Chapter V

Analytical Framework

5.0 Introduction

In the last chapter we discussed the research methodology followed to conduct this research work. We also discussed the relevance of the variable selection and the questionnaire design. We discussed sample selection and how the experiment was conducted by administering the questionnaire personally. This chapter discusses those techniques and concepts that are applicable to this research study. In this chapter, analytical frameworks of two multivariate statistical techniques that are applied to analyze the data of the study and related hypotheses are discussed. In section 5.1 detailed discussions on exploratory factor analysis are presented.

This is followed with discussions on multiple regression analysis in section 5.2. The theory and reasoning behind the hypotheses formulation between the relationships of organizational intelligence and causative factors are explained in the next chapter. Section 5.3 explains the conclusion of the analytical frameworks discussed in this chapter. The analytical frameworks of the above mentioned statistical techniques are discussed with their conceptual overviews and scientific context of the method. Interpretation of the findings and their significance are discussed in a theoretical perspective.

In the first level of data analysis, *Factor analysis*, a multivariate technique, is used to identify the structure of interrelationships between the variables and reveal functional units, thus forming the base of the change of variables. The Factor

Analysis model building is described in a sequential manner. The purpose of the study demands us to determine how Organizational Intelligence affects the financial performance of firms. There are various variables that account for Organizational Intelligence, as mentioned in Chapter 3. The groups of independent variables of Organizational Intelligence might be affecting the variables of Organizational Performance in multiple manners. Variables of similar behavior could be grouped by the Explanatory Factor Analysis method and the total number of variables can be reduced. This would also fulfill the requirement of adherence to one of the classical assumptions, namely Multicollinearity, for our next model, i.e., Multiple Regression. As we group the independent variables which are highly correlated, we can get rid of the multicollinearity problem, while applying Multiple Regression Analysis Technique.

In the next level, we shall estimate these grouped variables and certain unique variables (which could not be grouped under factors) as independent variables to predict Organizational Performance as captured by a single dependent variable. The equation will be a relationship model between OI (Organizational Intelligence) and OP (Organizational Performance). As mentioned earlier, Organizational Performance variables are represented by a single dependent variable and the factors of Organizational Intelligence are independent variables. Thus the second level of data analysis clearly demanded the application of *Multiple Regression Analysis*. Thus Organizational Performance factors are chosen to represent financial performance and other organic attributes of organizations are grouped suitably with factor analysis to represent organizational intelligence.

Part I

5.1 Exploratory Factor Analysis

As discussed earlier, we will first present the theoretical framework of Factor Analysis, which we have applied to reduce the problem of multicollinearity amongst the independent variables during the application of Multiple Regression. It also gives us a fair amount of justification into the process of

clubbing similar variables, or variables that behave identically. As Factor Analysis provides us with the necessary credence, the formation of the factors with proper rationale is in itself considered to be a good finding in the literature of OI and OP.

5.1.1 What is Factor analysis?

Social science often involves primary data collection using the questionnaire method. As this is a time consuming and costly affair, researchers often avoid taking the risk of having exactly the same number of questions that are required to address the variables considered in the study. Hence, it often leads to a pool of large number of variables, though the exact number of variables required for the study could be much less. Here comes the requirement of the application of Factor Analysis.

There are a large number of variables proposed, and hypotheses and theories are linked to each other to explain or describe the complex variety and interconnections of various relationships. Factor analysis can simultaneously manage more than a hundred variables, compensate for random error and invalidity, and disentangle complex interrelationships into their major and distinct regularities (Rummel, 1970)[260]. It is a good way of resolving the confusion of data complexity and identifying latent or underlying factors from an array of seemingly important variables (Nargundkar, 2004)[261].

Factor analysis techniques can achieve their objectives from either an exploratory or confirmatory perspective. *Exploratory Factor Analysis* (EFA) is useful in searching for structures among a set of variables or as a Data Reduction method. It is a widely utilized and broadly applied statistical technique in the social sciences (Osborne, 2005)[262]. Hair et al (Hair et al, 2006)[263] mention that Factor Analysis provides the tools for analyzing the structure of the interrelationships

[260] Rummel, R.J., Applied Factor Analysis, Evanston IL: Northwestern University Press; 1970
[261] Nargundkar, R., Marketing Research: Text and Cases, 2nd Ed, Tata McGraw-Hill Pb; 2004
[262] Osborne., By Best Practices in Quantitative Methods, Sage Publications Inc Pb, 2005
[263] Hair et al, Multivariate Data Analysis, 6th Ed, Printice Hall Pb; 2006

(correlations) among a large number of variables by defining sets of variables that are highly interrelated. Confirmatory factor analysis (Joreskog et al, 1993)[264] is used for analyzing the validity and reliability of actual structure of the data based on theoretical, latent constructs or prior research. In this study, EFA is discussed and termed as Factor Analysis.

The essential purpose of factor analysis is to describe, if possible, the covariant relationships among many variables in terms of few underlying, but unobservable, random quantities called *factors* (Johnson et al, 1992)[265] interpreted through weights of the variable called factor loadings, organized in a matrix of factor loadings.

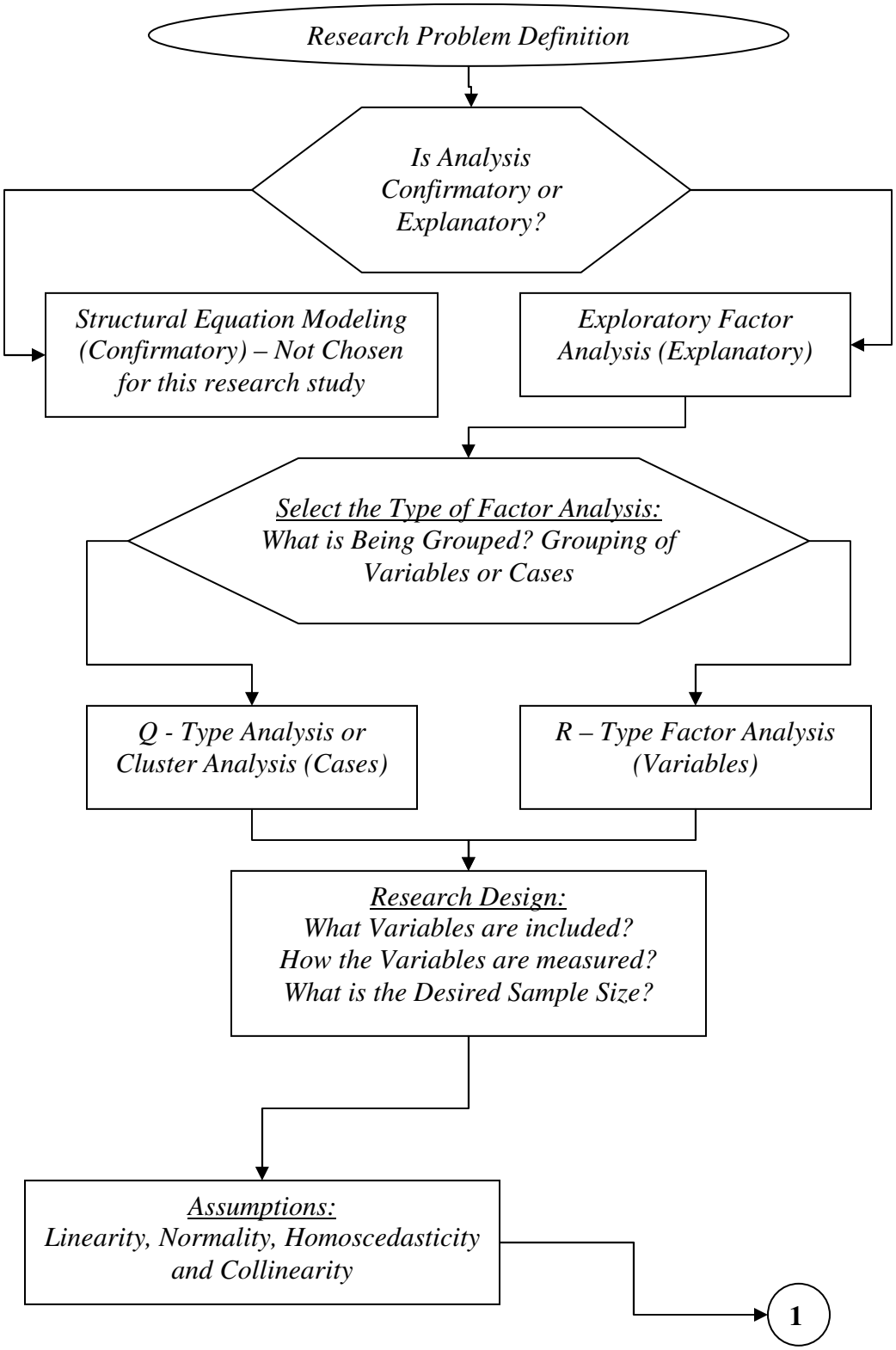
The factors, by definition, are highly inter-correlated and are assumed to represent dimensions within the data. By reducing the number of variables, the dimensions can guide in creating new composite measures (Hair et al, 2006)[266].

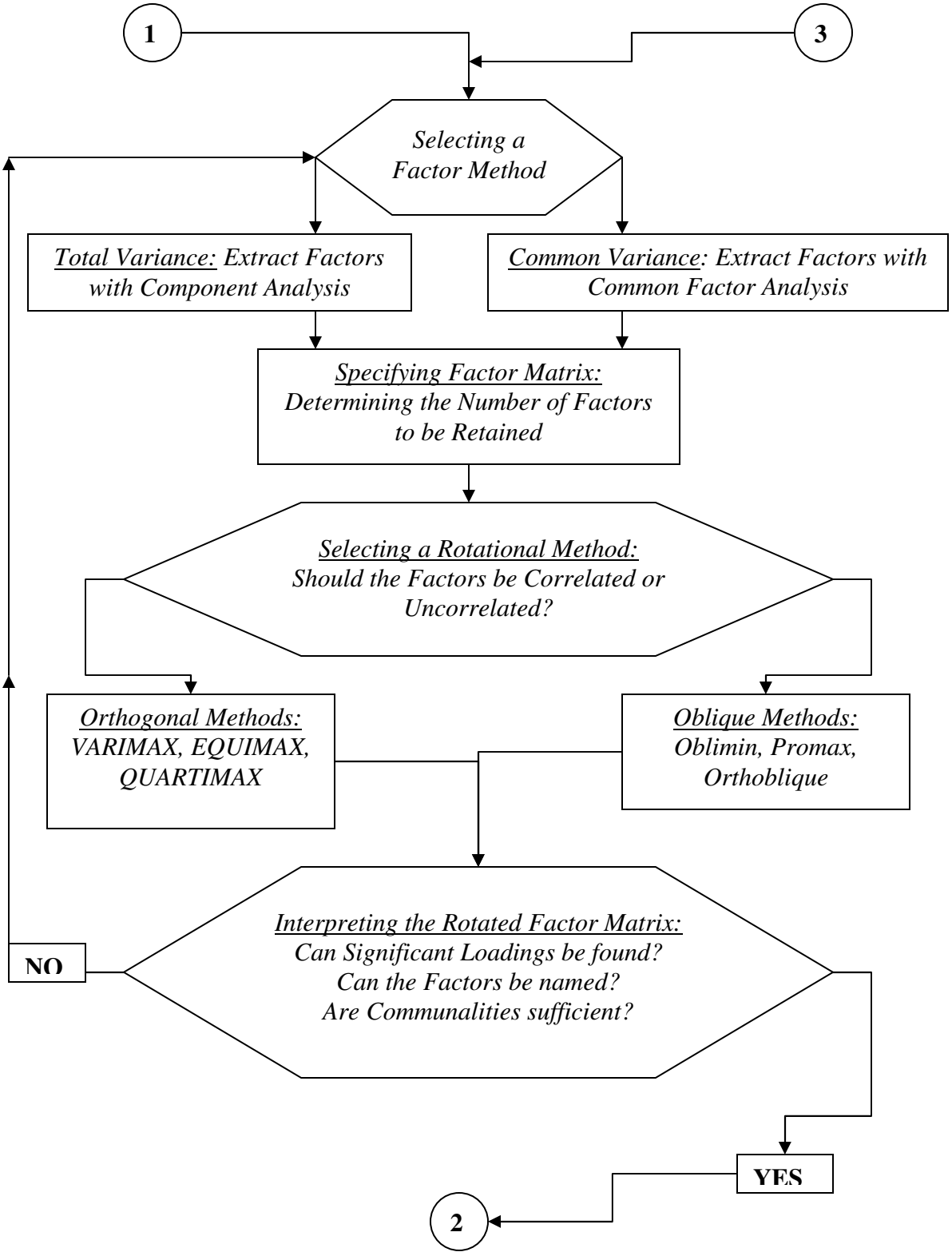
The Factor Analysis model is organized in such a way that all variables within a particular group are highly correlated among themselves, but have relatively small correlations with variables in another group. Typically, factors used for any further analysis should contain unique variables (Makhura *et al.*, 1997)[267].

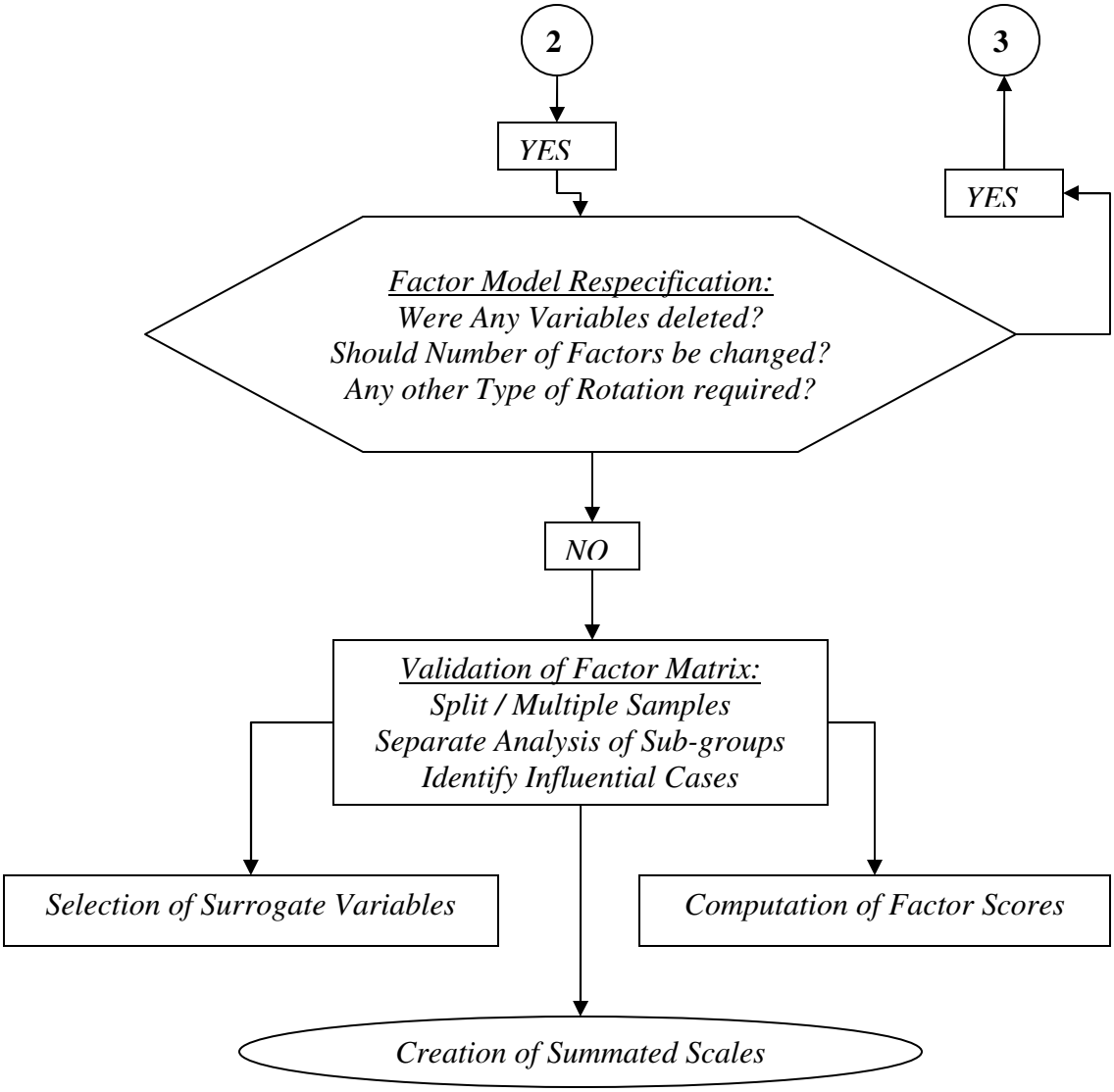
Flow chart depicting Exploratory Factor Analysis is given in Figure 5.1.

