CHAPTER 3
DATA AND METHODOLOGY

3.1 Introduction

There are varieties of studies on health and economic growth which measure 'health' differently. Macroeconomic studies have used a variety of health indicators and come out with different results. There is a difficulty in causality, i.e., whether income affects health or health affects income. Making the macroeconomic relationship between the two is difficult because health effects in the macro economy may have long time lags. There are issues with data in terms of access to health facilities, health care services, health outcome indicators, etc. Therefore, one has to use the data on economic growth and health which are available for different time periods and regions. This study tries to analyse the health and economic growth at macro level in India.

3.2 Data

There are difficulties in obtaining data on health for a long period of time for developing countries as the health sector is under-developed. In this study too, we did not find some indicators for the states of India before 1992. For the study, the data on health are obtained from different sources like National Family Health Survey (NFHS- 1, 2 and 3), Sample Registration Survey (SRS), Coverage Evaluation Survey (CES) and Census of India. The macroeconomic indicators are collected from Reserve Bank of India for four time periods 1992-93, 1998-99, 2005-06 and 2010-11.

The data on NSDP and PCNSDP (Per Capita Net State Domestic Product) have been obtained from *Handbook of Statistics on the Indian Economy*, Reserve Bank of India for the same period. Since the data were available on different base periods, all they have been transformed to 2004-05 basis by splicing method.

The data on female literacy rate are taken from the Economic Survey and they are based on the Census of India. Further, the data for IMR, UFMR and immunization are

taken from the National Family Health Survey Reports available for the periods 1992-93, 1998-99 and 2005-06. The data for IMR and UFMR for 2011 are taken from Sample Registration Survey. Data on immunization for 2009 are taken from Coverage Evaluation Survey, 2009, as they are not available for 2010-11.

The analyses of the time series data are conducted by using a composite database on health and economic growth collected from multiple sources in the time series framework for the period 1970-2010. The data on IMR and TFR are obtained from the Bulletin of Sample Registration System 2010. The variable gross domestic product at factor cost is taken at constant prices for the 2004-05 from the Handbook of Statistics of Indian Economy of Reserve Bank of India, 2011-12 and the values are in billion rupees. The data on health variable, Life Expectancy, as well as those for comparison with other countries are obtained from the World Development Index (WDI) of the World Bank.

3.3 Methodology

We adopted different methods and approaches for analysing the data in Chapters 4, 5, 6 and 7. For Chapter 5 on regional health inequality, the latter is measured to find the trend in disparities in health. For Chapter 6 a time series approach is used to understand the short run and long run relationships between growth and health. This chapter also addresses the causality of two indicators under the study. For Chapter 7, panel data approach is used to understand the impact of economic growth on health.

Regional Health Disparity:

Regional disparities cause significant impact on the standard of living and welfare in a society. As the gap between the rich and poor increases, the poor section of the society becomes worse off as it cannot afford the quality of life. We have used different measures to show the health inequality in India.

Gini Coefficient

Gini coefficient is commonly used as a measure of inequality of income. In Chapter 5 it is used to show the forms of inequality of health. It is mainly associated with

the descriptive approach to the measurement of equality. It measures inequality among values of a frequency distribution, for instance, income. It ranges mainly between 0 and 1. The value 0 expresses complete equality where all the values are identical. For example, we can say everyone in the society has the same level of income. On the other hand, the value 1 expresses complete inequality. We can say that all the income is in the hands of only the richer section and there is no income with the poorer section. A value greater than one may occur if some persons have negative income or wealth. For larger groups, values close to or above 1 are very unlikely in practice. Therefore, low Gini coefficient indicates a more equal distribution, with 0 (zero) corresponding to complete equality, while higher Gini coefficients indicate more unequal distribution, with 1 corresponding to complete inequality. We have used covariance method to calculate the Gini coefficient to measure inequality of income, female literacy rate and health in terms of IMR, UFMR and immunization.

