Chapter 2

Methodology

The study is cross-sectional, based on population data selected from the 2000 Population and Housing Census of Thailand. In this study, the analyses were divided in two sections. The first was **"**Demographic Trends Affecting education Completion in Pattani and Songkhla Provinces**"** and the second was **"**Demographic Factors Affecting Employment in Pattani and Songkhla Provinces**".**

2.1 Demographic Trends Affecting education Completion in Pattani and Songkhla Provinces

This first section is divided in two sub steps as follows:

2.1.1 Demographic factors on education completion

2.1.1.1 Variables and methodology

The determinants studied were demographic factors consisting of two categories of gender: male and female, two categories of religion: Islam and 'other', twelve categories of age groups: 0-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-64 and 64+, while housing area categories were 12 districts in Pattani and 16 districts in Songkhla. The outcome variables were 5 education levels. These consisted of a 'no education' group, which included the uneducated and persons completing only pre-elementary education, an 'elementary' group, a 'secondary' group, which consisted of those completing junior and/or senior high school general education, a 'high' education group, referring to those who studied at university or vocational school and an 'other/unknown' group, meaning those for whom no education information was available

and those whose education completion could not be grouped among others. To achieve clarity in relation to the demographic factors on education completion, the persons who were aged lower than 20 years old and those of unknown education status level were excluded. In analysis, there were three comparisons between groups: those who had completed elementary and more versus 'no education'; those who completed secondary or more versus elementary or less and those who had completed 'high level' education versus 'secondary or less'.

2.1.1.2 Statistical method

2.1.1.2.1 Univariate Analysis

Pearson's chi-square test and 95% confidence intervals for odds ratios are used to assess the association between the determinant variables and the outcome of this study. The formulas of contingency tables are as follows (X is the determinant of interest, Y is education completion, Z is a stratification variable).

2.1.1.2.2 2 x 2 table

X is the determinant and Y is the outcome. Each variable is binary (0 or 1). The odds ratio is a measure of the strength of an association between two binary variables (i.e., in which both the outcome and the determinant are dichotomous). That describes the degree of association between two variables in different factors associated with education completion in Pattani and Songkhla. To illustrate the definition of the odds ratio, a two-by-two table is constructed as follows.

Education completion

			Interest group	Other group
Determinant group	Interest group	1	a	b
	Other group	0	c	d

n = a + b + c + d

The ratio of these odds is referred to as the odds ratio (McNeil 1996). Thus the estimate the odds ratio is

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

One method of testing the null hypothesis of no association between determinant and outcome is using the z-statistic z = ln(OR)/SE, where SE is the standard error of the natural logarithm of the odds ratio. 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

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

Pearson's chi-square statistic is defined as

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

2.1.1.2.3 Non-stratified r×c tables

In this study, some of variables are multi-categorical. We use non-stratified r×c tables to compare them. For example, X is category of age group and Y is category of education completion.

Assume X is nominal (1, 2... r), and Y is ordinal (1, 2, 3, 4).

X	Y = 1	Y = 2	Y = 3	Y = 4
	a_{11}	a_{12}	a_{13}	a_{14}
	a_{21}	a_{22}	a_{23}	a_{24}
	..	..	..	..
	a_{r1}	a_{r2}	a_{r3}	a_{r4}

Thus the estimate of the odds ratio (OR) is

$$\mathrm{OR_{ij}} = \frac{a_{ij} d_{ij}}{b_{ij} c_{ij}} \ , \qquad (2.5)$$

where $\mathrm{b_{ij}} = \sum_{j=1}^{2} a_{ij} - a_{ij}$, $\mathrm{c_{ij}} = \sum_{j=1}^{r} a_{ij} - a_{ij}$, $\mathrm{d_{ij}} = n - a_{ij} - b_{ij} - c_{ij}$, $\mathrm{n} = \sum_{i=1}^{r} \sum_{j=1}^{3} a_{ij}$

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

The standard error is given by

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

A 95 % confidence interval is thus

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

Pearson's chi-square statistic for independence (i.e., no association) in an *r* x *c* table is defined as

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

Where $\hat{a}_{ij}$ is the expected value of a_{ij} assuming the null hypothesis of independence is true, that is $\hat{a}_{ij} = \frac{1}{n}\sum_{k=1}^{c} a_{ik} \sum_{l=1}^{r} a_{ij}$

