# Chapter 2

## Research Methods

This chapter describes the methods used in the study. The methodology comprises the following components.

1. Data Collection, Management and Data analysis
2. Graphical Methods
3. Statistical Methods

### 2.1 Data Collection, Management and Data analysis

*Data Collection*

The structured questionnaire modified from the Abuse Assessment Scale (AAS) was used to assess the nature of domestic abuse (Norton et al, 1995), using this structure questionnaire, primary data were collected from pregnant women in Pattani Province (Kuning et al, 2004). The sample comprised all pregnant women attending antenatal care clinics in Pattani Hospital between 1 July 2002 and 21 November 2002 and gave their consent to an interview. Relevant data were collected, verified and recorded in a separate data record form and used to investigate factors associated with the nature of domestic abuse. Determinants comprised two sets of factors as follows.

1. Risk factors of domestic abuse
2. Time of domestic abuse and duration of current pregnancy at the time of the interview.

*Data management*

The data were put in a Microsoft Excel spreadsheet file and analyzed using *Webstat* (a set of programs for graphical and statistical analysis of data stored in an SQL database, written in HTML and VBScript), *EcStat* (an add-in to Excel for graphing and analyzing data) and a program in Matlab for fitting the ordinal logistic regression model.

*Data analysis*

Statistics for descriptive analysis include percentages, for measuring prevalence or levels, and odds ratios for measuring associations. Pearson's chi-squared test is used to assess the association between two categorical variables.

Ordinal logistic regression is used to model the association between the domestic abuse (coded as an ordinal outcome) and the determinants.

## 2.2 Graphical Methods

The graphical methods are presented in the following steps.

*Odds Ratio Plot*

Graphs of odds ratios and 95% confidence intervals can be used to present the association between two nominal categorical variables. The association between an outcome variable and the determinants of interest is investigated by an odds ratio, which is similar to a relative risk. The graph of an odds ratio includes a 95% confidence interval. The confidence interval is graphed as a horizontal line containing a dot denoting the estimated odds ratio.

## 2.3 Statistical Methods

From the schematic diagram (Figure 1.1), this study focuses on the association between the outcome and each domestic abuse factor, and intervening variables including the time of the abuse and the duration of current pregnancy at the time of the interview. However, the outcome variable is ordinal and the determinant variables are complex.

*Descriptive Statistics*

The variables of interest are summarized by percentages. The determinants and intervening variables are nominal or ordinal categorical variables which are described by percentages.

*Univariate Analysis*

Pearson's chi-squared test and 95 % confidence intervals for odds ratio are used to assess the associations between the determinant variables and the outcome of this study. The formulas based on contingency tables (McNeil, 1998b) are as follows ($X$ is a determinant of interest, $Y$ is the outcome).

*A. 2 × 2 table*

$X$ is the determinant and $Y$ is the outcome. The odds ratio is a measure of the strength of an association between two binary variables (i.e., in which both the outcome and the determinant are dichotomous) (McNeil, 1998a, 1998b). To illustrate the definition of the odds ratio, assuming the variables takes values 1 and 2, a two-by-two table is constructed as follows.

| | | Y: 1 | Y: 2 |
|---|---|---|---|
| X | 1 | $a$ | $b$ |
| | 2 | $c$ | $d$ |

$n = a+b+c+d$

The estimate the odds ratio is

$$OR = \frac{ad}{bc} \tag{2.1}$$

One method of testing the null hypothesis of no association between the determinant and the outcome is to use the z-statistic $z = \ln(OR)/SE$, where $SE$ is the standard error of the natural logarithm of the odds ratio (McNeil, 1996). An asymptotic formula for this standard error is given by

$$SE(\ln OR) = \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}} \ . \tag{2.2}$$

A 95% confidence interval for the population odds ratio is thus

$$\text{OR} \times \exp\,(\pm\, 1.96\ \text{SE}\ [\ln \text{OR}]). \tag{2.3}$$

The more conventional method of testing the null hypothesis is based on Pearson's chi-square statistic, defined as

$$\chi^2 = \frac{(ad-bc)^2 n}{(a+b)(c+d)(a+c)(b+d)} \ . \tag{2.4}$$

The p-value is the probability that a chi-squared distribution with 1 degree of freedom exceeds this statistic.

*B. $r \times c$ tables*

In this study, some of variables are multi-categorical. We use $r \times c$ tables to compare them. For example: $X$ is occupation (coded as 1, 2, 3, 4, 5 and 6) and $Y$ is the type of domestic abuse (coded as 1, 2, 3, 4 and 5).

| | $Y$: 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $X$ | $a_{11}$ | $a_{12}$ | $a_{13}$ | $a_{14}$ | $a_{15}$ |
| | $a_{21}$ | $a_{22}$ | $a_{23}$ | $a_{24}$ | $a_{25}$ |
| | .. | .. | .. | .. | .. |
| | $a_{61}$ | $a_{62}$ | $a_{63}$ | $a_{64}$ | $a_{65}$ |

The estimate of the odds ratio associated with cell ($X = i$, $Y = j$) is obtained by collapsing the table into a two-by-two table with pivotal cell $a_{ij}$, that is,

$$OR_{ij} = \frac{a_{ij} d_{ij}}{b_{ij} c_{ij}} , \qquad (2.5)$$

where $b_{ij} = \sum_{j=1}^{c} a_{ij} - a_{ij}$, $c_{ij} = \sum_{j=1}^{c} a_{ij} - a_{ij}$, $d_{ij} = n - a_{ij} - b_{ij} - c_{ij}$, $n = \sum_{i=1}^{r} \sum_{j=1}^{c} a_{ij}$ .