[264] Jorescog et al, LISREL8: Structural Equation Modeling with the SIMPLIS Command Language, Mooresville, IN: Scientific Software International, 1993
[265] Jhonson et al, Applied Multivariate Analysis, 4th Ed, Prentice Hall Pb, 1998
[266] Hair et al, Multivariate Data Analysis, 6th Ed, Printice Hall Pb; 2006
[267] Makhura MN & Wasike WSK (2003). Patterns of Access to Rural Service Infrastructure: The Case of Farming Households in Limpopo Province. Agrekon Vol.42(2), p129-143.

Figure 5.1 - Flow Diagram of Factor Analysis




1
3
Selecting a Factor Method
Total Variance: Extract Factors with Component Analysis
Common Variance: Extract Factors with Common Factor Analysis
Specifying Factor Matrix: Determining the Number of Factors to be Retained
Selecting a Rotational Method: Should the Factors be Correlated or Uncorrelated?
Orthogonal Methods: VARIMAX, EQUIMAX, QUARTIMAX
Oblique Methods: Oblimin, Promax, Orthoblique
Interpreting the Rotated Factor Matrix: Can Significant Loadings be found? Can the Factors be named? Are Communalities sufficient?
NO
YES
2


2
3
YES
YES
Factor Model Respecification:
Were Any Variables deleted?
Should Number of Factors be changed?
Any other Type of Rotation required?
NO
Validation of Factor Matrix:
Split / Multiple Samples
Separate Analysis of Sub-groups
Identify Influential Cases
Selection of Surrogate Variables
Computation of Factor Scores
Creation of Summated Scales

5.1.2 Why Using Exploratory Factor Analysis for this Research Study?

Appropriateness of Exploratory Factor Analysis: Exploratory Factor Analysis can be a highly useful and powerful multivariate statistical technique for effectively extracting information from large bodies of interrelated data. When variables are correlated, the researcher needs to manage these variables, by grouping highly correlated variables together, labeling or naming the groups, and by creating a new composite measure that can represent each group of variables. The primary purpose of Exploratory Factor Analysis is to define the underlying structure among the variables in the analysis. As an interdependence technique, factor analysis attempts to identify grouping among variables, based on the relationships represented in a correlation matrix. It is a powerful tool to understand the structure of the data better. It is used to simplify analyses of large set of variables by replacing them with composite variables. When it works well, it points to interesting relationships that might not have been possible from examination of the raw data alone, or even the correlation matrix. Factor analysis provides the basis for data reduction through either summated scales or factor scores. The researcher can combine the variables within each factor into a single score that can replace the original set of variables with four new composite variables.

Difference between exploratory factor analysis and confirmatory factor analysis: Factor analysis used for this research work, which is primarily an exploratory technique, does not give enough control over the specification of the structure, such as number of factors and loadings on each variable etc. However an attempt to confirm the Factors will require Structural Equation Modeling.

Seven stages of applying Factor Analysis include (i) Clarifying the objectives of factor analysis (ii) Designing a factor analysis, including selection of variables and sample size (iii) Assumptions of factor analysis (iv) Deriving factors and assessing overall fit, including which factor model to use and the number of factors (v) Rotating the interpreting factors (vi) Validation of factor analysis

solutions (vii) Additional uses of factor analysis results, such as, selecting surrogate variables, creating summated scales or computing factor scores (Hair et al, 2006)[268]

Difference between R and Q Factor Analysis: The principal use of factor analysis is to develop a structure among variables, referred to as R factor analysis. Factor analysis can also be used to group cases, which is referred as Q factor analysis. Q factor analysis is similar to cluster analysis. The primary difference is that Q Factor analysis uses correlation as the measure of similarity whereas cluster analysis is based on a more distant measure.

Difference between component analysis and common factor analysis: 3 types of variance are considered when applying factor analysis; Common Variance, Unique Variance and Error Variance. They sum up to give the Total Variance. Component Analysis (principal component analysis), considers the Total Variance and derives the factors that contain small proportions of Unique Variance and in some instances Error Variance. Component analysis is preferred when data reduction is the primary goal. Common Factor Analysis is based only on Common Variance (Shared Variance) and assumes no importance to Unique and Error Variances in defining the structure of variables. It is more useful in identifying latent constructs and there is little information about Error and Unique variances. The 2 methods achieve essentially the same results in many situations.

Determining the number of factors to extract: The total number of factors extracted from Factor Analysis is retained for interpretation and further analysis. This decision on the number of factors depends on the questions such as, how many factors to extract, how many factors to retain in the structure?, and how many factors can be reasonably supported with empirical evidence? The research begins with some predetermined criteria such as the general number of factors and some general thresholds of practical relevance. These criteria are combined

[268] Hair et al, Multivariate Data Analysis, 6th Ed, Printice Hall Pb; 2006

with empirical measures of factor structure. An exact quantitative basis for deciding the number of factors to extract has not been developed. Stopping criteria for the number of factors to extract include latent root or eigen value, a priori, percentage of variance and scree test. These empirical criteria must be balanced against any theoretical bases for establishing the number of factors (Hair et al, 2006)[269]

Explaining the concept of rotation of factor: The most important tool in interpreting factors is Factor Rotation. The term rotation indicates the turning of the reference axes of factors about the origin until some other position has been reached. There are 2 types of rotation – orthogonal and oblique. Unrotated factor solutions extract factors in the order of their importance, with the first factor being general factor with almost every variable loading significantly and accounting for the largest amount of variance. The second and the subsequent factors are based on the residual amount of variance, with each accounting for successively smaller portions of variance. The ultimate effect of rotating the factor matrix is to redistribute the variance from earlier factors to later ones to achieve a simpler, theoretically more meaningful factor pattern. Factor Rotation assists in the interpretation of factors by simplifying the structure through maximizing the significant loadings of a variable on a single factor. In this manner, the variables most useful in defining the character of each factor can be easily identified (Hair et al, 2006)[270]

Naming the factor: Factors represent composite of many variables. When an acceptable factor solution had been obtained, all variables have a significant loading. The researcher attempts to find meaning out of the factor loadings. Variables with higher loadings are considered more important, for they have greater influence on the name or label selected to represent the factor. The significant variables for a particular factor are examined. Greater emphasis is on those variables with higher loadings. A name is assigned to a factor that reflects the variable loadings on that factor. The researcher identifies variables with the

[269] Hair et al, Multivariate Data Analysis, 6th Ed, Printice Hall Pb; 2006
[270] Hair et al, Multivariate Data Analysis, 6th Ed, Printice Hall Pb; 2006

greatest contribution to a factor and assigns a name to represent the factor's conceptual meaning (Hair et al, 2006)[271].

Uses of Factor analysis: The researcher can stop with the Factor Interpretation or further proceed to Data Reduction. If the objective is just to identify the logical grouping of variables through better understanding of the interrelationships among the variables, then the Factor Interpretation will suffice. If the objective is to identify appropriate variables for subsequent application of statistical techniques, then Data Reduction will be necessary. In the procedure of Data Reduction, the researcher would identify a single variable as the best representation of entire set of variables for further statistical analysis. Another option is to calculate the summation of the variables with highest factor loading. This is known as the summated scale. A single summated score represents a factor but only selected variables contribute to the composite score. A third option is to calculate the factor scores for each factor, where each factor contributes to the score based on its factor loading. This single measure is a composite variable that reflects the relative contributions of all the variables to the factor. If the summated scale is valid and reliable, it is probably the best of these 3 data reduction techniques.

Limitations of factor analysis technique: There are 3 most frequently cited limitations. (Hair et al, 2006)[272] There are many techniques available for performing Factor Analysis, although controversy exists over which technique is the best. The subjective aspects of Factor Analysis such as number of factors to be extracted, technique to be used to rotate the factor axes, and significant factor loadings are all subjected to many differences in opinion. The problem of reliability is real, and like any other statistical procedure Factor Analysis starts with a set of imperfect data. Changes in sample, data gathering procedures, and measurement errors affect the results of the analysis. The results of a single analysis are therefore not completely dependable. Factor analysis technique is

[271] Hair et al, Multivariate Data Analysis, 6th Ed, Printice Hall Pb; 2006
[272] Hair et al, Multivariate Data Analysis, 6th Ed, Printice Hall Pb; 2006

complex and plausible, though the fact that plausible solutions do not guarantee complete validity or stability remains unruffled.

5.1.3 Objectives of Factor Analysis

There are four key issues attached to the objectives of Factor Analysis: specifying the unit of analysis, Data Summarization and Reduction, variable selection, and using Factor Analysis results with other multivariate techniques. Each is briefly explained below.

There are several methods of factor analysis. However, the most commonly used are the R-Factor Analysis or Q-Factor Analysis (Thompson, 2000)[273]. These types refer to what is serving as variables and what is serving as the observations in the arrangement of data row and column wise. In R-Factor analysis, the variables are the columns of the data set and observations are the rows. In R-Factor analysis, we look for the latent factors that lie behind the variables, and the Q-Factor analysis condenses large number of people in distinctly different groups within a large population. There are other possible combinations of groups and variable types (Stewart et al, 1981[274], Thomson, 2000). The data analysis for the given study refers to R-Factor analysis.

There are 2 distinct, but interrelated outcomes of factor analysis: Data Summarization and Data Reduction. The concept of Data Summarization is to evolve the definition of structure, through the structures of the variables from most detailed levels to more generalized levels can be viewed. The goal is achieved by defining a small number of factors that adequately represent the original set of variables (Hair et al, 2006)[275]. The purpose of Data Reduction is to retain the nature and character of the original variables, but reduce their numbers to simplify subsequent Multivariate Analysis. The objective of applying

[273] Thompson.B. and J. Green, G. Camilli, & P.B. Elmore (Eds.), Research synthesis:Handbook of complementary methods in education research Washington, DC: American Educational Research Association., p. 583-603, 2006

[274] Stewart, D.W., (1981) "The application and the Misapplication of factor analysis in marketing research", Journal of Marketing Research, Vol.18, p 51-62

[275] Hair et al, Multivariate Data Analysis, 6th Ed, Printice Hall Pb; 2006

Factor Analysis in the study was to condense or summarize the variables, the building blocks of relationships, into smaller sets of new, composite dimensions called factors, with a minimum loss of information. The factors then created in new composite measures were applied in further analysis.

Factor Analysis is most efficient when conceptually defined dimensions can be represented by the derived factors. The quality and meaning of the derived factors reflect the conceptual underpinnings of the variables included in the analysis and judgment of the researcher. Factor Analysis still maintains the flavor of an art, and no single strategy should yet be 'chiseled into stone'. Factor Analysis should not be used in most practical situations (Chatfield et al, 1980)[276]. Heir et al (Hair et al, 2006)[277] mention that Factor Analysis provides a clear understanding of, which variables may act in concert and how many variables may actually be expected to have impact in the analysis. It is an excellent starting point for many other multivariate techniques.

5.1.4 Research Design for Factor Analysis

The research design of the Factor Analysis involves 3 decisions: (*i*) Calculation of a correlation matrix (input data); (*ii*) Design of the study in terms of number of variables, measurement properties of variables, and types of allowable variables and (*iii*) The necessary sample size. These decisions are discussed below.

i) ***Correlations among variables or respondents:*** The first decision focuses on calculating the input data for the analysis. Earlier we discussed R-type and Q-Type factor analyses. Hair at el (Hair et al, 2006)[278] posit in R-type factor analysis, the traditional correlation matrix specifying correlations among variables is used. In Q-Type factor analysis, the correlation matrix is derived from the correlations between the individual respondents. The resultant factor matrix identifies similar

[276] Chatfield et al, Introduction to Multivariate Analysis, CRC Press; 1980
[277] Hair et al, Multivariate Data Analysis, 6th Ed, Printice Hall Pb; 2006
[278] Hair et al, Multivariate Data Analysis, 6th Ed, Printice Hall Pb; 2006

individuals. R-Type factor analysis is used widespread and the discussion in this chapter continues on R-Type factor analysis.

ii) ***Variable Selection and Measurement:*** In factor analysis, correlations among variables is the only means of determining appropriateness and therefore the observed patterns have to be conceptually valid and appropriate to study. The primary requirement of the Factor Analysis is that a correlation value can be calculated among all variables. If the metric variables are used in factor analysis, they can be measured by several types of correlations. But non-metric variables can not use the same type of correlation measures that of metric variables. Therefore to include a non-metric variable, an approach of dummy variable (coded 0-1) is taken. If all the variables are dummy variables, then specialized forms of Factor Analysis such as Boolean Factor Analysis can be used (Hair et al, 2006)[279]. A rule of thumb for substantial correlation value is > 0.30. To find patterns among groups of variables, each proposed factor should include several variables (five or more). It is of little use in identifying factors composed of only a single variable (Hair et al, 2006)[280].

iii) ***Sample Size:*** The best method for standardizing sample size data is subject to item ratio. Anna Costello and Osborne conclude that a large percentage of Factor Analyses are done using relatively small sample sizes. Their research indicates that 14.7 percent studies were done with a subject to item ratio of 2:1 or less, 25.8 percent studies had a ratio of > 2:1, $\leq$ 5:1; 22.7 percent studies had the ratio of >5:1, $\leq$ 10:1. About 37 percent studies had the subject to item ratio $\geq$ 10:1. Past research has revealed that adequate sample size is partly determined by the nature of data. In general, the stronger the data in terms of uniformly high communalities without cross-holdings, plus several variables loading strongly on each factor, smaller is the sample. However, in practice these conditions can be rare (Osborne et al, 2005)[281].