2.1.1.2.4 Logistic Regression

Multiple logistic regression analysis is used for adjusting the association between determinant variables and education completion. Logistic regression is a method of analysis that gives a particularly simple representation for the logarithm of the odd ratio describing the association of an ordinal outcome with factors, and when fitted to data involving an ordinal outcome and multiple determinants, it automatically provides estimates of odds ratio and confidence intervals for specific combinations of the risk factors. For a set of predictor variables x_1, $x_{2...}$ $x_{p \text{ and}}$ a binary outcome Y the logistic regression model takes the form:

$$\ln(\frac{P}{1-P}) = \alpha + \sum_{i=1}^{p} \beta_i x_i \,, \tag{2.9}$$

where P denotes probability of occurrence of the specified outcome. The probability of the outcome Y = 1 can be expressed as

$$P[Y=1] = \frac{\exp(\alpha + \sum_{i=1}^{p} \beta_i x_i)}{1 + \exp(\alpha + \sum_{i=1}^{p} \beta_i x_i)} \,. \tag{2.10}$$

Using the logistic regression model for the data arising from a two-by-two table, we suppose x_i = 1 or 0 which the value of determinant X taken to be 1 (exposure) and 0 (no exposure).

Thus the logistic regression model can be written as follows

$$\ln\{\frac{P(Y=1/X=1)}{1-P(Y=1/X=1)}\} = \alpha + \beta \tag{2.11}$$

$$\ln\{\frac{P(Y=1/X=0)}{1-P(Y=1/X=0)}\} = \alpha \tag{2.12}$$

The equations (2.11) and (2.12) actually are the (natural) logarithms of the odds for the outcome given the exposed (X=1) and non-exposed (x=0), respectively. After exponentiating each equation, the odds for the exposure and non-exposure group can be written as $\exp(\alpha + \beta)$ and $\exp(\alpha)$, respectively. The odds ratio therefore is obtained from the simple formula.

$$\text{OR} = \frac{\exp(\alpha + \beta)}{\exp(\alpha)} = \exp(\beta) \ . \qquad (2.13)$$

For ordinal outcomes with more than two levels the logistic model takes a different form. The outcome categories are again coded as 0, 1, 2… c but p_k is now the probability that an outcome has value *at least k*. Thus for $0 < k \leq c$ these probabilities are given by

$$\ln\left(\frac{p_k}{1-p_k}\right) = \alpha_k + \sum_{j=1}^{m} \beta_j x_j \ . \qquad (2.14)$$

Logistic regression provides a further statistic, the deviance, which may be used to assess the statistical significance of a set of determinants in the model as follows. The deviance is defined as $-2 \ln L$, where L is the likelihood associated with the data for the fitted parameters. Two logistic regression models are fitted to the data, one containing all the determinants of interest, and the other containing all the determinants except for those being assessed. Asymptotically as the sample size gets large, the difference between the values of the two deviances has a chi-squared distribution, with the number of degrees of freedom equal to the number of parameters in the determinants being assessed. If p is the probability of a beneficial outcome,

$$\text{P} = \frac{1}{1 + \exp(-a - \sum_{j=1}^{p} b_j x_j)} \qquad (2.15)$$

And the risk of an adverse outcome is thus.

$$\text{Risk} = 1 - \frac{1}{1 + \exp(-a - \sum_{j=1}^{p} b_j x_j)} \qquad (2.16)$$

2.1.2 Trends in secondary education completion

2.1.2.1 Variables and methodology

The outcome variable is completion of secondary education. Persons aged less than 20 were omitted because they could still be in the process of completing their secondary education, and others who did not state their education completion status at the 2000 Census were also omitted from the study, giving a total study sample of 342,047 persons for Pattani and 793,927 for Songkhla. The determinants are gender, religion (Muslim or other), 10 age groups, and district. Following Cox and Wermuth (1996), the variables are shown as a path diagram in Figure 2.1.

