Chapter 2

Methodology

This chapter includes a description of the methods used in the study. Section 2.1 describes the computer program used. Section 2.2 describes data management. Section 2.3 describes graphical and statistical methods used.

2.1 Computer Programs

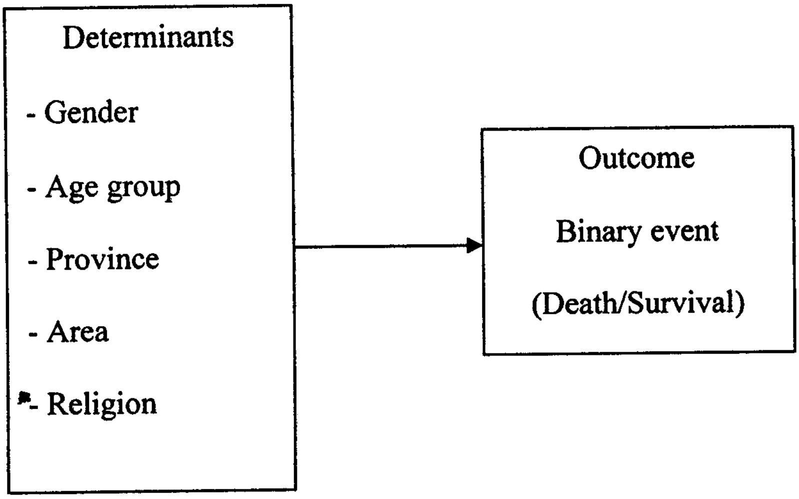
The following computer programs were used for data analysis and thesis preparation. *Microsoft Excel* was mainly used to manage the data used for this research. Some functions are helpful in plotting graphs and analysing data. *Microsoft Word* was mainly used to write and print the report of this research.

WebStat is a suite of web-database software engineering tools written in HTML and VBScript for graphing and analyzing statistical data stored in a SQL database. These programs use a web sever. It is used to perform regression modeling.

2.2 Data Management

Data used in this study is comprised of two sets. The data of population in Southern Thailand were collected from the 2000 Thai population and housing Census, National Statistical Office. The data on death in Southern Thailand were collected from the Registration Administration, December 2000. The two datasets were stored in a Microsoft Excel spreadsheet file by gender, 5-year age group and province.

Path Diagram



2.3 Graphical and Statistical Methods

Graphical Methods

Pyramid graphs are used to show the age distributions of populations. Line graphs can be used to compare age distribution by province, area and gender.

Life Table

The method for constructing a life table l_x for x in (0, 5, …, 85) by gender and province (see, for example, Pollard et al, 1974) is described as follows.

Denote the number of deaths and the population at risk in age group (x, x+5) by D_x and P_x, respectively. The age-specific death rate is $M_x = D_x/P_x$. The probability of dying between ages x and x+5 is $q_x = 5M_x/(1+5M_x/2)$ for $x < 85$ and $q_{85} = 1$. Now define $l_0 = 100,000$ and

$$l_{x+5} = (1 - q_x)\, l_x \tag{2.1}$$

for each value of x. As an illustration, Table 2.1 illustrates the results obtained for males and females in Pattani Province.