[279] Hair et al, Multivariate Data Analysis, 6th Ed, Printice Hall Pb; 2006
[280] Hair et al, Multivariate Data Analysis, 6th Ed, Printice Hall Pb; 2006
[281] Osborne., By Best Practices in Quantitative Methods, Sage Publications Inc Pb, 2005

As a thumb-rule, Factor Analysis requires minimum 50 observations as the sample size, and preferably 100 or larger sample. Another general rule is to have minimum ratio of observations to variable as 5:1. More acceptable ratio is 10:1. Stevens (Stevens et al, 2002)[282] summarizes some specific results backed by simulations as follows. The number of observations required for factors to be reliable depends on the data, particularly how well the variables load on the different factors. A factor is reliable if it has:

3 or more variables with loadings of 0.8 and any *n**

4 or more variables with loadings of 0.6 and any *n*

10 or more variables with loadings of 0.4 and $n \geq 150$

Factors with only a few loading require $n \geq 300$

* *n is the number of observations*

5.1.5 Assumptions in Factor Analysis

The critical assumptions underlying Factor Analysis are more conceptual than statistical. The character and composition of the variables included in the analysis require a strong theoretical foundation before meeting the statistical requirement of the multivariate technique. Given below are the assumptions that have to be met with for conducting Factor Analysis.

5.1.5.1 Conceptual and Statistical Aspects

The basic assumption of Factor Analysis is that some underlying structure does exist in the set of selected variables (Hair et al, 2006)[283]. The appropriateness of the technique is determined only by the correlations among variables, and therefore it is imperative that the observed patterns are conceptually valid and appropriate from the aspect of variables selection. Another assumption is that the sample is homogeneous with respect to the underlying factor structure. In case of 2 samples or sub-samples combined, the resulting correlations and factor

[282] Stevens et al, Steven's Handbook of Experimental Psychology, 3rd Ed, John Wiley and Sons Pb; 2002

[283] Hair et al, Multivariate Data Analysis, 6th Ed, Printice Hall Pb; 2006

structure gives a poor representation of the unique structure of each group. From statistical standpoint, some degree of multicollinearity is desirable, because the objective is to identify interrelated sets of variables.

5.1.5.2 Overall Measures of Intercorrelation

Hair et al posit data matrix of correlations should reveal substantial number of correlations greater than 0.30 to make factor analysis appropriate. If all of the correlations are low, or all correlations are equal, it implies that no structure exists to group variables and the application of factor analysis is questionable. The correlations among variables can also be analyzed by computing the partial correlations among variables. Partial correlation is the unexplained correlation when effects of other variables are taken into account. It should be small, i.e. less than .7, if the "true" factors exist in the data (Hair et al, 2006)[284].

Another method of determining the appropriateness of Factor Analysis is to examine the entire correlation matrix. The *Bartlett test of Sphericity* checks the null hypothesis that the original correlation matrix is an identity matrix (Andy Field, 2000)[285]. It provides the statistical significance that the correlation matrix has significant correlations among at least some of the variables. High significance ($p < .001$) of Bartlett's test indicates appropriateness of Factor Analysis.

The *Kaiser-Meyer-Olkin (KMO)* measure of sampling adequacy varies between 0 and 1. A value close to 1 indicates that patterns of correlations are relatively compact and so the factor analysis should yield distinct and reliable factors. Andy Field recommends acceptable values greater than .5. the values between .5 and .7 are mediocre, values between .7 and .8 are good, values between .8 and .9 are great, and values above .9 are superb (Andy Field, 2000)[286]. In Table 5.1, an

[284] Hair et al, Multivariate Data Analysis, 6th Ed, Printice Hall Pb; 2006
[285] Andy P. Field, Discovering Statistics Using SPSS for Windows: Advanced Techniques for the Beginner, Sage Pb; 2000
[286] Andy P. Field, Discovering Statistics Using SPSS for Windows: Advanced Techniques for the Beginner, Sage Pb; 2000

example of a SPSS output of the given study is depicted for KMO and Bartlett's test.

Table 5.1 - KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.906
Bartlett's Test of Sphericity	***Approx. Chi-Square***	1647.350
	Df	351
	Sig.	.000

KMO value is 0.906, slightly more than 0.9, and can be considered as a superb value, indicating that the patterns of correlations are compact and the factor analysis should yield distinct and reliable factors. Bartlett's test shows significance $P < 0.001$, and therefore the Factor Analysis is appropriate. The Measure of Sampling Adequacy (MSA) is the third measure to quantify the degree of interrelations among the variables and appropriateness of Factor Analysis. The measure can be interpreted with the following guidelines: 0.80 or above, meritorious; 0.70 or above, middling; 0.60 or above, mediocre; 0.50 or above, miserable; and below 0.50, unacceptable (Hair et al, 2006)[287].

5.1.6 Deriving Factors and Assessing Overall Fit

There are 2 decisions in applying Factor Analysis are concerned with: (1) the method of extracting the factors, and (2) the number of factors selected to represent the underlying structure in the data.

5.1.6.1 Criteria for Extracting the Factors

Factors are produced by common Factor Analysis (FA), while components are produced by Principal Components Analysis (PCA). They both are essentially Data Reduction techniques, differing in the variance of the observed variables

[287] Hair et al, Multivariate Data Analysis, 6th Ed, Printice Hall Pb; 2006

that is analyzed. In PCA, all the variance in the observed variables is analyzed whereas in FA, only shared variance is analyzed. Statistical theorists have disagreement about the applicability of each method. Some researchers favor FA as a true analysis method and propose severely restricted use of PCA, whereas others disagree, and point out either that there is almost no difference between PCA and FA, or that PCA is preferable (Osborne et al, 2005)[288].

The total variance of any variable consists of 3 types of variances: common, unique, and error. A variable's communality is the estimate of its shared or common variance among the variables as represented by the derived factors (Hair et al, 2006)[289]. The *communalities* represent the proportion of the variance for each of the variables included in the analysis that is explained or accounted for by the components in the factor solution. The derived components should explain at least half of each original variable's variance, so the communality value for each variable should be 0.50 or higher. If one or more variables have a value for communality that is less than 0.50, the variable with the lowest communality should be excluded and the Principal Component Analysis should be computed again.

Principal Components Analysis (PCA) considers the total variance and derives factors that contain small proportions of unique variance and, in some cases, error variance. The components are calculated using all of variance of the manifest variables, and all of that variance appears in the solution. As PCA does not discriminate between shared and unique variance, when the factors are uncorrelated and communalities are moderate, it can produce inflated values of variance accounted for by the components. However, researchers rarely collect and analyze data without an *a priory* idea about how the variables are related (Osborne et al, 2005)[290].

In Factor Analysis, only common or shared variance is considered with the assumption that both the unique and error variance are not of interest in defining

[288] Osborne., By Best Practices in Quantitative Methods, Sage Publications Inc Pb, 2005
[289] Hair et al, Multivariate Data Analysis, 6th Ed, Printice Hall Pb; 2006
[290] Osborne., By Best Practices in Quantitative Methods, Sage Publications Inc Pb, 2005

the structure of the variables. The aim of Factor Analysis is to reveal any latent variables that cause the manifest variables to co-vary. There are several factor extraction methods in Factor Analysis to choose from: unweighted least squares, generalized least squares, maximum likelihood, principal axis factoring, alpha factoring, and image factoring. However, information on their relative strengths and weaknesses is scare and often available in obscure references (Osborne et al, 2005)[291]. Probably because of this, Principal Component Analysis is the most preferred technique. PCA is the default method of extraction in many popular statistical software packages such as SPSS and SAS. The data for the study has been analyzed using PCA in SPSS.

5.1.6.2 Criteria for Selecting Number of Factors to be Retained

After extraction, the decision is to be made on how many factors to retain for rotation. (Mardia et al, 1980)[292] point out that there is a limit to the number of factors that can actually end up with a simpler model than the raw data. The minimum number of variables required to select the number of factors is given in Table 5.2.

Table 5.2 - Minimum Variables required for Factors Selection

Factors	**2**	**3**	**4**	**5**	**6**
Variables Required	**5**	**7**	**8**	**9**	**11**

This is a guideline and factor loadings on each variable also have to be assessed before actually deciding the meaningfulness of the factor. The decision on the number of factors to be retained from the extraction process is based on the several stopping criteria for the number of factors to extract. Usually in practice, more than one criterion is used to select the factors. The criteria available in SPSS software are discussed below.

[291] Osborne., By Best Practices in Quantitative Methods, Sage Publications Inc Pb, 2005
[292] Mardia et al, Multivariate Analysis, Academic Press; 1980

The *latent root criterion* is used with the rationale that any individual factor should account for the variance of at least a single variable if it is to be retained for interpretation. With component analysis each variable contributes a value of 1 to the total eigenvalue. One of the least accurate methods is retaining the number of factors having eigen values greater than 1 (Osborne et al, 2005)[293].

All the factors having eigen values > 1 can be retained for the correlation matrix. However, Hair et al., (Hair et al, 2006)[294] reports that establishing a cutoff is most reliable when the number of variables is between 20 and 50. Stevens (Stevens et al, 2002)[295] reports that if variables are greater than 40 and their communalities are around 0.40, they are considered to be too many.

A more accurate cutoff point is with 10-30 variables and their communalities are around 0.70. This criterion is also known as Kaiser's recommendation, and appears in SPSS as an option under the Extract box. In the '*a priory criterion'*, the number of factors to extract is decided before undertaking the Factor Analysis. This approach is used in testing a theory or a hypothesis about the number of factors to be extracted, or in replicating another researcher's work.

Percentage of variance is another criterion used to decide the number of factors to extract. This approach is based on achieving a specified cumulative percentage of total variance extracted by successive factors, by ensuring that they explain at least a specified amount of variance. In social sciences, where the information is less precise, it is common to consider a solution that accounts for 60 percent of the total variance, as there is no absolute threshold adopted for all application.

A variant of this criterion is to select the factors with communality of more than .50 for each of the variable. This approach is considered for not to neglect the degree of explanation for the individual variables (Hair et al, 2006)[296].

[293] Osborne., By Best Practices in Quantitative Methods, Sage Publications Inc Pb, 2005
[294] Hair et al, Multivariate Data Analysis, 6th Ed, Printice Hall Pb; 2006
[295] Stevens et al, Steven's Handbook of Experimental Psychology, 3rd Ed, John Wiley and Sons Pb; 2002
[296] Hair et al, Multivariate Data Analysis, 6th Ed, Printice Hall Pb; 2006

The scree test criterion is an alternate method for factor retention, and is available in most frequently used statistical software including SPSS. It is the method to identify the optimum number of factors that can be extracted before the amount of unique variance begins to dominate the common variance structure.

A graph is plotted for latent roots (eigen values) against the number of factors in their order of extraction. The graph is examined and at the point at which the curve first begins to straighten out or breaks from the natural bend, is considered the cutoff point. The number of data points above the "break" (not including the point at which the break occurs) is usually the number of factors to retain (Osborne et al, 2005)[297]. As a general rule, the scree test results in at least one or sometimes 2 or 3 more factors being considered for inclusion than does the latent root criterion (Hair et al, 2006)[298].

5.1.7 Interpretation of Factors

A strong conceptual foundation for the anticipated factor structure and its rationale is important, as there are no specific processes or guidelines for interpreting factors. In the study, the theoretical concepts of conflict typology and causative factors were related with the analytical framework of factor analysis to interpret factors and the structure lying underneath.

5.1.7.1 Factor Rotation

Factor interpretation is circular in nature. First, the initial unrotated factor matrix is computed, containing the factor loadings for each variable on each factor. Hair et al (Hair et al, 2006)[299] define factor loadings as the correlation of each variable and the factor. These are the means of interpreting the role each variable plays in defining each factor.

[297] Osborne., By Best Practices in Quantitative Methods, Sage Publications Inc Pb, 2005
[298] Hair et al, Multivariate Data Analysis, 6th Ed, Printice Hall Pb; 2006
[299] Hair et al, Multivariate Data Analysis, 6th Ed, Printice Hall Pb; 2006

The next decision is of selecting the rotation method. The goal of rotation is to simplify and clarify the data structure. It can not improve the basic aspects of the analysis such as the amount of variance extracted from the items (Osborne et al, 2005)[300].

The initial unrotated factor matrix does not provide enough information of the variables under observation. Ambiguities in the interpretation are found because the first factor tends to be a general factor with almost every variable loading significantly, accounting for the largest variance. Subsequent factors are based on the residual mount of variance. Therefore, factor rotation is used. Hair, et al (Hair et al, 2006) posit the ultimate effect of rotating the factor matrix is to redistribute the variance from earlier factors to later ones, to achieve a simpler, theoretically more meaningful factor pattern.

Two methods of rotation are used, orthogonal and oblique. Orthogonal rotations produce factors that are uncorrelated and oblique methods allow the factors to correlate. Varimax, quartimax, and equimax are commonly available orthogonal methods of rotation, while direct oblimin, quartimin, and promax are oblique methods (Osborne et al, 2005)[301].

Orthogonal rotation produces more easily interpretable results, and is commonly used method in research. The SPSS program gives five options for rotation (Ajai Gaur, 2006)[302]. The rotated factor matrix output is interpreted after orthogonal rotation; pattern matrix is examined for factor/item loadings in oblique rotation, and factor correlation matrix reveals any correlation between the factors. The substantive interpretations are essentially the same (Osborne et al, 2005).

[300] Osborne., By Best Practices in Quantitative Methods, Sage Publications Inc Pb, 2005
[301] Osborne., By Best Practices in Quantitative Methods, Sage Publications Inc Pb, 2005
[302] Ajai Gaur, Statistical Methods for Practice and Research: A Guide to Data Analysis Using SPSS, Response Books, 2006

5.1.7.2 Significance of Factor Loadings

In interpretation, it is essential to make the decision regarding the factor loadings that are worth considering. Practical significance of making a preliminary examination of factor loadings is important, as larger the absolute size of the factor loading, the more important the loading is in interpreting the factor matrix. Tabachnick and Fidell (Osborne et al, 2005)[303] suggest 0.32 as a good rule of thumb for the minimum loading of an item, which equates to approximately 10% overlapping variance with the other items in that factor.

Using practical significance as the criteria, factor loadings are assessed as follows (Hair et al, 2006)[304]:

- Factor loadings in the range of $\pm$ 0.30 to $\pm$ 0.40 are considered as the minimum level for interpretation of the structure.
- Loadings $\pm$ 0.50 or greater are considered practically significant.
- Loadings exceeding $\pm$ 0.70 are indicative of well-defined structure.

