CHAPTER 2

Methodology

This chapter describes the methodology including an overview of the statistical methods for data analysis aligned to the statistical models. Graphical and statistical analyses were carried out using R program (R Development Core Team 2008).

This chapter presents statistical methods adopted in the three papers contained in Appendix. These methods include chi-squared test, Kaplan-Meier survival analysis and Cox proportional hazards for MDR-TB, multivariate linear regression for TB mortality and forecasting in Thailand and log-transformed linear regression for spatial and temporal variations of TB incidence in Nepal.

2.1 Data Source and Management

Data for first and third studies were obtained from National Tuberculosis Center (NTC), Nepal. The reported TB cases for each year were available in computer files comprising characteristics of the disease, gender, address, and the severity of the illness. The MDR-TB data include individual records for disease cases, gender, age, religion, caste, address, year, MDR-TB registration group, sputum smear conversion, culture conversion status and treatment outcome. These data were obtained as excel format which were modified and entered into computer text files suitable for data cleaning and analysis.

Data for second study were provided by the Bureau of Health Policy and Strategy, Ministry of Public Health. The data were collected from death certificates across the whole country. Deaths certificate are issued by a physician or nurse when death

occurs in hospital and by head of village or health personnel when death occurs outside the hospital. This data is entered into the vital registration database that is maintained by Ministry of Interior. It is used by the Ministry of Public Health and coded cause of deaths using International Classification of Disease 10th edition (ICD-10). These data are available in computer files with individual records for disease cases and fields comprising characteristics of the subject and the disease, including dates of sickness and diagnosis, the subject's age, gender, address, and the severity of the illness, including date of death for mortality cases. Data were thus converted to a flat-file format for calculating descriptive statistics and modeling.

2.2 Variables

The variables for the studies are as follows:

Study I: Treatment Outcome for MDR-TB in Nepal

Determinants: Age, gender, religion, caste, year, region and MDR-TB registration group.

Outcome: Sputum smear conversion status and treatment outcome of MDR-TB cases

Study II: Forecasting TB Mortality in Thailand using Multivariate Linear Regression

Determinants: Gender, age, year and location

Outcome: Number of deaths of TB

Study III: Spatial and Temporal Variations of TB Incidence in Nepal

Determinants: Gender, year and location

Outcome: Number of cases of TB

2.3 Statistical methods

2.4 Descriptive statistics

The variables for the preliminary analysis are summarized by counts and percentage values.

2.5 Univariate Analysis

Pearson's chi-squared test and 95% confidence intervals for odds ratios are conventionally used to assess the association between the outcome and determinants. For the odds ratio, the null value is conventionally taken to be one, corresponding to equal risks of an outcome in two comparison groups. This corresponds to a null value of zero for the difference between two population proportions under the null hypothesis. The Pearson's chi-squared test gives the *p*-value for testing no relationship between the determinant and the outcome. The homogeneity test is used to tell if the association could be the same in different strata, small *p*-values providing evidence to the contrary (McNeil 1998a).

A 2×2 table

To illustrate the methods, a *2×2* contingency table is constructed as follows. Let *x* be the binary determinant and *y* the binary outcome coded as zero or one, and *a*, *b*, *c*, and *d* the cell counts (McNeil, 1998a, 1998b).

		y = 1	*y* = 0
x = 1		*a*	*b*
x = 0		*c*	*d*

$n = a + b + c + d$

The odds ratio is

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

Its asymptotic 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 is thus

$$95\%\ \text{CI} \quad = \text{OR} \times \exp\left(\pm 1.96\ SE\ [\ln \text{OR}]\right) \tag{2.3}$$

Pearson's chi-squared statistic is defined as

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

B Non stratified $r \times 2$ tables

In our first study, some of risk factors are multi-categorical, having more than two category levels. We use non-stratified $r \times 2$ tables to compare them. For example, x is age group and y is treatment outcome (1: treatment success, 0: treatment failure).

		y = 1	y = 0
x	1	a_{11}	a_{12}
	2	a_{21}	a_{22}
	…	..	..
	r	a_{r1}	a_{r2}

Thus the estimate of the odds ratio (OR) is

$$\mathrm{OR}_{ij} = \frac{a_{ij} d_{ij}}{b_{ij} c_{ij}} \ , \tag{2.5}$$

where $b_{ij} = \sum_{j=1}^{2} a_{ij} - a_{ij}$, $c_{ij} = \sum_{i=1}^{r} a_{ij} - a_{ij}$, $d_{ij} = n - a_{ij} - b_{ij} - c_{ij}$, $n = \sum_{i=1}^{r}\sum_{j=1}^{2} a_{ij}$

The standard error of the natural logarithm of the odds ratio is given by the same formula as for the *2×2* table. In general, the association is composed of $r \times c$ odds ratios, but only (r-1) ×(c-1) of them are independent.