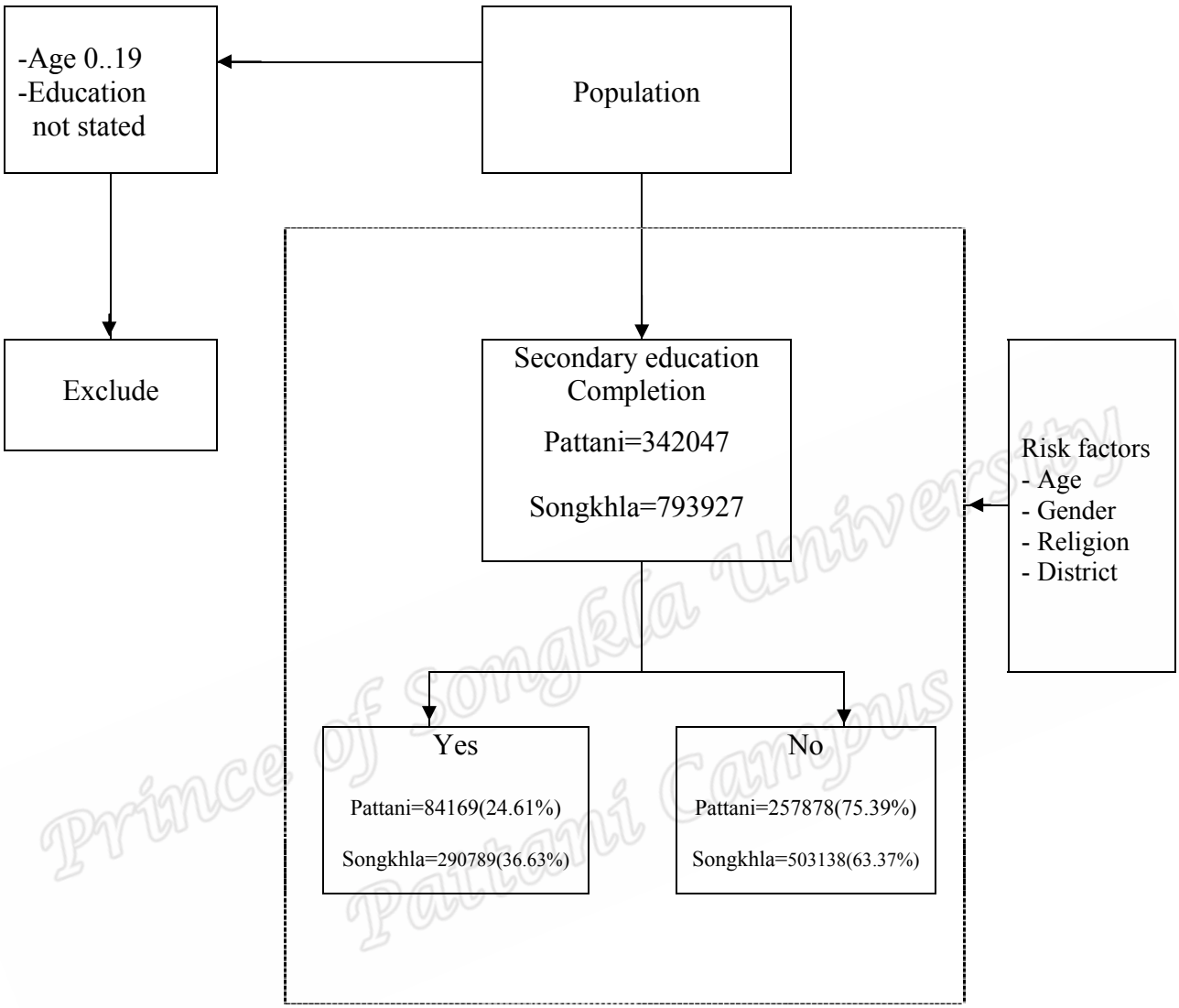

Figure 2.1: Path diagram of variables considered in study

Data from smaller geographically proximate similar districts were combined to avoid zero counts in the statistical analysis, reducing the number of regions from 12 to 7 for Pattani and from 16 to 11 for Songkhla.