The standard error of the natural logarithm of the odds ratio is given by the same formula as for the two-by-two table. In general, the association is comprised of $r \times c$ odds ratios, but only $(r-1)$ $(c-1)$ of them are independent.

Since the odds ratio in this case is obtained from a two-by-two table, Equation (2.2) gives the standard error, that is,

$$SE(\ln OR_{ij}) = \sqrt{\frac{1}{a_{ij}} + \frac{1}{b_{ij}} + \frac{1}{c_{ij}} + \frac{1}{d_{ij}}} , \qquad (2.6)$$

and an asymptotically valid 95% confidence interval is given by Formula (2.3).

Pearson's chi-squared statistic for independence (i.e., no association) in an $r \times c$ table is defined as

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(a_{ij} - \hat{a}_{ij})^2}{\hat{a}_{ij}} , \qquad (2.7)$$

where $\hat{a}_{ij}$ is the expected value of $a_{ij}$ assuming the null hypothesis of independence is true, that is

$$\hat{a}_{ij} = \frac{1}{n} \sum_{k=1}^{c} a_{ik} \sum_{l=1}^{r} a_{lj} . \qquad (2.8)$$

When the null hypothesis of the independence is true, the right-hand side of Equation (2.7) has a chi-squared distribution with $(r-1)$ $(c-1)$ degree of freedom (McNeil, 1998b).

*Logistic Regression*

Multiple logistic regression analysis is used for modeling the association between several determinant variables and type of domestic abuse. Logistic regression is a method of analysis that gives a particularly simple presentation for the logarithm of the odds ratio describing the association of an ordinal outcome with factors, and when fitted to data involving an ordinal outcome and multiple determinants, it automatically provides estimates of odds ratios and confidence intervals for specific combinations of the determinants (McNeil, 1996). For a set of predictor variables $x_1, x_2, ..., x_p$ and a binary outcome $Y$ the logistic regression model takes the form:

$$\ln\left(\frac{p}{1-p}\right) = \alpha + \sum_{i=1}^{p} \beta_i x_i , \qquad (2.9)$$

where $p$ denotes the probability of occurrence of the specified outcome. The probability of the outcome $Y = 1$ can be expressed as

$$P[Y=1] = \frac{\exp(\alpha + \sum_{i=1}^{p} \beta_i x_i)}{1+\exp(\alpha + \sum_{i=1}^{p} \beta_i x_i)} . \qquad (2.10)$$

Using the logistic regression model for the data arising from a two-by-two table, we suppose $x_i = 1$ or 0, that is, the values of determinant $X$ are taken to be 1 (exposure) and 0 (no exposure). Thus the logistic regression model can be written as

$$\ln\left\{\frac{P(Y=1/X=1)}{1-P(Y=1/X=1)}\right\} = \alpha + \beta ; \qquad (2.11)$$

$$\ln\left\{\frac{P(Y=1/X=0)}{1-P(Y=1/X=0)}\right\} = \alpha \qquad (2.12)$$

The equations (2.11) and (2.12) actually are the (natural) logarithms of the odds for the outcome given the exposure ($x = 1$) and non-exposure ($x = 0$), respectively. After exponentiating each equation, the odds for the exposed and non-exposed groups can be written as $\exp(\alpha + \beta)$ and $\exp(\alpha)$, respectively. The odds ratio is therefore obtained from the simple formula

$$OR = \frac{\exp(\alpha + \beta)}{\exp(\alpha)} = \exp(\beta). \tag{2.13}$$

For an ordinal outcome with more than two levels the logistic model takes a different form. The outcome categories are again coded as 0, 1, 2,..., $c$ but $p_k$ is now the probability that an outcome has value at *least k*. Thus for $0 < k \leq c$ these probabilities are given by

$$\ln\left(\frac{p_k}{1 - p_k}\right) = \alpha_k + \sum_{j=1}^{p} \beta_j x_j. \tag{2.14}$$

Logistic regression provides a further statistic, the deviance, which may be used to assess the statistical significance of a set of determinants in the model as follows. The deviance is defined as $-2 \ln L$, where $L$ is likelihood associated with the data for the fitted parameters. Two logistic regression models are fitted to the data, one containing all the determinants of interest, and the other containing all the determinants except for those being assessed. Asymptotically as the sample size gets large; the difference between the values of deviances has a chi-squared distribution, with the number of degrees of freedom equal to the difference in the number of parameters in the sets of determinants being assessed. If $p_k$ is the probability of an outcome of at least $k$,

$$p_k = \frac{1}{1 + \exp(-a_k - \sum_{j=1}^{p} b_j x_j)}. \tag{2.15}$$