CHAPTER 4

RESEARCH METHODS

4.1 Study Design

This study includes two sections:

- Empirical study.
- Methodological study.

The supply of doctors for health manpower and for the malaria control programme for the next six years will be forecasted by using a time series model with the data from the past 15 years. The relationship between the allocation of health manpower and the effectiveness of the malaria control programme will be analysed by using multiple regression with cross sectional data.

From the resuls of the study, some incentive policy options will be suggested. In order to make the possibility of applying those incentive policy options more practical, a survey will be carried out and primary data collected will be analyzed. However, because of time limitation, the study only devised a survey questionnaire and a method of analysis.

4.2 Population and Sample Size

The North Mountain region in the country has been chosen to carry out this study on the basis of having a chronic problem of shortage of medical doctors. The authorities in that region will have to produce assistant medical doctors to provide malaria care services at primary level.

In the North Mountain region, malaria morbidity rate is high, living conditions are poor in comparison with other regions. It is difficult to dispatch medical doctors to this region because they do not like to work there. That is why this region is chosen for the study.

Among 113 districts of the region, data from 40 districts were selected. Those districts are in a hyper-endemic area, the mortality and morbidity rates are high.

The criteria for the selection of the districts are:

- mortality rate higher than 0.2 per 10,000 population;

- morbidity rate higher than 200 per 10,000 population.

4.3 Data Collection

Time series data of the number of doctors supplied for the past 15 years (from 1980 to 1994) will be used to forecast the supply of doctors for the next six years; and cross sectional data will be used for regression analysis.

The number of doctors supplied from the whole country for the past 15 years will be used to forecast the supply of health manpower and health manpower for malaria control programme.

Mortality, morbidity rates, number of population protected and number of patients' visits per 10,000 population from district level (40 districts from 1985 to 1990) will be used to analyze the relationship between the distribution of health manpower and the effectiveness of the malaria control programme.

The number of doctors supplied, mortality, morbidity rates, number of patients protected, number of patients' visits for this study are secondary data. This is selected from the Statistics YearBooks of the Ministry of Health of Vietnam from 1980 to 1994 provided by the headquarters of the national malaria control programme.

4.4 Methods/ Tools to be Used

4.4.1 Time series model for forecasting the number of doctors for health manpower and for the malaria control programme

A *Time series Model* will be used to forecast the supply of health manpower for malaria control programme and the supply of total health manpower.

The function is: $Y = f(t)$

Where:

Y = the forecasted of number of doctors will be supplied to health manpower stock and to malaria control programme for the next six years.

t = time, and it can be measured as a continuous variable.

The period of six years is chosen to forecast because of two reasons:

- The data available is not enough for forecasting A longer period of time (the time series data used in the study are fifteen past years).

- The planning period of the government is from 1995 to the year 2000.

4.4.2 Regression function for regression analysis

The regression function to be used is: $Y = f(x)$

Where:

Y = the effectiveness of malaria control programme

X = the distribution of qualified health manpower

Multiple regression will be used to analyze the data and the relationship between number of patients treated per 10,000 population at public health facilities in the region and the number of medical doctors, number of assistant medical doctors, number of nurses per 10,000 population.

The regression equations are

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

Where:

Y = the effectiveness of malaria control programme can be measured by some indicators: mortality, morbidity, number of population protected and number of patients' visits per 10,000 population.

α = Constant

X_1 = Number of medical doctors at those health facilities per 10,000 population.

X_2 = Number of assistant medical doctors per 10,000 population.

X_3 = Number of nurses per 10,000 population.

ε = Error

4.4.3 Logit function for assessing the probability of applying some incentive policy options

1 Logit function

Some incentive policy options will be suggested by this study (see chapter V). In order to make the possibility of applying those incentive policy options more practical, a survey will be carried out and primary data collected will be analyzed, using logit function.

Logit Model: Logit analysis presents a complement to multiple regression in its ability to utilize a binary dependent variable.

Logit analysis does not predict only whether an event will occur or not (one or zero) but also, the probability of an event to happen. In this way, the dependent variable can be any value between zero and one. This also means that the predicted value must be bounded to fall within the range of zero and one.

The logit model based on the cumulative logistic probability function is given by the following equation:

$$P_i = F(Y_i) = \frac{1}{1 + e^{-Y_i}} = \frac{1}{1 + e^{(\beta o + \beta 1X1 + \beta 2X2 + \ldots + \beta nXn)}} \qquad (1)$$

Where P_i is the probability that an individual will make a certain choice, given knowledge of X_i.