Concentration Index

The other measure of inequality is the concentration index. It is based on the concentration curve which is used to found out if there is any socio-economic inequality in some health variable and whether it is prominent in one point of time or another, or one state than another. The concentration index is directly related with the concentration curve and it quantifies the degree of socio-economic inequality in a health variable. The important point to note is that it ranks individuals, states or countries on the basis of socio-economic features and then measures the health inequality. In doing so, it ranges between -1 and 1. We divide the sections based on the income levels, for instance, the value -1 shows that the poorer section is completely worse off as the concentration of ill-health is among them. The value 1 shows that the concentration of health variable is among the richer section. The value 0 shows that the health variable is concentrated equally between the rich and poor. The concentration index is positive when the concentration curve lies below the diagonal but negative when it lies above the diagonal. We can define the concentration index as cumulative proportions of health and so is insensitive to changes in the mean level of health. It can be generated by graphing the cumulative percentage of the population (along the X-axis) against the cumulative amount of health (along the Y-axis).

Formula for concentration index:

$$C = \frac{2}{\mu} \text{cov}(h,r)$$

Where h is the health variable, μ its mean and r the fractional rank by income.

Unconditional and Conditional Convergence Models

Apart from measuring inequalities by Gini coefficient and concentration index, we have employed the unconditional and conditional convergence models. They are mainly used to understand the steady path of per capita income and growth rates.

In Chapter 5 these models have been used to identify the steady paths of health and their rates of growth/decline. This is to understand whether over time the inequalities of health are decreasing across India? For this, we analysed unconditional beta convergence, conditional beta convergence and sigma convergence.

The early works on equilibrium theory predict that for the same amount of savings and investment, the growth of per capita income should be higher in poor regions than rich regions because of the assumption of diminishing returns of capital. This is the basic idea of unconditional beta convergence. However, it was questioned in the new growth theory in 1980s. The assumption of diminishing returns to capital was criticized as there are forces in economic system in the form of research and development that determine the productivity of capital. There was another criticism by Myrdal[10] as he believed that the orthodox theory of growth is static and ignores the dynamic consequences of factor migration and trade. Opposed to this, there is the theory of beta conditional convergence which says that we can determine the growth path by holding the constant variables which affect the growth of income other than the initial level of income.

[10]Karl Gunnar Myrdal was a Swedish economist, sociologist and politician. In 1974 he received the Nobel Memorial Prize in Economic Sciences.

Sigma Convergence

The sigma convergence is based on the standard deviation of coefficient of variation to see whether the dispersion of the levels of income has declined or not.

Using the above models of economic growth, we have tried to analyse the health disparity in different regions across India and see if it has reduced over time.

Time Series Study: Relationship between Growth and Health:

In Chapter 6, the analysis is done to understand the short run and long run relationships between growth and health. Additionally, the Chapter explains the causality of the two indicators under study. This section introduces the statistical approaches used and the analytical models estimated. Since the data are carried in the time framework for the period of 40 years from 1970 to 2010, this Chapter focuses on using the time series approach. The problem with this model is that when the series are measured in time, they tend to become non-stationary, or there exists the problem of autocorrelation. Therefore, the time series modelling is constructed in the following steps:

1. Testing for stationary of the series.
2. Testing of co-integration of the variables: Engle and Granger Approach, and Auto Regressive Distributed Lags (ARDL).
3. Application of ECM, if co-integration exists.
4. Causality test among the variables under study.

The description of these steps is given below in detail. Since, the data set contains the observations in time framework, the testing of hypothesis adopts the tests in following steps:

1. All the variables are tested for stationarity.
2. The second step involves the testing for co-integration of the variables.
3. Error correction method is applied to check how strong the disequilibrium is, if any,
4. Finally, causality of the variables is tested.

