CHAPTER 5

Methods of Data Analysis

5.0 Overview

The main purpose of this chapter is to provide some information on the statistical techniques to be used in this study. The discussion starts with the choice of statistical software to be used for this study and the factors that influence the choice of statistical techniques is next in the discussion. The procedures of statistical techniques used to analyse the survey data, in the form of univariate (descriptive statistics), bivariate (ANOVA) and multivariate (exploratory factor analysis, multivariate ANOVA) analyses are next in the sequence of discussion. The assumptions under the use of each technique applied are also become the part of the discussion.

5.1 Choice of a Statistical Package

The availability of excellent computer packages enhances the effectiveness and efficiency of statistical data analysis. The variety of typical statistical software available includes the Statistical Analysis System (SAS), BMDP, and the Statistical Package for the Social Science (SPSS), Systat and Minitab. These packages do not completely overlap Statistical programmes within and a specific problem can be handled better through one package than other. These programmes are continually being updated although not all improvements are immediately implemented at each facility.

The Statistical Package for the Social Science (SPSS) version 20 for Windows is chosen as the computer programme for data analysis in the present study. It is widely used by SPSS is the social scientists and other professionals for statistical analysis since it is most sophisticated software available among others. The large arrays of programmes for univariate, bivariate and multivariate statistical analysis are given by SPSS (Green and
Salkind 2003). SPSS has been considered as the most widely available and generally used comprehensive statistical computer package available for marketing research (Malhotra and Birks 1999; Zikmund 2000), and because of that it is used in this research as a statistical programme for data analysis.

5.2 Choice of Statistical Techniques

To serve the specific purpose of research, there is a variety of statistical tests available for the use. There are number of factors which need to be considered to determine the use of appropriate statistical test. The objectives of the analysis focus of the analysis, sample type and size, parametric versus non-parametric tests and the level of measurement are important consideration that guides the choice of statistical test (Afifi and Clark 1996; Diamantopoulos and Schlegelmilch 1997; Burns 2000; de Vaus 2002). The discussion on the factors influencing the choice of statistical techniques is presented below.

5.2.1 Objectives of the Analysis

The most critical factor that guides the choice of statistical techniques is the purpose of the analysis. There are three basic roles performed by objectives of the analysis: (i) help to ensure that only relevant analysis to be undertaken; (ii) to provide a check on the comprehensiveness of the analysis and (iii) to avoid redundancy in the analysis (Diamantopoulos and Schlegelmilch 1997).

For the purpose of achievement of the overall purpose of the research, the objective of the analysis should be linked to the overall aim of the research so that achievement of the former should contribute towards the achievement of the latter. It happens because it easier is to derive appropriate objectives of the analysis if the research objectives are better-specified (Diamantopoulos and Schlegelmilch 1997). The focus of this study is to examine the influence of religiosity as predictors of consumer shopping orientation which leads to the selection of appropriate statistical techniques to find out causal relationships between them.
5.2.2 Focus of the Analysis

The analytical stance or orientation to be adopted are better be described by the focus of the analysis. In the view of Diamantopoulos and Schlegelmilch (1997), the focus of the analysis has three basic forms: description; estimation and hypothesis testing. It can take any one of these three foci. The summary picture of the sample in terms of the variables of interest is provided by a descriptive focus, while an estimation focus is used to generalise the sample information on the population as a whole. To test specific propositions regarding the variables of interest and use the evidence to draw conclusions for the whole population is the focus of hypothesis testing.

Hypothesis testing is the main focus of the statistical analysis in this study in which hypothesis related to the influence religiosity on some aspects of consumer shopping orientation are tested to deduce conclusions based on the empirical findings. There are two types of hypothesis which are used in this study and it includes: (1) difference hypotheses between samples and (2) hypotheses of association between variables. The results obtained from the hypothesis testing are presented in the next two chapters.

5.2.3 Sample Type and Size

The third important consideration in the choice of statistical techniques for the data analysis is the type and size of samples. The use of inferential statistics is not legitimate if the samples are selected by using probability sampling because probability sampling makes use of the sampling error concept which cannot be assessed where nonprobability sampling methods are used (Diamantopoulos and Schlegelmilch 1997, p. 66). If the sample size is not sufficiently large, some statistical procedures do not work. It is recommended to have a sample size of at least 30 for a simple analysis using non-parametric statistics while a minimum sample of 100 is required for parametric statistics (Diamantopoulos and Schlegelmilch 1997). However the use parametric statistics with a sample size of less than 100 is commonly found in many non-experimental consumer research (see, for example, Shim and Kotsiopulos 1991; McDonald 1995; Slowikowski and Jarratt 1997; Emenheiser,
In the present study, the requirement for parametric statistics are satisfied since the sample size of 750 is taken. The size of sample used in a particular analysis is another important consideration in the choice of statistical techniques. The use of different statistical procedures require different numbers of participants as presented in Table 5.1 and for the selection of the appropriate statistical techniques it needs to be considered.

### TABLE 5.1 Rules of thumb for sample size selection

<table>
<thead>
<tr>
<th>Statistical analysis</th>
<th>Minimum size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chi-square</td>
<td>5 per cell</td>
</tr>
<tr>
<td>t-test, ANOVA, MANOVA</td>
<td>30 per cell</td>
</tr>
<tr>
<td>Factor analysis</td>
<td>50 - 100</td>
</tr>
<tr>
<td>Multiple regression analysis</td>
<td>50 - 300</td>
</tr>
</tbody>
</table>

5.2.4 The Level of Measurement

The fourth factor that contributes to the choice analytical technique is the level of measurement of variable. The level of measurement of variables is defined as how the categories of the variable relate to one another (de Vaus 2002). The level of sophistication in data analysis can be determined by the level of measurement. For more sophisticated analysis, the higher level of measurement is required. (Diamantopoulos and Schlegelmilch 1997). There are four main levels of measurement: ratio, interval (also called continuous), ordinal and nominal (also called categorical or qualitative). The application of Parametric statistics is limited to metric data (ratio and interval) while non-parametric statistics can be applied to both metric and non-metric (nominal) data. To analyse the data in the present study, parametric procedures are used because the data collected in this study are largely in the form of metric measurement.

The extent to which the variables differ in terms of their level of measurement and the number of variables to be analysed simultaneously determines the type of analysis to be applied. A method of univariate data analysis is used if only one variable is to be analysed. Bivariate analysis is used if two variables are to be analysed at a time. When multiple

---

dependent and independent variables are to be analysed simultaneously, one can use Multivariate analysis (Hair et al. 1998; Tabachnick and Fidell 2001). All three types of statistical analysis are used in the present study for the purpose of data analysis.

5.2.5 Distribution Pattern of the Data

The fifth factor that determines the type of statistical test to be used is the distribution pattern of the data. In general statistical test can be grouped in two major categories: parametric and nonparametric. Since data used in parametric tests are derived from interval and ratio measurements, it is considered more when the likelihood model (i.e. the distribution) is known, except for some parameters (Hair et al. 1998). Apart from interval and ratio measurements, non-parametric tests are also used, with nominal and ordinal data (Forza 2002). Experts on non-parametric tests (Hollander and Wolfe 1999) claim that nonparametric tests are as powerful as parametric tests.

The choice between these two is guided by ability of the measurement to fulfil the assumptions of parametric test. In the opinion of Burns (2000, p. 151-152), the use of parametric test is based on fulfilment of three assumptions. The first assumption is equality of interval for data collection, (e.g. Likert scale); second, data should be normally distributed or closely so, and third, the amount of random or error variance should be equally distributed among the different analyses. The non-parametric tests can be used to analyse the data if If these assumptions are not met (Forza 2002). The reason behind this is that application of non-parametric or distribution-free tests do not specify conditions about the shape or character of the distribution of the population from which samples are drawn.

Some statisticians strongly believe that parametric tests are comparatively robust. This means that “it is unlikely that the percentage probability will be very inaccurate unless the data do not meet the assumption at all, i.e. are not on an interval scale and/or are distributed in a very asymmetrical fashion” (Burns 2000, p. 152). In fact, it is common that data do not follow univariate normal distributions or much less multivariate normal distributions in data collected in the behavioural and social sciences (Micceri 1990). Many times the scales used by researchers are “dichotomous or ordered categories” rather than
truly continuous is the one of the reason suggested by West, Finch and Curran (1995, p. 57). Therefore, there is a probability of non normal data in the present study because the respondents give reply to items based on a 5-point Likert-type scale. So before applying any test on the data set in the present study, data are checked for normality by running a test of measures of central tendency (mean and standard deviation) is for each of the variables in the study. Further, to judge the normality of the distribution of the data, the skewness and kurtosis of each variable are also examined

The extent of symmetry of a distribution is described by Skewness and the mean of the skewed variable is not in the centre of the distribution for a given standard distribution (Norusis 1990). If the values for skewness and kurtosis are zero, the observed distribution is exactly normal (Hair et al. 1998; Coakes and Steed 2001) with a measure of skewness of +3.0 is usually regarded as a strong deviation from normality.

