CHAPTER 4

RESEARCH METHODOLOGY

In this chapter methods designed to achieve the objectives of this study is presented. Essentially, it is an extension of the preceding chapter in which the variables required for the model formulation were theoretically derived and the logic behind their selection also examined.

There are five sections in this chapter. The first two take a look at the measurement of the output and input variables and the collection of data used in the study. Sections 3 deals with the tool for the study and model specification. The initial test run of the regression to examine the linearity and multicollinearity among the selected variables and to finally arrive at the variables statistically significant for use in the study is the subject of section 4. Finally, techniques considered appropriate for the estimation of the model parameters are examined in Section 5.

4.1 Measurement of Output and Input Variables

4.1.1 Output Variable:

The output variable to be forecast is the quantity of demand for Malaria treatment services at sector office Malaria clinics. There is no typical unit of demand for health care services. Most studies reviewed expressed quantity demanded in terms of the number of admissions, the number of days spent in a hospital, the number of patients' visits to a physician (or physicians), the expenditures on a particular kind of services and so forth.

In fact, all treatments can be divided into two major types - outpatient and inpatient departments. The features associated with these two major divisions are so different that models on the demand of their services are dealt with separately.

In Thailand the Malaria clinics do not operate inpatient services - all of them deal with only walk-in-outpatient services. The more severe and complicated malaria cases are referred to the district hospitals. As a policy, patients are treated only after testing positive on microscopic examination. The cost of services offered to consumers are borne from central government budget - the services are free to consumers at the service points. Other costs to the patients were not taken into consideration in the study.

Consequently, the area of interest in the present study was centered on the patients perceived by the malaria staff to be truly infected with malaria parasite after a microscopic examination. Thus, the number of patients tested positive microscopically was chosen as a unit of measurement of demand for treatment services in each Malaria

clinics. But planning purposes we will develop another model as well which can be used to forecast total demand for the services by both malaria patients and other patients who seek the services (both of which are measured as the total annual blood slides examined in each Malaria clinic).

Thus, the operational definition for the output variable is the quantity of demand for treatment services at sector Malaria clinic, Q, measured by the total number of positive cases of malaria in each Malaria clinic.

4.1.2 Input Variables

These were theoretically derived as follows in the preceding chapter.

1. Incidence rate of malaria
2. Government budget
3. Household income
4. Travel distance
5. Number of blood slides

The logic behind the selection of the these variable was also presented. Here, only the review of their operational definitions is given.

4.1.2.1 Operational Definitions

The operational definitions of the input variables are as follows.

$$\text{1. Incidence rate} = \frac{\text{Total annual positive cases microscopically diagnosed in selected district}}{\text{Total population in the district}} \times 1{,}000$$

This is assumed to be homogeneous for all the malaria clinics in each district.

2. Government expenditure on Malaria control activities:

$$\text{per capita government expenditure} = \frac{\text{Total annual government expenditure spent on Malaria control activities a district}}{\text{Total population in the district}}$$

This is also assumed to be homogeneous for all Malaria clinics in each district.

3. Household Income is defined as the annual average household income of the selected district and assumed to be the same for all population in a district.

$$\text{4. Average Travel Distance} = \frac{\text{Total travel distance to clinic}}{\text{Total patient visits to clinic}}$$

$$\text{Total travel distance to clinic} = \sum_{i=1}^{N} X_i * P_i$$

where

X_i = travel distance of patients' village/town i to Malaria clinic

N = number of villages/ towns of patients

P_i = number of patients from village/town i to Malaria clinic

5. Number of Slides is the total annual blood slides microscopically examined at the Malaria clinics.

4.2 Data Collection

As said previously, the sampling method used in the selection of the malaria clinics and districts was purposive - the only criteria being high incidence of malaria and to a less extent the availability of data. However, the general effects of confounding factors on the data to be collected were also taken into consideration. Therefore the malaria clinics were grouped into geographic districts based on the assumption that there exists a sense of district homogeneity. In this case many of the confounding factors were considered to be constant for a particular geographical district.

