Chapter 2

Methodology

This chapter describes the research methodology used to identify patterns and fit a statistical model for road traffic accident mortality in southern Thailand from 1996 – 2006. First we describe the study design and data source and management. Conceptual framework is also described. Finally we identify data analysis and the statistical methods used.

2.1 Study design

We carried out analyses of the patterns and statistical models for road traffic accident mortality based on a retrospective data study. The study focus was on all road traffic accident deaths that occurred in southern Thailand during 1996 – 2006.

2.2 Data source and management

The mortality data for southern Thailand from 1 January 1996 to 31 December 2006 were obtained from the Bureau of Health Policy and Strategy, Ministry of Public Health. The data were collected from death certificates across the whole country. Death certificates are issued by a physician or nurse when death occurs in hospital and by head of village or health personnel when death occurs outside hospital. This data is entered into the vital registration database that is maintained by the Ministry of Interior. It is used by the Ministry of Public Health for coding cause of deaths and analyzing the data for the health statistical reports that they publish. These data contain information including gender, age, place of residence, year and cause of the death that is coded using ICD-10 (International Classification of Diseases 10th revision). For road traffic accident the ICD-10 is described with codes V01 – V89.

ICD (International Classification of Diseases) is the international standard diagnostic classification for all general epidemiological purposes for monitoring of the incidence and prevalence of diseases and other health problems of population. It relates to other variables such as the characteristics and circumstances of the individuals affected, reimbursement, resource allocation, quality and guidelines.

It is also used to classify diseases and other health problems recorded on death certificates and health records. These records also provide the basis for the compilation of national mortality and morbidity statistics (World Health Organization, 2009).

The numbers of mid-year population used as denominators was obtained from civil registration population of Thailand from 1996 to 2006.

Records for road traffic accident in southern Thailand were extracted. The data were checked for errors and missing records.

Ages were categorized into 17 groups, including 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, 56-69, 70-74 and 75 or more.

One outcome of this study was mortality rates calculated from number of road traffic accident deaths that was obtained by aggregating road traffic accident deaths according to variables including year, gender, age and place of residence (province).

2.3 Path diagram and variables

The path diagram of this study is shown in Figure 2.1. This study carried out statistical analyses for estimating road traffic accident mortality rates with the determinant variables comprising year, province, age and gender. This study covered the road traffic accident death for drivers and passengers of motor vehicles as well as pedestrians.

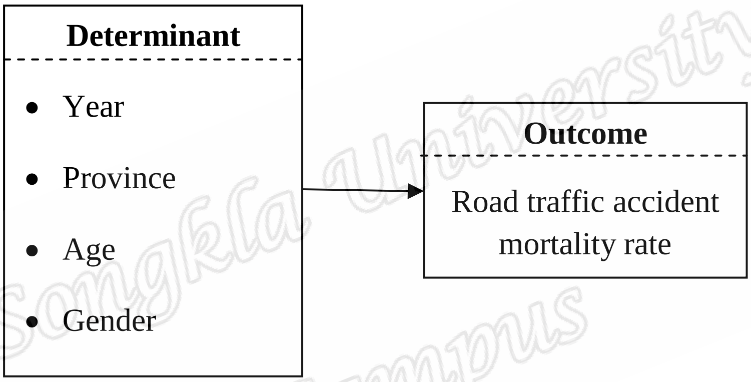

Figure 2.1 Path diagram of the study

2.4 Data analysis

Suppose that D_{ijkm} is a random variable denoted number of deaths for road traffic accidents in year i, province j, age-group k, and gender m, and estimated population P_{ijkm}. Thus the mortality rate can be computed by:

$$Y_{ijkm} = \frac{KD_{ijkm}}{P_{ijkm}} \qquad (1)$$

Where Y_{ijkm} is mortality rate for year i, province j, age-group k, and gender m, K is a scaling constant such as 1,000, 10,000 or 100,000.

In preliminary data analyses we identified road traffic accident mortality in the form of number of deaths and mortality rates. We identified the aggregated number of deaths and road traffic accident death are proportion of all causes of deaths conducted by dividing road traffic accidents number of deaths with number of “all causes of deaths” in corresponding age groups. We then computed age-specific mortality rates for road traffic accidents and “all causes of deaths” in order to identify the age patterns compared with road traffic accident and “all causes of death” for Japan in 2006 (Japan being the highest income country in Asia). In order to identify health inequality and the burden due to road traffic accidents death, we calculated excess death as well as years of life lost (YLL) compared with Japan and other regions of Thailand.

Excess death is the difference between the number of deaths in southern Thailand and the number of deaths that would have occurred in those particular age groups if it had the same mortality rates as Japan in 2006 and other regions of Thailand. Excess deaths were calculated by multiplying excess mortality rates for southern Thailand over Japan and other regions of Thailand with southern Thailand’s population per 1000.