5.3 Statistical Tests to be used in this Study

To analyse the survey data in the present study, the types of various statistical analysis applied are summarised in Table 5.2. These are in the form of descriptive statistics as a univariate analysis, analysis of variance (ANOVA) as a bivariate analysis and factor analysis, multivariate analysis of variance (MANOVA) as a multivariate analysis.

The selection of the statistical techniques is consistent with the research aims and objectives, characteristics of the data and properties of the statistical techniques (Malhotra and Birks 1999). The basic purpose behind the use of multiple techniques for data analysis is to achieve the objectivity, rigour and logical reasoning in examining the research problems. The various methodological issues and assumptions associated with each technique are discussed in the following sections.

**TABLE 5.2 Summary of statistical tests used for data analysis**

<table>
<thead>
<tr>
<th>Independent variables</th>
<th>Dependent variables Statistical procedures</th>
<th>Statistical procedures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demographic Characteristics</td>
<td>____</td>
<td>Descriptive</td>
</tr>
<tr>
<td>Religiosity Information sources</td>
<td>____</td>
<td>Exploratory factor analysis</td>
</tr>
</tbody>
</table>
Methods of Data Analysis

5.4 Univariate Analysis

If a single variable is analysed without reference to the other variable, it is known as univariate analysis (Tabachnick and Fidell 2001). In this research, univariate descriptive statistics are used to: (a) to know the distribution patterns of the data (e.g. normality of the data); (b) to describe the basic demographic characteristics of the samples obtained from the survey; and (c) to provide a descriptive analysis of responses. The demographic characteristics of the respondents are presented in the succeeding chapter while the Appendix C denotes descriptive analysis of responses.

5.5 Bivariate Analysis

In bivariate data analysis, two variables are analysed simultaneously to study the relationship between the variables (Tabachnick and Fidell 2001). Analysis of variance (ANOVA) is the main bivariate analysis carried out in this study. The subsequent section briefly describes various issues and assumptions related to the application of Analysis of variance (ANOVA).

5.5.1 Analysis of Variance (ANOVA)

Analysis of variance (usually abbreviated as ANOVA) is an extension of independent t test and is used to check whether there exist significant differences among two or more means. ANOVA is the most flexible and frequently used quantitative technique in marketing and consumer behaviour research (Malhotra, Peterson and Kleiser 1999). One-way ANOVA is applied in this study to test the hypotheses of mean differences in consumer behaviour (i.e.
use of information sources, shopping orientations) among those affiliated to religion and having different levels of religiosity.

The reason behind the use of ANOVA in the present study as the statistical methodology is that the researcher wants to compare mean differences of the constructs among the groups in which the constructs of interest (e.g. use of information sources and shopping orientation) are being measured on an interval scale and the groups (e.g. religious affiliations) are considered as the factors which are categorical. Also, for the purpose of plotting and interpretation, the predicted values of independent variables which are equal to mean values will also be desirable. The use of ANOVA has following advantages. First, it depicts mean differences of three or more groups without the reason for the cause of differences. Second, it provides a more sensitive test of a factor where the error term may be reduced (Cramer and Howitt 2004).

The key statistic use in ANOVA is the variance ratio (F), which measures whether the differences in the means of the groups formed by values of the independent variables (or combinations of values for multiple independent variables) are not occured by chance. ANOVA will give the same results as the t-test for independent samples if only two means are compared (Sirkin 1995). The differences between two estimates of variance is the base of the F-ratio. The first estimate which comes from variability among scores within each group is considered as a random or error variance and the second estimate which comes from variability in group means is considered as a reflection of group differences plus error. The numerator represents variance associated with differences among sample means and the denominator represents the variance associated with error (Tabachnick and Fidell 2001). The larger value of the F-ratio indicates bigger differences between the means of the groups making up a factor in relation to the differences within the groups and there are more chances that it is statistically significant (Cramer and Howitt 2004). On the other side, if the group means do not differ substantially, one can inferred that the dependent variables are not affected by independent variables. The subsequent section deliberates the assumptions that underlie ANOVA.

**ASSUMPTIONS:** ANOVA makes certain assumptions as in the case of all other parametric tests about data so that they can be analysed in this way. Application of
ANOVA requires three major assumptions to be met (Maxwell and Delaney 1990; Jaccard 1998; Roberts and Russo 1999). These are:

1. Individual differences and errors of measurement are independent from group to group;
2. Individual differences and errors of measurement must be normally or approximately normally distributed within each group; and
3. The size of variance in the distribution of individual differences and random errors is identical within each cell (i.e. homogeneity of variance).

ANOVA is a robust procedure as reported by many writers of statistical texts and that the above assumptions frequently can be violated with relatively minor effects by many researchers (Maxwell and Delaney 1990; Winer, Brown and Michels 1991; Hays 1994; Kirk 1995; Sirkin 1995; Hinton 1995; Diamantopoulos and Schlegelmilch 1997; Howell 1997; Jaccard 1998; Black 1999; Newton and Rudestam 1999; Roberts and Russo 1999; Everitt 2001; Cramer and Howitt 2004; Field 2005). Here the term Robust is used to designate the ability of the statistical method to produce correct results if the assumptions fail to hold.

Robustness of ANOVA is related with sample size. In non-normal distribution, the robustness of ANOVA is proportionate to the sample size; increase in the sample size reduces the influence of non-normality on the F-test (Hays 1994). According to Kirk (1995), when sample sizes are small, Kurtosis (or flatness) tends to have very little effects on Type I errors but can have effects on Type II errors. If the populations defined by the groups are homogeneous in forms, the robustness of the test can be assured, i.e., all groups show the same degree of skewness and kurtosis (Roberts and Russo 1999).

In case of moderate violations of homogeneity of variance, ANOVA is quite robust and because of that, in practice, this assumption is frequently violated (Maxwell and Delaney 1990; Jaccard 1998). If the ratio of largest to smallest group variances is less than 3.0 (Howell 1997; Roberts and Russo 1999) or the sample sizes are fairly close to one another, i.e. the larger group size divided by the smaller group size is less than 1.5 (van der Heijden 2003), then a violation of equal variance assumption has minimal impact. In other words, if
the sample sizes are equal, the F-test is highly immune and “strong enough” to withstand a violation of the equal variance assumption (Huck and Cormier 1996).

A number of different types of ANOVA designs are depicted in the Table 5.3. For conduct of this study, this table serves as a guide in selecting the appropriate type of ANOVA. The selection of the ANOVA design is guided by number of independent variable involved. The table shows that as we move from one independent variable to more than one, we change from one-way ANOVA to multiple ANOVA (factorial design). As shown in the table, in case of two independent variables, there are three effects to be evaluated, including one interaction while in case of three independent variables; there are seven effects to be evaluated, including three two-way and one three-way interactions.

TABLE 5.3 Selecting the appropriate method for ANOVA designs²

<table>
<thead>
<tr>
<th>Number of IVs</th>
<th>Number of categories of each IV</th>
<th>Type of design</th>
<th>Type of test</th>
<th>Effects</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>Two group (2 means)</td>
<td>t test or oneway ANOVA</td>
<td>1: Between groups</td>
</tr>
<tr>
<td>1</td>
<td>3+</td>
<td>Multigroup (3+ means)</td>
<td>One-way ANOVA</td>
<td>1: Between group</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2 x 2 factorial (4 means)</td>
<td>Factorial ANOVA</td>
<td>3: 2 main, 1 interaction</td>
</tr>
<tr>
<td>3</td>
<td>3,2,2</td>
<td>3 x 2 x 2 factorial (12 means)</td>
<td>Factorial ANOVA</td>
<td>7: 3 main, 3 two-way, 1 three-way interaction</td>
</tr>
<tr>
<td>4</td>
<td>3,3,4,2</td>
<td>3 x 3 x 4 x 2 factorial (72 means)</td>
<td>Factorial ANOVA</td>
<td>15: 4 main, 6 two way, 4 three-way, 1 four-way interaction</td>
</tr>
</tbody>
</table>

Note: All these designs assume one dependent variable, continuously distributed.