Due to the problem of communication gap and unfamiliarity of the chosen districts, the following procedures were adopted in order to achieve a reasonable level of quality control for the data collected:

1. A visit to Region 1 was made by the Researcher to acquaint himself with the real situations prevailing at the regional headquarters at Phrabuddabat, the selected districts and at the sector Malaria clinics during which a brief orientation about the research objective and methodology was given to the Malaria staff selected to collect the data.

2. All tables for data collection were translated into Thai and cross checked for clarification and correction by the Thesis advisor before the data were collected.
No attempt was made to follow up the sources of the data collected by the Researcher personally due to lack of time but the genuity of the data source was assured by the deputy Director of Malaria Region with whom the Researcher planned the data collection.

All the data collected were secondary. In all, 3 districts out of 8 in Tak Province were selected and consisted of 10 Sector Malaria clinics as previously stated. The frequent policy changes in connection with malaria control activities reflecting the prevailing malaria situation in Tak Province necessitated collection of secondary data from 1990 to 1993. The data came from three sources:

1. Malaria Zonal offices in Tak Province
2. Malaria sector offices at the district level and
3. District Administrative headquarters.

Data on Travel distance incidence rate, government expenses, slide examined and positive cases were obtained from statistical records at the Malaria Zonal and sector offices. The district administrative headquarters were the sources of the household income.

Unfortunately, some of the selected malaria clinics were established only a few years ago so data for some years were not found. However, the data found on these newly established clinics are added for the demand analysis.

Data on Myanmar patients who attended each of the clinics were also taken for analysis as their expenses at the clinics are also free of charge just like their Thai counterparts. But there were so many missing data that we decided to discard analyzing them. However, it is believed this would not affect the results of the study though the per capita government expenditures would be quite exaggerated as the Myanmar patients are also given free services.

4.3 Tool of Study and Model Specification

This study proposes to use quantitative method, multiple regression technique, to find the solutions, while the data are pooled cross-section time series of 10 sector Malaria clinics at Tak Province in the average of 3 years (1991-1993).

The data consists of incidence rate of malaria, per capita government expenditure on Malaria control activities, average household income, proportion of positive cases of patients on visit at the Malaria clinics, average travel distances to the Malaria clinics and number of blood slides examined - an indicator of the number of suspected patients on visit at the Malaria clinics. As stated before the study uses pooled cross-section and time series data, grouped in a natural way, by geographic districts. Our assumption is that although the relationship is structurally the same in each districts, the coefficients and residual variances may differ from district to district. The different variances constitute a case of heteroscedasticity that can be treated directly in the analysis.

The variable names and definitions appear in Table 4.1

Table 4.1 Description of Variables in Demand Analysis for Malaria clinic services

Variable	Description
Q_{ijt}	Quantity of demand for Malaria treatment services.
G_{ijt}	per capita government expenditure on Malaria Control activities at the district (in Bahts).
X_{ijt}	Average travel distance in km from village/town to Malaria clinic.
R_{ijt}	Incidence of malaria per 1000 population in a district.
I_{ijt}	Annual household income in a district (in Bahts).
S_{ijt}	Number of suspected patients (blood slides examined) in Malaria clinic.

There are alternative schemes by which cross-section and time-series data might be pooled. The first technique is to combine all cross-section and time-series data and perform ordinary least-squares regression on the entire data set. A second procedure involves the recognition that omitted variables may lead to changing cross-section and time-series intercepts. Covariance analysis involves the addition of dummy variables to the model to allow for these changing intercepts. A third polling technique improves the efficiency of the first least-squares estimation process by accounting for the existence of cross-section and time-series disturbances. This is called the error components pooling procedure - a variation of generalized least-squares estimation process. All of these techniques give unbiased and consistent parameter estimates. Since all the estimation techniques described will give unbiased and consistent parameter estimates, the central issue associated with pooling is one of efficiency. Error-components models are useful because they are estimated using a form of generalized least-squares regression and can be shown under reasonable conditions to be more efficient than the other two estimation processes (Pindyak and Rubinfeld, 1981).