Years of life lost (YLL) is a method that is used to estimate the number of years that are lost due to premature death in the population. It is useful for quantifying the burden of specific diseases (or all causes) that causes the premature death. YLL is the one of the components of the DALY (Disability Adjusted Life Years). DALY and YLL were used by World Health Organization to quantify the burden of premature death and disability for diseases or group of diseases in the publication “Global Burden of Disease” (Murray and Lopez, 1996). YLL can be computed by:

$$YLL[r,K] = \frac{KCe^{ra}}{(r+\beta)^2}\left\{e^{-(r+\beta)(L+\alpha)}\left[-(r+\beta)(L+\alpha)-1\right]-e^{-(r+\beta)\alpha}\left[-(r+\beta)\alpha-1\right]\right\}+\frac{1-K}{r}(1-e^{-rL}) \quad (2)$$

Where *K* is the age-weighting modulation constant, *C* is the adjustment constant for age-weighting constant, r is the discounting rate, α is the age at death, β is the age weighting constant and *L* is standard life expectancy at age α (Murray and Acharya, 1997). In order to compare YLL due to road traffic accident death in southern Thailand with that of Japan, we used the standard method that used in "Global Burden of Disease" with *C* defined as 0.1658, *β* as 0.04, using 3% discounting rate, and *K* assigned as 1 (use standard age weights). Additionally, we calculated YLL by using the standard life table West level 26 (Coale and Guo, 1989) in which life expectancy at birth is 80 and 82.5 years for males and females, respectively. All preliminary analyses are presented using appropriate graphs.

Additionally, this study constructed statistical models for estimating mortality rates. Linear regression our first choice for fitting death rates as the simplest modeling. However mortality rates have a positively skewed distribution. Thus mortality rates need to be transformed by taking the natural logarithm of the values. However, this solution is invalid when the mortality rate outcome includes zeros (i.e. when the number of deaths is zero) because the logarithm of zero is undefined. Therefore, a constant needs to be added to the mortality rates to avoid zeros. However, when mortality rates have an excessive number of zeros linear regression does not fit the data well, and is inappropriate. In this circumstance, Poisson regression can be our better choice.

Poisson regression is commonly used for modeling the number of deaths in a population within a certain time period. But there is a problem with Poisson regression that occurs when over-dispersion occurs. The negative binomial is the traditional alternative regression model for count data when over-dispersion occurs and we use negative binomial for fitting models when Poisson models have over-dispersion.

All graphical and statistical analyses were carried out using R (R Development Core Team, 2008) for entire preliminary data analysis and statistical modeling.

2.5 Statistical methods

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 regression, in which the regression line extends to the multiple linear relation represented as

$$Y = \beta_0 + \sum \beta_i x_i + \varepsilon \tag{3}$$

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 the data using least squares, which minimizes the sum of squares of the residuals.

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).

Poisson Regression

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

$$\text{Prob } (\text{Y} = y) \; \frac{e^{-\lambda}\lambda^{y}}{y!} \qquad (4)$$

Where λ is known as the Poisson parameter, which equals both the mean and the variance. 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 road traffic accident deaths in year *i*, province *j*, age group *k* and gender *m*. Then the Poisson regression model is taken the form:

$$\ln(\lambda_{ijkm}) = \ln(p_{ijkm}) + \mu + \alpha_i + \beta_j + \kappa_k \tag{5}$$

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.

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.

The negative binomial is the traditional alternative regression model for count data and it is the extension of Poisson regression. This distribution of observed counts *y* takes the form:

$$\text{Prob (Y} = y) = \frac{\Gamma(y+k)}{\Gamma(y+1)\Gamma(k)}\left(\frac{k}{k+\lambda}\right)^k\left(\frac{\lambda}{k+\lambda}\right)^y \tag{6}$$

Where Γ is the gamma function and *k* is known as the dispersion parameter, (k is greater than 0). Unlike the Poisson distribution where the mean must equal the variance, the negative binomial allows the variance greater than the mean. The variance of the negative binomial is $\lambda + \lambda^2/k$. Note that negative binomial is equivalence to the Poisson if *k* (the dispersion parameter) is equal to 0. Thus if *k* is equal to 0 Poisson regression model is appropriate, but the negative binomial is appropriate if *k* is significantly different from 0.

Coefficient of determination

Coefficient of determination or r^2 is a common test for evaluating goodness of fit. It relates to the correlation coefficient (the coefficient of determination is the square of the correlation coefficient). It gives the percentage of total variation in the dependent variable explained by the regression line (Schroeder *et al.*, 1986).

Suppose that *y* is the observed values and *f* is associated with the model predicted value. The variability of the data can be measured through different sums of squares:

$$SS_{tot} = \sum_i (y_i - \bar{y})^2 \tag{7}$$

$$SS_{err} = \sum_i (y_i - f_i)^2 \tag{8}$$

Where SS_{tot} is the total sum of squares and SS_{err} denoted as the sum of squared errors, also called the residual sum of squares. Thus coefficient of determination can be calculated as:

$$r^2 = 1 - \frac{SS_{err}}{SS_{tot}} \tag{9}$$

Likelihood ratio test

The likelihood ratio test is used to assess goodness of fit of two competing models. This statistical testing provides an evidence for support a full model over competing model that having a reduced number of model parameters. The likelihood ratio test is takes form:

$$X^2 = -2\left[LL(\beta_R) - LL(\beta_U)\right] \tag{10}$$

Where LL(β_R) is log likelihood at convergence of the “reduced” model (sometimes considered to have all parameters in β equal to 0, or just to include the constant term, to test overall fit of the model), and LL(β_U) is the log likelihood at convergence of the full model. This statistical testing has a distribution that is approximately χ^2 with degree of freedom equal to the difference in the numbers of parameters in the “reduced” and the full model. Therefore we can calculate p-value for testing hypothesis that the reduced number of model parameters is not needed. A small p-value indicates that the full model is fit better than the “reduced” model (Washington *et al*., 2003).