In the present study, as a follow-up analysis of multivariate analysis of variance (MANOVA) when MANOVA is significant; two types of univariate ANOVA procedures are used, namely one-way and two-way ANOVA. To obtain a simple effect of an independent variable on a dependent variable, one-way ANOVA is used while to test the

main and the interaction effects of the multiple independent variables on the dependent variables, two way ANOVA (i.e. factorial design) procedures is used. On obtaining significant ANOVA, Paired multiple comparisons will be conducted to know which pairs of group means significantly differed from one another. The issues related to the applications of univariate ANOVA are discussed in the succeeding paragraphs.

**ONE-WAY ANOVA:** Also known by various names such as univariate ANOVA, simple ANOVA, single classification ANOVA, or one-factor ANOVA, it is used to test the differences in a single interval dependent variable (for example, shopping orientation) among three or more groups (for example, Hindus, Muslims, Jains and Christians) formed by the categories of a single categorical independent variable (for example, religious affiliation) in the present study. Also, this design tests whether the groups formed by the categories of the independent variable seem similar on pattern of dispersion as measured by comparing estimates of group variances. If the groups differences are significant, then it is concluded that the dependent variable effected by an independent variable. The null hypothesis is that \( \mu_1 = \mu_2 = … = \mu_k \) with \( k \) equal to the number of means being compared.

To conduct a one-way ANOVA, the assumption of equality of variance among various categories of independent variable should be met. For this purpose, Levene’s test of homogeneity of variance is used in the present study. The reason for its use is that it is less impacted by the assumption of normality than most tests (Tabachnick and Fidell 2000). Equality of variance is met if the significance value of Levene statistic exceeds 0.05. if the assumption of equality is not met (i.e. significance value of Levene statistic is less than 0.05), Brown-Forsythe’s F ratio is used to compare the groups which does not assume equal variance (Huck and Cormier 1996; Cohen 2001; Field 2005).^3^

In the line with the practice of earlier studies, in the present study, \( p < 0.1 \) is set as accepted level of the probability for statistical significance of ANOVA which shows that the probability of occurrence of result by chance is 10%. Because of the exploratory nature of

---

^3^ Another alternative version of adjustment procedure for ANOVA with unequal variance is Welch statistic. It is to be noted however that neither procedure is consistently more accurate than the other and that there are no simple rules that suggest which procedure to use for each possible pattern of sample sizes and variances (Cohen 2001; Field 2005).
the study and the researcher’s desire to reduce the chance of committing a Type II error; that is, assuming no significant difference when a significant difference actually does exist, the researcher has made this choice. It implies that if the analysis reveals a small difference, the results have a good chance of being significant. To avoid committing a Type I error, more conservative probability level is ignored in this analysis. However, it is accepted that this may result in increased probability of detecting effect differences that may have occurred merely by chance (Bryman and Cramer 2001).

**TWO-WAY ANOVA:** Two-way ANOVA reveals two types of effect: main effects and interaction effects (Dancey and Reidy 2004). Main effect provides an individual effect of each independent variable by controlling for other variables. The combined effect of two or more independent variables on a dependent variable is termed as interaction effects. Present it in another way; when the effect of an independent variable on the dependent variable depends on the level of another independent variable an interaction effect would occur. Nevertheless, without interaction, two variables can both influence a dependent variable (Newton and Rudestam 1999). An interaction effect may be either quantitative or qualitative (Schaffer 1991). In a quantitative interaction, two effect occurs which are in the same direction but differ in strength, resulting a both significant main effect and interaction effect. It is also known as ordinal interaction. In qualitative interaction simple effects is in opposite directions and have a stronger emphasis in the reporting of results. It is sometimes also called a disordinal interaction.

Although ANOVA is robust and most appropriate for this study, major concern of applying this mode of analysis is the requirement of equal sample sizes. Though this is not an assumption of ANOVA, it simplifies calculations of sum of squares (Jaccard 1998). According to Jaccard (1998), some of the independent variables may be confounded if the sample sizes are not equal. It implies that proportion of variance in dependent variable which is explained by one factor is also explained by one factor may also be explained by other unknown factors.

**POST-HOC PROCEDURE:** The ANOVA is a method of determining accepting or rejecting null hypothesis on the basis of equality of group means but, but it is not possible to know exactly where the significant difference lies if there are more than two groups (Field 2005). To get the solution of this problem, method use is known as post hoc multiple
pair comparison test which is used to ascertain whether the means of the different groups that integrate each of the variables are significantly different. There are different types of post hoc test available but consensus on which tests are the most appropriate to use is not found (Cramer and Howitt 2004). Among the more common post hoc tests are the Tukey’s HSD method, Bonferroni, Tukey’s LSD approach, Newman-Kuels test, Scheffe test and the Duncan Multiple Range Test. Each test is used to identify which comparisons among groups (e.g. group 1 versus groups 2 and 3) have significant differences.

After getting ANOVA significant, to determine where the significant difference(s) lie, multiple comparison tests are conducted. In the present study two types of post hoc test are conducted. When the Levin test of equality of variance is insignificant (i.e equal variance assumption is tenable), Bonferroni post-hoc tests are used to determine the differences existed among the means. This procedure assumes equal variances and is preferred because it adjusts the observed significance level for the fact that multiple comparisons are being made. On the other hand, when the Levin test of equality of variance is significant (i.e equal variance assumption is not tenable), Tamhane’s T2 contrast is calculated because is robust against the violation of homogeneity of variance assumption. The significance level is set at p < 0.05 for all post-hoc comparisons.

5.6 Multivariate Analysis

It is a statistical procedure which is used to know the statistical significance involving simultaneously many independent variables and/or many dependent variables, all correlated to varying degrees (Hair et al. 1998; Tabachnick and Fidell 2001). It is an extension of univariate and bivariate statistics that analyses complicated datasets. Factor analysis and MANOVA are used as multivariate techniques in this study to analyse the data. The succeeding sections will discuss the relevancy and procedural aspects of these techniques in the context of present study.

5.6.1 Factor Analysis
To develop, interpret and validate the analytical tests, the researchers have to explain or predict behaviour in terms of constructs that are not directly observable. Such constructs are known as hypothetical or latent constructs (Ferguson and Cox 1993). The most common method used for this purpose is factor analysis which is used to identify and measure such construct. The underlying structure in a data matrix is determined by making use of factor analysis (Hair et al. 1998). It is an interdependence multivariate technique in which all variables are simultaneously considered.

In factor analysis, a larger set of variables are reduced to manageable number factors which are formed to explain the whole variable set and thus each factor is predicted by all of the others. In other words, in factor analysis intercorelated variables are grouped or combined together as a factor rather than a series of separate variables. With a minimum loss of information, the data are described by a much smaller number of variables than the original in this process (Hair et al. 1998; Cramer 2003). It is possible that in a matrix of correlation coefficients between a set of measures there are clusters of high correlation coefficients between subsets of the measures (Blaikie 2003). Factor analysis identifies how much variance clusters have in common and the extent to which each measure contributes to this common variance. So, by using this procedure, a small set of factors, or even just one factor can be identified from a large set of measures which explain the maximum amount of common variance in the bivariate correlations between them.

Factor analysis is the obvious choice to interpret the data used in this study since it is used to gain an overall understanding of the main dimensions underlying the variables. In SPSS data reduction command is used to perform factor analysis. In the present study factor analysis is conducted to determine the salient dimensions that make up the constructs of religious commitment, information sources, and shopping orientations. In addition, it also provides construct validity for this study.

In social sciences, factor analysis is used for two purposes. The first purpose is to know the underlying factor structures present in responses to a set of measures, and second purpose is to confirm whether a set of measures in the form specified in a model of their relationships. On the basis of the purpose for which it is used, the factor analysis may be either exploratory or confirmatory (Musil, Jones and Warner 1998). In exploratory factor
analysis, it is assumed that “everything is related to everything” and variables are grouped in factors on the basis of their intercorrelation. In confirmatory factor analysis (CFA), the researcher knows well in advance the latent variable model and the factors that make up this model. The CFA is used to test the hypothesis, that, in fact, the observed sample correlations are consistent with the factor structure proposed on the basis of hypothesised intercorrelations and the patterns of observed variable relationships to underlying factors (Musil et al. 1998). CFA allows the researcher to specify an exact factor model in advance and examine the goodness-of-fit between the hypothesised factor structure and the data and it has strong theoretical foundation. CFA is more of a theory testing procedure than is exploratory factor analysis (Cramer 2003).

In the present study, because of unavailability of exact factor structure, exploratory factor analysis is used. In this study, factor analysis is performed using five step procedure; pre-analysis checks, examination of the correlation matrix, factor extraction, factor rotation and interpretation of factor. The procedure followed in conducting exploratory factor analysis is discussed in the subsequent sections.