Therefore in the present study, an error-components model is proposed and formulated as:

$$Q_{ijt} = \beta_0 + \beta_1 S_{ijt} + \beta_2 I_{ijt} + \beta_3 X_{ijt} + \beta_4 G_{ijt} + \beta_5 R + \epsilon_{ijt} \qquad (4.1)$$

for $i = 1, 2 \ldots\ldots N$
$t = 1,2 \ldots\ldots T$
$j = 1, 2, 3,$ and
$\epsilon_{ijt} = U_{ij} + V_{tj} + W_{itj}$

Where

N = the number of cross-section units of sector Malaria clinics within the districts

T = the number of time periods (in years)
j = represent a district at Tak Province
ϵ_{ijt} = the error term of the model

$U_{ij} \sim N(0, \sigma^2_{2u})$ = cross-section error component
$V_{tj} \sim N(0, \sigma^{2v})$ = time-series error component
$W_{tj} \sim N(0, \sigma^{2v}_{w})$ = combined error component
$\beta_0 \ldots\ldots\ldots \beta_5$ = the regression coefficients

We assume also that the individual error-components are uncorrelated with each other and are not autocorrelated (across both cross-section and time-series units) that is

$$\text{Cov}(U_i, V_t) = \text{Cov}(U_i, W_{it}) = \text{Cov}(V_t, W_{it}) = 0$$
$$\text{Cov}(U_i, U_j) = 0 \quad i = j_1$$
$$\text{Cov}(V_t, V_{t1}) = 0 \quad t = t^1$$
$$\text{Cov}(W_{it}, W_{jt}) = 0 \quad i = j_1$$
$$\text{Cov}(W_{it}, W_{it1} = 0 \quad t = t^1$$
$$\text{Cov}(W_{it}, W_{jt1} = 0 \quad i = J_1, \; t = t^1$$

However, since the average annual household income, I, and the per capita government expenditure variables are measured in Bahts in their nominal values rather than their real values, we will want to pay close attention to specifying the relationship in our proposed error-components model (4.1) in such a way that handles the decline in the purchasing power as this has enormous effect on demand for the malaria services. The use of logarithms is almost mandatory in this problem (Hall, 1984). We therefore transform the income and per capita expenditure variables I and G by the use of logarithms and refomulate the model in equation (4.1) as thus:

$$Q_{itj} = \beta_0 + \beta_1 S_{itj} + \beta_2 \text{Log } I_{itj} + \beta_3 \text{Log } G_{itj} + \beta_4 X_{itj} + \beta_5 R + \epsilon_{ijt} \qquad (4.2)$$

This proposed error-components model will be compared with a pooled model using ordinary least squares which takes into account the districts differences. The data could be analyzed using dummy variables as done in covariance models to look for special effects associated with the districts or to formulate tests for the equality of regressions across districts. However, our objective here is to develop one relationship that can serve as the best representation for all districts and sector malaria clinics in Tak Province, Thailand. This goal is achieved by taking district differences into account through either an application of weighted ordinary least-squares regression to the pooled data or an error-components pooling procedure. Consequently, the application of covariance model technique to the pooled data is not considered in this study only the extension of the method of weighted least squares and error-components pooling procedure are compared in order to justify and confirm the selection of the technique proposed in equation (4.2).

The data is presented in Table 4.2. Malaria clinics are grouped into geographic districts based on the assumption that there

exists a sense of district homogeneity as stated before. The three broad geographic districts , (1) Tasongyang (2) Phop phra and (3) Mae Ramard are used to define the groups. The malaria clinics, MC, are represented by number 1-10.

Table 4.2 Demand analysis for Malaria Services at Sector Malaria Clinics in Tak Province.

