CHAPTER 2

METHODOLOGY

This chapter presents of the methods used in the study. This method include the following components.

(a) Study Design

(b) Population and Sample

(c) Variables

(d) Data Collection and Data Management

(e) Statistical Methods

(f) Graphical Methods

1. Study Design

A cross-sectional study is used in this study. It is most appropriate because the subjects are first sampled from a target population and subsequently classified with respect to both the outcome and exposure to determinants of interest (McNeil, 1996 page 6).

2. Population and Sample

The target population comprises all undergraduate students at Universities in Thailand which the demographic data similar to those in Pattani Campus. The sample selected for the study consists of the graduates of the 1993 class from the Faculty of Science & Technology, the Faculty of Humanities and Social Science, the Faculty of Education, and the Islamic Studies College, Prince of Songkla University Pattani Campus.

Students who were not registered, died, were expelled from the university, or were absent from the university during the study period (1993) were excluded. The method of selection can be described as follows.

In 1993, 777 students from Prince of Songkla University, Pattani were selected for the study. Of these, 150 students were excluded for the reasons indicated in Table 1.

Table 1: Subjects excluded

Exclusion	Science and Technology	Humanities and Social Science	Education	College of Islamic Studies	Total
No registered	2	50	39	14	105
Died	-	-	1	-	1
Expelled	4	12	4	2	22
Absent	3	16	2	1	22
Total	9	78	46	17	150

The total number of students in this study is thus 627. All of them have complete data for analysis. The number of students in each faculty is shown in Table 2.

Table 2: Subject selected

Faculty	Total Enrollment	Total Exclusion	Total Eligible	%Excluded
Education	247	46	201	5.9
Humanities and Social Science	394	78	316	10.0
Science and Technology	70	9	61	1.2
College of Islamic Studies	66	17	49	2.2
Total	777	150	627	19.3

The sample size needed to obtain a specified precision can be estimated using the following equation (McNeil, 1996, page 267).

$$n = Z^2_{\alpha/2}\,\sigma^2/d^2$$

where $Z_{\alpha/2}$ is the critical value for the standardized normal distribution corresponding to a two-tail probability α, σ is the population standard deviation of the outcome, and d is half of the width of the $100(1-\alpha)\%$ confidence interval. Thus to estimate the university grade point average of a group of students to within 0.03, with $\alpha = 0.05$, the sample size of the study should be

$$n = 1.96^2(\sigma^2)/(0.03)^2$$

In practice σ is not known in advance but evidence suggests that university grade point average have a standard deviation close to 0.4, in which case

$$n = 3.84(0.4^2)/(0.03)^2$$
$$= 683$$

Thus, the sample size n = 627 should be sufficient to give an accuracy close to 0.3.

3. Variables

The stratification variables included are demographic status (gender, religion and age) and socioeconomic status (family income, father's occupation, and mother's occupation). The predictor variables are university enrollment (entrance, and the university entrance examination score) and education background (formal school, non-formal school, and school grade point average). Both of these predictor variables are measured in categories. The categories of each variable are shown in Table 3. The outcome variable, achievement, was measured by university grade point average at the end of senior year.

Table 3: Categories of each variable

Variables	Categories
Demographic Status	
Gender	male, female
Age	16-17, 18, 19, and 20'
Religion	not Muslim, and Muslim
Residence	local, near local, and others
Socioeconomic Status	
Family status	couple, single parent, and separated
Family income	<5,000 baht/month, 5,000-15,000 baht/month, and >15,000 baht/month
Father's education	no degree, and degree
Father's occupation	Government, private employee, private owner, and others
Mother's education	no degree, and degree
Mother's occupation	Government, private employee, private owner, and others
University Enrollment	
Entrance mode	pooled, and direct
University Entrance-Examination Score	<200 marks, 200-300 marks, and >300 marks
Education Background	
School	formal, and nonformal
School GPA	<2, 2-3, and >3
Faculty	Education, Humanities and Social Science, Science and Technology, and College of Islamic Studies
Basic education	Science, language, language&math, and others

4. Data Collection and Data Management

The university entrance examination score and the university grade point average after 4 years were collected from the Education Services Division, Prince of Songkla University, The Pattani Campus. Planning Division, Prince of Songkla University, provided this information in the form that the students completed before registration.

The data were stored in Microsoft Access, and recoded, where necessaries using the Spida package (Gebski et al, 1992). Matlab version 4 (Hanselman et al, 1995) and Asp (McNeil et al, 1997) were used for graphical presentation and statistical analysis. The information form was coded. Cross tabulation techniques and frequency distributions were used to check for inconsistent or invalid data values. Any dubious data were rechecked from the original information form (questionnaire) and re-entered correctly into the database.