	males					females				
x	D_x	P_x	$1000M_x$	q_x	l_x	D_x	P_x	$1000M_x$	q_x	l_x
0	90	32165	2.798	0.01389	100000	74	30567	2.419	0.01202	100000
5	19	34188	0.556	0.00278	98611	10	32354	0.309	0.00154	98798
10	29	31801	0.912	0.00455	98337	16	31776	0.504	0.00251	98645
15	36	28437	1.266	0.00631	97890	15	28736	0.522	0.00261	98397
20	37	26262	1.409	0.00702	97272	21	26174	0.802	0.00400	98141
25	87	23783	3.658	0.01812	96589	23	24589	0.935	0.00467	97748
30	97	21540	4.503	0.02226	94839	34	23057	1.475	0.00735	97292
35	90	20052	4.488	0.02219	92727	37	21672	1.707	0.00854	96577
40	67	16497	4.061	0.02010	90669	39	17163	2.272	0.01130	95756
45	66	13991	4.717	0.02331	88846	46	14641	3.142	0.01559	94674
50	73	10940	6.673	0.03282	86775	38	11044	3.441	0.01706	93198
55	87	9365	9.290	0.04540	83928	50	9579	5.220	0.02576	91609
60	129	8793	14.671	0.07076	80118	86	10251	8.389	0.04109	89249
65	161	6183	26.039	0.12224	74449	155	7422	20.884	0.09924	85582
70	218	5084	42.880	0.19364	65348	189	5607	33.708	0.15544	77089
75	167	2645	63.138	0.27265	52694	176	2917	60.336	0.26214	65106
80	158	1676	94.272	0.38146	38327	171	2052	83.333	0.34483	48039
85	215	1203	178.720	1.00000	23707	300	1759	170.551	1.00000	31474

Table 2.1: Life table calculation for males and females in Pattani Province

Logistic regression

In the simplest case, when there is a single continuously varying determinant x, the regression model logistic for the probability, p that a binary outcome takes a specified value (the "adverse" event) takes the form

$$\ln\left(\frac{p}{1-p}\right)=\alpha+\beta\,x\,, \tag{2.2}$$

Equation (2.2) can be inverted to give an expression for the probability of the event as

$$p=\frac{1}{1+\exp(-\alpha-\beta\,x)}\,. \tag{2.3}$$

The functional form of Equation (2.3) ensures that its values are always between 0 and 1, as they should be given that they are probabilities.

This model is easily extended to handle multiple determinants. For *m* continuous or binary determinants ($x_1, x_2, \ldots, x_m$), it may be written as

$$\ln\left(\frac{p}{1-p}\right) = \alpha + \sum_{j=1}^{m} \beta_j x_j \,. \tag{2.4}$$

Nominal determinants are handled by separating them into their binary components, giving $k-1$ such components for a determinant with *k* categories.

Logistic regression provides an appropriate statistical method for modelling a set of life tables. Since males and females have essentially different life tables (Intachat et al, 2005) we fit separate models for the two sexes. In this method, the outcome is the binary event denoting the death or survival of a male or female at risk in a specific demographic group indexed by 5-year age group and province. The risk of death M_{xj} to such a person in age group (x, x+5), and province j is defined in terms of its logit as

$$ln\{M_{\mathrm{xj}} /(1-M_{\mathrm{xj}})\} = a_{\mathrm{x}} + b_{\mathrm{j}} \,, \tag{2.5}$$

where a_{x} is an age effect and b_{j} is a province effect. To avoid overparametrisation we can force the province effects to have zero mean, i.e., $\Sigma b_{\mathrm{j}} = 0$.

The model life table for province j is now obtained by substituting the values of M_{xj} given by Equation (2.5) into Equation (2.1).

Asymptotic results using statistical theory provide estimates based on maximum likelihood fitting of the model, together with confidence intervals and p-values for testing relevant null hypotheses (see, for example, Kleinbaum & Klein, 2002).

Goodness-of-fit of model

For each cell corresponding to a combination of nominal determinants, the Pearson residual is defined as

$$z = \frac{p - \hat{p}}{\sqrt{\hat{p}(1-\hat{p})/n}}, \tag{2.6}$$

where p is the proportion of outcomes observed in the cell, $\hat{p}$ is the corresponding probability given by the model, and n is the total number of cases in the cell. The goodness-of-fit of the model can be assessed visually by plotting these z-values against corresponding normal scores. The fit is adequate if the points in this plot are close to a straight line with unit slope. A p-value for the goodness-of-fit is obtained by subtracting the deviance associated with the saturated model from the model deviance and comparing this difference R_g with a chi-squared distribution having degrees of freedom equal to $n_g - m$, where n_g is the number of cells and m is the number of parameters in the model.