Obs.	M.C.	Dist.	Time	Q_{ijt} * 100	G_{ijt}	X_{ijt}	R_{ijt} /1000	I_{Ijt} *1000	S_{ijt} *1000
1	1	1	1991	0.06	70.80	6.6	232	4.00	7.41
2	1	1	1992	11.39	72.94	6.6	309	4.00	64.42
3	1	1	1993	7.79	85.85	6.6	291	4.00	60.39
4	2	1	1991	2.53	70.80	5.2	232	4.00	25.96
5	2	1	1992	2.27	72.94	5.2	309	4.00	23.68
6	2	1	1993	2.00	85.85	5.2	291	4.00	21.39
7	3	1	1991	7.26	70.80	5.3	232	4.00	40.22
8	3	1	1992	6.47	72.94	5.3	309	4.00	44.35
9	3	1	1993	4.86	85.85	5.3	291	4.00	44.31
10	4	2	1992	39.31	28.50	4.5	193	17.50	145.05
11	4	2	1993	23.86	26.10	4.5	152	17.85	87.07
12	4	2	1990	20.32	46.20	4.5	131	18.05	82.23
13	4	2	1991	11.39	48.08	4.5	82	21.50	65.93
14	5	2	1992	16.62	28.50	5.3	193	17.50	81.23
15	5	2	1993	50.78	26.10	5.3	152	17.85	172.46
16	5	2	1990	46.42	46.20	5.3	131	18.05	150.50
17	5	2	1991	35.60	48.08	5.3	82	21.50	143.90
18	6	2	1992	3.30	46.20	3.3	131	18.05	30.90
19	6	2	1993	8.80	48.08	3.3	82	21.50	49.51
20	7	3	1989	15.36	46.84	4.9	208	9.80	65.50
21	7	3	1990	23.07	87.74	4.9	261	10.30	82.67
22	7	3	1991	11.28	71.34	4.9	157	11.00	58.34
23	7	3	1992	7.91	74.12	4.9	152	12.00	42.49
24	7	3	1993	6.09	78.79	4.9	106	12.50	34.67
25	8	3	1991	0.04	71.34	3.7	157	11.00	1.49
26	8	3	1992	0.34	74.12	3.7	152	12.00	11.61
27	8	3	1993	0.23	78.79	3.7	106	12.50	8.49
28	9	3	1991	0.86	71.34	2.4	157	11.00	8.64
29	9	3	1992	1.28	74.12	2.4	152	12.00	11.57
30	9	3	1993	1.03	78.79	2.4	106	12.50	9.33
31	10	3	1993	0.27	78.79	1.7	106	12.50	7.62

Districts
1 = Tasongyang District; 2 = Phop Phar district;
3 = Mae Ramard District

Clinics
1 = Mae tarn clinic; 2 = Mae Laryang clinic;
3 = Mae usu clinic 4 = Phop Phra clinic;
5 = Sawoh clinic; 6 = Chongkab clinic
7 = Mae Ramard clinic; 8 = Banpae clinic;
9 = Bankanejeo clinic 10 = Mongwa clinic.

4.4 Initial Test Run of Multiple Regression

The purpose in this section is to check the linearity of all the data on the independent and dependent variables and to detect the presence of multicollinearity among the variables.

The first assumption in regression analysis is that a linear relationship exists. This assumption states that the dependent variable is linearly related to each of the independent variables. When the assumption of linearity is not met, the usual way of achieving linearity is to transform the variables into new variables that do exhibit linear relationship with Q, the independent variable. To achieve this, the relationships between the dependent variable Q and each of the independent variable S, I, X, R, G are graphed to determine whether the linearity assumption has been met. An individual graph for each pair of variables helps to identify any nonlinearities.

Multicollinearity can develop when two or more of the independent variables are highly correlated. If multicollinearity exists, the result is extremely large numbers that cannot be handled by the computer. The regression coefficient and all other output from the computer may, therefore, be erroneous. The problem is very difficult to defect.

In this study, we choose to use the rule of the thumb proposed by Makridakis et. al. 1989. This consists of studying the relationship with a simple correlation coefficient. The rule of the thumb is that a simple correlation coefficient exceeding +0.7 or less than -0.7 indicates the possibility of multicollinearity if the variables concerned are used. This then provides a clue of the problem of multicollinearity which can be investigated later on.