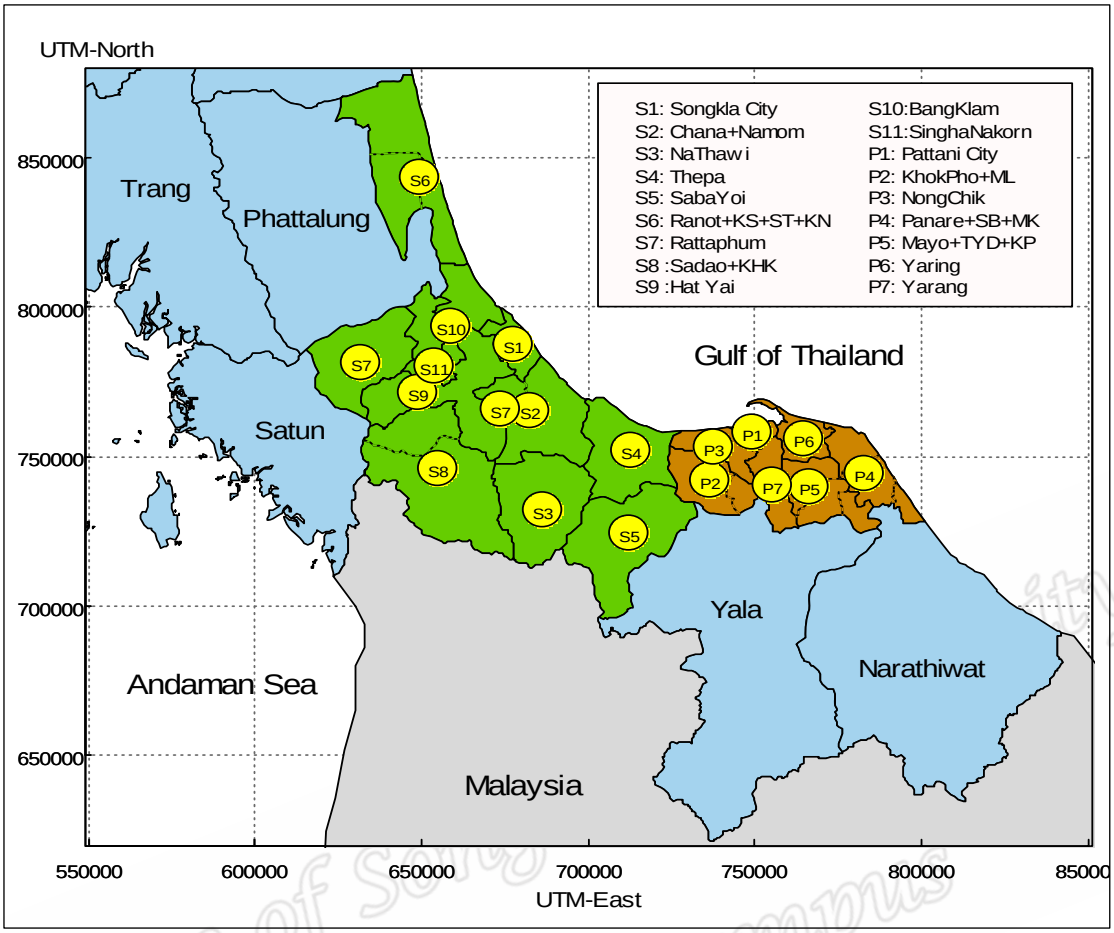

Figure 2.2: Districts in Songkhla and Pattani Provinces with study regions defined by aggregation

Notice: For Songkhla; KS=Kraseasin, ST=SathingPra, KN=KuanNiang and KHK=KlongHoiKong

For Pattani; ML=Maelan, SB=Saiburi, MK=MaiKean, TYD=ThungYangDeang and KP=KaPhor

2.1.2.2 Statistical methods

In preliminary data analysis we compare the secondary education completion rates within the seven district-based regions of Pattani and eleven district-based regions of Songkhla Provinces by plotting these proportions against age group separately for each combination of gender and religion. Since persons in a specified age group were all born within the

same 5-year period, these plots will show how the secondary education completion rates within each district have changed over the period from 1960 to 2000 for each gender-religion group. The secondary school completion rates can be modeled using logistic regression, which provides a method for modeling the association between a binary outcome and multiple determinants. In the simplest case, when there is a single continuously varying determinant x, the model takes the form

$$\ln\left(\frac{p}{1-p}\right) = \alpha + \beta x\,, \tag{2.17}$$

where p is the probability that the outcome is in the specified category. Equation (2.17) can be inverted to give an expression for the probability of the event as

$$p = \frac{1}{1+\exp(-\alpha - \beta x)}\,. \tag{2.18}$$

The functional form of Equation (2.18) 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, …, x_m), it may be written as

$$\ln\left(\frac{p}{1-p}\right) = \alpha + \sum_{j=1}^{m} \beta_j x_j\,. \tag{2.19}$$

Nominal determinants are handled by separating them into their binary components, giving $\tilde{k}1$ such components for a determinant with k categories. Asymptotic results based on 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). For our study we use a special case of Equation (2.19) where the model contains district as a nominal determinant together with its multiplicative interactions with x and x^2, where x denotes age coded as a continuous determinant, with

values 1 for age group 20-24, 2 for age group 25-29, …, 10 for age group 65+. This model can be written as

$$\ln\left(\frac{p}{1-p}\right) = \alpha_i + \beta_i x + \gamma_i x^2 . \qquad (2.20)$$