The equation (1) can be rewritten as follows:

$$\text{Ln} \frac{P_i}{1 - P_i} = \beta o + \beta_1 X_1 + \beta_2 X_2 + \beta_n X_n \qquad (2)$$

2 Survey Implemention

i) Sample size of the survey

The target population for the survey is the total number of doctors for the malaria control programme and the number of doctors for total health manpower.

The sample size for the survey is N, calculated as follows:

$$N = \frac{Z^2_\alpha * p * q}{d^2}$$

where:

- N is the desired sample size (when population > 10,000);
- p is the variability of characteristic to be measured in the population; here, it is the maximum expected proportion of health staff who want to continue to work for the malaria control programme.
- q is the pessimatic value, and is obtain by 1 minus p;
- d is the maximum margin of error tolereted or degree of accuracy required;
- Z_α is the degree of confidence (odds ratio) which is required to be within the special range ± d. A value of Z = 2 or Z = 1.96 is often chosen since this provides a degree of confidence equal to odds of 19 to 1. This is technically termed a 95 % confidence interval.

With:

- p = 0.5 (50 %); that value is chosen because there is no reasonable estimate; so to maximize the sample size, we chose 50 % ;
- q = 1 - p = 0.5; as there is no reasonable estimate of p, we used 50 % , then q = 50 %. This maximizes the expected variance (pq = 0.25) and ensures the sample size to be large enough for the purpose;
- d = ± 5 % = 0.05;
- Z_α = 1.96;

N = 384 doctors of total health manpower.

ii) Sampling technique

From seven regions of the country, some regions where malaria is endemic, poor living conditions and doctors/ population ratio low will be chosen for the survey by a purposive (non probability) sampling. The reason is that in those regions, malaria control programme staff need to be encouraged by some incentive policy of the government. Those are regions, where it is very difficult to allocate malaria control programme staff to work for the programme.

iii) Survey questionnaire concerning the possibility of applying the incentive policy for reallocation of health manpower for the malaria control programme.

The survey questionnaire is given in Appendix 1.

3 Data analysis by logit model

After the survey, the collected data will be analyzed to appreciate the probability of applying those incentive policy options by using logit function.

i) Data process

Data processing was shown in Appendix 2: an example of the summary table form of survey result for running logit model).

ii) Data analysis

The survey data will be used to run the logit model to find the regression Y_i related to the probability of malaria control programme's staff, who are willing to continue to work for the malaria control programme with the given incentive policy options.

Using a software programme to evaluate the regression, the equation is as follows

$$Y_i = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_{11} X_{11} \qquad (3)$$

Where:

Y_i is the willingness of the malaria control programme' staff or health staff to continue to work for the malaria control programme or to go to work for the malaria control programme

β_o is constant

$\beta_1, \beta_2, \ldots \beta_{11}$ are the coefficients of independent variables $X_1, X_2, \ldots X_{11}$ respectively

X_1 is age of the malaria control programme staff or other health staff. It is continuous variable.

X_2 is policy option 1, the answering is coded into two groups (with numbering in the parentheses):

Yes (1)
No (0)

X_3 is policy option 2, the answering is coded as following:

Yes (1)
No (0)

X_4 is policy option 3, the answering is coded:

Yes (1)
No (0)

X_5 is sex of the malaria control programme staff or other health staff and it is coded:

- Male (1)
- Female (0)

X_6 is marital status of the malaria control programme staff or other health staff and it is coded:

- Single (1)
- Others (0)

X_7 is number of children of the malaria control programme staff or other health staff and it is coded into two groups. According to the family planning of the country, each family should have not more than two children so, if people have got more than two children, thre will be more difficulties for them and they might want to change their job, that is why they are grouped into two groups:

- The number of children is equal or less than two will be coded (1)
- The number of children is more than two will be coded as (0)

X_8 : number of dependent people of the malaria control programme staff or other health staff is coded into two groups. In general, often there will be three generations staying together in a family: grandfather and grandmother, parents and two children. In that case, grandfather, grandmother and children are dependents. If the number of dependents of malaria control programme staff is high, the malaria control programme staff will have more difficulties in life. This might influence their job, they may be more likely to change their job. This is also a quantitative variable but here it is coded into two groups because the purpose of the analysis is also to test the different probability of willing to continue the job between two groups of people (one has equal or less than 4 dependent people and the other has more than 4). Suppose that group (0) want to change their job, there should be some more incentive policy for that group.