Multicollinearity is not a problem of misspecification. Therefore, the empirical investigation of problems that result from a collinear data set should begin only after the model has been satisfactorily specified. However, there may be some indications of multicollinearity that are encountered during the process of adding, deleting and transforming variables or data points in search of the good model. Indication of multicollinearity that appear as instability in the estimated coefficients are as follows:

1. large changes in the estimate coefficients when a variable is added or deleted,
2. large changes in the coefficients when a data point is altered or dropped.

Once the residual plots indicate that the model has been satisfactorily specified, multicollinearity may be present if

3. The algebraic signs of the estimated coefficients do not confirm to prior expectations or
4. Coefficients of variables that are expected to be important

have large standard errors.

A statistically significant variables that will not pose a problem of multicollinearity in the subsequent analyses are selected in this initial run.

4.5 Estimation of Model Parameters

As stated above, to develop one relationship that can serve as the best representation for all the sector malaria clinics in all the districts of Tak Province the district differences are taken into account through an extension of the method of weighted least squares and an error-components model technique.

It is assumed that there is a unique residual variance associated with each of the three districts. The variances are denoted as $(C_1\sigma)^2$, $(C_2\sigma)^2$ and $(C_3\sigma)^2$, where σ is the common part and the C_i's are unique to the districts. These differences in variances can be treated either by assuming that the error term of the model (4.2) consists of a single combined disturbance and that the corrective procedure is to apply weighted ordinary least-squares regression to the pooled data or by using an error-components pooling procedure which takes into account the existence of cross-section and time-series disturbances as well as proposed in the model (4.2). Each of these methods involve two-stage estimation procedures as described in the next sections.

4.5.1 Estimation of Weighted Ordinary Least-Squares Regression

According to the principle of weighted least squares, the regression coefficients should be determined by minimizing the weighted sum of squared residuals,

$$S_w = S_1 + S_2 + S_3$$

Where

$$S_j = \sum_{i=1}^{N_j} \frac{1}{C^2} (Q - \beta_0 - \beta_1 S - \beta_2 \log I - \beta_3 \log G - \beta_4 X - \beta_5 R)^2$$

$$j = 1, 2, 3$$

Each of S_1 through S_3 corresponds to a district and the sum is taken over only those sector malaria clinics that are in the district. The factors $\{1/C^2_j\}$ are the weights that determine how much influence residual has in estimating the regression coefficients. The weighting scheme is intuitively justified by arguing that observations that are most erratic (large residual variance) should have little influence in determining the coefficients.

The weighted least squares estimates can also be justified by a second argument. The objective is to transform the data so that the

parameters of the model are unaffected, but the residual variance in the transformed model is constant. The prescribed transformation is to divide each observation by the appropriate C_j resulting in a regression of

$$Q = \beta_0 \frac{1}{C_j} + \beta_1 \frac{S}{C_j} + \beta^2 \frac{\log I}{C_j} + \beta_3 \frac{\log G}{C_j} + \beta_4 \frac{X}{C_j} + \beta_5 \frac{R}{C_j} + \mu \quad (4.3)$$

where the variance of μ is σ^2.

Then, the residual term, in concept, is also divided by C_j, the resulting residuals have a common variance, σ^2, and the estimated coefficients have all the standard least squares properties.

The values of the C_j's are unknown and must be estimated in the same sense that σ^2 and the β's must be estimated. We proposea two-stage estimation procedure. In the first stage, ordinary least squares is run on the entire pooled sample. The ordinary least-squares regression residuals are grouped by districts to compute an estimate of district variance as follows: for Tasongyang district (1) we compute

$$S_1 2 = \sum e^2_i /n \quad (4.4)$$

Where the sum is taken over the n residuals corresponding to n individual observations in Tasongyang district. S_2^2 and S_3^2 for Phop Phra and Mae Ramard districts respectively are computed in a similar manner. S^2 is obtained as weighted average of S_1^2, S_2^2 and S_3^2 as thus:

$$S^2 = \frac{S_1^2 + S_2^2 + S_3^2}{3} \quad (4.5)$$

Then C_j is estimated as $\left(\frac{S_j^2}{S^2}\right)^{\frac{1}{2}}$

The C_j's thus estimated are used in the second stageto derive the transformed equation (4.3) as shown above. Thus the equation (4.3)

