

Chapter 2

Methodology

This chapter describes the research methodology used for the study of mortality from infectious diseases and the study of length of stay of patients dying in hospital. Maps of the study areas are included to assist readers. Data collection, data management and data analysis are described separately for each study in this thesis. Diagrams are used to assist the explanation of conceptual frameworks.

2.1 Conceptual framework

Mortality data in Thailand has been computerized since 1917 and kept by the Department of Local Administration, Ministry of Interior. The records of causes of death, based on the ICD-10 system, are then coded by the Bureau of Policy and Strategies, Ministry of Public Health. Previously, numbers of deaths were recorded as counts according to region, gender, age group and cause of death. In 1996, the database system was changed to record the data for each case. This kind of data record provides more opportunities for analysis than does aggregated data, however, the quality of data is still a problem. Porapakkham (1986) and Rukumnuaykit (2006) suggested that disease groups with large number of deaths can be used for statistical analysis and provide more useful results than data involving low counts.

There have been only limited studies attempting to forecast of numbers of deaths according to infectious disease.

Another source of mortality data is the in-patient database from every hospital in Thailand. This data source is more reliable than the civil registration mortality database because the causes of death are given by physicians. All in-patient data from every hospital in Thailand are sent to the National Security Health Office (NSHO). The NSHO uses these data to calculate the expense incurred by each hospital and then calculate the budget that should be allocated to each hospital. Because of the financial implications there are checks of this data and so the data, including data on length of stay, can be assumed to be relatively reliable. There are few studies of length of hospital stay according to different diseases, region, hospital size and patient characteristics and so there is a need for this study.

This thesis consists of two parts: part I, studying mortality from infectious diseases and part II, LOS of patients dying in hospital. All of the mortality data used in this thesis was from 14 provinces in Southern Thailand. The map of provinces of Southern Thailand, with borders of districts included, is shown in Figure 2.1.