   A broad description of these steps is as follows:

Stationarity Test

Stationarity is the first fundamental statistical property tested for in time series analysis. A stochastic process is said to be stationary if:

1. It has a constant mean and the series tend to return to its mean (mean reversion).
2. It has constant variance. It has a theoretical correlogram that diminishes as the lag length increases.
3. The value of covariance between the two time periods depends only on the distance, gap or lag between the two periods and not the actual time at which the covariance is computed.

Therefore, a variable y_t is said to be stationary if:

a) $E(y_t)$ = constant for all t;

b) $Var(y_t)$ = constant for all t; and

c) $Cov(y_t, y_{t+k})$ = constant for all t and all $k \neq 0$.

In short, if a time series is stationary, its mean, variance and autocovariance (at various lags) remain the same no matter at what point we measure them, that is, they are time invariant. There are many tests of stationarity. The variables in this analysis are tested by 'unit root test'. To understand this broadly, consider the following auto-regressive model:

$$y_t = \phi y_{t-1} + u_t \quad (1)$$

Where u_t is a white noise process and the stationarity condition is $|\phi| < 1$.

In general we have three possible cases:

***Case 1*:** if $|\phi| < 1$, the series is stationary.

***Case 2*:** if $|\phi| > 1$, the series explodes.

***Case 3*:** if $|\phi| = 1$, the series contains a unit root and is non-stationary.

Testing for the order of integration

A test for the order of integration is for the number of unit roots, and it follows the steps described below:

Step 1: Test y_t to see if it is stationary. If yes, then $y_t \sim I(0)$; if no then $y_t \sim I(n)$; n>0.

Step 2: Take first differences of y_t as $\Delta y_t = y_t - y_{t-1}$, and test Δy_t to see if it is stationary. If yes, then $y_t \sim I(1)$; if no, then $y_t \sim I(n)$; n>0.

Step 3: Take second differences of y_t as $\Delta^2 y_t = \Delta y_t - \Delta y_{t-1}$, and test $\Delta^2 y_t$ to see if it is stationary. If yes, then $y_t \sim I(2)$; if no, then $\Delta^2 y_t = \Delta y_t - \Delta y_{t-1}$; n>0. Etc… till we find that it is stationary and then we stop. So, for example, if $\Delta^3 y_t \sim I(0)$, then $\Delta^2 y_t \sim I(1)$ and $\Delta y_t \sim I(2)$, and finally $y_t \sim I(3)$; which means that y_t needs to be differenced three times in order to become stationary.

Simple Dickey-Fuller Test for Unit Roots

The early and pioneering work on testing for a unit root in time series was done by Dickey and Fuller (Fuller 1976; Dickey and Fuller 1979) and devised a procedure to formally test for non-stationarity. The key insight of their test is that testing for non-stationarity is equivalent to testing for the existence of a unit root. Thus, the test is based on the simple AR(1) model of the form and is the following:

$$y_t = \phi y_{t-1} + u_t$$

The null hypothesis is $H_0 : \phi = 1$, and the alternative hypothesis is $H_1 : \phi < 1$.

A different or more convenient version of the test can be obtained by subtracting y_{t-1}. From both sides of above AR (1) model.

$$y_t - y_{t-1} = \phi y_{t-1} - y_{t-1} + u_t \quad (2)$$

$$\Delta y_{t-1} = (\phi - 1) y_{t-1} + u_t \quad (3)$$

$$\Delta y_{t-1} = \gamma y_{t-1} + u_t \quad (4)$$

Where $\gamma = (\phi - 1)$. Then, the null hypothesis is $H_0 : \gamma = 0$ and the alternative hypothesis is $H_1 : \gamma < 0$, where if $\gamma = 0$ then y_t follows a pure random walk model.

