Normal Distribution: - The dependent variable should be normally distributed within groups. Overall, the *F* test is robust to non-normality, if the non-normality is caused by skewness rather than by outliers (outliers are values that are very low or very high as compared to the most values in the

data set). Tests for outliers should be run before performing a MANOVA, and outliers should be transformed or removed.

Linearity - MANOVA assumes that there are linear relationships among all pairs of dependent variables, all pairs of covariates, and all dependent variable-covariate pairs in each cell. Therefore, when the relationship deviates from linearity, the power of the analysis will be compromised.

Homogeneity of Variances: - Homogeneity of variances assumes that the dependent variables exhibit equal levels of variance across the range of predictor variables. Remember that the error variance is computed (SS error) by adding up the sums of squares within each group. If the variances in the two groups are different from each other, then adding the two together is not appropriate, and will not yield an estimate of the common within-group variance.

Homogeneity of Variances and Covariances: - In multivariate designs, with multiple dependent measures, the homogeneity of variances assumption described earlier also applies. However, since there are multiple dependent variables, it is also required that their intercorrelations (covariances) are homogeneous across the cells of the design. There are various specific tests of this assumption.

Special Cases

Two special cases arise in MANOVA, the inclusion of within-subjects independent variables and unequal sample sizes in cells. Unequal sample sizes - As in ANOVA, when cells in a factorial MANOVA have different sample sizes, the sum of squares for effect plus error does not equal the total sum of squares. This causes tests of main effects and interactions to be

correlated. SPSS offers and adjustment for unequal sample sizes in MANOVA.

Within-subjects design - Problems arise if the researcher measures several different dependent variables on different occasions. This situation can be viewed as a within-subject independent variable with as many levels as occasions. Or, it can be viewed as a separate dependent variables for each occasion.

Additional Limitations:

Outliers - Like ANOVA, MANOVA is extremely sensitive to outliers. Outliers may produce either a Type I or Type II error and give no indication as to which type of error is occurring in the analysis. There are several programs available to test for univariate and multivariate outliers. Multicollinearity and Singularity - When there is high correlation between dependent variables, one dependent variable becomes a near-linear combination of the other dependent variables. Under such circumstances, it would become statistically redundant and suspect to include both combinations.

4.2.7. DISCRIMINANT FUNCTION ANALYSIS:

Discriminant function analysis is used to determine which variables discriminate between two or more naturally occurring groups. For example, an educational researcher may want to investigate which variables discriminate between high school graduates who decide (1) to go to college, (2) to attend a trade or professional school, or (3) to seek no further training or education. For that purpose the researcher could collect data on

numerous variables prior to students' SSLC. After SSLC, most students will naturally fall into one of the three categories. *Discriminant Analysis* could then be used to determine which variable(s) are the best predictors of students' subsequent educational choice.

Stepwise Discriminant Analysis:

Probably the most common application of discriminant function analysis is to include many measures in the study, in order to determine the ones that discriminate between groups. For example, an educational researcher interested in predicting high school students' choices for further education would probably include as many measures of personality, achievement motivation, academic performance, etc. as possible in order to learn which one(s) offer the best prediction. Put another way, we want to build a "model" of how we can best predict to which group a case belongs. In the following discussion we will use the term "in the model" in order to refer to variables that are included in the prediction of group membership, and we will refer to variables as being "not in the model" if they are not included.

Forward stepwise analysis: In stepwise discriminant function analysis, a model of discrimination is built step-by-step. Specifically, at each step all variables are reviewed and evaluated to determine which one will contribute most to the discrimination between groups. That variable will then be included in the model, and the process starts again.

Backward stepwise analysis: One can also step backwards; in that case all variables are included in the model and then, at each step, the variable that contributes least to the prediction of group membership is eliminated. Thus, as the result of a successful discriminant function analysis, one would only

keep the "important" variables in the model, that is, those variables that contribute the most to the discrimination between groups.

***F* to enter, *F* to remove.** The stepwise procedure is "guided" by the respective *F* to enter and *F* to remove values. The *F* value for a variable indicates its statistical significance in the discrimination between groups, that is, it is a measure of the extent to which a variable makes a unique contribution to the prediction of group membership.