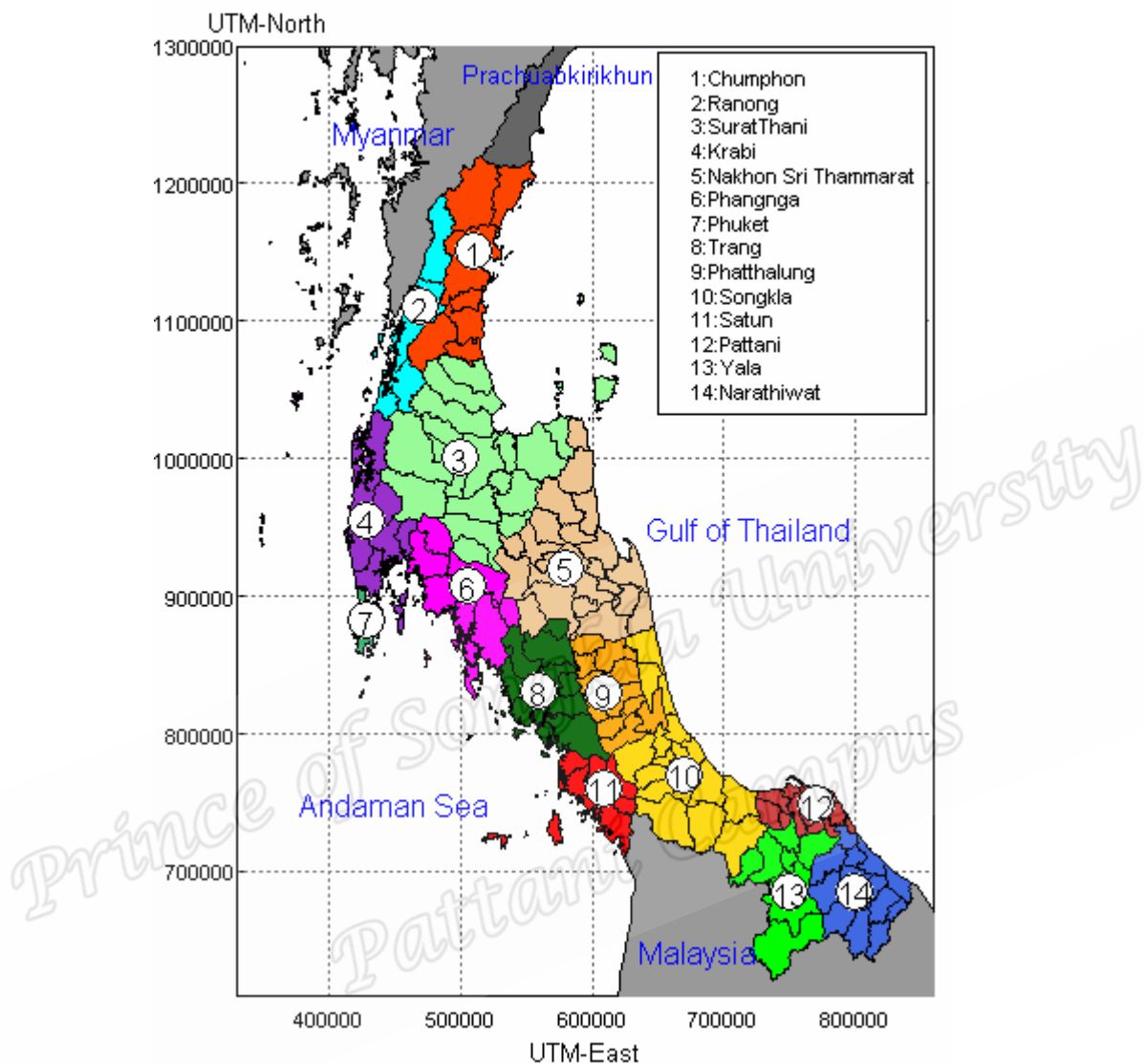


Figure 2.1 Map of provinces with district borders of Southern Thailand

Part I: Deaths from infectious diseases were separated into two groups: deaths from HIV/AIDS and deaths from other infectious diseases. The outcomes for these two groups are death rates. Determinants for the HIV/AIDS incidence rate are gender, location, seasonal period (two-month), place of death and autoregressive terms. Region is defined as the combination of 151 districts from 14 provinces to 38 regions called “super-district”. The two-month period is defined as a combination of two

months together, starting from January and February, March and April and so on.

Autoregressive terms are lag death rates from previous periods. Spatial correlation arises from the death rates between adjacent districts during the preceding two-month period. Gender correlation arises from the correlation between death rates of males and females in the same region during the preceding period. The determinants for other infectious disease death rates are place of death, two-month period and autoregressive terms. Autoregressive terms are included for the determinants of death rates for both HIV/AIDS and other infectious diseases.

The conceptual framework of this part, depicted as path diagrams in Figures 2.2 – 2.3, is used to summarise the variables considered in the study and their roles.