**PRE-ANALYSIS:** The distribution pattern of the data and the sample size are two important considerations in performance of factor analysis and to get robust solution (Hutcheson and Sofroniou 1999). So, it is required to make a pre-analysis check before using factor analysis to ensure that: (1) a stable population factor structure can emerge from the sample; (2) items are properly scaled and free from biases, and (3) the data set is appropriate for the application of exploratory factor analysis (Ferguson and Cox 1993).

**Stable factor structure**
To ensuring a stable factor structure in exploratory factor analysis, four types heuristic are proposed by statistician that needs to be satisfied for. These heuristics are outlined in Table 5.4.

**TABLE 5.4** Type of heuristic for stable factor structure

---

Chapter 5

<table>
<thead>
<tr>
<th>Rule</th>
<th>Range</th>
<th>Advocate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject-to-variables ratio (N/p ratio)</td>
<td>Between 2:1 and 10:1</td>
<td>Kline (1986); Gorsuch</td>
</tr>
<tr>
<td>Absolute minimum number of subjects (N)</td>
<td>100 to 200</td>
<td>(1983); Nunnally (1978)</td>
</tr>
<tr>
<td>Relative proportions of: variables to expected factors (p/m ratio), and subjects to expected factors (N/m ratio)</td>
<td>between 2:1 and 6:1</td>
<td>Kline (1986); Comrey (1978); Cattell (1978)</td>
</tr>
</tbody>
</table>

The most important consideration among these four heuristics is sample size is regarded as. According to Hutcheson and Saforinou (1999), factor analysis is based on correlation coefficients, which tend to be most reliable when computed for large samples. Nevertheless, there is no agreement as to what constitutes large (Pedhazur and Schmelkin 1991). The guidelines given by Comrey and Lee (1992) for sample sizes is; samples size of 50 as very poor, 100 as poor, 200 as fair, 300 as good, 500 as very good and 1000 as excellent. In the opinion of Guadagnoli and Velicer (1988), mean factor loadings for a factor (factor saturation) is also a critical parameter and N become irrelevant if four or more items load on each emergent factor >0.6. However, if both the factor saturation and the ratio of variables to expected factors (p/m) are low, the importance of N increases. An N of at least 300 is required in such case (Tabachnick and Fidell 2001).

Contrary to the above argument, in the argument of other writers, smaller Ns might be acceptable. To obtain an accurate solution in exploratory factor analysis, Pedhazur and Schmelkin (1991) suggest N = 50; Ferguson and Cox (1993) and Hair et al. (1998) suggest N = 100 while Hinkin, Tracey and Enz (1997) suggest N = 150. When each factor is overdetermined (i.e. at least 3 or 4 variables represent each component) and the communalities are high (average 0.7 or higher), MacCallum et al. (1999) propose samples as small as 100 to obtain accurate estimates of population parameters. If all components are not over determined and communalities are of around 0.5, under such a more moderate conditions, samples of 200 would be more appropriate (Fabrigar et al. 1999). The sample size taken for this study is 750 which fulfil the minimum sample requirement for factor analysis.

The number of variables to be analysed is another important aspects of sample size determination. As a general rule, the observation should be e at least five times as many to
the variables to be analysed, and a more acceptable size would have a ten-to-one ratio (Hair et al. 1998, p. 99). In the present study, this requirement is also getting fulfilled since the final sample size of 750 included in the analysis meets this criterion. Specifically, $5 \times 30 = 150$ for religiosity measure; $5 \times 8 = 40$ for information sources measure; and $5 \times 33 = 165$ for shopping orientation measure.

**Item scaling**

In factor analysis, the Pearson correlation coefficient $r$ is used to determine the factors which require data to be measured on a true continuous scale (i.e. interval or ratio). However, in practice, these requirements are rarely satisfied and Likert-type scales (e.g. five-point scale) are often deemed adequate (Ferguson and Cox 1993). According to Hutcheson and Saforinou (1999), the relaxation of the requirement for continuous data can be justified for EFA since the interpretability of the factors is taken as a base to determine the usefulness of factor analysis. In this study, all variables are measured on 5 point Likert-type scale which is assumed to provide interval-level data (Mitchell 1994; Blaikie 2003) and therefore the variables fit for factor analysis.

**Appropriateness of dataset**

Another requirement for factor analysis is appropriateness of dataset which requires the variables used should demonstrate univariate normality; that is, it is assumed that each variable conforms to the normal distribution curve (when the mean is in the centre of the distribution). The univariate normality of each variable can be determined by using the coefficients of skewness and kurtosis. In the opinion of Muthen and Kaplan (1985), some degree of univariate skew and kurtosis is acceptable, for the majority of variables, if neither coefficient exceed $+2.0$ (where zero indicates no kurtosis). If there are at least 60% low correlations ($<0.2$) in the initial correlation matrix, then greater skew is acceptable. In this study, the coefficients of skewness and kurtosis of each variable are within the range of $0+2.58$ as specified by Hair et al. (1998), suggesting a relatively normal distribution (see Appendix C). Therefore the data sets are deemed appropriate for factor analysis.
EXAMINATION OF THE CORRELATION MATRIX: The suitability of set of variables for the selection in the factor analysis is determined by observing a systematic covariation among the variables under consideration (Ferguson and Cox 1993). The correlation matrix, Bartlett’s test of sphericity and the Kaiser-Meyer-Olkin measure of sampling adequacy (MSA) is used for this purpose. Some degree of multicollinearity is desirable in factor analysis because the objective of factor analysis is to identify interrelated sets of variables. In the opinion of Hair et al. (1998), factor analysis is appropriate if visual inspection of the correlation matrix reveals a substantial number of correlations greater than 0.30. The partial correlations among variables can be used to analyse the correlations among variables. The values of partial correlation should be small if “true” factors exist in the data.

The presence of correlations among variables is determined by using the Bartlett test. It provides the statistical probability that the correlation matrix has significant correlations among at least some of variables. Thus, a significant Bartlett’s test of sphericity is required. The KMO index indicates the degree to which each variable in a set is predicted without error by the other variables which ranges from 0 to 1. Each variable is perfectly predicted by the other variables without error if the KMO index reaches 1. A value below 0.50 is unsatisfactory, a value of 0.70 or more is generally considered sufficiently high, while the value above 0.90 is outstanding (Hair et al. 1998). SPSS provides the overall KMO value as a single statistic, whilst the anti-image correlation matrix gives KMO values for individual variables. The anti-image correlation matrix contains the negative values of the partial correlations among variables; smaller anti-image correlations are indicative of a data matrix suited to factor analysis (Hair et al. 1998).

FACTOR EXTRACTION: After the screening and selection of variables for the factor analysis is done, the factors needed to represent the data can be determined. Factor extraction is performed to identify and retain those factors which are necessary to reproduce adequately the initial correlation matrix (Ferguson and Cox 1993). This subsequent paragraph will describes the choice of extraction method and the rules followed to be followed in obtaining the optimum number of factors to be extracted.

Method of factor extraction
The factor solution can be obtained by making use of either principal component analysis or common factor analysis (Hair et al. 1998). The purpose of using principal component is to summarise most of the original information (variance) in a minimum number of factors for prediction purposes while common factor analysis (CFA) is used primarily to identify underlying factors or dimensions that reflect what the variables share in common. Principal component analysis assumed that all variability in an item should be used in the analysis. The objectives of the factor analysis and the amount of prior knowledge the researcher has about the variance in the variables determine the kind of factor model to be used (Mitchell 1994). Regardless of which factor model is used by the researcher, the results of extraction are similar in most of the cases (Fava and Velicer 1992). However, if the purpose of the analysis is data principal component analysis method is preferred but if the purpose to detect structure then common factor analysis is preferred (Jackson 1991).

Since the purpose of this study is to identify the factors which may explain the relationships within the data, principal component (PC) analysis is used in this study which is in conformity with the recommendation made by Ferguson and Cox (1993). To extract the factors in marketing research, principal component analysis is widely used. The principal component (PC) analysis is popular because it leads to unique, reproducible results which are not supported by some of the less structured factor analytical procedures. Principal component analysis relies upon the total variance to derive the factors with small proportions of unique variance. Since the main concern of this study is to predict the minimum number of factors that are required to account for the maximum portion of the variance represented in the original set of variables and there is a priori set of variables (Mitchell 1994), this method is deemed appropriate for the present analysis. The principal component method extracts a linear combination of variables (a component) that accounts for as much variation in the original variables as possible. Further, it finds the next component that accounts for as much of the remaining variation as possible and it is not correlated with the previous component, continuing in this way until there are as many components as original variables (Hair et al. 1998).