The terms α_i, β_i and γ_i represent constant, linear and quadratic effects of age for district *i*. Since a goal of the analysis to compare the district effects with each other, the model is more appropriately expressed as

$$\ln\left(\frac{p}{1-p}\right) = \alpha_1 + \beta_1 x + \gamma_1 x^2 + \delta\left(\alpha_i + \beta_i x + \gamma_i x^2\right), \qquad (2.21)$$

where $\delta = 0$ if $i = 1$, $\delta = 1$ otherwise. With this reformulation, the parameters α_i, β_i and γ_i represent the differences between the effects for district *i* and district 1. Since any district may be specified as district 1, this makes it possible to compare any district with any other district.

2.2 Demographic Factors Affecting Employment in Pattani and Songkhla Provinces of Thailand

2.2.1 Variables and methodology

The study aims to examine and compare demographic factors affecting employment in Pattani and Songkhla Provinces. Because of the substantial proportions of young persons in education and of older persons in retirement, persons aged 0-24 and 59+ years were excluded from the analysis. Commencing with younger persons might have distorted results because not all would have had the opportunity to complete their education. Also excluded were those whose level of education completed was not recorded and others whose occupational category was recorded as ‘not stated’. The exclusion of these categories means that people such as full-time students, early retires, unpaid careers within a family and others who are not employed but do not define themselves as unemployed, are excluded from the data. As a

result, the focus of this study is very close to the usual definitions of 'labour force', those people seeking employment, whether unemployed or employed in any of the three occupational categories used (agriculture; elementary, professional). Binary employment status is the outcome variable. For each province, gender or other category, the number of unemployed as a percentage of the total in the labour force gives an 'unemployment rate'. The total study sample was 232,220 persons for Pattani and 551,695 persons for Songkhla. Data from smaller geographically proximate similar districts were combined to avoid zero counts in the statistical analysis, reducing the number of regions from 12 to 7 for Pattani and from 16 to 11 for Songkhla. The determinant variables are demographic factors consisting of gender, religion (Islam and Other), 7 age groups (25-29, 30-34, ….. , 54-59), education level (none, primary, secondary, high) and district (7 regions for Pattani and 11 regions for Songkhla). Following Cox and Wermuth (1996), the variables are shown as a path diagram in Figure 2.3.