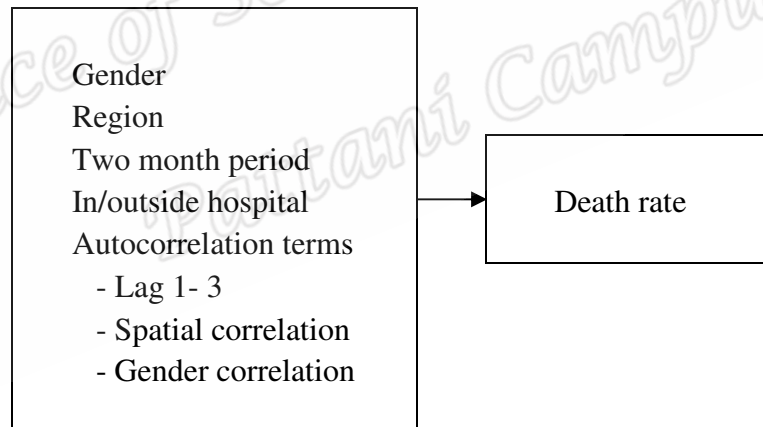


Figure 2.2 Path diagram for HIV/AIDS

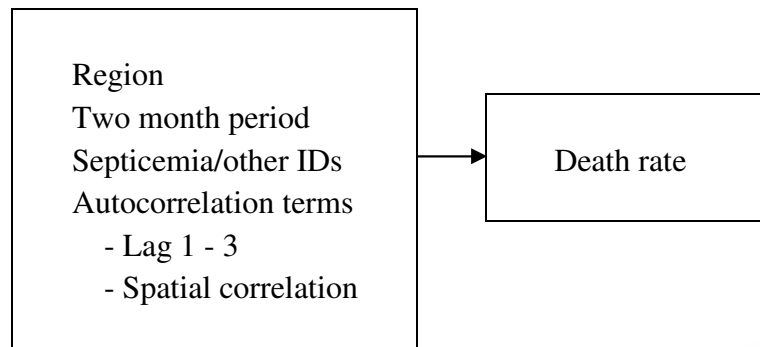
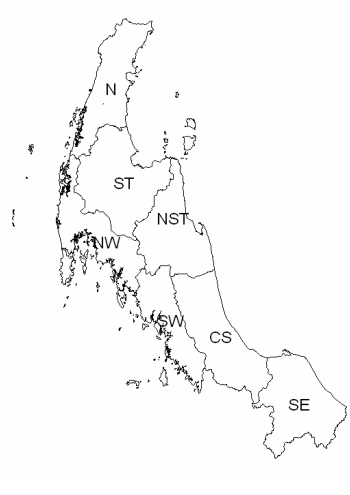


Figure 2.3 Path diagram for other infectious diseases

Part II: The outcome is length of hospital stay. Two nominal explanatory factors were defined: (1) the combination of diagnosis, gender and age group (54 categories), and (2) the combination of region and hospital size (19 categories). Age and gender were grouped into six categories by dividing age into three groups: 0-59 years, 60-74 years and 75 and over. Principal diagnosis using ICD 10 was regrouped into nine broad classes: injuries, digestive diseases (DD), unspecified septicemia (ICD-10 code A41.9), other infectious diseases (ID), chronic obstructive pulmonary diseases (COPD), respiratory infection (RI), cardiovascular diseases (CVD), cancer, and other diseases. The 14 provinces were reduced to seven regions as follows: Chumphon and Ranong (N), Surat Thani (ST), Phangnga, Phuket and Krabi (NW), Nakhon Si Thammarat (NST), Satun and Trang (SW), Songkhla and Phattalung (CS) and Pattani, Yala and Narathiwat (SE). These seven regions were then combined with hospital size (60 or fewer beds, 61-499 beds, 500 beds or more) to give the second factor as shown in Table 2.1.

Table 2.1 Classification of hospitals by region and size in southern Thailand

	Region	Number of beds			Total
		≤ 60	61 - 499	≥ 500	
	North	14	2	1	17
	Surat Thani	17	5	2	24
	North West	16	4	1	21
	Nakhon Si Thammarat	14	5	1	20
	South West	16	2	-	18
	Central South	26	5	2	33
	South East	27	7	-	34
	Total	130	30	7	167

The conceptual framework for this part is depicted in Figure 2.4.