Criteria for selection of factors
The trouble in the use of factor analysis is in taking decision on how many factors to be extracted in the final solutions which account for the greatest amount of the total variance. The most common and widely used method is Keiser’s rule of latent root criterion among several other methods available and it is used in this study. The rule is quite simple to apply, but its applicability is dependent on which factor model has been chosen. An eigenvalue of greater than or equal to 1.0 are considered to be significant for the principal component analysis. An eigenvalue gives an estimate of the amount of variance associated with any factor, so that the rule involves retaining those factors which account for above average variance for interpretation (Ferguson and Cox 1993). This criterion is based on theoretical rationales developed using true population correlation coefficients. It is commonly thought to yield about one factor to every three to five variables. It appears to correctly estimate the number of factors when the communalities are high and the number of variables is not too large (Afifi and Clark 1996). Tabachnick and Fidell (2001) suggest that (a) if the number of components with eigenvalues greater than one is a reasonable number of components for the data (i.e. somewhere between the number of variables divided by 3 and the number of variables divided by 5), (b) if the number of variables is 40 or fewer, and (c) if the sample size is large, the number of components indicated by this criterion is probably correct.

**FACTOR ROTATION:** The initial principal components which explain most of the variance in the variables should be rotated to a simple structure, defined as each variable having a high loading on one of the factors, and zero or small loadings on the others (Ferguson and Cox 1993). Factor rotation simplifies the factor structure and improves the interpretation by removing the ambiguities which often accompany initial unrotated factor solutions (Mitchell 1994). Factor rotation can be done in two ways. In an orthogonal extraction, factors are extracted in such a way that the factor axes are maintained at 90 degrees which results in each factor being completely independent of all other factors. In an oblique extraction, factors are not completely independent and some commonality is maintained. According to Hair et al. (1998), if the objective is to reduce the number of original variables regardless of how meaningful the resulting factors may be, or when the objective is to reduce a large set of variables to a smaller number of uncorrelated variables which can be used in subsequent predictive techniques such as multiple regression, Orthogonal rotation is mathematically simpler to handle and should be used.
There are number of ways to perform factor rotations. The best and most commonly used method is the varimax procedure of orthogonal rotation among the available various methods of rotation (Mitchell 1994; Afifi and Clark 1996; Fabrigar et al. 1999) applied on the principal component solutions. The Varimax procedure gives solution on the basis of a number of the squared loadings for the variables. Thus a rotation position is sought that maximising the sum of variances the squared factor loadings within each factor in the matrix. Further, these factor loadings are adjusted by dividing each of them by the communality of the corresponding variable. This adjustment is known as the Keiser normalisation, which tends to equalise the impact of variables with varying communalities (Afifi and Clark 1996). This means that varimax rotation gets over the problem of a general factor. Thus this procedure is used in this study because it minimises correlation across factors and maximises within the factors. As discussed by Hair et al. (1998), varimax procedure clearly separate the factors and has proved very successful as an analytic approach to obtaining an orthogonal rotation of factors.

**INTERPRETATION OF FACTORS:** The fifth section factor analysis describes the interpretation of factor loadings and the labelling of factors. The factor loadings represent the correlation between the original variable and its factor. The contribution an item makes to a particular factor is measured by the factor loadings. It is ideal if an item should have a high loading on only one factor. The greater the loading, the more the variable is a pure measure of the factor (Tabachnick and Fidell 2001). However, there is a difference in the opinion of statisticians on what constitutes a high loading. Sample size is an important consideration in determining factor loading. The minimum loading for a sample of 50 is 0.72, for 100 is 0.51, for 200 is 0.36, for 300 is 0.30, for 600 is 0.21 and for 1000 is 0.16 for a level of significance of 0.01 (two-tailed)(Stevens 1992). The more precise and comprehensive specification of loadings is given by Comrey and Lee (1992) and they described that the loading in excess of 0.71 (50% overlapping variance) as excellent, 0.63 (40% overlapping variance) as very good, 0.55 (30% overlapping variance) as good, 0.45 (20% overlapping variance) as fair and 0.32 (10% overlapping variance) as poor. These are merely the guidelines and selection of the “cutoffs” for loadings is a matter of a researcher preference (Tabachnick and Fidell 2001), based on consideration of a theoretically sound solution.
A loading of is 0.30 and above is commonly recommended (Kline 1994; Hair et al. 1998; Tabachnick and Fidell 2001). In the view of Stevens (1992), loadings of 0.4 and above should be taken seriously. A loading of 0.4 indicates 16 percent of the item’s variance contributes to the factor (arrived at by squaring the loading and multiplying by 100) while a factor loading of 0.3 indicates an item accounts for 9 percent of an item’s variance. Inclusion of such an item in a scale means 91 percent of unrelated variance, thus producing very ‘muddy’ and imprecise scales (Blaikie 2003). A criterion of 0.4 is adopted as a meaningful factor loading in interpreting the analyses reported in this study which follows Stevens’s (1992) recommendation. A factor which has a loading below 0.4 and/or consisted of only one item is dropped. It is practically significant to consider a factor loading of 0.4 as conservative cut-off point as recommended by Hair et al. (1998) when the sample size is 200 or greater.

After the determination of the factors, the next task is assigning a label to each factor in some meaningful way on the basis of what the factors actually represent. The labelling is decided on basis of common properties of the set of statements loaded within each factor. According to Hair et al. (1998), the label given should be the accurate representation of factor. Thus, the past literature should be used as a reference in labelling the factors. Although labelling the factor is a matter of subjective interpretations of the general nature of the construct based on its component items, this step can be regarded as the most important part of factor analysis process. If the selected factors are not interpretable, then the factor analysis presents little value to the researcher is the main reason behind this.

DATA REDUCTION: In all, the process of exploratory factor analysis ends at the factor interpretation stage as discussed in the earlier section but the objective for applying exploratory factor analysis in the present study is to use the factors for further application to other statistical analysis, some form of data reduction must be made. According to Hair et al. (1998), direct measurement of latent variable cannot be possible but it can be represented by one or more indicator variables which represent the theoretical concept in a better way and simultaneously increase the reliability of the measure. According to Hair et
al. (1998), data reduction can be done by making use of three different methods, namely (1) selecting surrogate variables, (2) computing factor scores and (3) creating summated scales.

The first method envisages the selection of the variable with the highest factor loading on each factor to act as a surrogate variable to represent that factor. The application of this method is possible only when one variable has a factor that is substantially higher than all other factor loadings. However, if other variables have loadings that are fairly close to the surrogate variable, the process of selection becomes more cumbersome. Also, use of this method is risky because the selected surrogate variable may not address the issue of measurement error encountered when using single measures and thus mislead the subsequent analyses (Hair et al. 1998).

Calculation of factor scores is the second method of data reduction. Conceptually, the factor score represents the degree to which each individual score high on the group of items that have high loadings on a factor. Thus, the higher value of variable loading, the higher the factor scores. It represents the total score of different variables loading on a factor. However, the problem of using this method is that all variables will have some degree of influence in computing the factor scores and make interpretation more difficult (Hair et al. 1998).

To create a smaller set of variables which replace the original set by combining several individual variables into a single composite measure is the third method of data reduction which is known as summated approach. In summated approach, all variables loading highly on a factor are combined and the average score of the variables is used as a replacement variable. This method has two specific benefits. First, since this method uses multiple variables to calculate the replacement variable, to some extent; it overcomes measurement error that might occur in a single question. Second, since it represents the multiple aspects of a concept in a single measure, it facilitates the subsequent analysis (Hair et al. 1998).
Summated scales are created in this study in which scores on the items within each of the factors are summed for every respondent and used as replacement variables for each factor. The sum is then divided by the number of items summed to create the scale in order to retain the original meaning of the numbers and to provide for comparison between scales. Thus this score may range from 1.00 to 5.00 where 1.00 indicates of strong disagreement and 5.00 indicate strong agreement. Since this method is a compromise between the surrogate variable and the factor scoring method, it is chosen in this study. Because this method has the advantage of a composite measure, it reduces measurement error and it considers those variables which have high loadings on each factor and excludes those having little impact (Hair et al. 1998).

Cronbach’s coefficient alpha is used to test an internal reliability of the summated scales for each resulting factor. The information about the relationships among individual items and their internal consistency as well as examined the properties of a measurement scale and the questions that make it is provided by reliability analysis procedure.