Dickey and Fuller (1979) also proposed two alternative regression equations that can be used for testing the presence of a unit root. The first contains a constant in the random-walk process as in the following equation:

$$\Delta y_{t-1} = \alpha_0 + \gamma y_{t-1} + u_t \tag{5}$$

This is an extremely important case, because such processes exhibit a definite trend in the series when $\gamma = 0$, which is often the case for macroeconomic variables. The second case is also to allow, a non-stochastic time trend in the model, so as to have:

$$\Delta y_{t-1} = \alpha_0 + a_2 t + \gamma y_{t-1} + u_t \tag{6}$$

The Dickey Fuller test for stationarity is then simply the normal 't' test on the coefficient of the lagged dependent variable y_{t-1} from one of the three models. Dickey and Fuller provided their own critical values for each of the three models above.

a) Augmented Dickey-Fuller (ADF) Test for Unit Roots

As the error term is unlikely to be white noise, Dickey and Fuller extended their test procedure suggesting an augmented version of the test which includes extra lagged terms of the dependent variable in order to eliminate auto-correlation. The lag length on these extra terms is either determined by the Akaike Information Criterion (AIC) or Schwartz Bayesian Criterion (SBC), or more usefully by the lag length necessary to whiten the residuals (i.e., after each case we check whether the residuals of the ADF regression are auto-correlated or not through LM tests and not the DW test). The three possible forms of the ADF test are given by the following equations:

$$\Delta y_t = \gamma y_{t-1} + \sum_{i=1}^{p} \beta_i \Delta y_{t-i} + u_t \tag{7}$$

$$\Delta y_t = \alpha_0 + \gamma y_{t-1} + \sum_{i=1}^{p} \beta_i \Delta y_{t-i} + u_t \tag{8}$$

$$\Delta y_t = a_0 + \gamma y_{t-1} + a_2 t + \sum_{i=1}^{p} \beta_i \Delta y_{t-i} + u_t \quad (9)$$

The difference between the three regressions again concerns the presence of the deterministic elements a_0 and $a_2 t$. The critical values for the ADF tests are the same as those for the Dickey-Fuller (DF) test.

b) Phillips-Perron Test (PP)

The distribution theory supporting the DF test is based on the assumption that the error terms are statistically independent and have a constant variance. So when using the ADF methodology, we have to make sure that the error terms are uncorrelated and that they really have a constant variance. Phillips and Perron (1988) developed a generalization of the ADF test procedure that allows for fairly mild assumptions concerning the distribution of errors. The present study, therefore, involves detection of stationarity by ADF and Phillips-Perron tests. The non-stationary series then involves the first difference method to render them stationary and proceed further.

Co-integration (Engle-Granger Approach)

One way of resolving the problem of non-stationarity in time series data is to difference the series successively until stationarity is achieved and then use the stationarity series for regression analysis. However, this solution is not ideal. The desire to have models which combine both short-run and long-run properties and which at the same time maintain stationarity in all of the variables has led to a reconsideration of the problem of regression using variables that are measured in their levels.

In most cases, if two variables that are *I*(1) are linearly combined, then the combination will also be *I*(1). Generally if variables with differing orders of integration are combined, the combination will have an order of integration equal to the largest. This linear combination of *I*(1) variables will itself be *I*(1), but it would obviously be desirable to obtain residuals that are *I*(0), so that the variables are co-integrated.

According to Engle and Granger (1987), a set of variables is defined as co-integrated if a linear combination of them is stationary.

Many time series are non-stationary but 'move together' overtime – that is, there exist some influences on the series, which implies that the two series are bound by some relationship in the long run. A co-integrating relationship may also be seen as a long term or equilibrium phenomenon, since it is possible that co-integrating variables may deviate from their relationship in the short run, although their association would return in the long run. In order to understand this approach (often called EG approach), the first step is to fit a regression:

$$Y_t = \beta_o + \beta_1 X_t + €_t \quad (10)$$

This regression is known as the 'potential co-integration regression'. Then, find the residual:

$$e_t = Y_t - b_0 - b_1 X_t \quad (11)$$

Where b_0, b_1 and e_t are the estimates of the coefficients β_o, β_1 and $€_t$ respectively.