The significance level for the interpretation of loadings can be determined in the similar way of determining the statistical significance of correlation coefficients. However, researchers have demonstrated that factor loadings have substantially larger standard errors than typical correlations (Hair et al, 2006)[305]. Therefore factor loadings have to be evaluated at a considerably stricter level.

Anna Costello et al (Osborne et al, 2005)[306] caution that Factor Analysis is a large-sample procedure in which generalizable or replicable results are unlikely if the sample is too small. Hair et al (Hair et al, 2006)[307] present the guidelines for identifying significant factor loadings based on sample size, as follows in Table 5.3.

[303] Osborne., By Best Practices in Quantitative Methods, Sage Publications Inc Pb, 2005
[304] Hair et al, Multivariate Data Analysis, 6th Ed, Printice Hall Pb; 2006
[305] Hair et al, Multivariate Data Analysis, 6th Ed, Printice Hall Pb; 2006
[306] Osborne., By Best Practices in Quantitative Methods, Sage Publications Inc Pb, 2005
[307] Hair et al, Multivariate Data Analysis, 6th Ed, Printice Hall Pb; 2006

Table 5.3 - Guidelines for Identifying Significant Factor Loadings Based on Sample Size

Factor Loading	*Sample Size Needed for Significance*[a]
.30	350
.35	250
.40	200
.45	150
.50	120
.55	100
.60	85
.65	70
.70	60
.75	50

[a]*Significance is based on a .05 significance level (α), a power level of 80 percent, and standard errors assumed to be twice those of conventional correlation coefficients.*

Source: Computations made with SOLO Power Analysis, BMDP Statistical Software, Inc., 1993.

5.1.7.3 Factor Matrix

To identify the most indicative factors of the underlying structure, all the factor loadings are sorted and a five step process is applied. In the first step, the factor matrix of loadings is examined. It contains factor loading on each variable. In the rotated factor loading analysis, the factors are arranged as columns, and each column of numbers represents the loadings of a single factor. The factor pattern matrix had loadings that represent the unique combination of each variable to the factor. A factor with less than 3 variables is generally weak and unstable; five or more variables, with loadings >0.50 in a factor are desirable and indicate a

solid factor. It may be possible to reduce the number of variables and maintain a strong factor in large samples with further analysis (Osborne et al, 2005)[308].

The second step is of identifying the significant loading(s) for each variable. The interpretation starts with the first variable on the first factor, from left to right, looking at the highest loading for that variable on any factor. When the highest loading is identified and is significant as per the criteria discussed earlier, it is underlined. The process of selecting highest loading per variable continues till all the loadings are sorted. When a variable is found to have more than one significant loading, it is known as *cross-loading*. Different rotation methods can be used to eliminate cross-loadings and simplify the data.

Third step is to assess the communalities of the variables. In case of any variables that are not adequately accounted for by the factor solution, one approach is to identify any variable(s) lacking at least one significant loading. Another approach is to examine communality of each variable, which represents the amount of variance accounted for by the factor solution for each variable. Variable communalities are considered 'high' if they are .80 or greater.

However, it is unlikely to occur in real data. More common magnitudes in the social sciences are low to moderate communalities of 0.40 to 0.70. A variable having < 0.40 communality is either not related to other variables, or suggest an additional factor that should be explored (Osborne et al, 2005)[309]. As a general guideline, all the variables with communalities less than 0.50 are identified as variables not having sufficient explanation (Hair et al, 2006)[310].

The fourth step is to re-specify the factor model, if needed. In case of a variable having no significant loadings, or its communality is deemed too low, or a variable having cross-loading, several ways can be taken. These are either to ignore those problematic variables and interpret the solution as it is; or to

[308] Osborne., By Best Practices in Quantitative Methods, Sage Publications Inc Pb, 2005
[309] Osborne., By Best Practices in Quantitative Methods, Sage Publications Inc Pb, 2005
[310] Hair et al, Multivariate Data Analysis, 6th Ed, Printice Hall Pb; 2006

employ alternative rotation methods; or to increase/decrease the number of factors retained, or modify the type of factor model used.

The fifth step is to label the factors. The labels have to be developed intuitively based on their appropriateness for representing the underlying dimensions of a particular factor. Each extracted factor is given a name or a label that represents each of the derived factors as accurately as possible.

5.1.8 Creation of Factor Scores

The objective of the study is not only Data Reduction, but also is to identify appropriate variables for subsequent application to other statistical techniques. Hair et al (Hair et al, 2006)[311] elaborate 2 methods of data reduction and creation of new factors. In one method, the variable with the highest factor loading is selected as a surrogate representative for a particular factor dimension, and in another method the original set of variables are replaced with an entirely new, smaller sets of variables created from factor scores. Creation of factor scores is discussed in detail, as it is the technique used for the Factor Analysis of the data of the study.

Factor scores are used for diagnostic purposes and also as inputs to the subsequent analysis. They are smaller sets of variables that replace original set. Conceptually factor score represents the degree to which each case (individual) scores high on the group of items with high loadings on a factor. Thus, higher values on the variables with high loadings on a factor will result in a higher factor score (Hair et al, 2006)[312].

Factor score represents all variables loading on the factor, and is used for complete data reduction. By default, the factor scores are orthogonal and can avoid complications caused by multicollinearity. Factor scores are the scores of

[311] Hair et al, Multivariate Data Analysis, 6th Ed, Printice Hall Pb; 2006
[312] Hair et al, Multivariate Data Analysis, 6th Ed, Printice Hall Pb; 2006

each case (row) on each factor (column). To compute the factor score for a given case, for a given factor, the case's standardized score is taken on each variable, is multiplied by the corresponding factor loading of the variable for the given factor, and these products are summed up.

The objectives of the study as mentioned earlier were: data reduction; identification of the variables that construct the factors; and replace the variables with the factors in the original data. This reduces number of independent variables in the statistical model and makes the process parsimonious. Of course, in the process of replacing the variables with factors a certain degree of explanatory power is lost, as the percentage of variance explained be the factors is generally not more than 70 percent.

5.1.8.1 Methodology of Calculating Factor Scores

The methodology of calculating factor scores which will replace the independent variables with new factors for further analysis is as follows.

The process starts with the rotated factor loadings of the variables. For example, n_1 variables construct factor 1. The rotated factor loadings of the variables have to be converted into relative loadings by dividing the factor loading of the variable by the sum of the factor loadings of all the n_1 variables. As a result, all n_1 variables that construct factor 1 lead to a sum-total of 1, when relative factor loadings are considered. These values are considered as the coefficient of the n_1 variables that construct factor 1. If the relative factor loadings are represented as $\beta_1, \beta_2, ..., \beta_{n_1}$ and n_1 variables are denoted as $X_1, X_2, ...X_{n_1}$ then factor 1 can be represented as, $factor1 = \beta_1 X_1 + \beta_2 X_2 + ... + \beta_{n_1} X_{n_1}$. Similarly, other factors are formed. In this context, it is crucial to check that that factors are not created mechanically just by observing at the rotated factor loadings. It is equally important to interpret and understand the relevance of the factor created. If the factor created does not make adequate sense, it is wise to drop it. Factor score for each factor calculated in the above manner is transferred to SPSS data sheet. For

each individual respondent (row), a new data file of seven factor scores (columns) is created. The factor scores now become the starting point for the second multivariate technique of multiple regression.

5.1.9 Software used for Factor Analysis Technique

The most widely used statistical software package SPSS is used in the Data Analysis for the study. Although SPSS incorporates statistical and mathematical processes for Factor Analysis as described above, it has a specific terminology and commands to be applied for conducting the Data Analysis. Data Analysis with SPSS software is discussed in chapter 6. SPSS 16.0 is selected for factor analysis for this research work as it incorporates Principal Component Analysis such as SAS. SPSS is preferred over SAS for the simplicity, usability and availability.

Part II

5.2 Multiple Regression Analysis

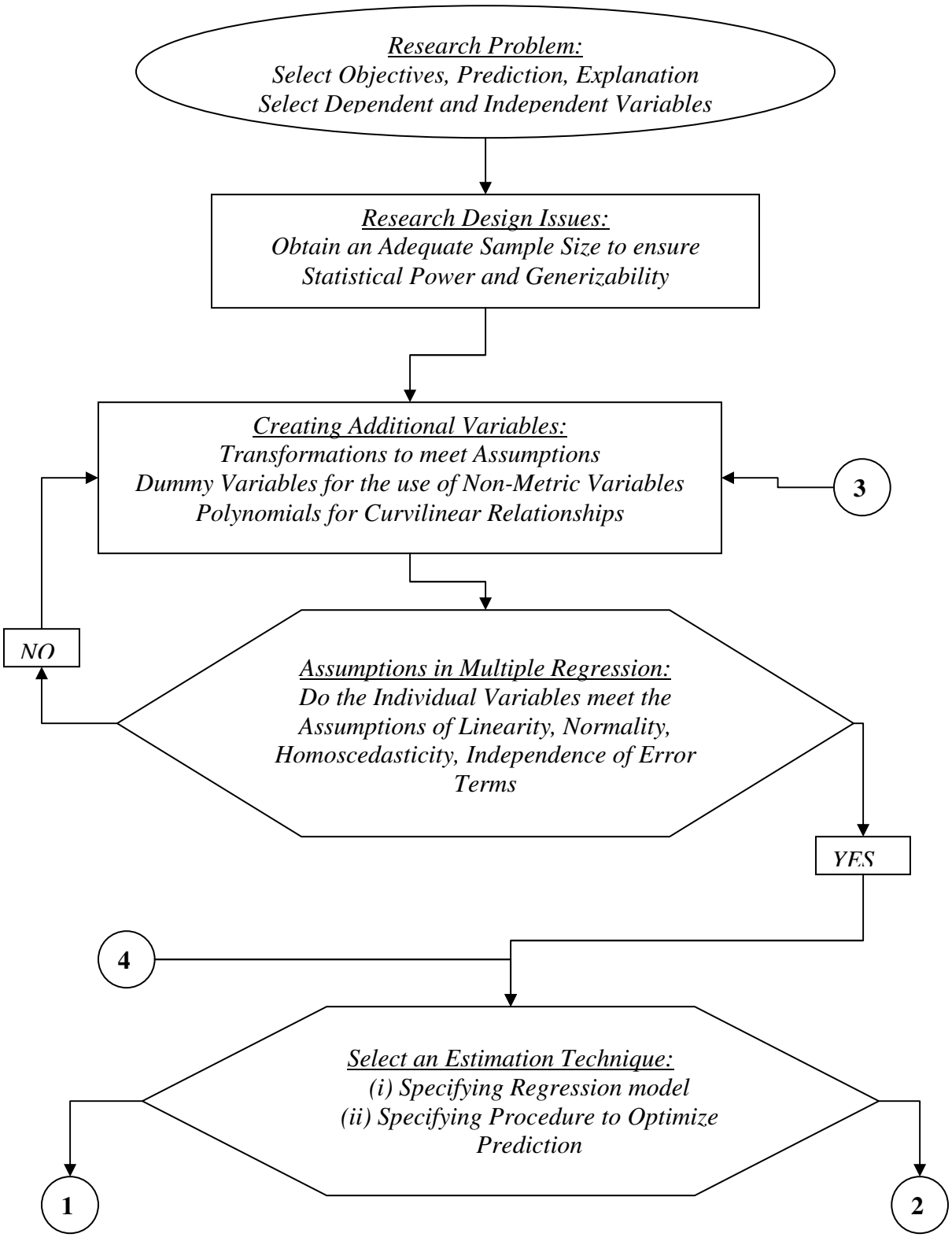
After the completion of Factor Analysis, we start our discussion on linear multiple regression. Multiple regression analysis is a statistical technique that can be used to analyze the relationship between a single metric dependent variable and several independent variables which could be either metric or dichotomous. It is a dependence technique. The objective of this technique is to form a regression variate – a linear combination of independent variables that predict the dependent variable the best. The regression variate is also known as regression equation or regression model. This technique is used when both dependent and independent variables are metric. Under special circumstances, it is possible to include non metric data either as independent variables (by transforming either ordinal or nominal data with dummy variable coding) or the dependent variable (by the use of a binary measure in logistic regression). To apply Multiple Regression Technique they must be transformed and before

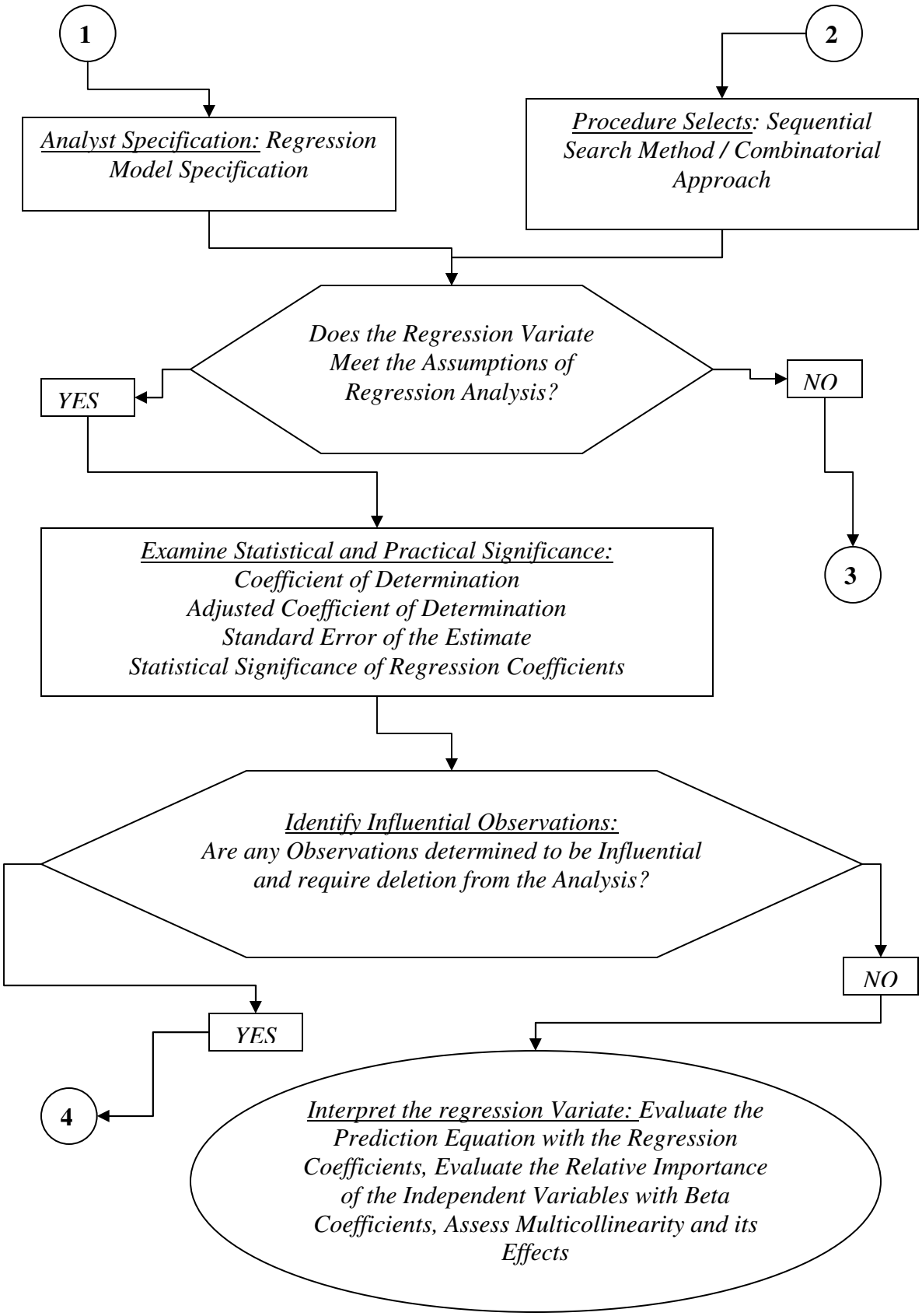
formulating the regression equation, the dependent and independent variables have to be segregated.