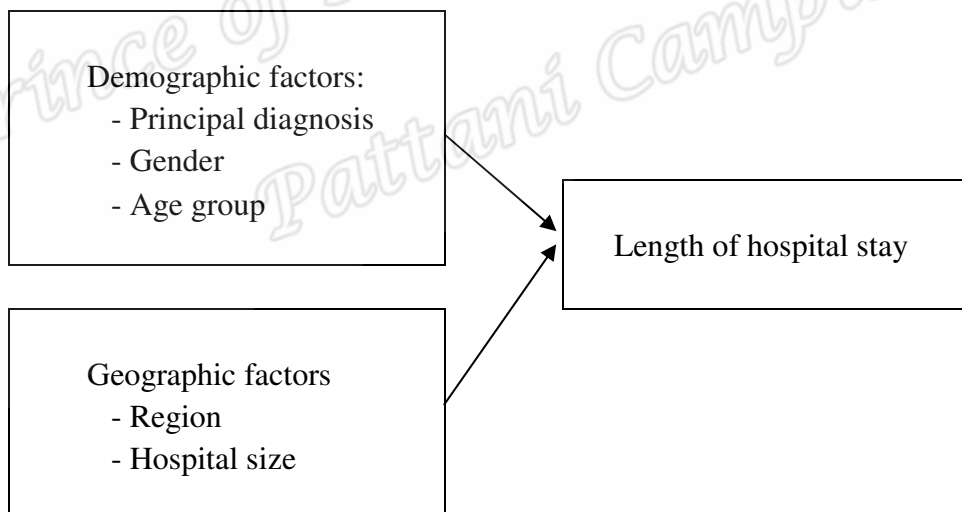


Figure 2.4 Path diagram for length of hospital stay

2.2 Methodology Part I: Studying of mortality from infectious diseases

2.2.1 Study design

A retrospective longitudinal study was conducted. Data on mortality from infectious disease and length of hospital stay were gathered from the Bureau of Policy and Strategy and National Health Security Office (NHSO), Ministry of Public Health.

2.2.2 Study population

The study population comprised persons who died from infectious diseases in Southern Thailand during 1999 – 2004.

2.2.3 Inclusion criteria

All persons dying from infectious diseases in 14 Southern Thailand provinces during the 6 year period from 1999 – 2004.

Persons residing in, and dying in, provinces of Southern Thailand.

2.2.4 Exclusion criteria

Persons dying from HIV/AIDS aged below 15 years.

Persons dying from other infectious diseases (excluding HIV/AIDS) who died outside hospital and age under 40 years.

2.2.5 Variables

Determinants: gender, region, two-month period, place of death, cause of death (septicaemia or other IDs) and autoregressive terms.

Outcomes: Mortality rate (bi-monthly incidence).

2.2.6 Data collection

Data from death certificates in Thailand were obtained from the Ministry of Public Health in collaboration with the Ministry of Interior as part of the Vital Registration System. This database provides information on registered deaths from 1996 to 2003 with approximately 300,000 to 400,000 deaths a year and a total of almost three million observations. This rich data set provides information on age, gender, date of death, cause of death and place (province) of occurrence.

Mortality data from 1999 to 2004 in 14 provinces of southern Thailand were collected from the Bureau of Policy and Strategy, Ministry of Public Health. For these data, the principal diagnosis and demographic information are given on the death certificates including gender, age, occupation, marital status, place of residence, place of death, death date, diagnosing officer and principal diagnosis. The principal diagnosis is coded using the International Classification of Disease in its 10th revision (ICD10). Population denominators were obtained from the Population and Housing Census of Thailand in year 2000 undertaken by the National Statistics Office of Thailand.

2.2.7 Data management

Mortality data from the Ministry of Public Health are kept in database format. Data cleaning was undertaken for correct coding and dealing with missing values by using MySQL. Data were converted to a flat-file format for calculating descriptive statistics, modelling and forecasting.

2.2.8 Statistical analysis

To simplify the effect of location of residence when calculating death rates, one or more contiguous districts in each province were grouped together to form “super-districts” containing populations of 200,000 on average (Table 2.2 and Figure 2.5), where they are listed in order of geographical locations from the northernmost to the southernmost of the region (keeping super-districts within the same province together) with their 2000 census populations.

Table 2.2 Populations (2000 census) of super-districts in Southern Thailand

Super-district	Code	Population
Chumphon North	1	246,279
Chumphon South	2	199,927
Ranong	3	161,210
Surat Thani NW	4	243,238
Surat Thani City	5	241,373
Surat Thani East	6	168,801
Surat Thani South	7	215,998
Phang-nga	8	234,188
Nakhon Si Thammarat North	9	176,496
Nakhon Si Thammarat Northwest	10	163,187
Nakhon Si Thammarat North Coast	11	212,903
Nakhon Si Thammarat Central	12	164,324
Nakhon Si Thammarat City	13	267,560
Nakhon Si Thammarat South Coast	14	238,059
Nakhon Si Thammarat Southwest	15	297,282
Krabi North	16	130,564

Table 2.2 Cont.

