Chapter 2

Methodology

This chapter describes the methods for selecting the subjects from the target population, and for collecting the data from the subjects, and the statistical methods used to analyse the data. Factor analysis is used to reduce the dimensionality of the multivariate outcomes. Since the outcomes are continuous we use linear regression to answer the questions of interest and to measure the association between the outcomes and the determinants.

2.1 Data Collection

The data were collected by the Graduate School at Prince of Songkla University. Questionnaires were administered to graduating students in 2002, and the data from these questionnaires were linked to records in the Registrar's Office.

2.2 Graphical and Statistical Methods

The data are stored in Excel, imported to Microsoft SQL Server and analyzed using *Webstat*, the new statistical expert system technology written in HTML and VBScript. The Excel add-in *Ecstat* and imported to *SPSS* for windows version 11.0 are also used for statistical data analysis. This is a suite of functions for graphing and analysis statistical data, as follow: histogram and numerical summaries for data from all variables, factor analysis, two-sample t-test, one way analysis of variance and multiple regression analysis of the variables described by box plots and 95% confidence intervals of means.

Factor Analysis

Since we have multivariate outcomes, factor analysis is used to reduce the dimensionality of these outcomes. Factor analysis is a data reduction technique. It is a group of procedures designed for removing duplicated information from a set of correlated variables and representing the variables with a smaller set of derived variables or factors. There are three procedures involved. The first stage is obtaining the original data matrix. A set of subjects O_1, $O_{2,......,}O_n$ are measured with a different number of variables V_1 with V_2, V_1 with V_{3} ,etc., according to the following formula: If x_i is the observation from subject i on V_1 and y_i is the observation from subject i on V_2, then the correlation between V_1 and V_2 is given by

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

where s_x and s_y are the sample standard deviations of V_1 and V_2, and n is the number of pairs of observations.

The last stage involves the factor loadings. These reveal the extent to which each of the variables contributes to the meaning of each of the factors. Within any one column of the factor matrix, some of the loadings will be high and some will be low. The variables with a high loading on a factor will be the ones that provide the meaning of the factor (Kachigan, 1991).

There are many ways to determine the number of factors. Educational researchers often use methods based on eigenvalues. Since the aim of factor analysis is to reduce the number of variables, eigenvalues less than one indicate that this factor contributes less than the original variable and therefore should not be retained.

While methods based on sizes of eigenvalues have some popular appeal, the most statistically valid method is based on maximum likelihood estimation of the coefficients in the factor analysis decomposition.

Maximum likelihood factor analysis is a widely used method. This method enables us to carry out test of the goodness of fit of a solution comprising k factors. It provides a test of the null hypothesis that k common factors are sufficient to describe the data. The algorithms for this method are given as follows.

Suppose we have p variables and want to fit k factors. Let R be the p × p correlation matrix of the variables, L the p × k matrix of factor loadings, and ψ the vector of length p containing the unique variances. Then we need to find values for L and ψ that maximise the likelihood function, $f(L,\psi)$.

For the fixed value of ψ, we maximize $f(L,\psi)$ with respect to L. The value of L is then substituted into $f(L,\psi)$. Now f can be reviewed as a function of ψ. A transformation of this function gives

$$M(\psi) = \sum_{m=k+1}^{p} \left[\log \gamma_m + \frac{1}{\gamma_m} - 1 \right]$$

where $\gamma_1 \leq \gamma_2 ... \leq \gamma_p$ are the eigenvalues of $\psi R^{-1} \psi$. We then minimize $m(\psi)$. This gives an estimate of ψ, which is then put into the likelihood $f(L, \psi)$. Then the likelihood is again maximized with respect to L. Then a new value for $m(\psi)$ is computed and so on.

After making the decision on how many factors to extract from the original set of variables we can redefine the factors so that the explained variance is redistributed among the new factors. This technique is used to sharpen the distinction in the meaning of the factors. A redefinition of the factors, with the loading on the various factors either very high or very low, and then eliminating as many medium sized loading, aids in the interpretation of factors.

Varimax rotation is one of many types of rotation and is regarded as the standard approach. This approach places more emphasis on the simplification of the factors. It tends to avoid a general factor. Using the comprehensibility method to select a number of factors, suppose that three factors are retained. Table 2.1 shows the factor loadings before and after using a rotation of the factors.