Sometimes the independent variables exhibit a quality of multicollinearity (correlation among 3 or more independent variables amongst themselves). The impact of multicollinearity is to reduce any single independent variable's predictive power to the extent to which it is associated with the other independent variables. As collinearity increases, the unique variance explained by each independent variable decreases and the shared prediction percentage rises. Because the shared prediction can account only once, the overall prediction increases much more slowly as independent variables high multicollinearity are added. To maximize the prediction power of the model, from the given set of independent variables, the researcher should look for independent variables that have low multicollinearity with the other independent variables and have high correlation with the dependent variable. In this case, as discussed before, factor analysis has taken care of the multicollinearity problem. The figure below diagrammatically explains the flow of research design of Multiple Regression Analysis.

Flow chart depicting Multiple Regression Analysis is given in Figure 5.2.

Figure 5.2 - Flow Diagram of Multiple Regression Analysis




1
2
Analyst Specification: Regression Model Specification
Procedure Selects: Sequential Search Method / Combinatorial Approach
Does the Regression Variate Meet the Assumptions of Regression Analysis?
YES
NO
3
Examine Statistical and Practical Significance:
Coefficient of Determination
Adjusted Coefficient of Determination
Standard Error of the Estimate
Statistical Significance of Regression Coefficients
Identify Influential Observations:
Are any Observations determined to be Influential and require deletion from the Analysis?
NO
YES
4
Interpret the regression Variate: Evaluate the Prediction Equation with the Regression Coefficients, Evaluate the Relative Importance of the Independent Variables with Beta Coefficients, Assess Multicollinearity and its Effects

5.2.1 Why using Multiple Regression Analysis for this Research Study?

Appropriateness of Multiple Regression Analysis: We decided to use Multiple Regression Analysis to predict and explain Financial Performances of small and medium business organizations. Multiple Regression Analysis can describe the relationship among 2 or more intervally scaled variables and is much more powerful than simple regression with a single independent variable. Multiple Regression Analysis is used to analyze the relationship between a single dependent (criterion) variable and several independent (predictor) variables. The objective of Multiple Regression Analysis is to use several independent variables whose values are known to predict the single dependent variable. Multiple Regression Analysis is a dependence technique.

To use this technique effectively, both dependent and independent variables must be distinct from each other and they must be metric. Under certain circumstances, it is possible to include non-metric data either as independent variables (by transforming either ordinal or nominal data with dummy variable coding) or the dependent variable (by the use of binary measure in the specialized technique of logistic regression). Thus to apply Multiple Regression Analysis, the data must be metric and appropriately transformed as well the depending and independent variables from the groups have to be decided.

In this research work, the data collected is sorted and suitably transformed with 5- level Likert scale and those questions varying from this 5-level Likert scale are with seeking ordinal answers and 3-level answers are suitably transformed to yield uniformity for regression analysis purposes.

One of the objectives of the research is to establish the relationship between OP and OI. There are variables such as, Financial Returns, Market Share Growth, Business Valuation, Profit Growth and Rate of Business Expansion as the measures of Organizational Performance. They are variables capturing Financial Performance of the Firm. Financial Performance is a universally accepted

standard measure of high performing organizations (Jeffrey et al., 1997)[313]. 'Financial returns' captures the perceptions of the Business owner on 'Return on Equity, Return on Assets, Financial Growth'; Market share growth captures perceptions on 'Growth rate of market share over a period of 1 year'; Economic Value Added captures 'Business Value'; Profit Growth measures 'Growth of profit before tax'; and Rate of Business expansion captures 'Increase in business verticals and diversification' (Jeffrey et al., 1997)[314]. These are variables that depend on other independent variables listed in Appendix 4 - Variable selection from the Literature. Each of these dependent variables can be predicted with a set of independent variables with the group of factors evolved from exploratory factor analysis and a few unique variables which could not be grouped.

Ordinary Least Square Method and Accuracy: Before estimating the regression equation, we must calculate the baseline against which we will compare the predictive ability of our regression models. The baseline should represent our best prediction without the use of any independent variables. In regression, the baseline prediction is the simple mean of dependent variable. Because the mean will not predict each value of the dependent variable, we must have a way to assess predictive accuracy that can be used with both the baseline prediction and the regression models we create. The customary way to assess the accuracy of any prediction is to examine the errors in predicting the dependent variable. Although we might expect to obtain a useful measure of prediction accuracy by simply adding the errors, this approach is not possible, because the errors from using a mean value always sum to zero. To avoid this problem, we can sum up the squares of all the errors – known as sum of squared errors - provides a measure of prediction accuracy that will vary according to the amount of prediction errors. The objective is to obtain the smallest possible sum of squared errors as our measure of prediction accuracy. Hence the concept of least squares

[313] Jeffrey et al., (1997), "The Search for the Best Financial Performance Measure", Financial Analysts Journal, p11-20.
[314] Jeffrey et al., (1997), "The Search for the Best Financial Performance Measure", Financial Analysts Journal, p11-20.

helps us achieve highest accuracy possible. This method is also known as Ordinary Least Squares method (OLS) (Hair et al, 2006)[315].

Interpreting Dummy Variable: Usually researchers desire to utilize non-metric independent variables. Many multivariate techniques assume metric measurement of both independent and dependent variables. When dependent variable is measured as a dichotomous variable (0,1), either discriminant analysis or a specialized form of regression – logistic regression – is appropriate. The 'Business Valuation' variable captures dichotomous value and hence we have eliminated it from Multiple Regression Analysis. The other 4 variables are taken as dependent variables for the model equation. When the independent variables are non-metric, and have 2 or more categories, we can create dummy variables that act as replacement independent variables. Each dummy variable represent one category of non-metric independent variable, and any non metric variable with *k* categories can be represented as *k-1* dummy variables. Thus non-metric variables can be converted to a metric format for use in most multivariate techniques (Hair et al, 2006)[316].

Assumptions in Multiple regression analysis: Improvements in predicting the dependent variable are possible by adding independent variables and transforming them to represent non linear relationships. To do so, we must make several assumptions about the relationships between the dependent and independent variables that affect the least square procedure used for multiple regressions. The basic issue is to know whether in the course of calculating the regression coefficients and predicting the dependent variable, the assumptions of regression analysis have been met. We must know whether the errors in predictions are the results of the absence of a relationship among the variables or caused by some characteristics of the data that are not accompanied by the regression model.

[315] Hair et al, Multivariate Data Analysis, 6th Ed, Printice Hall Pb; 2006
[316] Hair et al, Multivariate Data Analysis, 6th Ed, Printice Hall Pb; 2006

The assumptions to be examined include linearity of the phenomenon measured, constant variance of the error terms, and normality of the error term distribution. The assumptions underlying Multiple Regression Analysis apply both to the individual variables, (dependent and independent) and to the relationship as a whole. Once the variate has been derived, it acts collectively in predicting the dependent variable, which necessitates assessing the assumptions not only for individual variables but also for variate. The principal value of prediction error for the variate is the residual – the difference between the observed and predicted values for the dependent variable. Plotting the residuals versus the independent or predicted variables is a basic method of identifying assumption violations for the overall relationship (Hair et al, 2006)[317].

Usually, statistical inferences from classical linear regressions are based on several assumptions in addition to the above mentioned assumptions on interrelationships between independent variables error distributions of the predictors. These assumptions are listed below.

(i) The regression model is linear in parameters.

(ii) The values of regressors are fixed in repeated sampling.

(iii) For a given set of independent variables, the mean value of the disturbances is zero.

(iv) For a given set of independent variables, the variance is constant or homoscedastic.

(v) For a given set of variables there is no autocorrelation in the disturbances.

(vi) If the independent variables are stochastic, the disturbance term and the independent variables are uncorrelated.

(vii) The number of observations must be greater than the number of independent variables.

(viii) There must be sufficient variability in the values taken by the regressors.

(ix) The regression model is correctly specified.

(x) There is no exact linear relationship in the regressors. (presence of multicollinearity) .

(xi) The stochastic disturbance term is normally distributed.

[317] Hair et al, Multivariate Data Analysis, 6th Ed, Printice Hall Pb; 2006

When a relationship between *several* independent variables and a dependent variable is turned into a multivariate model, it is known as a Multiple Regression Model. Most theoretical results developed for the simple regression model naturally extend to Multiple Regression. Such a model has the general form

$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \ldots + \beta_k x_{ki} + e_i$

In this model also the subscript t denotes sample number, there being i total number of samples available for parameter estimation and analysis. As is the case with simple regression, to make the Multiple Regression Models complete and acceptable for forecasting and other applications, certain assumptions must hold for the errors or residuals $\{e_i\}$. They are;

- $E[e_i] = 0$, implying that each random error has a probability distribution with zero mean.
- $var(e_i) = \sigma^2$. Each random error has a variance equal to σ^2. Such errors that have equal variance are called *homoscedastic*.
- $cov(e_i, e_j) = 0$ for $i \neq j$, implying that the covariance between two random errors corresponding to any two different observations marked by i and j is zero.
- Sometimes it is further assumed that errors $\{ e_i\}$ are normally distributed. That implies $e_i \sim N(0, \sigma^2)$

Since the general linear Multiple Regression Model is developed by following procedures similar to that for the simple regression model, the OLS parameter estimation procedure is again used here. This is valid provided the above assumptions are met by the random errors $\{e_i\}$.

The goodness of fit of a regression model—simple or multiple—is given by a measure R^2, which expresses the fraction of the variability in the endogenous variable y that may be "explained" by the exogenous terms (X_1, X_2, X_3, etc.) of the regression model. The "significance test" of a regression model tests the relevance of all the explanatory variables included in the model. This test hypothesizes that all the model parameters $\{\beta_i\}$ are zero, except the intercept β_0, and then checks the acceptability of this statistical hypothesis by performing an F test.

Regression models are built based on data collected in which each observation consists of the set values of the independent variables, and the corresponding observed value of the dependent variable y. Parameter estimation in multiple regression analysis procedurally requires matrix algebra to manipulate the several simultaneous equations derived form the least squares criteria.

When data are collected from uncontrolled experiments, many of the "independent" variables may move together in systematic ways. Such variables are called collinear and when several such variables are involved, the system is said to have the problem of multicollinearity. In this case even if several independent variables are involved, the data collected may not be "rich in information". In such cases it is not possible to isolate the relationship between the dependent and the independent variables reliably. Such situations are handled by special analytical approaches. We note again, that a key assumption of Multiple Regression Model building based on the least squares or OLS criteria is that the values of the explanatory variables are not random and are not exact linear functions of the other explanatory variables.

Prediction problems with Multiple Regression Models are similar to the simple regression case: we first need to reliably estimate all model parameters (coefficients) β_0, β_1, β_2, β_3, etc. and also establish an acceptable goodness of fit for the model. Then it is possible not only to estimate the dependent variable given certain specified values of the explanatory variables, but also the variance and the confidence interval of the prediction. Note that even categorical or discrete variables (white, male, graduate, etc.) can be incorporated into regression models. Also, nonlinear relationships can be modeled, with suitable mathematical transformations of the variables, such as taking log, to convert the relationships into linear relationships, so that the technique of regression may be applied to develop a model. Interactions between the independent variables also can be used as contributing terms in a multiple regression model. Furthermore, polynomial terms may be used in a regression model.

We briefly mention some other problems in successfully developing Multiple Regression Models. We have already mentioned the issue of collinearity. This is detected by observing the covariance matrix, especially the co-variances estimated between the different explanatory variables. The solution is to drop a few explanatory variables from the model in order for the OLS algorithm to work, which requires solving simultaneous and linearly independent equations to deliver the estimated model parameters b_1, b_2, b_3, etc. The other way to take care of the multicollinearity problem is applying factor analysis on the independent variables before going for Multiple Regression. This is what we are doing in this research.

The other critical issue is that of autocorrelation (among errors over different time periods) when one is developing a multiple regression model using time series data. As our data is cross-section data, free from any time series analysis, autocorrelation is absent. In this situation also the standard OLS procedure cannot be directly applied. The solution requires one to use an extended procedure known as the *generalized least squares* procedure (Hill et al, 2001)[318]. A similar problem that is faced by cross-section data is procedure is heteroscedasticity, which is applicable when error variances are not constant, i.e., is present among errors $\{e_i\}$. Hence, when we use Multiple Regression as a model we need to make sure that all the above classical assumptions regarding the behavior of the error term $\{e_i\}$ are met.

Selection of an Estimation Technique: In a Multiple Regression, a researcher may chose from a number of possible independent variables for inclusion the regression equation. Sometimes the set of independent variables are exactly specified and the regression model is essentially used in a confirmatory approach. This approach referred to as a simultaneous regression, includes all variables at the same time. In other instances the researcher may use the estimation technique to pick and chose among the set of independent variables with either sequential search methods or combinatorial processes. The most popular sequential search method is stepwise estimation which enables the

[318] Grifiths et al, Learning and Practicing Econometrics, John Wiley Pb, 2001.

researcher to examine the contribution of each independent variable to the regression model. The combinatorial approach is a generalized search process across all possible combinations of independent variables. The best known procedure is all possible subsets regression which is exactly as the name suggests. All possible combinations of independent variables are examined and the best fitting set of variables are identified. Each estimation technique is designed to assist the researcher in finding the best regression model using different approaches. In this research study, there were initially 165 variables which reduced to 153 by eliminating similar types of variables and finally reduced to 40 variables as found from the literature. So the estimation techniques were not used, instead the entire set of independent variables are used against each of the dependent variable prediction (Hair et al, 2006)[319].

Interpreting the Results of Regression: The regression variate must be interpreted by evaluating the estimated regression coefficients for their explanation of the dependent variable. The researcher must evaluate not only the regression model that was estimated but also the potential independent variables that were omitted if a sequential search or a combinatorial approach was employed. In those approaches, multicollinearity may substantially affect the variables ultimately included in the regression variate. Thus, in addition to assessing the estimated coefficients, the researcher must also evaluate the potential impact of omitted variables to ensure that the managerial significance is evaluated along with statistical significance. The estimated regression coefficients, or beta coefficients represent both the type of relationship (positive or negative) and the strength of the relationship between independent and dependent variables in the regression variate. The sign of the coefficient denotes whether the relationship is positive or negative, while the value of the coefficient indicates the change in the dependent value each time the independent variable changes by one unit.