**5.6.2 Multivariate Analysis of Variance (MANOVA)**

A technique which test the significance of mean difference the groups of two or more dependent variable is known as Multivariate analysis of variance (MANOVA). It is an extension of the analysis of variance (ANOVA) technique. Univariate ANOVA is used to test the significance of group differences along a single facto whereas MANOVA is used to test the significance of group differences along a combination of factors. According to Tabachnick and Fidell (2001), MANOVA has a number of advantages over univariate analysis of variance (ANOVA). Firstly, the researcher improves the chance of discovering what it is that changes as a result of different treatments and their interactions by measuring several dependent variables at a time instead of only one,. Secondly, the chances of committing Type I error is minimised with the use of MANOVA, i.e. a significant effect being identified when none exists. Thirdly, some differences that are not shown in separate ANOVAs may be revealed by the use of MANOVA.

**ASSUMPTIONS:** in order to carry out a meaningful analysis, there are certain assumptions associated with MANOVA which needs to be met. The normal distribution of the dependent variables is the first assumption. MANOVA assumes multivariate normality;
that is, the normal distribution of the dependent variables and all linear combinations of
them (Tabachnick and Fidell 2001; Field 2005). As such, multivariate normality is not
directly tested and therefore univariate normality is generally used (Hair et al. 1998).
Although it is preferred to have a multivariate normality, it should be noted that
“MANOVA is still a valid test even with modest violation of the assumption of
multivariate normality, particularly when we have equal sample sizes and a reasonable
number of participants in each group” (Dancey and Reidy 2004, p. 488). As noted by
Tabachnick and Fidell (2001, p. 329), if the differences are due to skewness, as opposed to
outliers, violations of this assumption can be tolerated for larger sample sizes or moderate
sample sizes. They have further suggested that a sample size of about 20 in the smallest
cell should ensure robustness even if the sample size is unequal.

The second assumption is related with equality variance-covariance matrices of the
dependent variables across the groups. To check the equality of variance covariance
matrices in this study, The Box’s M test is used in this study. If Box’s M test reveals
significant results, it indicates that the covariance matrices are significantly different and
so the assumption of homogeneity would have been violated (Field 2005). However, if the
sample sizes across groups are approximately equal, disregard the outcome of Box’s M test
(Hair et al. 1998; Field 2005). According to Hair et al. (1998), groups are regarded as
approximately equal if the ratio of largest group to smallest group is less than 1.5. If Box’s
M test is significant at p < 0.001 and sample sizes are unequal, then robustness cannot be
assumed, (Tabachnick and Fidell 2001; Field 2005). The Levene test at p < 0.05 is used to
assess the equality of variance for a single variable across groups.

**CRITERIA FOR SIGNIFICANCE TESTING:** To judge multivariate differences across
groups, there are four most popular criteria: Roy’s greatest characteristic root (gcr); Wilks’
lambda (also known as the U statistics); Hotelling’s trace and Pillai’s criterion (also known
as Pillai-Bartlett trace, V). In making choice from these four statistics, Hair et al. (1998)
suggest that “the measure to use is one most immune to violations of the assumptions
underlying MANOVA that yet maintains the greatest power” (p. 351). When sample size
decreases, unequal cell sizes appear or assumption of homogeneity of covariance is
violated, Pillai’s criterion is the most powerful to be used (Hair et al. 1998).
Pillai’s trace and Wilks’ lambda are used in this study as a criterion for significance testing. This is consistent with Essoo and Dibb (2004) and Mokhlis (2006). The significance level for this study is set at \( p < 0.1 \) level.

**FOLLOW-UP ANALYSIS:** It is required to examine which dependent variables are responsible for the statistically significant MANOVA results when MANOVA turned out to be significant. To serve this purpose, univariate ANOVA is carried out.

### 5.6.3 Multiple Linear Regression Analysis

The objective of this study is to examine the influence of religious variables on consumer use of information sources and shopping orientation. The significant differences in consumer behaviour among respondents from different religious groups and having different level of religiosity is determined by using ANOVA procedure but it could not predict the direction and the magnitude of the linear relationship between consumer behaviour variables (dependent variables) and the religious variables (independent variables). So to predict the direction and the magnitude of the linear relationship between consumer behaviour variables (dependent variables) and the religious variables (independent variables), multiple linear regression analysis which is also known as Ordinary Least Square (OLS) regression, has been applied in the present study.

Regression analysis is a statistical technique that explains the change in one variable (dependent variable) as a result of change in a set of other variables (independent or explanatory variables) (Studenmund 2001). It describes the degree of relationship between a single dependent (criterion) variable and several independent (predictor) variables. Thus regression analysis measures the collective influence of the independent variables on a dependent variable. The values of independent variables are known which is used to predict the single dependent values selected by the researcher. It is assumed that all variables observable and they have no measurement error (i.e. perfect measurement of variables).

Multiple regression analysis is used for two broad categories of research problems: prediction and explanation (Hair et al. 1998). On one end, it is used to predict the
dependent variable with a set of independent variables. In doing so, multiple regressions fulfil one of the two objectives (Hair et al. 1998). To maximise the overall predictive power of the independent variables as represented in the variate is the first objective. It is explicitly designed to make errors of prediction as small as possible using the least squares criterion for overall smallness (Allison 1999). To compare two or more sets of independent variables to assess the predictive power of each variate is the second objective. The second purpose of multiple regression analysis is to provide an assessment of the degree and direction (positive or negative) of the linear relationship between independent and dependent variables by forming the variate of independent variables (Hair et al. 1998).

To control the variance, the most powerful method is multiple linear regression analysis. From multiple linear regression analysis, on the basis of analysis of the variable intercorrelations, one is able to estimate the magnitudes of different sources of influence on the dependent variable. It also indicates the dependency of the dependent variable on the independent variables. The combined and individual effect of independent variables on the dependent variable is given by multiple linear regression tests. Thus, multiple linear regression analysis can be used to examine the effects of some independent variables on the dependent variable while “controlling” (i.e. held constant) for other independent variables (Allison 1999).

The regression technique seeks to establish a rectilinear relationship between the variables concerned in order to calculate statistical predictions,. Subsequently, the equation of a straight line is of important value, and is denoted by $Y = bX + a$ where $Y =$ predicted score; $b =$ slope; $X =$ X intercept and $a =$ Y intercept. To predict a criterion variable on the basis of number of predictor variables the following equation is therefore used in multiple linear regressions:

$$Y = a + b_1X_1 + b_2X_2 + \ldots + b_kX_k + e$$

where $Y$ represent the predicted value of the dependent variable, $a$ is the Y intercept (the value of $Y$ when all the $X$ values are 0), $X$ represents the various independent variables and $b$ is the coefficients assigned to each of the independent variables during regression (Tabachnick and Fidell 2001).
MEASUREMENT OF VARIABLES: The major concern of multiple regression analysis is the measurement of the dependent and independent variables for the analysis. Only quantitative explanatory variables, measured on an interval or continuous scale is the prerequisite of multiple regression analysis. If the independent variables are categorical (i.e. not interval) and if they have two or more categories, they need be coded into dummy variables. Dummy variables are dichotomous variables which act as replacement independent variables. Indicator or binary coding is the most common form of dummy coding in which the category is represented by either 1 or 0 (Pedhazur 1997). Instead of defining its level, the use of the values of 0 and 1 merely describe the presence or absence of a particular attribute. For example, if gender is coded 1 for male and 2 female, the indicator conversion of the variable into a dummy variable is 1 for male and 0 to female.

However, the use of dummy variables is not straightforward in an OLS regression because the inclusion of all of them at the same time leads to a situation where perfect multicollinearity exists (Hutcheson and Sofroniou 1999). To resolve this problem, one of the categories must be omitted. So, if we have j categories, a maximum of j – 1 dummy variable can be entered into the model. For example, in the case of gender, either the male or the female category must be eliminated, leaving only one dummy variable for the analysis. The dummy variable which is omitted is called the reference category and is the category against which other dummy variables are compared (Hutcheson and Sofroniou 1999). It is to be noted that the model fit is not affected by the choice of reference category as this remains the same no matter which category is designed as the reference.

In this study, dummy variables are created to represent categories of grouped demographic variables for the regression analyses. Five independent variables namely religious affiliation, gender, marital status, occupation, income, area of residence, age and education converted to dummy variables. Gender is represented by a dummy variable with a value of 0 indicating male and a value of 1 indicating female. Marital status is represented by a dummy variable with a 1 for married and a 0 for not married. Ethnic is represented by two dummy variables with a value of 1 on the first (Ethnic 1) indicating Chinese, a value of 1 on the second (Ethnic 2) indicating Indian and a 0 on both representing Malay. Religious affiliation is represented by three dummy variables with a value of 1 on the first (Religion 1) indicating Buddhist, a value of 1 on the second (Religion 2) indicating Hindu, a value of
Methods of Data Analysis

1 on the third (Religion 3) indicating Christian and a 0 on the three variables representing Muslim. Work status is represented by a dummy variable with a value of 0 indicating working and a value of 1 indicating non-working. Age, education attainment and monthly income are originally a grouped variable with multiple categories. For the regression analyses, these variables are treated as “ordered categorical data” by using a scoring method (in integer coding, 1, 2, 3…) where the ordered nature of the data is retained (Hutcheson and Sofroniou 1999).