Recoding of data, to combine those categories, which had too many values, was done to facilitate analysis. For example, some categories had very few numbers and thus contained little information.

5. Statistical Methods

Analysis of variance and t-tests are used for preliminary analysis. Since the outcome variable is continuous and the determinants comprise more than one variable, multiple regression analysis is the appropriate method for statistical modeling.

(a) A p-value for a dichotomous determinant is obtained by using the two-sample t-test. In this case the null hypothesis states that the population means for the two groups are the same, and the t-statistic is obtained by dividing the difference between the sample means by the standard error. This gives

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

where s is the pooled sample standard deviation. The p-value is then obtained by computing the area in the two tails of the t-distribution with n_1+n_2-2 degrees of

freedom. The statistical assumptions are that the data arise from normally distributed populations with common standard deviation.

(b) One -way Analysis of Variance is the method used for the analysis of data in which the outcome is continuous and the determinant is categorical. The null hypothesis states that the samples have arisen from the same population. This null hypothesis can be 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 defined as

$$F = \frac{(S_0 - S_1)/(c-1)}{S_1/(n-c)}$$

where S_0 is the sum of squares of the data after subtracting their overall mean, while S_1 is the sum of squares 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 square of the population standard deviation and the *p*-value is the area in the tail of the *F*-distribution with *c*-1 and *n-c* degrees of freedom. The statistical assumptions are that the data arise from normally distributed population.

(c) A sample of size n from the population might be represented by the set $\{(x_{1i}, x_{2i}, \ldots, x_{ri}, y_i); i = 1, 2, \ldots, n\}$. The value y_i is a value of a random variable Y_i. The theoretical equation takes the form

$$\mu_{Y/x_1,x_2,\ldots,x_r} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_r x_r,$$

where $\beta_o, \beta_1, \beta_2, \ldots, \beta_r$ are parameters to be estimated from the data. Denoting these estimates by $b_o, b_1, \ldots, b_r$, respectively, then the sample regression equation becomes

$$y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_r x_r$$

The constants b_1, b_2,..., b_r are called the regression coefficients. The regression coefficients measure the associations between the determinants and the outcome variable. There are three assumptions for multiple regression, the first assumption is that the relationship between the outcome and the determinants is linear in the population. The second assumption is that the standard deviation is the same in each subgroup (the *homogeneity* assumption). Finally, it is assumed that the errors from the regression model are normally distributed, and mutually independent.

The relation between a response variable and several predictor variables cannot be determined without entering all the predictors into the model (Lunn, and McNeil, 1991, 134). A stepwise procedure is used to combine groups that are not statistically distinguishable. The Asp, (McNeil et al, 1997) suite of Matlab programs, was used for this analysis.

A regression analysis may have two different goals (Kleinbaum, 1998): to predict the dependent variable by using a set of independent variables; and to quantify the relationship between one or more independent variables and a dependent variable. The first of these goals focuses on finding a model that fits the observed data and predicts future data as possible, whereas the second pertains to producing accurate estimates of one or more regression coefficients in the model. The second goal is of particular interest when the research question concerns disease etiology, such as trying to identify one or more determinants of a disease or other health-related outcomes.

Confounding and interaction are two methodological concepts that pertain to assessing a relationship between independent and dependent variables.

Interaction, which takes precedence over confounding, exists when the relationship of interest differs at different levels of extraneous (control) variables. In linear regression, interaction is evaluated by using statistical tests about product terms involving basic independent variables in the model.

Confounding, which is not evaluated with statistical testing, is present when the effect of interest differs depending on whether an extraneous variable is ignored or retained in the analysis. In regression terms, comparing crude versus adjusted regression coefficients from different models assesses confounding.

6. Graphical Method

The graphical method is presented in the following steps.

1. Histograms and statistics of raw data for all variables.

2. One way analysis of variances of each determinant described by using box plot and 95% confidence intervals of mean. Box plots are abbreviated histograms invented by Tukey (1977). The box covers the central 50% of the distribution of the data between the lower quartile to the upper quartile. It could be denoted by a rectangular box, with the median given special attention. Unusually high or low measurements (outliers) could also be plotted separately, with the bulk of the data represented simply by continuous lines. An outlier is defined as an observation that is more than distance D, where D is the interquartile range or midspread, that is

$$D = \text{Upper Quartile} - \text{Lower Quartile}$$

3. Time series plots of the variables.

4. Scatter plots show the relation between actual scores and predicted scores.