Prediction is an integral element in regression analysis, both in the estimation process as well as forecasting situations. Regression involves the use of a variate

[319] Hair et al, Multivariate Data Analysis, 6th Ed, Printice Hall Pb; 2006

to estimate a single value for the dependent variable. This process is used not only to calculate the predicted values in the estimation procedure, but also with additional samples for validation or forecasting purposes. The researcher often is interested not only in prediction, but also explanation. Independent variables with larger regression coefficients make a greater contribution to the predicted value. Insight into the relationship between independent and dependent variables is gained by examining the relative contributions of each independent variable. Thus for explanatory purposes, the regression coefficients become indicators of relative impact and importance of independent variables in their relationship with the dependent variable (Hair et al, 2006)[320].

Assessing Influential Observations: Influential observations include all observations that have a disproportionate effect on the regression results. The three basic types of influentials are,

(i) ***Outliers****:* Observations that have large residual values and can be identified only with respect to a specific regression model.

(ii) ***Leverage Points****:* Observations that are distinct from the remaining observations based on their independent variable values.

(iii) ***Influential*** *Observations:* all observations that have disproportionate effect on the regression results.

These 3 aspects depend on 4 conditions;

(a) ***An error in observations or in data entry****:* This can be corrected by correcting the data or deleting the data.

(b) ***A valid but exceptional observation that is explainable by an extraordinary situation:*** This can be corrected by deleting the case unless variables reflecting the extraordinary situation are included in the regression equation.

(c) ***An exceptional observation with no likely explanation****:* This is a special problem because the researcher has no reason for deleting the case, but its inclusion cannot justify either, suggesting analyses with and without observations to make a complete assessment.

[320] Hair et al, Multivariate Data Analysis, 6th Ed, Printice Hall Pb; 2006

(d) ***An ordinary observation in its individual characteristics but exceptional in its combination of characteristics:*** This indicates modifications to the conceptual basis of regression model and should be retained.

The researcher should delete truly exceptional observations but avoid deleting observations that, although different, are representative of the population (Barnett et al, 1994)[321].

5.2.2 Objective of Multiple Regression Analysis

The objective of Multiple Regression Analysis is to predict the dependent variable with the help of the independent variables. While doing so, the analysis fulfills couple of objectives which are discussed as below.

5.2.2.1 Research Problems Appropriate for Multiple Regression

The first problem is due to those assumptions about the specification of the model and about the disturbances. The second issue is due to the assumptions about the data (Barrie et al, 1986)[322].

Under these assumptions listed earlier in this chapter, the variables are selected to be,

$$Y = C + \beta 1X1 + \beta 2X2 + B3X3 \ldots + BnXn$$

In this specified model C represents the disturbance term, Y the dependent variable and Xi the independent variables and β - the regression coefficients.

Multiple Regression Technique is used for prediction and explanation. Prediction involves the extent to which the regressors can predict the dependent variable. Explanation examines the regression coefficients for each independent variable.

[321] Barnett et al, Outliers in Statistical Data, 3rd ed., NewYork: Wiley Pb; 1994

[322] Barrie Wetherill, Regression Analysis with Applications, Chapman and Hall, NewYork; 1986, p 14-15

Attempts are made to develop a theoretical reason to understand the behavior of the relationship between Xi and Y.

Prediction with Multiple Regression has 2 key objectives. One is, to maximize the overall predictive power of the independent variables as represented in the variate. Predictive accuracy is always crucial to ensure the validity of the set of independent variables. Measures of predictive accuracy are developed and statistical tests are used to assess the significance of predictive power.

While considering the applications of prediction alone, the interpretations from beta coefficients are relatively less important. Predictive accuracy is improved at the cost of beta coefficient interpretations. Next objective is, to compare 2 or more sets of independent variables to ascertain the predictive power of each variate. The predictive power of more models are studied and compared to judge about the dependent variables.

Explanation with Multiple Regression provides a means of objectively assessing the degree and character of the relationship between dependent and independent variables by forming the variate of independent variables and then examining the magnitude and direction as well statistical significance of regression coefficient for each independent variable. The independent variables collectively as well individually predict dependent variable and their beta coefficients will explain their relationship with dependent variable individually.

Interpretation of the variate will rely on 3 perspectives; the importance of the independent variables, the types of relationships found, the types of interrelationships among the independent variables. (i) The most direct interpretation of the regression variate is a determination of relative importance of each independent variable in the prediction of dependent measure. (ii) In addition to assessing the importance of each variable, Multiple Regression Analysis also affords the researcher a means of assessing the nature of the relationships between the independent variables and the dependent variable. (iii) The multiple Regression Analysis also provides insight into the relationships

among the independent variables in their prediction of the dependent measure (Hair et al, 2006)[323].

These interrelationships are important for 2 reasons. First, the correlation among the independent variables may make some variables redundant in the predictive effort. In such instances, the independent variable having strong relationship with dependent variable which is diminished due to the presence of relationships of other independent variables with the dependent variable. Then the researcher must guard against determining the importance of independent variables based solely on the derived variate, because the relationships among the independent variables may mask or co found relationships that are not needed for predictive purposes but represent substantive findings nonetheless.

The interrelationships among the variables can extend not only to their predictive power but also to the interrelationships among their estimated effects, which is best seen when the effect of one independent variable is contingent on another independent variable. Multiple Regression Analysis provides diagnostic analyses that can determine whether such effects exist based on empirical or theoretical rationale. Indications of high degree of interrelationships (multicollinearity) among the independent variables will suggest the use of summated scales (Hair et al, 2006)[324].

5.2.2.2 Selecting Dependent and Independent Variables & Specifying the Model.

Functional relationship calculates the exact value whereas a statistical relationship estimates an average value. In predicting the dependent variable accurately, it is important to define the assumptions made while formulating the relationship model. Predictive power of Multiple Regression Analysis depends on the assumptions made and the validation of interpretations of the independent variable. The success of a Multiple Regression Analysis depends on

[323] Hair et al, Multivariate Data Analysis, 6th Ed, Printice Hall Pb; 2006
[324] Hair et al, Multivariate Data Analysis, 6th Ed, Printice Hall Pb; 2006

the selection of dependent and independent variables and the specification of the model. *Strong Theory* says that the selection of variables must be based on conceptual or theoretical grounds even when the objective is solely for prediction. The researcher must select variables indiscriminately or allow the selection of independent variables to be solely based on empirical bases.

The other aspect of variable selection is measurement Error. *Measurement error* refers to the degree to which the variable is an accurate and consistent measure of the concept being studied. If the dependent variable has substantial measurement error, then the best set of independent variables will not be able to achieve higher levels of predictive accuracy.

Measurement error can be addressed by the usage of summated scales of independent variables or by structural equation modeling. Summated scales can be directly incorporated into Multiple Regression by replacing either dependent or independent variables with the summated scale values, while structural equation modeling requires the use of an entirely different technique generally regarded as s difficult analysis to implement.

Thus, summated scales are recommended as the first choice as a remedy for measurement error. Another error that occurs in variable selection is *Specification error. Specification error* is due to the inclusion of irrelevant variables or the omission of relevant variables from the set of independent variables. Inclusion of irrelevant variable impacts regression variate. It reduces model parsimony, which might be critical in the interpretation of results. It can mask and replace the effects of more important variables if some sequential form of model estimation is used. It can reduce the precision of the multiple regression models and reduce the significance of the entire analysis.

Similarly, the exclusion of relevant variables can bias the results and misdirect the interpretation considerably. If there is no correlation between the excluded and the included variables then the model accuracy will be reduced. If there is a correlation between them, then the prediction will be biased t the extent of the

correlation between the excluded and included variables. Model interpretation will suffer from precision and accuracy (Hair et al, 2006)[325].

5.2.3 Research Design of Multiple Regression Analysis

Research Design of Multiple Regression analysis primarily means the design of *sample size* as this technique maintains the necessary levels of statistical power and significance across broad range of sample sizes. The design includes the decision of *unique elements of the dependence relationship.* It is assumed that dependent variable and the independent variables share a linear relationship. Additional variables can be added to this relationship to represent special aspects of the relationship. Multiple Regression accommodates metric independent variables that are assumed to be fixed in nature as well as those with the random component. *Nature of Independent Variables* also decides the research design.

5.2.3.1 Sample Size

In multiple regression *power* refers to the probability of detecting a significant R-square. Sample size plays a role in assessing the power of current analysis as well proposed analysis (Mason et al, 1991)[326].

Table 5.4 illustrates the interplay among the sample size, the significance level (α) chosen, and the number of independent variables in detecting significant R-square (Hair et al, 2006)[327]. The table values are minimum R-square that the specified sample size will detect as statistically significant at the specified alpha (α) level with the power (probability) of 0.80.

[325] Hair et al, Multivariate Data Analysis, 6th Ed, Printice Hall Pb; 2006
[326] Mason et al, (1991), "Collinearity, Power, and Interpretation of Multiple Regression Analysis", Journal of Marketing research, Vol.28, p268-280
[327] Hair et al, Multivariate Data Analysis, 6th Ed, Printice Hall Pb; 2006

Table 5.4 – Minimum Significant R-Square

	Significance level (a) = 0.01 Number of Independent Variables				*Significance level (a) = 0.05 Number of Independent Variables*			
Sample size	**2**	**5**	**10**	**20**	**2**	**5**	**10**	**20**
20	45	56	71	NA	39	48	64	NA
50	23	29	36	49	19	23	29	42
100	13	16	20	26	10	12	15	21
250	5	7	8	11	4	5	6	8
500	3	3	4	6	3	4	5	9
1000	1	2	2	3	1	1	2	2
NA = Not applicable								

(Minimum R-square that can be found statistically Significant with a Power of 0.80 for Varying Numbers of Independent Variables and Sample Sizes)
Source: (Hair, 2006)[328]

The researcher must be aware of the anticipated power of any proposed Multiple Regression Analysis. The researcher can determine the sample size needed to detect effects for individual independent variables given the expected effect size (correlation), the α level, and the power desired (Cohen et al, 2002)[329]. The general rule is, the ratio of independent variables and sample size should not fall below 1:5. The maximum can be 1:20. When this level of samples is obtained, the results are generalizable as the samples become representative of population. A stepwise procedure can be employed to increase the ratio to 1:50, however this

[328] Hair et al, Multivariate Data Analysis, 6th Ed, Printice Hall Pb; 2006
[329] Cohen et al,Applied Multiple Regression / Correlation Analysis for the Behavioral Sciences, 3rd ed, Hillsdale, NJ: Lawrence Erlbaum Associates; 2002

ratio can lead to a tendency towards the results being sample specific (Wilkinson, 1975)[330].

When the ratio falls below 1:5, there is a risk of over fitting the variate to the sample, making the results too specific to the sample and thus lacking generalizability. Each observation represents a separate and independent unit of information (i.e., one set of values for each independent variable). Ideally the researcher should dedicate a single variable to perfectly predicting only one observation, second variable to another observation and so forth. If the sample is relatively small, then predictive accuracy could be quite high and many of the observations could be perfectly predicted. The number of estimated parameters (regression coefficients and the constant) equals the sample size, perfect prediction will occur even if all the variable values are random numbers. This scenario is totally unacceptable and it is extreme over fitting as the stimated parameters relate only to the sample data and no generalizability is possible. Whenever a variable is added to the regression equation, R-square value will increase.

The degree of generalizability is represented by the degrees of freedom.

Degrees of Freedom (df) = sample size – Number of estimated Parameters

Or

Degrees of freedom (df) = N – (Number of independent variables + 1)

The larger the degree of freedom, the better is the generalizability. Degrees of freedom increases for a given sample if the number of independent variables reduces. The objective is to achieve highest predictive accuracy with large degrees of freedom. When there is perfect prediction, with the number of estimated parameters equaling the sample size, zero degrees of freedom appears. The researcher is advised to reduce the number of independent variables to improve predictive accuracy. Degrees of freedom indicate the generalizability of the results for a given size of samples. There thumb rules are, (i) Simple

[330] WilKinson. L., (1975), tests of Significance in stepwise regression, Psychological Bulletin, Vol.86 , p168-174

regression can be effective with the sample size of 20, but maintaining power at 0.80 in Multiple Regression requires a minimum sample of 50 and preferably 100 for most of the research situations. (ii) The minimum ratio of observations to variables is 5:1. Preferred ratio is 15:1 or 20: 1, which would increase further if stepwise estimation is used. (iii) Maximizing the degrees of freedom improves generizability and addresses both model parsimony and sample size concerns.

5.2.3.2 Creating Additional Variables

Problems appear when a non-metric data such as gender or occupation had to be incorporated into a regression equation. Regression is meant for metric data. This introduction of non-metric data will lead to non linear equations of regression. In such situations new variables are created by transformations. *Variable transformation* methods (Box et al, 1964)[331] are used primarily to improve or modify relationship between dependent and independent variables and to enable the use of non-metric variables in the regression variate. *Data transformations* are achieved by trail and error, t make the analysis to best represent the actual data set. All these transformations are carried out by the statistical software used for regression analysis.

When dependent variable is measured as a dichotomous (0, 1) variable, either discriminant analysis or logistic regression is appropriate. When independent variables are non-metric, dummy variables are introduced. If there are non-metric variables in k categories k-1 dummy variables are introduced in multiple regression analysis. The most common format of dummy variable coding is '*indicator coding*', where each category of the non-metric variable is represented by either 1 or 0. The regression coefficients of dummy variables represent differences on the dependent variable for each group of respondents from the reference category (the omitted group that received all zeros). These group differences can be assessed directly because the coefficients are in the same units as the dependent variable. This form of coding is most appropriate when a

[331] Box et al, (1964), "Analysis of Transformations", Journal Of Royal statistical society, Vol.B.26, p 211-243

logical reference group is present, as in the case of an experiment. An alternative method of dummy variable coding is termed as '*effects coding*'. It is the same as indicator coding except that the comparison or omitted group (the group that got all zeros) is given a value of -1 instead of zero for the dummy variables. The coefficients represent differences for any group from the mean of all the groups rather than from the omitted group. Both the forms of coding give same predictive results, coefficient of determination and regression coefficients for the continuous variables. Interpretation of results will depend on the coding of dummy variables. There are thumb rules for *variable transformations.* They are, (i) Non-metric can only be included in regression analysis by creating dummy variables. (ii) Dummy variables can only be interpreted in relation to their reference category (Hair et al, 2006)[332].

The estimation procedures for models using both types of independent variables are the same except for the error terms. In the random effects models, a portion of the random error comes from the sampling of the independent variables. The statistical procedures based on the fixed model are quite robust. Using the statistical analysis as if a fixed model is being dealt with will be appropriate as a reasonable approximation.