ASSUMPTIONS: Although “it has been demonstrated that regression analysis is generally robust in the face of departures from assumptions” (Pedhazur 1997, p. 34), it is required to the assumptions to build a basic knowledge on how to obtain the best linear unbiased estimators from the regression analysis. The purpose is to point to some of the steps taken to ensure the validity of the results. As Pedhazur (1997) notes, “knowledge and understanding of the situations when violations of assumptions lead to serious biases, and when they are of little consequence, are essential to meaningful data analysis” (p. 33). The assumptions of the regression analysis are discussed below.

The First assumption of regression analysis is of normal distributions. The relationships and significance tests are distorted by the variables which are non-normally distributed (highly skewed or kurtotic variables, or variables with substantial outliers). The normal probability plot of regression residuals is inspected visually to examine the assumption of normality. The normal distribution makes a straight diagonal line and the plotted residuals are compared with the diagonal. If the standardised residual line closely follows the diagonal, it indicates the normal distribution. However, in the statistical literature, it is agreed that if the sample size is sufficiently large, slight departures from this assumption do not appreciably alter our inferences (Afifi and Clark 1996, p. 109).

The linearity of the relationship between the dependent and the independent variables in the model is the second assumption of regression analysis. Linearity implies that this relationship is constant across the range of values for the independent variables. It means that there is similar change in the mean value of the dependent variable as a result of a unit...
increase in the independent variable. Furthermore, correlation (Pearson’s r) can capture only the linear association between variables.

The third assumption of regression analysis is the equality of variance of the residuals (homoscedasticity). Homoscedasticity states that at different levels of the independent variables, the variance of the dependent variable is approximately the same (Hair et al. 1998). Afifi and Clark (1996, p. 109) claim that this assumption is not crucial for the resulting least squares line. This is because the least squares estimates of \( \alpha \) and \( \beta \) are unbiased whether or not the assumption is valid. However, violation leads to serious distortion of findings and seriously weakens the analysis which will increase the possibility of a Type I error (Tabachnick and Fidell 2001).

In the present study, to check the assumptions of linearity and homogeneity, a residual scatterplot (plot of the standardised residuals as a function of standardised predicted values) is constructed. Ideally, residuals are randomly scattered around zero (the horizontal line) which indicates a relatively even distribution with no strong tendency to be either greater or less than zero. According to Hair et al. (1998), if the residuals are not evenly scattered around the line, a violation is indicated.

The independence of error term is the fourth assumption of regression analysis. Regression analysis assumes each predicted value to be independent, that is, serially uncorrelated. When the error terms for two or more independent variables are correlated, it is called autocorrelation or serial correlation (Lewis-Beck 1993). However, this problem is found in time-series studies when the errors associated with observations in a given time period carry over into future time periods (Pindyck and Rubinfeld 1991; Lewis-Beck 1993). This assumption is a less serious problem for the present study because this is a cross-sectional study.

Additionally, outliers and multicollinearity are two issues that can arise during the analysis, that strictly speaking are not assumptions of regression, are none the less, of great concern to the researcher.

**OUTLIERS:** According to Hair et al. (1998), outliers are the observations that are substantially different from the remainder of the data set (i.e. has an extreme value). Hair et
al. (1998) has the opinion that if it is not evident that outliers are truly deviant and not representative of any observations in the population, they should be retained to ensure the generalizability to the entire population. The case wise subcommand in the regression procedure is used to identify the outlier in the present study. Cases that proved to have standardised residuals in excess of three are eliminated from the analysis (Hutcheson and Sofroniou 1999).

**MULTICOLLINEARITY:** The possibility of multicollinearity is existed when large numbers of possibly highly correlated explanatory variables are used. In its simplest form, multicollinearity is defined as a situation where an independent variable is highly correlated to one or more of the other independent variable which leads to unstable model when deleting or adding variables to the model (Hutcheson and Sofroniou 1999). In the opinion of Lehmann, Gupta and Steckel (1998), although multicollinearity does not violate any assumption (the independent variables do not have to be independent of each other), not does it affect the overall predictive capabilities of the model, it does make the estimates of the regression coefficients unreliable because the effect of the predictor variables are mixed or confounded. The regression equation cannot even be formulated If one independent variable can be precisely predicted from one or more of the other independent variables (perfect multicollinearity). When a relationship is strong but not perfect (high multicollinearity), the regression equation can be formulated, but the parameters may be unreliable (Hutcheson and Sofroniou 1999). However, in most research, multicollinearity is present to some degree, but it should be ignored if the correlation coefficient between any two variables is not too large (Pedhazur 1997). As a general rule of thumb, multicollinearity might be present if any of the following situations exists (Mueller 1996; Grapentine 1997):

1. Absolute values of one or more of the zero-order correlation coefficients between independent variables are relatively high
2. One or more of the metric or standardized regression coefficients have theory contradicting signs. For example, the coefficients take on negative values when theory or common sense suggests a positive relationship exists between the independent and dependent variable
3. One or more of the standardized regression weights are very large
4. The standard errors of the beta regression coefficients are unusually large
5. The regression equation has a large overall R2 with several insignificant independent variables.

For the purpose of assessing multicollinearity problem in the present study, in the initial stage, to detect the presence of high correlation, Pearson product-moment correlation matrix among the independent variables is examined. The high correlation values of about 0.8 or higher indicate a level of multicollinearity that may prove to be problematic (Hutcheson and Sofroniou 1999). In the similar line, according to Hair et al. (1998), whilst no limit has been set that defines high correlations, values exceeding 0.9 should always be considered, and many times correlation exceeding 0.8 can be indicative of problems. In the present study, there is an absence of high intercorrelations as can be seen from Appendix D.

Nevertheless, it cannot be decided that there is a lack of collinearity on the basis of the absence of high bivariate correlations because in case of more than two variables, the correlation matrix may not reveal collinear relationships (Mason and Perreault 1991). Therefore, in case of the stepwise regression analyses, it is required to measure the tolerance values and variance of inflation factors (VIF). The amount of variability of the selected independent variable not explained by the other independent variables is represented by Tolerance values. On the other hand, VIF is an indicator of the effect that the other predictor variables have on the variance of a regression coefficient. A very small tolerance values and thus large VIF values denote high collinearity. According to Hair et al. (1998), it is an indication of high multicollinearity if a tolerance value below 0.1, which corresponds to VIF greater than 10. In the present study, to indicate high multicollinearity, these values are used as cut-off thresholds.

Condition index is another useful measure which can be used as an indicator of multicollinearity in multiple regressions. Belsley (1991) and Mason and Perreault (1991) has suggested that large condition indices be scrutinised to identify those associated with large variance proportions for two or more coefficients. Specifically, collinearity is indicated for the variables whose coefficients have large variances associated with a given
large condition index (Pedhazur 1997). According to Belsley (1991), “weak dependencies are associated with condition indexes around 5-10, whereas moderate to strong relations are associated with condition indexes of 30-100” (p. 56). Hair et al. (1998) stated that when a condition index that exceed the threshold value of 30 accounts for a substantial proportion of variance (0.9 or above) for two or more coefficients, it is an indication of a collinearity. In the present study, tolerance value and VIF values are nearer to 1 and also the condition index does not exceed 20 which indicate there is an absence of multicollinearity problem.

**AUTOMATED MODEL SELECTION:** Once it is found that variables met the assumptions of regression analysis, the next step is to select the procedure of the independent variables to be included in the model. Forward selection, backward elimination and stepwise selection are three common automated selection procedures which are described by Hutcheson and Sofroniou (1999) that can be used to find the “best” regression model. In forward selection, independent variables are added one at a time provided they meet the statistical criteria entry (usually at p = 0.05) starting with empty equation. The opposite of the forward selection procedure is the backward elimination. In backward elimination, all independent variables are entered in the equation first and if they do not contribute significantly to regression, they are deleted one at a time. The equation is re-calculated and those variables left in the model is re-examined after each variable is removed to see if any contribute less than the criterion level. This process continues until no more variable reach the selection for removal. A compromise between forward and backward elimination is stepwise selection procedure which is one of the most frequently used methods of automated variable selection in marketing research. In the stepwise procedure, mathematical maximisation procedure is used to determine the order in which the predictor variables enter a regression equation. It means the independent variable which a maximum correlation with the dependent variable is entered first and the second variable which is entered is the predictor with the largest semi-partial correlation, and so on. However, a test is made of the least useful predictor and the importance of each predictor is constantly reassessed at each stage of the stepwise procedure. As a result of this procedure, a predictor becomes superfluous in the later stage that was deemed earlier to be the best entry candidate. This method is advantageous because the order of the
independent variables in a regression equation is determined on the basis of their significance for the predicted characteristic. Hair et al. (1998) described stepwise procedure as a method of variable selection in which variables are considered for inclusion in the regression model and it selects the best predictors of the dependent variable. Hutcheson and Sofroniou (1999) described the stepwise procedure as a screening procedure which purifies the target independent variables prior to their entering into the model.