Before rotation			
Variable	F_1	F_2	F_3
V_1	M	M	L
V_2	H	L	L
V_3	M	M	L
V_4	M	L	H
V_5	H	M	M
V_6	H	M	M
V_7	L	H	M
V_8	M	M	H
V_9	M	M	L

After rotation			
Variable	F_1	F_2	F_3
V_1	M	M	H
V_2	H	L	L
V_3	L	H	L
V_4	L	L	H
V_5	H	L	M
V_6	H	M	L
V_7	M	H	L
V_8	L	M	H
V_9	H	L	L

Factor loading H: high, M: medium and L: low

Table 2.1 Factor rotation

Two sample t-test

The two sample t-test is used to test the null hypothesis that the population means are the same, and the t-statistic is obtained as follows

$$t = \frac{\overline{y}_1 - \overline{y}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

If s_1 and s_2 denote the standard deviations of the two samples, respectively, it may be shown that the pooled sample standard deviation is given by the formula

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

A p-value is now obtainable from the table of the two-tailed *t* distribution with $n_1 + n_2 - 2$ degree of freedom. This statistical procedure is called the two sample t-test (McNeil, 2000).

One-way analysis of variance

Considering the analysis of data in which the outcome is continuous and the determinant is categorical, this leads to a procedure called the (one-way) analysis of variance (anova). The null hypothesis is that the population means of the outcome variable corresponding to the different categories of the determinant are the same, and this hypothesis is tested by computing a statistic called the *F-Statistic* and comparing it with an appropriate distribution to get a *p-value*. Suppose that there are n_j observation in sample *j*, denoted by y_{ij} for i = 1, 2,..., n_j. The F-statistic is

$$F = \frac{(s_0 - s_1)/(c-1)}{s_1/(n-c)}$$

where $$s_0 = \sum_{j=i}^{c}\sum_{i=1}^{n_j}(y_{ij} - \overline{y})^2, s_1 = \sum_{j=1}^{c}\sum_{i=1}^{n_j}(y_{ij} - \overline{y}_j)^2$$

and $$y_j = \frac{1}{n}\sum_{i=1}^{n_j} y_{ij}, y = \frac{1}{n}\sum_{j=1}^{c}\sum_{i=1}^{n_j} y_{ij}, n = \sum_{j=1}^{c} n_j$$

s_0 is the sum of squares of the data after subtracting their overall mean, while s_1 is the sum of squared of the residuals obtained by subtracting each sample mean. If the population means are the same, the numerator and the denominator in the F-statistic are independent estimates of the square of the population standard deviation (assumed the same for each population). The p-value is the area in the tail of the F-distribution with *c-1* and *n-c* degrees of freedom (McNeil, 1996).

Multiple Linear Regression Analysis

Linear regression analysis is used to analyze data in which both the determinants and the outcome are continuous variables. In the simplest case involving a single determinant, it can summaries the data in the scatter plot by fitting a straight line. In conventional statistical analysis the line fitted is the least squares line, which minimizes the distances of the points to the line, measured in the vertical direction. If there is more than one determinant, the method generalizes to multiple linear regression, in which the regression line extends to the multiple linear relation represented as (McNeil, 1998).

$$Y = \beta_0 + \sum \beta_i x_i + \varepsilon$$

where Y is the outcome variable, β_0 is a constant, $\{\beta_i\}$ is a set of parameters ($i = 1$ to p, the number of determinants), and $\{x_i\}$ is a set of determinants ($i = 1$ to p).

The model is fitted to data using least squares, which minimizes the sum of squares of the residuals.

Linear regression analysis resets on three assumptions as follows.

(1) The association is linear.

(2) The variability of the error (in the outcome variable) is uniform.

(3) These errors are normally distributed.

If these assumptions are not met, a transformation of the data may be appropriate. Linear regression analysis may also be used when one or more of the determinants is categorical. In this case the categorical determinant is broken down into *c*-1 separate binary determinants, where *c* is the number of categories. The omitted category is taken as the baseline or referent category.

Correlation Coefficient

The correlation coefficient is a measure of the linear or straight-line, relationship between variables and level of relation. The model of correlation coefficient is defined as (McNeil, 1998).

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 (y_i - y)^2}}$$

It may be shown that r ranges from a minimum of -1 to maximum value of 1. A correlation coefficient equal to 0 indicates no linear relationship between the two variables.