5.2.4 Assumptions in Multiple Regression Analysis

To improve the predictive accuracy of the model, the researcher needs to lay down a few assumptions about the relationship between the dependent and independent variables that affect the least square procedure used for Multiple Regression. There are 4 types of assumptions made. (i) Linearity of the phenomenon measured. (ii) Constant variance of the error terms. (iii) Independence of error terms. (iv) Normality of the error term distribution.

In Multiple Regression once the variate is derived, it acts collectively in predicting the dependent variable, which necessitates assessing the assumptions

[332] Hair et al, Multivariate Data Analysis, 6th Ed, Printice Hall Pb; 2006

not only for individual variables but also for the variate itself. Testing assumptions occur before as well after predicting the model. The principal measure of prediction error is the *residual* – the difference between the observed and the predicted values of the dependent variable. Some form of standardizations is recommended to make the residuals comparable while predicting dependent variable. Studentized residual – the most widely used values that correspond to t-values. Plotting residuals Vs Independent variables is a basic method of identifying assumption violations for the overall relationship. They are also plotted against predicted dependent values. These plots are compared with null plot where all the assumptions are completely met. The patterns are compared to understand the error of the variate (Hair et al, 2006)[333].

5.2.4.1 Linearity of the Phenomenon

The linearity of the relationship between dependent and independent variables represents the degree to which the change in the dependent variable is associated with the independent variable. The regression coefficient is constant across range of values for the independent variable. The concept of correlation is based on the linear relationship, thus making it a critical issue in regression analysis. Linearity of a bivariate relationship is examined through residual plots. Any consistent curvilinear pattern in the residuals indicates that the corrective action will increase both predictive accuracy of the model and the validity of the estimated coefficients. The corrective actions could be; transforming the data values (logarithm, square root etc.) of one or more independent variables to achieve linearity; Directly including non linear relationships in the regression model, such as creation of polynomial terms; Using specialized methods such as nonlinear regressions specifically designed to accommodate the curvilinear effects of independent variables or more complex nonlinear relationships (Hair et al, 2006)[334].

[333] Hair et al, Multivariate Data Analysis, 6th Ed, Printice Hall Pb; 2006
[334] Hair et al, Multivariate Data Analysis, 6th Ed, Printice Hall Pb; 2006

The residual plot reveals the combined effects of all independent variables. Determining the independent variable for a corrective action from the pattern seen from such a plot could not help. So, partial regression plots are prepared where, relationship of individual independent variables with dependent variable are plotted separately controlling the other variables. So, the unique relationship between a specific independent variable and the dependent variable can be come obvious. These plots when superimposed on a residual plot reveal whether the variable violates the linearity assumption or not.

5.2.4.2 Constant Variance of the Error Term

The presence of heteroscedasticity (unequal variances) is one of the most common assumption violations. Diagnosis is made with residual plots or simple statistical tests. Plotting the residuals against the predicted dependent values and comparing them to a null plot shows a consistent pattern if the variance is not constant. Many a times a number of violations occur simultaneously such as non linearity and heteroscedasticity. All statistical softwares provide tests for homogeneity of variance which measures the equality of variances. If heteroscedasticity is present, two remedies are available. One is, if the violation can be attributed to a single independent variable through analysis of residual plots, then the procedure of weighted least squares can be employed; the other is to execute variance stabilizing transformations that allow transformed variables t exhibit homoscedasticity(equality of variance) (Hair et al, 2006)[335].

5.2.4.3 Independence of the Error Terms

In regression, researchers assume that each predicted value is independent, which means that the predicted value is not related to any other prediction (i.e., they are not sequenced by any variable). This occurrence can be identified by plotting the residuals against any possible sequencing variable. If the residuals are independent the pattern should appear random and similar to the null plot of

[335] Hair et al, Multivariate Data Analysis, 6th Ed, Printice Hall Pb; 2006

residuals. Violations will be identified by a consistent pattern in the residuals. They can be identified when there is a dependence of error term with time and when there is a dependence of error term with respect to the occurrence of events (Hair et al, 2006)[336]

5.2.4.4 Normality of the Error Term Distribution

The most frequently encountered assumption is the normality of the error term distribution or the violation of the non normality of the dependent or independent variables or both (Seber, G.A, 2004)[337]. Simplest diagnosis is to plot the independent variables against dependent variables and obtain a histogram ideally. For smaller samples this method is ill formed while plotting. A better method is the use of normal probability plots. They differ from residual plots in that the standardized residuals are compared with the normal distribution. The normal distribution makes a straight diagonal line, and the plotted residuals are compared with the diagonal. If a distribution is normal, the residual line closely follows the diagonal. The same procedure can compare the dependent and or independent variables separately to the normal distribution (Daniel et al, 1999)[338].

The rules of Thumb for assessing statistical assumptions are,

(i) Testing assumptions must be done not only for each dependent and independent variable, but for the variate as well.

(ii) Graphical analyses (i.e., Partial regression plots, residual plots, normal probability plots) are the most widely used methods of assessing assumptions for the variate.

(iii) Remedies for problems found in the variate must be accomplished by modifying one or more independent variables.

[336] Hair et al, Multivariate Data Analysis, 6th Ed, Printice Hall Pb; 2006
[337] Seber, G.A., Multivariate Observations. NewYork:Wiley; 2004
[338] Daniel et al, Fitting Equations to Data, 2nd Ed, NewYork: Wiley-Interscience; 1999

5.2.5 Estimating the Regression Model and Assessing the Overall Model Fit

The researcher is expected accomplish 3 basis tasks; To select a method for specifying the regression model to be estimated; To assess the statistical significance of the overall model in predicting the dependent variable; and To determine whether any observations exert any undue influence on the results. In '*confirmatory specification*', the researcher chooses the exact set of independent variables. It should be noted that the selection should not be empirical but based on theoretical justification. The other approach is '*sequential search*', which employs stepwise selection or forward addition and backward elimination techniques to select variables one after another to bargain for better predictive accuracy.

Stepwise Estimation: This method of estimation has a framework as given below (Hair et al, 2006)[339].

(i) Start with simple regression model by selecting the one independent variable that is the most highly corrected with the dependent variable. The equation would be, Y= b0 + b1X1.

(ii) Examine the partial correlation coefficients to find an additional independent variable that explains the largest statistically significant portion of the unexplained – error – variance remaining in the first regression equation

(iii) Recomputed the regression equation using the two independent variables and examine the partial F value for the original variable in the model to see whether it still makes a significant contribution, given the presence of new independent variable. If it does not, eliminate the variable. This ability to eliminate variables already in the model distinguishes the stepwise model form the forward addition/backward elimination models. If the original variable still makes a significant contribution, the equation would be, Y = b0 + b1X1 + b2X2.

(iv) Continue this procedure by examining all independent variables not in the model to determine whether one would make a statistically significant addition

[339] Hair et al, Multivariate Data Analysis, 6th Ed, Printice Hall Pb; 2006

to the current equation and thus should be included in the revised equation. If a new independent variable is included, examine all the independent variables previously in the model to judge whether they should be kept.

(v) Continue adding independent variables until none of the remaining candidates for inclusion would contribute a statistically significant improvement in the predictive accuracy. This point occurs when all the remaining partial regression coefficients are non-significant

A potential bias in the stepwise procedure results from considering only one variable for selection at a time. Multicollinearity among the independent variables can substantially affect all sequential estimation methods. Examining ten different factors stepwise with five different dependent variables is cumbersome. The factors are grouped from individual variables that may be collinear and due to multi collinearity issues this method does not suit our research purpose of finding the strongest fit to explore the linkage between the factors of Organizational Intelligence and five different variables of Organizational Performance.

Forward Addition and Backward Elimination: The procedures of forward addition and backward elimination procedures are largely trial and error processes for finding the best regression estimates. The forward addition model is similar to the stepwise procedure in that it builds the regression equation starting with the single independent variable. The backward elimination procedure starts with the regression equation including all the independent variables and then deletes independent variables that do not contribute significantly.

The primary distinction between stepwise procedure and the forward-addition and backward-elimination procedure is, in stepwise method, addition or deletion of a variable at each stage is possible where in, in forward-addition, backward elimination procedures addition of variables in a later stage is not possible. This flexibility makes stepwise procedure preferred method for researchers. The procedures of variables addition and elimination would not be suitable for our

analysis as we are interested in finding the strongest regression fit model to establish the linkage between Organizational Intelligence and Organizational Performance which is obtained by regressing the factors of Organizational Intelligence with the five dependent variables chosen to represent Organizational Performance.

Caveats to the above Sequential Search Methods*:* there are three key caveats to the sequential search methods of estimations discussed above.

(i) The multicollinearity among independent variables has substantial impact on model specification. Although the sequential search approaches will maximize the predictive ability of the regression model, the researcher must be careful in using these methods in establishing the impact of independent variables without considering multicollinearity among independent variables.

(ii) All sequential search methods create a loss of control for the researcher. Though the researcher specifies the variables to be considered for the regression variate, it is the estimation technique, interpreting the empirical data specifies the final regression model.

(iii) In stepwise procedure, multiple significance tests are carried out in the model estimation process. To ensure the overall error rate across all significance tests is reasonable, the researcher should employ more conservative thresholds (e.g., 0.01) in adding or deleting variables.

Combinatorial Approach: *T*his approach suggests regression of all possible subsets of independent variables and the best fitting set of variables is chosen. This procedure is not preferred as it does not consider multicollinearity, identification of outliers and influentials and the interpretability of results in this research. The rules of the thumb of estimation techniques are; Irrespective of the estimation techniques, theory must be the guiding factor for evaluating the final regression model; Confirmatory specification method allows direct testing of pre-specified model. This is also the most complex from the perspective of specification error, model parsimony and predictive accuracy; Sequential search methods make the estimation fully automated leaving the researcher with out any control on the selection of variables; Combinatorial approach removes

control from the researcher, however gives an understanding of parallel models of predictive accuracy. Using more than one method in combination may provide a balanced perspective.

Thus we proposed to regress all the ten independent factors collected from exploratory factor analysis with the five different dependent variables of financial performance and study the stronger fit of the models from R-Square value and chose the strongest fit as the best explaining model of IO-OP relationship.

5.2.5.1 Testing the Regression Variate for Meeting the Regression Assumptions

With independent variables selected and regression coefficients estimated, the researcher must now assess the estimated model for meeting the assumptions underlying multiple regression. The individual variables as well the variate must meet the assumptions of linearity, constant variance, independence and normality. If substantial violations are found the researcher must take corrective actions on independent variables and re-estimate the regression model.

5.2.5.2 Examining the Statistical Significance of the Model

If Researchers take random samples of respondents and estimate regression equation for the sample, the regression coefficient values will differ for each set of sample and the sampling error will cause this situation. Researchers usually chose only one sample set and estimate the regression model. This approach demands the tests of the random variation explained – coefficient of determination – and regression coefficient.

Testing the Coefficients of Determination: To test the hypotheses that the amount of variation explained by the regression model is more than the baseline prediction (i.e., the R-square is significantly greater than zero). The *F Ratio* is

calculated as the ratio of the ratios of the sum of squares per degree of freedom for regression and residuals respectively.

F ratio = E1/E2;

where,
E1 = Sum of squares / degrees of freedom: (from regression model);
E2 = sum of squares / degrees of freedom: (from unexplained variance - the residual).

Intuitively, if the ratio of the explained variance to the unexplained is high, the regression variate must be significant in explaining the dependent variable. Larger the R-square values, higher the F values. Statistical significance is the impact of sampling error. Statistically significant values are all practically significant. It is to be noted that for larger samples smaller R-square can be of high significance.

Adjusting the Coefficients of Determination: Addition of a variable in the regression model will increase R-square value. Generalizability of the model should be depending on R-square value as R-square value may increase even if a non-significant predictor variable is introduced. This demands an adjustment based on the number of independent variables and sample size combination. Adding non significant variables in the regression model will change R-square and this is adjusted R-square – adjusted coefficient of determination. The adjusted R-square is useful in comparing across the regression equations involving different numbers of independent variables or different sample sizes because it makes allowances for the degree of freedom for each model (Hair et al 2006)[340].

Significance Tests of Regression Coefficients: Significance testing of a regression co-efficient is a statistically based probability estimate of whether the estimated

[340] Hair et al, Multivariate Data Analysis, 6th Ed, Printice Hall Pb; 2006

coefficients across a large number of samples of a certain size will be different from zero. To make this judgment, a confidence level must be established around the estimated co-efficient. If the confidence interval does not include the value of zero, then it can be said that the coefficient's difference from zero is statistically significant. To make this judgment, the researcher relies on 3 concepts. (i) Establishing significance level (α) denotes the chance the researcher is willing to take of being wrong about whether the estimated coefficient is different from zero. A value typically used is 0.05. as the researcher desires a smaller chance of being wrong, and sets the significance level smaller (0.01 or 0.001), the statistical test becomes more demanding. Increasing the significance level to a higher value (0.1) allows for a larger chance of being wrong, but makes it easier to conclude that the coefficient is different from zero. (ii) The sampling error is being the cause for variation in the estimated regression coefficients for each sample drawn from a population. For small sample sizes, the sampling errors are larger and the estimated coefficients will most likely vary widely from sample to sample. As the size of the sample increases, the samples become more representative of the population (i.e., sampling error decreases), and the variation in the estimated coefficients for these large samples become smaller. This relationship holds true until the analysis is estimated using the population. Then the need for significance testing is eliminated as the sample size is equal to population and thus exact representative of the population (i.e., no sampling error). (iii) The standard error is the expected variation of the estimated coefficients (both the constant and regression coefficients) due to sampling error. The standard error acts like the standard deviation of a variable by representing the expected dispersion of the coefficients estimated from repeated samples of this size.

With the significance level selected and the standard error calculated, we can establish a confidence interval for a regression coefficient based on the standard error. There are 3 key angles to be looked at while checking the confidence interval. They are; (i) the researcher sets the significance level from which the confidence interval is derived (e.g., a significance level of 5% for a large sample establishes the confidence interval at ±1.96×standard error). A coefficient is deemed statistically significant if the confidence interval does not include zero.

(ii) if the sample size is small, sampling error may cause the standard error to be so large that the confidence interval includes zero. However if the sample size is larger, the test has greater precision because the variation in the coefficients become less (i.e., the standard error is smaller). Larger samples do not guarantee that the coefficients will not equal zero, but instead make the test more precise. (iii) a coefficient being statistically significant does not guarantee the practical significance. Evaluating the sign of the coefficient is thus crucial (Hair et al, 2006)[341].

A simple regression model implies hypotheses about 2 estimated parameters; the constant and regression coefficient. To assess the significance level, the appropriate test is t-test which is available in all regression analysis programs. The t value of the coefficient is the coefficient divided by the standard error. T value represents the number of standard errors that the coefficient is from zero. For example, a regression coefficient of 2.5 with the standard error of 0.5 would have a t value of 5.0 (i.e., the regression coefficient is 5 standard errors from zero). To determine whether the coefficient is significantly different from zero the computed t value is compared with the table value for the sample size and the confidence interval selected. If our value is greater than the table value, we can be confident that the coefficient has a statistically significant effect in the regression variate for the selected confidence level.