Interpreting a Two-Group Discriminant Function:

In the two-group case, discriminant function analysis can also be thought of as (and is analogous to) multiple regression. If we code the two groups in the analysis as *1* and *2*, and use that variable as the dependent variable in a multiple regression analysis, then we would get results that are analogous to those we would obtain via *Discriminant Analysis*. In general, in the two-group case we fit a linear equation of the type:

Group = a + b1*x1 + b2*x2 + ... + bm*xm

where*a* is a constant and *b1* through *bm*are regression coefficients. The interpretation of the results of a two-group problem is straightforward and closely follows the logic of multiple regression: Those variables with the largest (standardized) regression coefficients are the ones that contribute most to the prediction of group membership.

Interpreting the discriminant functions: As before, we will get *b* (and standardized *beta*) coefficients for each variable in each discriminant (now also called *canonical*) function, and they can be interpreted as usual: the larger the standardized coefficient, the greater is the contribution of the respective variable to the discrimination between groups. However, these

coefficients do not tell us between which of the groups the respective functions discriminate. We can identify the nature of the discrimination for each discriminant (canonical) function by looking at the means for the functions across groups. We can also visualize how the two functions discriminate between groups by plotting the individual scores for the two discriminant functions

Significance of discriminant functions: One can test the number of roots that add *significantly* to the discrimination between group. Only those found to be statistically significant should be used for interpretation; non-significant functions (roots) should be ignored.

Classification (Predictive Discriminant Analysis): Another major purpose to which discriminant analysis is applied is the issue of predictive classification of cases. Once a model has been finalized and the discriminant functions have been derived, how well can we *predict* to which group a particular case belongs?

***A priori* and *post hoc* predictions:** Before going into the details of different estimation procedures, we would like to make sure that this difference is clear. Obviously, if we estimate, based on some data set, the discriminant functions that best discriminate between groups, and then use the *same* data to evaluate how accurate our prediction is, then we are very much capitalizing on chance. In general, one will *always* get a worse classification when predicting cases that were not used for the estimation of the discriminant function. Put another way, *post hoc* predictions are always better than *a priori* predictions. Therefore, one should never base one's confidence regarding the correct classification of future observations on the

same data set from which the discriminant functions were derived; rather, if one wants to classify cases predicatively, it is necessary to collect new data to "try out" (cross-validate) the utility of the discriminant functions.

Classification functions. These are not to be confused with the discriminant functions. The classification functions can be used to determine to which group each case most likely belongs. There are as many classification functions as there are groups. Each function allows us to compute *classification scores* for each case for each group, by applying the formula:

Si = ci + wi1*x1 + wi2*x2 + ... + wim*xm

In this formula, the subscript *i* denotes the respective group; the subscripts *1, 2, ..., m* denote the *m* variables; *ci* is a constant for the *i*'th group, *wij*is the weight for the *j*'th variable in the computation of the classification score for the *i*'th group; *xj*is the observed value for the respective case for the *j*'th variable. *Si* is the resultant classification score. We can use the classification functions to directly compute classification scores for some new observations.

Classification of cases: Once we have computed the classification scores for a case, it is easy to decide how to classify the case: in general we classify the case as belonging to the group for which it has the highest classification score. Thus, if we were to study high school students' post-school career/educational choices (e.g., attending college, attending a professional or trade school, or getting a job) based on several variables assessed one year prior to graduation, we could use the classification functions to predict what each student is most likely to do after SSLC.

However, we would also like to know the *probability* that the student will make the predicted choice. Those probabilities are called *posterior* probabilities, and can also be computed. However, to understand how those probabilities are derived, let us first consider the so-called *Mahalanobis* distances.

Mahalanobis distances: In general, the Mahalanobis distance is a measure of distance between two points in the space defined by two or more correlated variables. For example, if there are two variables that are uncorrelated, then we could plot points (cases) in a standard two-dimensional scatter plot; the Mahalanobis distances between the points would then be identical to the Euclidean distance; that is, the distance as, for example, measured by a ruler. If there are three uncorrelated variables, we could also simply use a ruler (in a 3-D plot) to determine the distances between points. If there are more than 3 variables, we cannot represent the distances in a plot any more. Also, when the variables are correlated, then the axes in the plots can be thought of as being *non-orthogonal*; that is, they would not be positioned in right angles to each other. In those cases, the simple Euclidean distance is not an appropriate measure, while the Mahalanobis distance will adequately account for the correlations.