The standard error is given by

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

A 95 % confidence interval is thus

$$95\ \%\ \mathrm{CI} = \mathrm{OR} \times \exp\left(\pm 1.96\ SE\ [\ln \mathrm{OR}]\right) \tag{2.7}$$

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

$$\chi^2_{(r-1)(c-1)} = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{\left(a_{ij} - \hat{a}_{ij}\right)^2}{\hat{a}_{ij}} \tag{2.8}$$

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

When the null hypothesis of the independence is true, this has a chi-squared distribution with (r-1)×(c-1) degrees of freedom.

2.6 Survival analysis

Survival analysis is concerned with measuring the risk of occurrence of an outcome event as a function of time. It thus focus on the duration of time elapsed from when a subject enters a study until the event occurs, and uses the survival curves to describe its distribution. The outcome of interest is the time until the events occurs and the duration can be measured in days, weeks, months or years from the beginning of observed follow-up on a subject. As usual the event could be death, disease incidence or relapse, recovery or partial recovery, or generally any designated occurrence to individual. Survival analysis is also concerned with the comparison of survival curves for different combinations of risk factors, and uses statistical models to facilitate this comparison (McNeil, 1996).

In general, survival analysis allows for the proper treatment of incomplete data due to subjects dropping into or out of the study. It give rise to *censored* (more precisely, *right-censored*) data. In fact survival data may be censored for any of the following reasons.

(a) The subject withdraws from the study for any reason before experiencing the event (this includes what is called ‘loss to follow-up’).

(b) An intervening event occurs (such as a failure from an unrelated cause), prohibiting further observation on the subject.

(c) The subject does not experience the event before the study ends (or before an analysis of the results is required).

When the event of interest occurs, the survival time is conventionally called a failure time (even though the event might be a 'success', such as recovery from some disease).

The Kaplan-Meier survival curve is defined as the proportion of subjects surviving beyond a given duration of time *t*. For a large population in which the survival times range continuously over an interval, this curve will be a smooth function of *t* that decreases from a maximum value 1 when *t* is 0. In practice the survival curve estimated from a sample of data is a step function that decreases only at the failure times.

Survival function *S(T):* The probability that a subjects survives longer than time *t*.

$S(t) = P$(surviving longer than time t)

$= P(T > t)$

$$\hat{S}(t) = \frac{number\ of\ \ patients\ \ surviving\ longer\ \ than\ \ t}{total\ \ number\ of\ \ patients\ \ in\ \ the\ \ study}$$

then the survival probability at time t can be estimated as

$$S(t_j) = \frac{n_j - d_j}{n_j} \qquad (2.9)$$

$$= 1 - \frac{d_j}{n_j} \qquad (2.10)$$

where d_j is the number of events (deaths) at time *t* and n_j is the number alive just before *t*.

A useful summary of survival that can be computed directly from a survival curve is the *median survival time*. This is the survival time exceeded by 50% of the subjects, and is obtained simply by finding where the survival curve has the value 0.5.

2.7 The Logrank Test

Peto and Peto (1972) derived a *p-value* for testing the null hypothesis that two survival functions are identical, and called this test procedure the logrank test.

The log-rank test assesses the significance of differences between survival curves. It is similar to a chi-squared test for comparing two proportions, but is more complex, in the sense that it is based on a sum of components, where each component corresponds to a different failure time.

In our first study, Kaplan-Meier survival analysis was used to determine overall times to sputum smear conversion, cure, and failed/died, respectively, with other outcomes classified in each case as censored data.

2.8 Cox proportional Hazards

Cox Proportional Hazard Model is one of the most popular tools used in the study of Survival Analysis. Mathematically, the hazard rate $h=h(t)$ is a function of (or depends on) say, n independent covariates $\boldsymbol{X}$, where $\boldsymbol{X}$ denotes the vector $X_1, X_2, X_3 \ldots, X_n$ each of which is Xi, $i = 1, 2, 3, \ldots n$, and t is time. The hazard function can also be written as $h(t, \boldsymbol{X})$. This denotes that the summation of influences of one group over the other is a fixed proportion.