The next step is to apply ADF test on the series:

$D(e) = \lambda e_t + \sum\lambda D(e_{t-1}) + v_t$

Where D = change

Now to test stationarity:

Null hypothesis of e_t= no-cointegration

Alternative hypothesis = co-integration.

Error Correction Model (ECM)

The presence of a co-integrating relationship forms the basis of error correction specification. The ECM was first used by Sargan (1984) and later popularised by Engle and Granger (1987). An important theorem known as Granger Representation Theorem states that if two variables *Y* and *X* are co-integrated, then the relationship between the two can be expressed as ECM. The ECM takes the following form of equation:

$$D(Y_t) = \beta_1 D(X_t) + \lambda(e_{t-1}) + V_t \quad (12)$$

In this equation, (e_{t-1}) is known as the error correction term. If Y_t and X_t are co-integrated with co-integrating coefficient, then (e_{t-1}) will be *I*(0) even though the constituents are *I*(1). It is, thus, valid to use Ordinary Least Squares (OLS) and standard procedures for statistical inference on (12). The error correction term appears with a 'lag'.λ defining the long run relationship between *x* and *y*, while β_1 describes the short run relationship between changes in *x* and *y*. Broadly, it describes the speed of adjustment back to equilibrium, and its strict definition is that it measures the proportion of the last period's equilibrium error that is corrected for.

Granger Causality Test

Granger (1969) developed a relatively simple test that defined causality as follows: a variable Y_t is said to Granger-cause X_t, if X_t can be predicted with greater accuracy by using past values of the Y_t variable rather than not using such past values, all other terms remaining unchanged. Granger causality test for the case of two explanatory variables Y_t and X_t involves as a first step the estimation of the following VAR model:

$$y_t = a_1 + \sum_{i=1}^{n} \beta_i x_{t-i} + \sum_{j=1}^{m} \gamma_j y_{t-j} + e_{1t} \quad (13)$$

$$x_t = a_2 + \sum_{i=1}^{n} \theta_i x_{t-i} + \sum_{j=1}^{m} \delta_j y_{t-j} + e_{2t} \quad (14)$$

Where it is assumed that both e_{2t} and e_{1t} are uncorrelated white-noise error terms. In this model we can have the following different cases:

Case 1: The lagged x terms in (13) may be statistically different from zero as a group, and the lagged y terms in (14) not statistically different from zero. In this case x_t causes y_t.

Case 2: The lagged y terms in (14) may be statistically different from zero, and the lagged x terms in (13) is not statistically different from zero. In this case y_t causes x_t.

Case 3: Both sets of x and y terms are statistically different from zero in (13) and (14) so that we have bi-directional causality.

Case 4: Both sets of x and y terms are not statistically different from zero in (13) and (14), so that x_t is independent of y_t .

According to the results of the variable deletion tests, one may conclude about the direction of causality based on the four cases mentioned above.

A Panel Data Analysis:

In Chapter 7 the objective of the study is carried out using the panel data analysis. There are many advantages of the panel data over others like cross sectional because it relates the individual observations (states of India in our case) over time. Therefore, by combining time series of cross-section observations, the panel data give "more informative data, more variability, less co-linearity among variables, more degrees of freedom and more efficiency" (Gujarati 2004, 4th ed.). In addition to this, the panel data sets are better to study complex issues of dynamic behaviour, which contain repeated measures of the same variable. They can better detect and measure effects that simply cannot be observed in pure cross-sectional form. Therefore, in order to study the effect of economic growth on health, the data can give more enriched analysis.

Within and Between Variables

Dependent variables and repressors can potentially vary over both time and individuals. Variation over time or a given individual is called within variation and variation across individuals is called between variations. This distinction is important because estimators differ in their use of within and between variations. In particular, in the fixed effects(FE) model the coefficient of a regressor with little within variation will be imprecisely estimated and will be not identified if there is no within variation at all.