Super-district	Code	Population
Krabi South	17	205,646
Phuket	18	249,446
Trang North	19	184,815
Trang City	20	190,340
Trang South	21	219,955
Pattalung City	22	251,029
Pattalung West	23	247,442
Songkhla North Coast	24	149,706
Songkhla West	25	205,607
Songkhla City	26	162,700
HatYai	27	324,596
Songkhla South East Coast	28	177,396
Songkhla South	29	235,657
Satun	30	247,875
Pattani City-West	31	253,567
Pattani Central	32	219,932
Pattani East	33	122,486
Yala City	34	228,042
Yala South	35	187,495
Narathiwat Coast	36	250,997
Narathiwat Central	37	234,441
Narathiwat South West	38	176,912

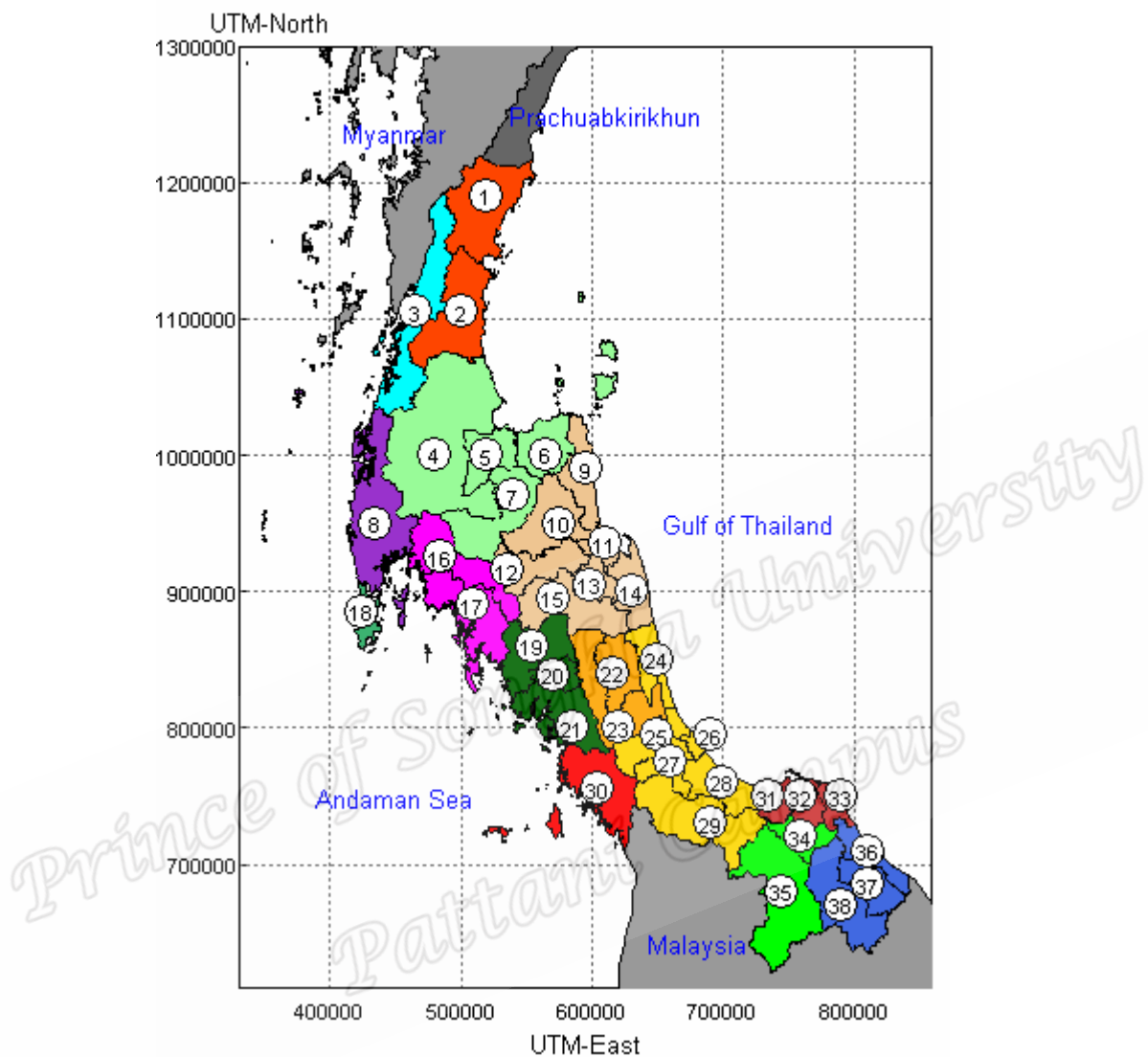


Figure 2.5 Map of combining districts in each province to form super-districts

Since age was included as a demographic determinant, it was divided into three groups: below 5 years of age, 5 to 39 years and 40 and over. The number of deaths in each demographic group defined by age and/or gender and super-district of residence of the deceased was aggregated in intervals of two months: January-February, March-April, etc, giving six annual seasonal periods.