Under the proportional hazards assumption:

$$h(t_i, x) = h_0(t) \exp\left(\sum \beta_i x_i\right) \qquad (2.11)$$

The left-hand side of the equation says that the hazard is influenced by time and the covariates. The right-hand side of the equation contains $h_0(t)$, which is the baseline hazard function when all the X_i are zero. This baseline hazard function is multiplied by *e* to the power of the summation of all the covariates weighted by the estimated coefficients, β_i.

Consequently,

$$\frac{h(t_i x)}{h_0(t)} = \exp\left(\sum \beta_i x_i\right) \qquad (2.12)$$

The left-hand side is the proportion, or ratio, between the hazard of the group with exposure of ***X*** against the baseline hazard. The right-hand side is the exponentiation of the sum of products of estimated coefficients and the covariate vector, X_i, which is now independent of time, i.e. assumed constant over time. Thus $e^{\beta i X i}$ is the increment of the hazard, or hazard ratio, due to the independent effect of the i^{th} variable. Cox (1972) suggested this model. It is called the proportional hazards model, because the relative risk of an event for two subjects depends only on their determinants, and not on their duration of survival.

Both univariate and multivariate Cox proportional hazards models were used to generate estimates of the associations between demographic factors and treatment and the time to cure, with other outcomes censored in the first study.

2.9 Regression analysis

Regression analysis is the method for estimating values of one or more response variables from a set of predictor variables. The purpose of regression analysis is to assess the effects of the predictors on the response variable(s).

2.9.1 Log-transformed linear regression model

The conventional model for handling data where the outcome is continuous is linear regression, assuming independent error terms, each following an identical Gaussian distribution.

Let Y be a log-normally distributed random variable, that is, a random variable whose (natural) logarithm is Normal with mean μ and variance σ^2. This implies that the probability density function of ln(Y) is the density function of the normal distribution, namely

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(z-\mu)^2}{2\sigma^2} \right] \tag{2.13}$$

In our third study, the incidence rates of TB generally have positively skewed distributions so it is conventional to transform them by taking logarithms to obtain the outcome as

$$y_{ij} = \ln\left(1000 \times \frac{n_{ij}}{P_i} \right). \tag{2.14}$$

Thus an additive linear model is fitted to the logarithms of the log-transformed TB incidence rates, namely

$$\ln\left(\frac{n_{ij}}{P_i} \right) = y_{ij} = \mu + \alpha_i + \beta_j \tag{2.15}$$

In this model, P is the corresponding population at risk in 1000s and the terms α_i and β_j represent super-districts and gender- year effects that sum to zero so that μ is a constant encapsulating the overall incidence.

Graphical assessment of normality

A scatter plot of standardized residuals versus their normal scores is commonly used to assess normality of errors in a linear regression model. If these errors are normally distributed, then the standardized residuals should match the normal scores. The plot should then resemble a nearly straight line with intercept zero and slope one.

Homoskedasticity

A scatter plot of the standardized residuals against the fitted response variable can be used to check the homoskedasticity assumption. If the dots on the plot are randomly scattered evenly within a horizontal band around zero, then the homoskedasticity assumption is plausible.

Sum contrasts

Sum contrasts (Venables and Ripley 2002, Tongkumchum and McNeil 2009) are used to obtain confidence intervals for comparing adjusted incidence rates within each factor with the overall incidence rate. An advantage of these confidence intervals is that they provide a simple criterion for classifying levels of a factor into three groups according to whether each corresponding confidence interval exceeds, crosses, or is below the overall mean.

Methods for creating geographical maps

A thematic map is a type of map that uses different colours or shades to graphically display information about the underlying data representing estimated values of a variable at different locations on the map. The thematic map using data in regions might show one region in dark red to indicate that the region has high values, while showing another region in very pale red to indicate that the region has low values. A

range map is a type of thematic map that displays data according to ranges set by the users. The ranges are shaded using colors or patterns. These types of maps are used to show the geographical distribution of the adverse outcome and to identify areas of high risk. Appropriate graphs are used for exploratory data analysis, visualizing the pattern of the data and highlighting possible errors in the data that could cause problems in further analysis.

Since the confidence intervals for factor-specific incidence rates obtained from a model divide naturally into three groups according to their location entirely above the mean, around the mean, or entirely below the mean, we used this trichotomy to create thematic maps of districts according to their estimated incidence rates.