Most computer programs calculate the significance level for each regression coefficient's t value, showing the significance level at which the confidence interval would include zero. The researcher can then assess whether this level meets the desired level of significance. For example, if the statistical significance of the coefficient is 0.02, then we can say that it was significant at the 0.05 level because it is less than 0.05, but not significant at 0.01 level. It is to be noted that the estimated parameters would be different from zero within specified level of acceptable error (Hair et al, 2006)[342].

[341] Hair et al, Multivariate Data Analysis, 6th Ed, Printice Hall Pb; 2006
[342] Hair et al, Multivariate Data Analysis, 6th Ed, Printice Hall Pb; 2006

5.2.5.3 Identifying Influential Observations

There are generally sets of observations that influence the model by staying outside the data set. They have disproportionate effect on the model. *Outliers* are the observations that have large residual values and can be identified only with respect to a specific regression model.

Outliers were traditionally the only form of influential observation considered in regression models, and specialized regression methods (e.g., robust regression) were even developed to deal specifically with outlier's impact on the regression results (Rousseeuw et al, 2003)[343]. The key aberration in the observation is the presence of heteroscedasticity due to outliers. *Leverage points* are observations that are distinct from the remaining observations based on their independent variable values. Their impact is particularly noticeable in the estimated coefficients for one or more independent variables. Influential observations are the broadest category, including all observations that have a disproportionate effect on the regression results. Influential observations potentially include outliers and leverage points but may include other observations as well. Also, not all outliers and leverage points influence observations (Batnett et al, 1994)[344]. *Identifying Influential Observations* are difficult many a time through traditional analysis of residuals for outliers. Their patterns of residuals go undetected because the residual for the influential points (the perpendicular distance from the line of regression) would not be as large as to be classified as an outlier. Thus, focusing only on large residuals would generally ignore these influential observations. Reinforcing, conflicting and shifting of the regression lines will occur due to influential observations. Table 5.5 shows the aberrations and the remedy (Hair et al, 2006)[345].

[343] Rousseeuw et al, 'Robust regression and Outlier Detection', NewYork: Wiley; 2003.
[344] Barnett et al, Outliers in Statistical Data, 3rd ed., NewYork: Wiley Pb; 1994
[345] Hair et al, Multivariate Data Analysis, 6th Ed, Printice Hall Pb; 2006

Table 5.5 - Influential Observations and Remedies

Item No	*Influential Observation*	*Remedy*
1	An error in the observations or in data entry	Correcting the data or deletion of the case.
2	A valid exceptional observation that is explained by an extraordinary situation	Remedy by deletion of the case unless variables reflecting the extraordinary situation are included in the regression equation.
3	An exceptional observation with no likely explanation	Presents a special problem as it doesn't permit the deletion of the case. Inclusion of it cannot be justified either. Analyzing the entire data set with and without the inclusion of this observation for assessment is suggested.
4	An ordinary observation in its individual characteristics but exceptional in its combination of characteristics	Indicates modifications to the conceptual basis of the regression model and should be retained.

In all of the situations the observations are to be deleted. Each case should be individually studied by the researcher before the deletion as in some outliers cannot be deleted as well. The thumb rules of statistical significance and influential observations are; (i) Always ensure practical significance while using large sample sizes, because the model results and regression coefficients could be deemed irrelevant even when statistically significant due just to the statistical power arising from large sample sizes; (ii) Use the adjusted R-square as an overall measure of model's predictive accuracy; (iii) Statistical significance is required for a relationship to have validity, but statistical significance without theoretical support does not support validity; (iv) Although outliers may be easily identifiable, the other forms of influential observations requiring more specialized diagnostic methods can be equal to or even more impacting on the results.

It is to be noted that in this research study, the presence of outliers and leverage points are eliminated by closed end questionnaire measured with Likert scales.

5.2.6 Interpreting the Regression Variate

Prediction and Explanation are the integral parts of interpreting regression variate and independent variables. While explaining a regression model, the regression coefficients become indicators of the impact of the independent variables on the dependent variable. Most of the time, the regression coefficients do not explain the relationship completely. To avoid this issue of regression coefficients pretending to explain the variate with higher accuracy, it is necessary to make the independent variables in comparable scales and variability. The coefficient thus obtained after this is called beta coefficient by research arena.

Standardizing Regression coefficients: the variation in the response scale and variability across variables makes direct interpretation problematic. Standardization converts variables to a common scale and variability – the most common being a mean of zero and standard deviation of one. Thus all variables become comparable. Multiple regression not only gives regression coefficients but also coefficients resulting from the analysis of standardized data termed beta (β) coefficients. The problems of dealing with different units are eliminated in these coefficients. The relative impact on the dependent variable by one standard deviation in either variable is reflected better. However beta coefficients are used with 2 cautions. They are; (i) beta coefficients are used as a guide to understand the relative importance of individual independent variable only when collinearity is minimal; (ii) beta values can be interpreted only in the context of other variables in the equation (Hair et al, 2006)[346].

[346] Hair et al, Multivariate Data Analysis, 6th Ed, Printice Hall Pb; 2006

5.2.7 Assessing Multicollinearity

The key issue in interpreting regression variate is the correlation among the independent variables. This problem is due to data and not because of the model specification. The ideal situation for a best model interpretation would be high collinearity between independent and independent variables and less correlation among independent variables. Use of factor scores from factor analysis fixes the problem of multicollinearity among the independent variables. There are 3 key tasks to be done by the researchers to handle the issue of multicollinearity. They are; (i) Assess the degree of multicollinearity; (ii) Determine the impact of results; (iii) Apply the necessary remedies if needed.

Identifying Multicollinearity: The simplest and most obvious means of identifying collinearity is the examination of correlation matrix for the independent variables. The presence of high correlations (.90 and above) is the first indication of substantial collinearity. Lack of high correlation values, does not ensure lack of collinearity. Collinearity may be due to the combined effect of 2 or more other independent variables. To assess multi collinearity, we need a measure expressing the degree to which each independent variable is explained by the set of other independent variables. In simple terms, each independent variable becomes dependent variable and regressed against the remaining independent variables. The 2 most common measures of assessing both pair-wise and multiple variable collinearity are tolerance and its inverse, the variance inflation factor (hair et al, 2006)[347].

A direct measure of multicollinearity is '*Tolerance*' – the amount of variability of the selected independent variable not explained by the other independent variables. For any regression model with 2 or more independent variables, the tolerance can be simply defined in 2 steps. Step1: take each independent variable, one at a time, and calculate R-square. This is the amount that the independent variable is explained by all of the other independent variables in the regression model. In this process, the selected independent variable is made a dependent

[347] Hair et al, Multivariate Data Analysis, 6th Ed, Printice Hall Pb; 2006

variable predicted by all the other remaining independent variables. Step2: tolerance is then calculated as (1 – 'R-square'). For example, if the other independent variables explain 25% of the independent variable X1, (R-square = 0.25), then the tolerance value of X1 is 0.75 (i.e., 1.0 - 0.25 = 0.75) (hair et al, 2006)[348].

Another measure of Multicollinearity is '*Variance Inflation Factor*' (VIF) which is calculated simply as the inverse of the tolerance value. In the preceding example with a tolerance of 0.75, the VIF would be 1.33 (1.0 / 0.75 = 1.33). Thus instances of higher degrees of multicollinearity are reflected in lower tolerance values and higher VIF values. The VIF gets its name from the fact that the square root of the VIF is the degree to which standard error has been increased due to multicollinearity. VIF translates the tolerance value which directly expresses the degree of multicollinearity, into an impact of estimation process. As the standard error is increased, it makes the confidence intervals around the estimated coefficients larger, thus making it harder to demonstrate that the coefficient is significantly different from zero(Hair et al, 2006)[349].

The Effects of Multicollinearity: The effects of multicollinearity can be categorized from the point of view of estimation or explanation. In either case the underlying reason is the same. Multicollinearity creates 'shared' variance between variables, thus decreasing the ability to predict the dependent measure. The ability to ascertain the relative roles to the independent variables for predicting dependent variable is also reduced.

Impacts of Estimation: Multicollinearity can have substantial effects not only on the predictive ability of the model but also on the estimation of the regression coefficients and their statistical significance tests. The extreme case multicollinearity is that, two or more variables are perfectly correlated, termed singularity, prevents the estimation of any coefficients. Although singularities may occur naturally among the independent variables, many times they are the

[348] Hair et al, Multivariate Data Analysis, 6th Ed, Printice Hall Pb; 2006
[349] Hair et al, Multivariate Data Analysis, 6th Ed, Printice Hall Pb; 2006

results of an error of including all the dummy variables used to represent a non metric variable, rather than omitting one as the reference category. Also actions such as including a summated scale along with the individual variables that created it will result in singularities. These singularities must be removed earlier than the estimation proceedings. As multicollinearity increases, the ability to demonstrate that the estimated regression coefficients are significantly different from zero can become markedly impacted due to increase in the standard error. This is a serious issue with smaller sample sizes, where the standard errors are larger due to sampling error (Hair et al, 2006)[350].

Apart from affecting statistical tests of the coefficients or the overall model, high degrees of multicollinearity can also result in regression coefficients being incorrectly estimated and even having the wrong signs. In some instances, the reversal of signs is expected and desirable. This is suppression effect. It denotes instances when the true relationship between the dependent and the independent variable has been hidden in the bivariate correlations (e.g., the expected relationships are non-significant or even reversed in sign). By adding more independent variables and including multicollinearity some unwanted shared variance is accounted for and remaining unique variance allows for the estimated coefficients to be in the expected direction (Cohen et al, 2002)[351].

Theoretically supported relationships are reversed due to multicollinearity demanding explanations from the researcher on the findings. In these instances, the researcher needs to revert to bivariate correlations to describe the relationship rather than the estimated coefficients that are impacted by multicollinearity. The reversal of signs may be encountered in all of the estimation procedures, but is seen more often in confirmatory estimation processes where a set of variables is entered into the regression model and the likelihood of weaker variables being affected by multicollinearity increased.

[350] Hair et al, Multivariate Data Analysis, 6th Ed, Printice Hall Pb; 2006
[351] Cohen et al,Applied Multiple Regression / Correlation Analysis for the Behavioral Sciences, 3rd ed, Hillsdale, NJ: Lawrence Erlbaum Associates; 2002

Impacts of Explanation: The effects of explanation are concerned primarily with the ability of regression procedure and the researcher to represent and understand the effects of each independent variable in the regression variate. As multicollinearity occurs (even at a relatively low levels of 0.30 or so), the process for identifying the unique effects of independent variables becomes increasingly difficult. Remember that the regression coefficients represent the amount of unique variance explained by each independent variable. As the multicollinearity results in larger portions of shared variance and lower levels of unique variance, the effects of the individual independent variables become less distinguishable. It is even possible to find those situations in which multicollinearity is so high that none of the independent regression coefficients are statistically significant, yet the overall regression model has a significant level of predictive accuracy (Hair et al, 2006)[352].

How much Collinearity is too much? – is a question addressed by all researchers who come across multicollinearity issue. Because the tolerance value is the amount of a variable unexplained by other independent variables, small tolerance values (high VIF values, VIF = 1 / Tolerance) denote high collinearity. A common cut off threshold is a tolerance value of 0.10 which corresponds to a VIF value of 10. Particularly when sample sizes are smaller, the researcher may wish to be more restrictive due to the increases in the standard errors from multicollinearity. With a VIF threshold of 10, this tolerance would correspond to standard errors being 'inflated' more than 3 times (square root of 10 = 3.16) what they would be without multicollinearity. Each researcher must determine the degree of collinearity that is acceptable, because most defaults or recommended thresholds still allow for substantial collinearity. For example, the suggested cut off for the tolerance value of 0.1 corresponds to a multiple correlation of 0.95. Moreover, a multiple correlation of 0.9 between one independent variable and all others will result in a tolerance value of 0.19. Thus any variable with tolerance value below 0.19 (or above a VIF of 5.3) would have a correlation of more than 0.9

[352] Hair et al, Multivariate Data Analysis, 6th Ed, Printice Hall Pb; 2006

5.2.8 Managing Heteroscedasticity

In classical liner equations, there is an equal spread of the disturbance term throughout. This is *Homoscedasticity.* When the spread is unequal, it becomes *Heteroscedasticity.* A critical assumption of the classical linear regression model is that the disturbances *'Ui'* have all the same variance '*σ-square*'. If this assumption is not satisfied, there is heteroscedasticity. Heteroscedasticity does not destroy the unbiasedness and consistency properties of OLS estimators. But these estimators are no longer minimum variance or efficient (i.e., they are not BLUE – best linear unbiased estimator). The BLUE estimators are provided by the method of weighted least squares, provided the heteroscedastic error variances (σ-square) are known. In the presence of heteroscedasticity, the variances of OLS estimators are not provided by the usual OLS – ordinary least squares formulae. But if we persist in using the usual OLS formulae, the t and the F tests based on them can be highly misleading, resulting in erroneous conclusions. Documenting the consequences of heteroscedasticity is easier than detecting it. There are several diagnostic tests available, but one cannot tell for sure which will work in a given situation. Even if heteroscedasticity is suspected and detected, it is not easy to correct the problem. If the sample is large, one can obtain white's heteroscedasticity corrected standard errors of OLS estimators and conduct statistical inference based on the standard errors. Otherwise, on the basis of OLS residuals, one can make educated guesses of the likely pattern of heteroscedasticity and transform the original data in such a way that in the transformed data there is no heteroscedasticity (Hair et al, 2006)[353].

5.2.9 Software for Multiple Regression Analysis Technique

'E-views' is one of the most widely used statistical software package for Multiple Regression Analysis. Like every Multiple Regression Analysis Software, this also takes care of issues of data transformation and heteroscedasticity, we preferred

[353] Hair et al, Multivariate Data Analysis, 6th Ed, Printice Hall Pb; 2006

E-views. Availability, usability and the familiarity with E-views are the reasons of choosing this software. Data analysis with 'E-views' software is discussed in chapter 6.

5.2.10 Hypotheses

Once the model is constructed, the hypotheses related to the factors that construct the model will be proposed. The null and the alternative hypotheses will be defined and validated from the model. This is discussed in chapter 6 in detail.

5.2.11 Validity of Multiple Regression Analysis

The Multiple Regression Models are usually validated with a new set of samples and the results are compared to establish the accuracy of the instrument. R-square value of the model will reveal predictive power of the model and hence the accuracy. The predictive power and the accuracy of the model are determined by the error term in the model equation. Error term or the Residual determines the model validity and the fit. Standard Error Estimate is also considered for determining the Accuracy of the Model.

5.3 Conclusion

In this chapter we discussed the selection of Factor Analysis and the dependence technique - Multiple Regression Analysis and the reason behind selecting them. We also discussed the thumb rules for decision making on factor selections, interpretations of results. Entire research design and execution plans of these analytical techniques are also conversed. In the next chapter we would discuss the Data Analysis and the Findings.