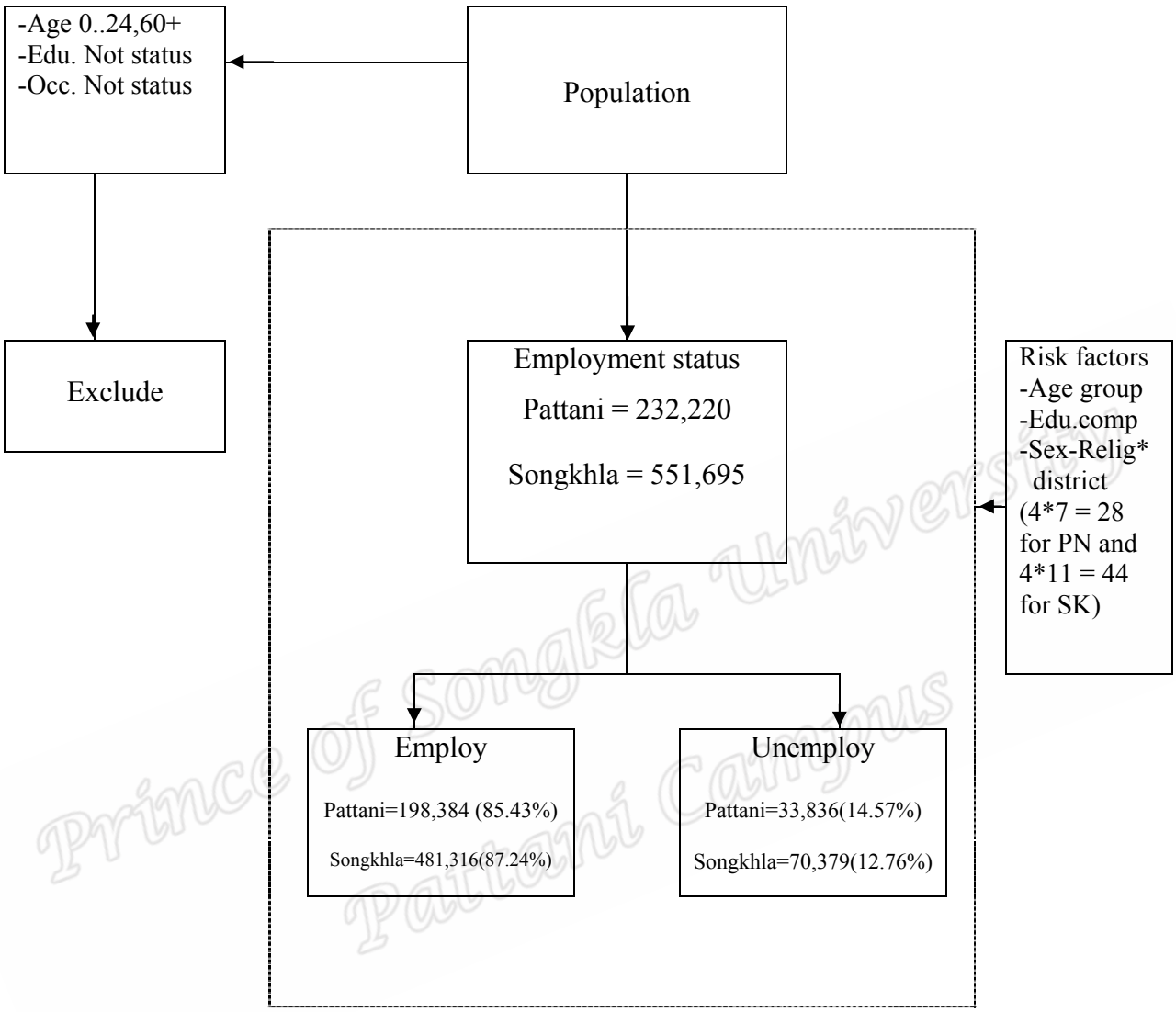

Figure 2.3: Path diagram of variables considered in study.

2.2.2 Statistical methods

The specific aim of the study was to examine the association between unemployment status and completion of education, and to compare these associations in the two selected provinces, after taking into account demographic factors (age group, gender, religion, and location of residence). Since the unemployment outcome is binary and the education

completion determinant is a categorical variable with four factors, this association may be described by a set of odds ratios, and logistic regression (see, for example, Hosmer and Lemeshow 2000) may be used to adjust these odds ratios for the demographic factors. To facilitate comparisons with respect to different levels of education completion, we chose completion of elementary education as the referent level. This allows the results to be expressed in terms of three odds ratios: none versus elementary, secondary versus elementary, and high versus elementary.

Preliminary analysis involved using the logistic regression model to compute separate age-group adjusted odds ratios for each combination of gender, religion and location in each province, giving 44 (2 x 2 x 11) such odds ratios for Songkhla province and 28 (2 x 2 x 7) for Pattani province. Further analysis involved using meta-analysis to combine these odds ratios within each province, and thus obtain overall odds ratios showing the associations between unemployment and each level of education completion. Since odds ratios are more symmetrically distributed when expressed on a logarithmic scale, and logistic regression routinely provides estimates and standard errors of natural logarithms of odds ratios, we did the meta-analysis on the logarithms of the odds ratios, using a method described in McNeil (1996) as follows.

Denote the estimated log odds ratio in stratum g by y_g, its standard error by σ_g and define the weight $w_g = 1/(\sigma_g)^2$. Then the overall (combined) estimate of the log odds ratio is the weighted mean

$$\bar{y} = \frac{\sum w_g y_g}{\sum w_g}, \tag{2.22}$$

and its standard error is

$$\sigma_{\bar{y}} = \frac{1}{\sqrt{\sum w_g}}. \tag{2.23}$$

The overall odds ratio estimate is thus obtained by exponentiation as $\exp(\bar{y})$ and its corresponding 95% confidence interval is given by $\left(\exp(\bar{y} - 1.96\sigma_{\bar{y}}), \exp(\bar{y} + 1.96\sigma_{\bar{y}})\right)$. Confidence intervals for the individual odds ratios may be plotted together with the combined results as a meta-analysis plot (see, for example, Moja et al 2007). The statistical analysis was performed using R (R Development Core Team, 2007).