- The number of dependent people is equal or less than four (1)
- The number of dependent people is more than four (0)

X_9 : having extra work of the malaria control programme staff or other health staff and it is coded:

Yes (1)
No (0)

X_{10}: income of the malaria control programme staff or other health staff, is coded into two groups as follows because the average salary of the doctor is 600,000 VN dong so, if the income of the malaria control programme staff or other health staff is $\cong$ 600,000 VN dong, it is average income. If it is more than 600,000 VN dong, it is over average income. This is also a quantitative variable but here it is coded into two groups because the purpose of the analysis is also to test the different probability of willingness to continue the job between two groups of people (one has average income and the other has higher average income). Suppose that group (1) want to change their job, there should be some more incentive policy for that group.

- The income per month is less than 600,000 VN dong (1)

The income per month is more than 600,000 VN dong (0)

X_{11}: duration of working for the malaria control programme (for the malaria control programme staff) or for health sector (for others). It is a continuous variable.

From equation (3);

$$\text{Ln} \frac{P_i}{1 - P_i} = Y_i = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n \quad (4)$$

From equation (4);

$$\frac{P_i}{1 - P_i} = \text{anti-Log}_e \ (Y_i) \quad (5)$$

Where P_i is probability of applying the incentive policy option.

After estimating the value of Y_i, anti-Log_e for two sides of equation (4), we can rewrite the equation (4) to equation (5).

iii) Testing the statistical significance of the slope coefficient

Logit analysis can be tested by using t-test. Null and alternative hypotheses are:

- Ho : $\beta_i = 0$, there is no relationship between independent variables and dependent variable.
- Ha : $\beta_i \neq 0$, there is at least a relationship between independent variables and dependent variable.

Then compare the t-statistic computed and the critical value of t-statistic in the statistical tables. With the larger t-statistic, the null hypothesis (Ho: $\beta_i = 0$) will be rejected and we can therefore conclude that there is a relationship between independent variables and dependent variable.

4.5 Hypothesis Testing

Some statistic tests are needed to test the significance of statistic of the study:

4.5.1 Hypothesis testing for forecasting model

i) Test for Auto-correlation

The more widely used method is the Durbin-Watson test. The Durbin-Watson statistic d always lies between 0 and 4. Values close to 4 indicate negative auto-correlation, values close to 0 indicate positive auto-correlation.

Durbin-Watson have provided lower (dL) and upper (dU) bounds such that a more exact test for auto-correlation can be made after the d value has been computed. The decision rules are as follows:

- If d < dL or d > (4 - dL), reject the null hypothesis Ho: p = 0; conclude that there is auto-correlation.
- If dU < d < 4 - dU, do not reject Ho: p = 0; and conclude that there is no auto-correlation.
- If d falls between dL and dU or between 4 - dU and 4 - dL, the test is inconclusive and usually more data is needed.

ii) Testing the statistical significance of the slope coefficient

One method for evaluating the regression equation is to test the slope of the regression line. If the slope is statistically equal to zero, the time-series data does not contain a linear trend. The statistical test used for testing whether the slope equals 0 is the t-test. Null and alternative hypotheses that can be used are:

Ho: $\beta = 0$, the slope is equal to zero.

Ha: $\beta \neq 0$, the slope is statistical significally difference from zero.

Then compare the t-statistic computed and critical value of t-statistic in the statistical tables. With the larger t-statistic, the null hypothesis (Ho: $\beta = 0$) will be rejected and we can therefore conclude that there is a statistically meaningful linear relationship between Yt and t.

iii) Testing the goodness of fit of a regression model (testing the overall linear relationship)

Null and alternative hypotheses used are:

Ho: there is no linear relationship between Y_t and time.

Ha: there is a linear relationship between Y_t and time.

Critical table value:, F with k - 1 and n - k degrees of freedom ($\alpha = 0.05$).

Where:

k = the number of estimated coefficient plus the intercept estimated in the regression.

n = the number of time periods.

Then compare the F-statistic computed and critical value of F-statistic in the statistical tables. With the larger F-statistic, the null hypothesis will be rejected and we can therefore conclude that there is a linear relationship between Yt and t.

iv) The Coefficient of Determination

The coefficient of determination is a measure of the goodness of fit of the regression equation to the data. The better the fit of the line, the closer R^2 will be to 1. In other words, if the regression line provides a perfect fit, the variance (change) in the data (Y_t) is completely explained, and R^2 exactly equal to 1. If R^2 is equal to 0.750, we can say that 75 % of the variance in Y_t is explained by the model.

4.5.2 Hypothesis testing for multiple regression

i) Testing the statistical significance of the slope coefficient

Hypothesis:

- Ho: $\beta_1 = \beta_2 = 0$, there is no relationship between independent variables and dependent variable.