satisfy the necessary assumption of constant variance. Regression of Q/C_j against the six new variables consisting of $1/C_j$ and the five transformed explanatory variables, S/C_j ,log I/C_j ,log $G(-1)/C_j$, X/C_j and R/C_j using OLS produces the desired estimates of the regression coefficients and their standard errors. The regression with the transformed variables are carried out with the constant term constrained to be zero. That is, β_0, the intercept of the original model is now the coefficient of $1/C_j$. The equation (4.3) has no intercept. The OLS estimates are obtained using Micro TSP computer programme.

4.5.2 Estimates of the Error-Components Model

The estimation of the error-components model is a generalization of the weighted least-squares technique because it weights observations in inverse relationship to their variances. The only difference is in the error term which in this case is made up of three component parts; the cross-section error component, σ^2_u time-series error component, σ^2_v and the combined error component σ^2_w i.e.

$$Var(\epsilon_{it}) = \sigma^2_u + \sigma^2_v + \sigma^2_w \qquad (4.6)$$

4.5.2.1 Estimate of the Variance Components

If the ϵ_{it} were observable, the best quadratic unbiased estimators of the variance components (maximum likelihood if normality is assumed) are, from Graybill (1961),

$$\delta_w^2 = \frac{1}{(N-1)(T-1)} \sum_i \sum_r \left(\varepsilon_{it} - \frac{1}{T}\varepsilon_i + \frac{1}{NT}\varepsilon_{..} \right)^2, \qquad (4.7)$$

$$\hat{\sigma}_u^2 = \frac{1}{T}\left[\sum_{i=1}^{N} \frac{\left(\frac{\hat{\varepsilon}_i}{T} - \frac{\hat{\varepsilon}_{..}}{NT}\right)^2}{(N-1)} - \hat{\sigma}_w\right], \tag{4.8}$$

$$\hat{\sigma}_v^2 = \frac{1}{N}\left[\sum_{i=1} \frac{\left(\frac{\hat{\varepsilon}_t}{N} - \frac{\hat{\varepsilon}_{..}}{NT}\right)^2}{(T-1)} - \hat{\sigma}_w\right], \tag{4.9}$$

However, the ϵ_{it} are not observable. Therefore we turn to estimates of ϵ_{it}, say $\hat{\epsilon}_{it}$, which are observed residuals obtained by least squares, applied directly to equation (4.2). Thus

$$S_{jw}^2 = \frac{1}{(N-1)(T-1)} \sum_i \sum_j \left(\hat{\varepsilon}_{it} - \frac{1}{T}\hat{\varepsilon}_t - \frac{1}{N}\hat{\varepsilon}_t\right)^2, \tag{4.10}$$

$$S_{ju}^2 = \frac{1}{T}\left[\sum_{i=1} \frac{\hat{\varepsilon}_i^2}{T(N-1)} - S_{jw}^2\right], \tag{4.11}$$

$$S_v^2 = \frac{1}{N}\left[\sum_{t=1}^{T}\frac{\bar{\varepsilon}_t^2}{N(T-1)} - S_w^2\right]. \qquad (4.12)$$

Where $S_2 2$, $S_u 2$ and $S_v 2$ are estimates of $\hat{\sigma}^2_w$, $\hat{\sigma}^2_u$ and $\hat{\sigma}^2_v$ respectively.

In the present study, the residual variance S^2_j are calculated for each of the districts as done in section 4.5.2.1. Thus for Tasongyang districts (1) we compute $S_1^2 = S_{1w}^2 + S_{1u}^2 + S_{1v}^2$. We do the same for S_2^2 and S_3^2 and the rest of the calculations are the same as for the method described in section 4.5.2.1. Thus in the first stage, ordinary least squares is run in the entire pooled sample. The ordinary least-squares regression residuals are then used to calculate sample estimates of the variance components for each district from which the C_j 's are computed for the weighting in equation (4.3). The estimated results in these two techniques are then compared.