From the preceding review, four important points can be made in favour of the stepwise selection in regression analysis. First, stepwise is a compromise between the other two procedures, that is, forward selection and backward deletion in which the equation starts out empty and independent variables are added one at a time as long as they meet statistical criteria, and deleted at any step if they no longer make a significant contribution to the regression model. Second, it is a screening procedure that skims out the redundant independent variables and selects the best predictors for the dependent variable. Third, it determines whether additional independent variables make any contribution compared to the other variables already included in the equation. Fourth, this procedure helps selection of the best variables where the researcher has selected variables on the basis of a strong grounded theory for the purpose of analysis.

Stepwise selection procedure is used in the current study to find the best regression model without testing all possible regressions. Since there is no theoretical a priori assumption regarding the importance of each variable, the use of this procedure is justified for this study. Additionally, it helps the researcher to examine the contribution of each independent variable to the explained variance of the dependent variable. In this procedure, the independent variable with the greatest contribution is added first and before the main equation is developed, each independent variable is considered for inclusion. Independent variables are in the equation on the basis of their incremental contribution over the variable(s) already in the equation (Hair et al. 1998). In this study, an alpha level of 0.1 is used as the entry cut-off value because it is believed that some of the variables would be excluded if the lower level of significance is used and the researcher tried to minimise the effect of collinearity as far as possible through variable selection following the advice of Speed (1994).
However, there are some controversies in the use of stepwise selection because the independent variables that are entered into the regression equation are based solely on statistical rather than theoretical or logical considerations (Menard 1995; Polit 1996). In step regression, it is the computer which determines the order of entry of the variables and thus the results obtained are idiosyncratic and difficult to replicate in any sample other than the sample in which they are originally obtained (Menard 1995, p. 54). So in the present study stepwise regression is used as an exploratory regression technique without specifically hypothesising which variables are most predictive of the criterion variable.

**COEFFICIENT OF MULTIPLE DETERMINATIONS ($R^2$):** The predictive power of the regression model can be judged on the basis of the coefficient of determination ($R^2$). $R^2$ is a measure of the proportion of the total variance of the dependent variable that is “explained” or “accounted for” by the independent variable (Lewis-Beck 1993). The value of $R^2$ ranges between 0 and 1 and it is the squared product-moment correlation coefficient. Thus, the fit of the model is determined on the basis of how close the value of $R^2$ is to 1. The closer the value of $R^2$ is to 1, the better the fit of the model (i.e. the independent variables accounts for or explain the variation in dependent variable in a better way) because if $R^2$ is 1 then the regression model is accounting for all the variation in the outcome variable.

In the words of Hair et al. (1998), if the regression model is properly applied and estimated, it can be assumed that the higher the value of $R^2$, the greater the explanatory power of the regression equation, and therefore the better the prediction of the dependent variable. However, Uncles and Page (1998) stated that there is no hard-and-fast statistical argument for deciding what level of $R^2$ is “high enough”. The value of $R^2$ value is directly attributed to the number of variables in the model and it can be improved by adding more variables to the model, though their contribution is very small or accidental.

An alternative to the standard $R^2$ is adjusted- $R^2$ which measure the fit considering the number of independent variables and the sample size. Although the value of $R^2$ rises with the addition of predictor variables, the value of adjusted $R^2$ may fall if the added predictor
variables have little explanatory power and are not statistically significant (Hair et al. 1998; Newton and Rudestam 1999). Hence, adjusted $R^2$ is used in this study for the interpretations of explanatory power because it is a less biased measure for the variance explained by the model.

It is to be point out that the sample size included in the regression analysis directly influences the predictive power of the regression. Hair et al. (1998) has prepared a table which shows the relationship between the sample size, the significance level chosen and the number of independent variables in detecting a significant $R^2$ which is presented in Table 5.5. The table indicates that the higher the sample size, the lower the $R^2$ for a given number of independent variables at a given significance and power levels. This guide is taken into consideration assessing the overall model fit for the regression equations in this study.

**TABLE 5.5** Minimum $R^2$ that can be found statistically significant with a power of 0.8 for varying numbers of independent variables and sample sizes\(^5\)

<table>
<thead>
<tr>
<th>Sample size</th>
<th>Significance level 0.01</th>
<th>Significance level 0.05</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No. of independent variables</td>
<td>No. of independent variables</td>
</tr>
<tr>
<td>20</td>
<td>2 5 10 20</td>
<td>2 5 10 20</td>
</tr>
<tr>
<td>50</td>
<td>45 56 71 n.a.</td>
<td>39 48 64 n.a.</td>
</tr>
<tr>
<td>100</td>
<td>23 29 36 49</td>
<td>19 23 29 42</td>
</tr>
<tr>
<td>250</td>
<td>13 16 20 26</td>
<td>10 12 15 21</td>
</tr>
<tr>
<td>500</td>
<td>5 7 8 11</td>
<td>4 5 6 8</td>
</tr>
<tr>
<td>1000</td>
<td>3 3 4 6</td>
<td>3 4 5 9</td>
</tr>
<tr>
<td></td>
<td>1 2 2 3</td>
<td>1 1 2 2</td>
</tr>
</tbody>
</table>

n.a. = not applicable

**INTERPRETATION OF REGRESSION VARIATE:** Once the model is derived and its power is estimated, the final task is to interpret the regression parameter by evaluating the estimated regression coefficients for their explanation of the dependent variable. The regression coefficient captures the effect of one variable while controlling for (i.e. holding

---

constant) the other variables in the model. There are two coefficient values; the unstandardized (b) and standardised regression coefficients (beta) which is of our interest for interpretation purpose. While controlling for all other independent variable, the amount of change in the dependent variable due to a one-unit change in that independent variable is represented by the unstandardised regression coefficients (b-coefficients) (Newton and Rudestam 1999).

Since, the b-coefficient cannot reveal which independent variable is a more important predictor of the dependent variable (Hair et al. 1998; Newton and Rudestam 1999), in this study, a modified b-coefficient, called beta-coefficient, is used for the purpose of variate interpretation which is a standardised regression coefficient that allows for a direct comparison between coefficients as to their relative explanatory power of the dependent variable and it is commonly termed as beta-weight with a β symbol. Since b-coefficients are expressed in terms of the units of the associated variable, it makes the comparisons inappropriate while beta coefficients use standardised data which can be directly compared (Hair et al. 1998).

This allow the comparison of two independent variables which are measured on very different scales, for example, education measured on ordered categorical scale and religiosity measured on a 5-point scale. The beta coefficients is interpreted in term of the expected change in the dependent variable, expressed in standard scores, associated with a change of one standard deviation in an independent variable, while holding the remaining independent variables constant (Newton and Rudestam 1999). For example, if an independent variable has a beta weight of 0.3, it means that the dependent variable will increase by 0.3 a standard deviation when other independents are held constant.

5.7 Summary

To conduct a successful data analysis, well-specified statistical procedures are essential prerequisites. A careful attention is required to be paid to determine the suitability of the statistical techniques chosen to conduct a sound empirical analysis.
In this chapter, an attempt is made to shed lights on statistical techniques applied to analyse the data obtained from the survey. The chapter starts with the discussion of the choice of statistical package for the present study and it is decided that the Statistical Package for the Social Science (SPSS) version 20 is the most suitable statistical programme to be used for this study. Subsequently, various factors that guide the choice of statistical techniques are reviewed. The factors include the objectives of the analysis, focus of the analysis, sample type and size, the level of measurement and the distribution pattern of the data. The discussion on the method of statistical analyses used in the present study follows the discussion. There are a variety of statistical analyses available which can be employed to analyse the survey data. These are descriptive statistics in the form of univariate analysis, bivariate analysis which consist of univariate analysis of variance (ANOVA), and finally multivariate analysis which consist of exploratory factor analysis, multivariate analysis of variance (MANOVA) and multiple linear regression analysis.