Chapter 2

Methodology

This chapter describes the methods used in the study and study design. Data collection and management, path diagram and variables, and statistical methods are also described. Graphical and statistical analyses were performed by using R program.

2.1 Study design

A Cross sectional study was conducted. Data on mortality from committed suicide were gathered from Bureau of Health Policy and Strategy, Ministry of Public Health.

2.2 Study population

The study population comprised persons who died from committed suicide in Southern Thailand during 1996-2006.

Prince of Songkla University
Pattani Campus

2.3 Data collection and management

Gender-age-specific mortality for 14 provinces of southern Thailand in years 1996-2006 were obtained from the vital registration database. This database is provided by the Bureau of Health Policy and Strategy, Ministry of Public Health and contains cause-of-death data based on the tenth International Classification of Diseases (ICD10) according to the National Cause of Death Register 4,588 persons committed suicide during the study period. This data set provides information on age, gender, cause of death and place of resident (province). Since age was included as a demographic determinant, it was divided into 17 groups: 0, 1-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-64, 65-69, 70-74, and 75 years and over. All suicide methods were classified into four groups using the tenth

International Classification of Diseases (ICD10): poisoning by drugs and other means, hanging, firearms, and other methods.

The population denominators by gender, age group and provinces of Thailand during 1996-2006 were estimated by Ministry of Interior of Thailand.

Mortality data were obtained from the Ministry of Public Health. The data were checked for correcting wrong coding and dealing with missing value by using open source statistical package R for windows version 2.8.1. Graphical and statistical modeling was also performed by using this software.

2.4 Path diagram and variables

The outcome was suicide mortality rates in Southern Thailand during 1996-2006. The mortality rates per 1,000 were computed from number of death divided by corresponding populations at risk.

The determinants in the study were age, gender, year and province. The schematic diagram for this study is shown in Figure 2.1

Prince of Songkla University
Pattani Campus

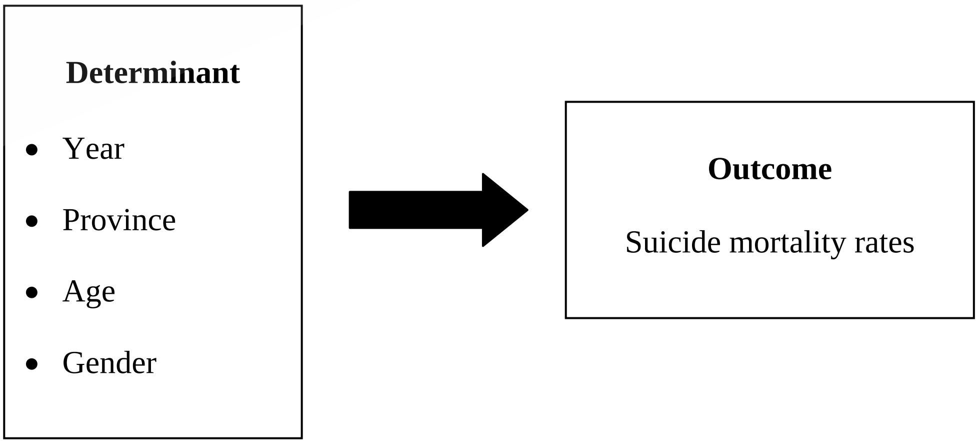

Figure 2.1 Path diagram

2.5 Statistical models

Suicide mortality rates were computed as the number of cases per 1,000 residents in the province, given by

$$y_{ijkn} = \frac{Kn_{ijkn}}{P_{ijkn}} \tag{1}$$

Where n is the number of deaths, y_{ijkn} is mortality rate for year i, province j, age-group k, and gender n, P_{ijkn} is the population at risk and K is a specified constant, here equal to 1,000.