Mahalanobis distances and classification: For each group in our sample, we can determine the location of the point that represents the means for all variables in the multivariate space defined by the variables in the model. These points are called group *centroids*. For each case we can then compute the Mahalanobis distances (of the respective case) from each of the group

centroids. Again, we would classify the case as belonging to the group to which it is closest, that is, where the Mahalanobis distance is smallest.

Posterior classification probabilities: Using the Mahalanobis distances to do the classification, we can now derive probabilities. The probability that a case belongs to a particular group is basically proportional to the Mahalanobis distance from that group centroid. In summary, the posterior probability is the probability, based on our knowledge of the values of other variables that the respective case belongs to a particular group.

Summary of the prediction: A common result that one looks at in order to determine how well the current classification functions predict group membership of cases is the *classification matrix*. The classification matrix shows the number of cases that were correctly classified (on the diagonal of the matrix) and those that were misclassified.

4.2.8 CLUSTER ANALYSIS:

The purpose of cluster analysis is to reduce a large data set to meaningful subgroups of individuals or objects. The division is accomplished on the basis of similarity of the objects across a set of specified characteristics. Outliers are a problem with this technique, often caused by too many irrelevant variables. The sample should be representative of the population, and it is desirable to have uncorrelated factors. There are three main clustering methods: hierarchical, which is a treelike process appropriate for smaller data sets; nonhierarchical, which requires specification of the number of clusters a priori; and a combination of both. There are four main rules for developing clusters: the clusters should be different, they should be

reachable, they should be measurable, and the clusters should be profitable (big enough to matter). This is a great tool for market segmentation.

4.2.9 MULTIDIMENSIONAL SCALING (MDS):

The purpose of MDS is to transform consumer judgments of similarity into distances represented in multidimensional space. This is a decompositional approach that uses perceptual mapping to present the dimensions. As an exploratory technique, it is useful in examining unrecognized dimensions about products and in uncovering comparative evaluations of products when the basis for comparison is unknown. Typically there must be at least four times as many objects being evaluated as dimensions. It is possible to evaluate the objects with non-metric preference rankings or metric similarities (paired comparison) ratings. Kruskal's Stress measure is a "badness of fit" measure; a stress percentage of 0 indicates a perfect fit, and over 20% is a poor fit. The dimensions can be interpreted either subjectively by letting the respondents identify the dimensions or objectively by the researcher.

4.2.10 CORRESPONDENCE ANALYSIS:

This technique provides for dimensional reduction of object ratings on a set of attributes, resulting in a perceptual map of the ratings. However, unlike MDS, both independent variables and dependent variables are examined at the same time. This technique is more similar in nature to factor analysis. It is a compositional technique, and is useful when there are many attributes and many companies. It is most often used in assessing the effectiveness of advertising campaigns. It is also used when the attributes are too similar for

factor analysis to be meaningful. The main structural approach is the development of a contingency (crosstab) table. This means that the form of the variables should be nonmetric. The model can be assessed by examining the Chi-square value for the model. Correspondence analysis is difficult to interpret, as the dimensions are a combination of independent and dependent variables.

4.2.11. CONJOINT ANALYSIS:

Conjoint analysis is often referred to as "trade-off analysis," since it allows for the evaluation of objects and the various levels of the attributes to be examined. It is both a compositional technique and a dependence technique, in that a level of preference for a combination of attributes and levels is developed. A part-worth, or utility, is calculated for each level of each attribute, and combinations of attributes at specific levels are summed to develop the overall preference for the attribute at each level. Models can be built that identify the ideal levels and combinations of attributes for products and services.