death rates per 1,000 persons for causes of death categories, and per 100,000 persons for HIV/AIDS deaths, in the demographic group resident in each super-district, were calculated to analyze the level and pattern of mortality.

Poisson and negative binomial lagged observation-driven regression models (Venables and Ripley, 2002) for mortality incidence were fitted to the data.

Correlations with death rates in preceding periods, for demographic groups, are incorporated in these models using autoregressive terms. The simplest model is based on linear regression taking the outcome variable as the death rate in a cell indexed by super-district and two-month period, with super-district, season, and autoregressive terms as categorical determinants. More complex models include further demographic components separating in-hospital and outside-hospital deaths, and gender, together with corresponding autoregressive terms.

Incidence rates generally have positively skewed distributions so we transformed them by taking logarithms. To handle zero counts we defined the outcome as

$$y = \ln\left(1 + \frac{Kn}{P}\right), \quad (2.1)$$

where n is the number of disease cases in the cell, P is the population at risk, and K is a specified constant. Suppose that N_{ijt} is a random variable denoting the reported number of disease cases in demographic group i , super-district j and two-month period t and n_{ijt} is the corresponding number observed. An observation-driven linear regression model with m lagged variables may be defined by the equation

$$Y_{ijt} = \mu + \alpha_i + \beta_j + \eta_s + \sum_{k=1}^m \gamma_k y_{ij,t-k} + \rho y_{ij,t-1}^{(\alpha)} + \delta y_{ij,t-1}^{(\beta)} + \varepsilon_{ijt}, \quad (2.2)$$

where $y_{ijt}^{(\alpha)}$ and $y_{ijt}^{(\beta)}$ denote the observed (transformed) incidence rates in all demographic groups other than i and in all super-districts other than j . Y_{ijt} is the outcome variable specified in Equation (2.1) and y_{ijt} the corresponding number observed, ε_{ijt} comprises a set of zero-mean independent Gaussian random variables, and $s = \text{mod}(t, 6)$. We assume $\alpha_1 = 0$, $\beta_1 = 0$ and $\eta_1 = 0$.

Davis et al (2003) suggested observation-driven generalized linear models (GLMs) for time series counts N_t based on the Poisson distribution with mean λ_t , where $\ln(\lambda_t)$ is expressed as an additive function of determinants and lagged observations on N_t . These models are not appropriate for disease epidemics because they express the mean of the process at time t as an exponential function of lagged observations on the same process and are thus numerically unstable when substantial variations occur. However, they become stable when the lagged observations are log-transformed incidence rates using Equation (2.1), and a suitable generalized linear model based on the Poisson distribution is

$$\ln(\lambda_{ijt}) = \ln(p_{ij}) + \mu + \alpha_i + \beta_j + \eta_s + \sum_{k=1}^m \gamma_k y_{ij,t-k} + \rho y_{ij,t-1}^{(\alpha)} + \delta y_{ij,t-1}^{(\beta)}, \quad (2.3)$$

where λ_{ijt} is the mean of N_{ijt} .

Poisson models for disease counts are often over-dispersed due to spatial or temporal clustering of cases, in which case the negative binomial distribution is more appropriate. This distribution has an additional parameter γ and takes the form

$$\text{Prob} [N_t = n] = \frac{\Gamma(n + \gamma)}{\Gamma(n + 1)\Gamma(\gamma)} \left(\frac{\gamma}{\gamma + \lambda_t} \right)^\gamma \left(\frac{\lambda_t}{\gamma + \lambda_t} \right)^n. \quad (2.4)$$

The conditional expected value of N_i is λ_i as in the Poisson model, but the conditional variance is now $\lambda_i + \lambda_i^2/\gamma$ (see for example, Venables and Ripley, 1999, page 233). The parameter γ is inversely related to the over-dispersion, with the Poisson model being the limit as $\gamma \rightarrow \infty$.