2.5.1 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 describe 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 squares of 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 regressions, in which the regression line extends to the multiple linear relation represented as

$$Y = \beta_0 + \sum \beta_i x_i + \varepsilon \tag{2}$$

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

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

Prince of Songkla University Pattani Campus

There are three assumptions that have to be checked when using linear regression analysis. First, the association between dependent and independent variables is linear. Second, the variability of the error (in the outcome variable) is uniform and 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 are 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 (McNeil, 1996).

2.5.2 Poisson Regression

Poisson regression is appropriate for fitting models with count data (non-negative integer-values). Suicide death is the count data, being the number of people who died from complete suicide which are the non-negative integer-values. The probability function for the Poisson distribution with observed counts of *y* is given by:

Prince of Songkla University
Pattani Campus

$$\text{Prob (Y} = y) \; \frac{e^{-\lambda}\lambda^{y}}{y!} \tag{3}$$

Where

e is the base of the natural logarithm (e = 2.71828...)

y is the number of occurrences of an event - the probability of which is given by the function

λ is a positive real number, equal to the expected number of occurrences that occur during the given interval.

Poisson regression model can be fitted by using the generalized linear models (GLMs) equation with the log link function (McCullagh and Nelder, 1989). Suppose that Y_{ijkm} is a random variable denoting the number of suicide deaths in year i, province j, age group k and gender m. Then the Poisson regression model is takes the form:

$$\ln(\lambda_{ijkm}) = \ln(p_{ijkm}) + \mu + \alpha_i + \beta_j + \kappa_k \qquad (4)$$

Where λ is the mean of Y_{ijkm}, p_{ijkm} is the population in year i province j age group k and gender m, α is the effect of year, β is the effect of province and κ is the effect of age. We assume $\alpha_1 = 0$, $\beta_1 = 0$ and $\kappa_I = 0$.

A problem with the Poisson regression model occurs when we encounter over-dispersion. This means that the variance is greater than mean and thus an assumption of the Poisson distribution is broken.

2.5.3 Over- dispersion

A characteristic of the Poisson distribution is that its mean is equal to its variance. In certain circumstances, it will be found that the observed variance is greater than the mean; this is known as over-dispersion and indicates that the model is not appropriate. A common reason is the omission of relevant explanatory variables. Another common problem with Poisson regression is excess zeros: if there are two processes at work, one determining whether there are zero events or any events, and a Poisson process determining how many events there are, there will be more zeros than a Poisson regression would predict.

An example would be the distribution of cigarettes smoked in an hour by members of a group where some individuals are non-smokers. A method of dealing with such

Prince of Songkla University
Pattani Campus

over-dispersion is to use the more general negative binomial regression in stead of the simple Poisson model.

2.5.4 Goodness of Fit

A measure of discrepancy between observed and fitted values is the deviance.

We show that for Poisson responses the deviance takes the form

$$\mathrm{D} = 2\sum\left\{y_i \log\left(\frac{y_i}{\hat{y}_i}\right) - \left(y_i - \hat{y}_i\right)\right\} \qquad (6)$$

The first term is identical to the binomial deviance, representing "twice a sum of observed times log of observed over fitted". The second term, a sum of differences between observed and fitted values, is usually zero, because Poisson models has the property of reproducing marginal totals, as noted above. For large samples the distribution of the deviance is approximately a chi-squared with *n-p* degrees of freedom, where *n* is the number of observations and *p* the number of parameters. Thus, the deviance can be used directly to test the goodness of fit of the model. An alternative measure of goodness of fit is Pearson's chi-squared statistic, which is defined as

$$\chi_p^2 = \sum\left(\frac{y_i - \hat{y}_i}{\hat{y}_i}\right)^2$$

The numerator is the squared difference between observed and fitted values, and the denominator is the variance of the observed value. The Pearson statistic has the same form for Poisson and binomial data, namely a sum of squared observed minus expected over expected.

Prince of Songkla University Pattani Campus

In large samples the distribution of Pearson's statistic is also approximately chi-squared with *n-p* degree of freedom one advantage of the deviance over Pearson's chi-squared is that it can be used to compare nested models.

Prince of Songkla University
Pattani Campus