2.9.2 Multivariate linear regression

Multivariate linear regression is the extension of multiple linear regression to allow for several correlated outcome variables. Multivariate regression estimates the same coefficients as one would obtain using separate univariate regression models (Mardia 1979). In addition, multivariate regression, being a joint estimator, also estimates the between-equation covariance. This means that it is possible to test coefficients across equations.

Suppose that data are available for *n* observations, and the response variables are arranged into a *matrix* whose columns are *p* outcome variables and rows correspond to the *n* observations. The model (Mardia et al 1979) is defined in matrix form as

$$\mathbf{Y}_{(n \times p)} = \mathbf{X}_{(n \times q)} \mathbf{B}_{(q \times p)} + \mathbf{E}_{(n \times p)} . \qquad (2.16)$$

In this formulation $\mathbf{Y}_{(n \times p)}$ is an observed matrix of *p* response variables on each of the *n* observations, $\mathbf{X}_{(n \times q)}$ is the matrix of *q* predictors (including a vector of 1s) in

columns and n observations in rows, $\mathbf{B}_{(q \times p)}$ contains the regression coefficients (including the intercept terms), and $\mathbf{E}_{(n \times p)}$ is a matrix of unobserved random errors with mean zero and common covariance matrix $\mathbf{\Sigma}$. Thus, the error terms associated with different response variables may be correlated.

In our second study, each response variable (TB mortality) for each gender and region is assumed to follow its own regression model, so that

$$y_{xt} = \log(m_{xt}) = a_x + b_x t \qquad (2.17)$$

where y_{xt} is the log-transformed central death rate (per 100,000) in age group x and year t, a_x is the level of the age-specific mortality rate, b_x describe annual increase of the age-specific mortality rate and t is the year.

Note that the model for each response variable takes the same form as that given by a univariate model for this response based on the common determinants. The coefficients and their standard errors given by the two methods are precisely the same, but the multivariate method also gives the covariances between the estimated parameters. This model has the additional advantage is that it takes account of correlations between data in different age groups.

Handling zeroes

If any count m_{xt} is zero, Equation (2.17) needs to be modified to give a finite result, so that m_{xt} is replaced by a positive value m_{xt}.

Various methods may be considered for this data modification. Zero counts simply could be omitted, and the fitted model then used to impute counts for these cases before refitting the model (Ardkeaw and Tongkumchum 2009). This method has advantages in situations where under-reporting is known or suspected.

Another method involves adding a constant c to all counts so that $m^*_{xt} = m_{xt} + c$. A third method involves replacing the zeroes by a suitably chosen constant d without changing any values of m_{xt} greater than 0.

2.9.3 Factor analysis

Factor analysis is a mathematical model that tries to explain the correlation between a large set of variables in terms of a small number of underlying factors. A major assumption of the analysis is that it is not possible to observe these factors directly: the variables depend upon the factors but are also subject to random errors (Mardia et al 1979).

In our second study, factor analysis is performed on the age groups with the aim of substantially reducing correlations between them that could mask their associations with the outcome variables. Each factor identifies correlated groups of variables. Ideally each group (which must contain at least two variables to contribute to the factor analysis) contains variables with small correlations with variables in other groups. To achieve this, any variable uncorrelated with all other variables is omitted from the factor analysis. Each factor comprises weighted linear combinations of the variables, and these factors are rotated to maximize the weights of variables within the factor group and minimize the weights of variables outside the group. The resulting weights are called “loadings”. Variables omitted from the factor analysis due to low correlation with all other variables (high “uniqueness”) are treated as separate predictors, so predictors include single variables as well as factors.

In particular, the multivariate linear regression can be extended to factor analysis model by involving the weight sum of factors to data covariance matrix and

minimizing the correlations between the factors for specified number of factors (substantially less than the number of variables). The factor model formulation with *p* common factors is:

$$y_{xt} = \mu_x + \sum_{k=1}^{p} \lambda_x^{(k)} \phi_t^{(k)} \qquad (2.18)$$

where $\lambda_x^{(k)}$ is the loadings and $\phi_t^{(k)}$ is the factors. We used the covariance matrix of estimated slopes in the regression model to fit the factor model.

The number of factors selected was based on obtaining an acceptable statistical fit using the chi-squared test, and these factors were fitted using maximum likelihood with promax rotation in preference to varimax, which requires the rotation to be orthogonal (Browne 2001, Abdi 2003).