- Ha: At least one of $\beta i \neq 0$, there is at least a relationship between independent variables and dependent variable.

Then compare the t-statistic computed and critical value of t-statistic in the statistical tables. With the larger t-statistic, the null hypothesis will be rejected and we can therefore conclude that there is a relationship between independent variables and dependent variable.

ii) Testing the overall linear relationship

Null and alternative hypotheses are:

- Ho: there is no linear relationship between dependent variable and independent variable.
- Ha: there is a linear relationship between dependent variable and independent variable.

Critical table value:, F with k - 1 and n - k degrees of freedom ($\alpha = 0.05$).

Where:

k = the number of estimated coefficient plus the intercept estimated in the regression.

n = the number of observation.

Then compare the F-statistic computed and critical value of F-statistic in the statistical tables. With the larger F-statistic, the null hypothesis (Ho) will be rejected and we can therefore conclude that there is a linear relationship between dependent variable Y and multiple variables X.

iii) The Coefficient of Determination

The coefficient of determination is a measure of the goodness of fit of the regression equation to the data. The better the fit of the line, the closer R^2 will be to 1. In other words, if the regression line provides a perfect fit, the variance (change) in the data (Y) is completely explained, and R^2 exactly equal to 1. If R^2 is equal to 0.850, we can say that 85 % of the variance in Y is explained by the multiple variables X.

4.5.3 Hypothesis testing for linear regression

i) Testing the statistical significance of the slope coefficient

One method for evaluating the regression equation is to test the slope of the regression line. If the slope is statistically equal to zero, the time-series data does not contain a linear trend. The statistical test used for testing whether the slope equals to zero is the t-test.

The alternative hypothesis that can be used are:

Ho: $\beta = 0$, the slope is equal to zero.

Ha: $\beta \neq 0$, the slope is statistical significally difference from zero.

Then compare the t-statistic computed and critical value of t-statistic in the statistical tables. With the larger t-statistic, the null hypothesis will be rejected and we can therefore conclude that there is a statistically meaningful linear relationship between dependent variable and independent variable.

ii) Testing the overall linear relationship

- Null and alternative hypotheses that can be used are:

Ho: there is no linear relationship between dependent variable and independent variable.

Ha: there is a linear relationship between dependent variable and independent variable.

Critical table value:, F with k - 1 and n - k degrees of freedom ($\alpha = 0.05$).

Where:

k = the number of estimated coefficient plus the intercept estimated in the regression.

n = the number of observations.

Then compare the F-statistic computed and critical value of F-statistic in the statistical tables. With the larger F-statistic, the null hypothesis will be rejected and we can therefore conclude that there is a linear relationship between dependent variable and independent variable.

iii) The Coefficient of Determination

The coefficient of determination is a measure of the goodness of fit of the regression equation to the data. The better the fit of the line, the closer R^2 will be to 1. In other words, if the regression line provides a perfect fit, the variance (change) in the data (Y) is completely explained, and R^2 exactly equal to 1. If R^2 is equal to 0.950, we can say that 95 % of the variance in Y is explained by the variable X.

4.6. *Table 4.2* **Outcome Measurement**

Specific Objective	*Variables*	*Measurement Method*	*Sources o[f] data*
1. To estimate the supply of total health manpower and health manpower for M.C.P. for the next six years in order to analyze the different trends between them.	1)Estimate supply of health manpower. 2)Times	- Time series model	- Health record
2. To identify the existing distribution situation of health manpower on malaria control programme in the North mountain region of Vietnam.	1)No. of Medical doctors. 2)No. of Assistant Medical doctors. 3)No. of Nurses.	-Counting	-Health records. -MOH statistics
3. To identify the difference between the cost for doctors and assistant medical doctors in the malaria control programme	1)Cost for supply doctors. 2) Cost for supply assistant medical doctors	- Counting and calculate	-Health records. -MOH statistics
4. To analyze how the effectiveness of malaria control programme in this region can be affected by the distribution of qualified Health manpower.	1) No. of M.D 2) No. of A.M.D 3) No. of Nurses. 4) No. of malaria patients were treated 5) No. of population were protected. 6) Mortality rate. 7) Morbidity rate.	- Modelling through two steps: Use "Multiple regression with discrete explanatory variables model" to analyze the relationship between the distribution of qualified health manpower and No. of Patients care; No. of population protected. -Analyzing	-Health record - M O H statistics
5. To provide some policy options of the distribution and reallocation of qualified H.M. to improve the effectiveness of MCP.	1) The effectiveness of M.C.P. 2) The distribution of qualified Health Manpower.		- From the result of above steps.