4.2.12 TIME SERIES ANALYSIS: Time series has a long history in social sciences, especially in economics and finance. As it is well known, much of economics and finances are concerned with modeling dynamics and systematization of data over time was a subject that appeared early. The cumulated historical data permitted to applied statistical methods in order to find evidence of causation between social variables, finding some support to social theories. Considering the non-experimental nature of the social sciences, this also encourages the development of statistical techniques. For this reason, much of the statistical effort, in particular econometric effort

was focused on developing powerful statistical tests, considering the availability of small samples. The basic time series modeling, covering univariate time series analysis utilizing the Box-Jenkins approach, along with time series analysis within a classical regression framework, multivariate time series analysis, focusing first on vector autoregression had gained momentum over a period of time. Recently, several advanced techniques like ARIMA models, time-series regression, unit-root diagnosis, vector autoregressive models, error-correction models, intervention models, fractional integration, ARCH models, structural breaks, and forecasting has been playing a lead role in determining the dynamics of any time series data.

ARIMA (p,d,q) forecasting equation: ARIMA models are, in theory, the most general class of models for forecasting a time series which can be made to be “stationary” by differencing (if necessary), perhaps in conjunction with nonlinear transformations such as logging or deflating (if necessary). The acronym ARIMA stands for Auto-Regressive Integrated Moving Average. Lags of the stationarized series in the forecasting equation are called "autoregressive" terms, lags of the forecast errors are called "moving average" terms, and a time series which needs to be differenced to be made stationary is said to be an "integrated" version of a stationary series. Random-walk and random-trend models, autoregressive models, and exponential smoothing models are all special cases of ARIMA models. A nonseasonal ARIMA model is classified as an "ARIMA (p,d,q)" model, where:

- **p** is the number of autoregressive terms,

- **d** is the number of non-seasonal differences needed for stationarity, and
- **q** is the number of lagged forecast errors in the prediction equation.

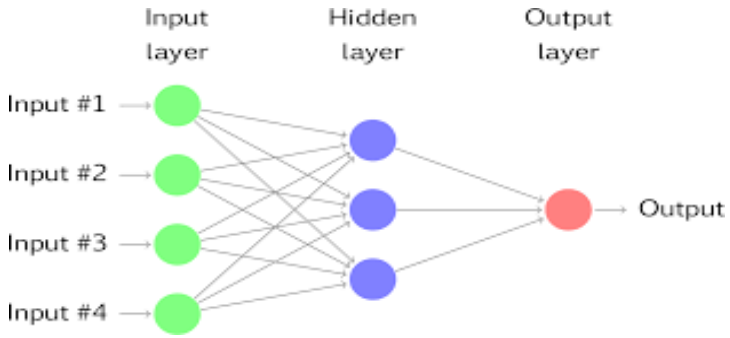
In terms of y, the general forecasting equation is:

$\hat{y}_t = \mu + \phi_1 y_{t-1} + \ldots + \phi_p y_{t-p} - \theta_1 e_{t-1} - \ldots - \theta_q e_{t-q}$

Here the moving average parameters (θ's) are defined so that their signs are negative in the equation, following the convention introduced by Box and Jenkins. Some authors and software (including the R programming language) define them so that they have plus signs instead. When actual numbers are plugged into the equation, there is no ambiguity, but it's important to know which convention your software uses when you are reading the output. Often the parameters are denoted there by AR(1), AR(2), …, and MA(1), MA(2), … etc..To identify the appropriate ARIMA model for Y, you begin by determining the order of differencing (d) needing to stationarize the series and remove the gross features of seasonality, perhaps in conjunction with a variance-stabilizing transformation such as logging or deflating. If you stop at this point and predict that the differenced series is constant, you have merely fitted a random walk or random trend model. However, the stationarized series may still have autocorrelated errors, suggesting that some number of AR terms ($p \geq 1$) and/or some number MA terms ($q \geq 1$) are also needed in the forecasting equation.