The data used in Chapter 7 are short panel and a balanced data. We have considered the following model:

$$Y_{st} = \gamma_s + X_{st}\beta + €_{st} \qquad (1)$$

Where X_{st} is regressor, $\boldsymbol{\gamma_s}$ is random-specific effects, and ϵ_{st} is an idiosyncratic error.

It is evident from the data that there are two different models for the γ_s that are fixed effects and random effects. The distinction between these two models is that in the FE model, the γ_s in equation 1 is correlated with the regressors X_{st}. This allows limited form of endogeneity. On the other hand, in the random effects (RE) model, it is assumed that γ_s in equation 1 is purely random. This is a strong assumption implying that γ_s is uncorrelated with the regressors. Also, it is important to recognize that if effects are fixed, then the pooled OLS and RE estimators are inconsistent, and instead the within (of FE) estimator needs to be used. The within estimator is otherwise less desirable, because using only within variation leads to less-efficient estimation and inability to estimate coefficients of time-invariant regressors.

Since it is difficult to choose among these models as to which one is more suitable for our analysis, Hausman test is used that checks a more efficient model against a less efficient but consistent model to make sure that the more efficient model also gives consistent results.

Hausman test is based on the null hypothesis that coefficients estimated by the efficient random effects estimator are the same as the ones estimated by the consistent fixed effects estimator.

However, there is a shortcoming in the standard Hausman test as it requires the RE estimator to be efficient. This in turn requires that the γ_s and ϵ_{st} are independently and identically distributed (i.i.d) which could be an invalid assumption if cluster-robust standard errors for the RE estimator differ from the default standard errors. Thus, we have tried to figure out the best suitable model using both Hausman test and Robust Hausman test.

Random Effects (RE) Estimator

In the analysis, we have used RE model after testing its efficiency by using both Hausman test and Robust Hausman test. The RE estimator is the feasible generalised

least- Squares (FGLS) estimator in the RE model (as shown in equation 1) under the assumption that the random effect γ_s id i.i.d and the idiosyncratic error $€_{st}$ is i.i.d. the RE estimator is consistent if the RE model is appropriate and is inconsistent if the FE model is appropriate. The RE model is the individual-effects model (equation 1):

$$Y_{st} = X_{st}\beta + (\gamma_s + €_{st}) \quad (2)$$

With $\gamma_s \sim (\gamma, \sigma^2\gamma)$ and $€_{st} \sim (0, \sigma^2_€)$

Where $u = \gamma_s + €_{st}$(combined error)

Then, the combined error is correlated with ***t*** for the given ***s*** with

$$Cor(u_{st}, u_{si}) = \sigma^2\gamma/(\sigma^2\gamma + \sigma^2_€), \text{ for all } t \neq i. \quad (3)$$

The RE estimator is the FGLS estimator of β in (2) given (3) for the error correlations.

In different settings such as heteroskedastic errors and AR(1) errors, the FGLS estimator can be calculated as the OLS estimator in a model transformed to have homoskedastic uncorrelated errors. This is also possible here. Algebra shows that the RE estimator can be obtained by OLS estimation in the transformed model:

$$(Y_{st} - \hat{\theta}_s Y) = (1 - \theta_s)\gamma + (X_{st} - \theta_s X_s)'\beta + \{(1 - \theta_s)\gamma_i + (€_{st} - \theta_s €_{st})\} \quad (4)$$

The RE estimator is consistent and fully efficient if the RE model is appropriate. It is inconsistent if the FE model is appropriate, because then correlation between X_{st} and γ_s implies correlation between the regressors and the error in equation 4. Also, if there are no fixed effects but the errors exhibit within-panel correlation, then the RE estimator is consistent but inefficient, and cluster-robust standard errors should be obtained.

The RE estimator uses both between and within variation in the data and has special cases of pooled OLS ($\hat{\theta}_s = 0$) and within estimation ($\hat{\theta}_s = 1$). The RE estimator approaches the within estimator as T gets large and as $\sigma^2\gamma$ gets large relative to $\sigma^2_€$, because in those cases $\hat{\theta}_s = 1$.