We used maximum likelihood estimation to fit these GLMs. Deviance residuals (Venables and Ripley, 1999, page 217) can be plotted against normal scores with points close to a line with unit slope indicating that the fit is satisfactory.

It should be noted that, as in all time series regression models that include lagged observations, the regression coefficients reflect the effects of the predictor variables on the outcomes after these outcomes have been adjusted for the autocorrelations, rather than the direct effects of the predictors on the outcomes.

All analyses in this study were undertaken using R software (R Development Core Team, 2007).

2.3 Methodology Part II: Study of length of stay of patients dying in hospital

2.3.1 Study design

A retrospective longitudinal study was conducted. Data on mortality from infectious disease and length of hospital stay are gathered from the Bureau of Policy and Strategy and National Health Security Office (NHSO), Ministry of Public Health.

2.3.2 Study population

Patients who died in hospitals in southern Thailand during 2002 – 2004.

2.3.3 Inclusion criteria

Persons dying in hospital in the 14 southern Thailand provinces during 2002 – 2004.

2.3.4 Exclusion criteria

We excluded 636 cases with no information on patient demographics or principal diagnosis.

2.3.5 Variables

Determinants: Principal diagnosis, age group, gender, region and hospital size

Outcome: Length of stay in hospital

2.3.6 Data collection

Statistics of patients who died in hospital were obtained from the National Health Security Office. This database provides information on in-patient deaths from 2002-2004. Information in this dataset consists of age, gender, date of birth, date of admission, date of death, principal diagnosis and place of residence.

2.3.7 Data management

Mortality data from the National Security Health Office are kept in database format. Data cleaning was undertaken for correct coding and dealing with missing values by using MySQL. Data were converted to a flat-file format for calculating descriptive statistics and modelling.

2.3.8 Statistical analysis

Two statistical models were used to examine the influences of the two factors on length of stay for patients who died in hospital.

In the first model LOS was treated as a binary variable with patients staying in hospital at least 7 days as the outcome of interest. Logistic regression (Hosmer & Lemeshow, 2000; Kleinbaum & Klein, 2002) was then used to estimate the proportion p_{ij} of these outcomes in diagnosis-demographic group i and region-hospital size group j using the model

$$\ln\left(\frac{p_{ij}}{1-p_{ij}}\right) = \mu + \alpha_i + \beta_j. \quad (2.5)$$

To avoid over-specification of the parameters, for each factor the category having the largest group size was taken as the referent with corresponding parameter 0. To calculate the adjusted prevalence $p_{i\bullet}$ for category i of the first factor, the term β_j in equation (1) was replaced by a constant β_0 , chosen to make the sum of the expected number of outcomes equal to the sum of the observed number, that is,

$$\sum_{i=1}^m p_{i\bullet} n_i = \sum_{i=1}^m p_i n_i, \quad (2.6)$$

n_i being the sample size in category i . The constant β_0 was computed using an iterative procedure. Similarly, to calculate the adjusted prevalence $p_{\bullet j}$ for category j of the second factor, the term α_i in equation (2.5) was replaced by a constant α_0 , again chosen to ensure that the sum of the expected number of outcomes equaled the total observed.

In the second model LOS was treated as a continuous outcome, by taking its natural logarithm after adding a constant d to handle zero days stay, giving the transformed outcome y . The linear regression model is thus similar to equation (2.7), namely

$$y_{ij} = \mu + \alpha_i + \beta_j. \quad (2.7)$$

Estimates of LOS for different levels of the first factor after adjusting for the second factor were calculated by replacing β_j in equation (2.5) by a constant β_0 to give fitted values $y_{i\bullet}$ and then reversing the transformation to give $\exp(y_{i\bullet}) - d$, with β_0 chosen to match the overall fitted mean LOS with the overall observed mean.

Confidence intervals for these parameters were obtained by using the standard errors obtained through fitting each model.

All maps, graphs, data manipulation and statistical analysis were carried out using the R program (R Development Core Team, 2007). In this thesis, we used R because it is an open source program and it is very powerful in producing graphs.