4.2.13 ARTIFICIAL NEURAL NETWORKS: ANNs are computing systems inspired by the biological neural network that constitute animal brains. Such systems learn (progressively improve performance on) tasks by considering examples, generally without task-specific programming. An

ANN is based on a collection of connected units or nodes called artificial neurons (analogous to biological neurons in an animal brain). Each connection (analogous to a synapse) between artificial neurons can transmit a signal from one to another. The artificial neuron that receives the signal can process it and then signal artificial neurons connected to it. In common ANN implementation, the signal at a connection between artificial neurons is a real number, and the output of each artificial neuron is calculated by a non-linear function of the sum of its inputs. Artificial neurons and connections typically have a weight that adjusts as learning proceeds. The weight increases or decreases the strength of the signal at a connection. Artificial neurons may have a threshold such that only if the aggregate signal crosses that threshold is the signal sent. Typically, artificial neurons are organized in layers. Different layers may perform different kinds of transformations on their inputs. Signals travel from the first (input), to the last (output) layer, possibly after traversing the layers multiple times. The original goal of the ANN approach was to solve problems in the same way that a human brain would. Over time, attention focused on matching specific mental abilities, leading to deviations from biology. ANNs have been used on a variety of tasks, including computer vision, speech recognition, machine translation, social network filtering, playing board and video games and medical diagnosis.



4.2.14 STRUCTURAL EQUATION MODELING:

Unlike the other multivariate techniques discussed, structural equation modeling (SEM) examines multiple relationships between sets of variables simultaneously. This represents a family of techniques, including LISREL, latent variable analysis, and confirmatory factor analysis. SEM can incorporate latent variables, which either are not or cannot be measured directly into the analysis. For example, intelligence levels can only be inferred, with direct measurement of variables like test scores, level of education, grade point average, and other related measures. These tools are often used to evaluate many scaled attributes or to build summated scales.

Reference

[1] Betsy J. Becker (499), College of Education, Michigan State University, Anderson. T W (1958) An Introduction to Multivariate Statistical Analysis. John Wiley & Sons, London.

[2] Borg I. & Groenen P. (1997).Modern multidimensional scaling. New York: Springer-Verlag.] Escoer, B., &Pagµes, J. (1988).Analyses factorielles multiples. Paris: Dunod.

[3] EBetsy J. Becker (499), College of Education, Michigan State University, East Lansing, Michigan 48824.

[4] Fisher. R A (1936) The use of multiple measurements in taxonomic problems. Annals of Eugenics **7.**179-188.

[5] Flury,Bernhard and Riedwyl, Hans (1988) Multivariate Statistics:A Practical *Approach.*Chapman and Hall, London.

[6] Gnanadesikan, R (1977), Methods for Statistical Data Analysis of Multivariate Observations.

[7] Gould, S J (1981) The Misrneasure of Man.W ***W***Norton & Co, New York.

[8] Johnson R.A. &Wichern D.W. (2002).Applied multivariate statistical anal-ysis. Upper Saddle River (NJ): Prentice-Hall.

[9] Kaciak, Eugene and Koczkodaj, Waldemar W (1989) A spreadsheet approach to principal components analysis. Journal of Microcomputer Applications 12**,** 281-291.

[10] LisabethF. DiLalla (439), School of Medicine, Southern Illinois University, Carbondale, Illinois 62901

[11] Mark L. Davison (323), Department of Educational Psychology, University of Minnesota, Minneapolis, Minnesota 55455

[12] Michael T. Brown (209), Graduate School of Education, University of California, Santa Barbara, Santa Barbara, California 93016

[13] Naes T., Risvik E. (Eds.) (1996). Multivariate analysis of data in sensory science. New York: Elsevier.

[14] Rao, C R (1964) The use and interpretation of principal component analysis in applied research. ***Sankya A*** 26, 329-358.

[15] ReneV. Dawis (65), University of Minnesota, Minneapolis, Minnesota 55414; and The Ball Foundation, Glen Ellyn, Illinois 60137

[16] Robert Cudeek (265), Department of Psychology, University of Minnesota, Minneapolis, Minnesota 55455

[17] Roessner, D. (2002). *Outcome Measurement in the United States: State of the Art.* Paper presented at the Annual Meeting of the American Association for the Advancement of Science, Boston, MA.

[18] Scheirer, M.A. (2000). Getting more "bang" for your performance measures "buck". *American Journal of Evaluation, 21*, 139-149.

[19] Steven D. Brown (3), Department of Leadership, Foundations, and Counseling Psychology, Loyola University of Chicago, Wilmette, Illinois 60091

[20] Weller S.C. & Romney A.K., (1990). Metric scaling: Correspondence analysis. Thou