ignou
THE PEOPLE'S
UNIVERSITY

Indira Gandhi
National Open University
School of Social Sciences

MEC-109
Research Methods in Economics

Block

3

QUANTITATIVE METHODS-I

UNIT 9

Two Variable Regression Models 5

UNIT 10

Multivariable Regression Models 37

UNIT 11

Measures of Inequality 59

UNIT 12

Construction of Composite Index in Social Sciences 93

Expert Committee

Prof. D.N. Reddy
Rtd. Professor of Economics
University of Hyderabad, Hyderabad

Prof. Harishankar Asthana
Professor of Psychology
Banaras Hindu University
Varanasi

Prof. Chandan Mukherjee
Professor and Dean
School of Development Studies
Ambedkar University, New Delhi

Prof. V.R. Panchmukhi
Rtd. Professor of Economics
Bombay University and Former
Chairman ICSSR, New Delhi

Prof. Achal Kumar Gaur
Professor of Economics
Faculty of Social Sciences
Banaras Hindu University, Varanasi

Prof. P.K. Chaubey
Professor, Indian Institute of
Public Administration, New Delhi

Shri S.S. Suryanarayana
Former Joint Advisor
Planning Commission, New Delhi

Prof. Romar Korea
Professor of Economics
University of Mumbai, Mumbai

Dr. Manish Gupta
Sr. Economist
National Institute of Public Finance and Policy
New Delhi

Prof. Anjila Gupta
Professor of Economics
IGNOU, New Delhi

Prof. Narayan Prasad **(Convenor)**
Professor of Economics
IGNOU
New Delhi

Prof. K. Barik
Professor of Economics
IGNOU
New Delhi

Dr. B.S. Prakash
Associate Professor in Economics
IGNOU, New Delhi

Shri Saugato Sen
Associate Professor in Economics
IGNOU, New Delhi

Course Coordinator and Editor: **Prof. Narayan Prasad**

Block Preparation Team

Unit	Resource Person	IGNOU Faculty (Format, Language and Content Editing)
9	Dr. Anoop Chatterjee Associate Professor in Economics ARSD College (University of Delhi), Delhi	Prof. Narayan Prasad Professor of Economics IGNOU, New Delhi
10	Shri B.S. Bagla Associate Professor in Economics PGDAV College (University of Delhi), Delhi	Prof. Narayan Prasad Professor of Economics IGNOU, New Delhi
11	Prof P.K. Chaubey IIPA, New Delhi	Prof. Narayan Prasad Professor of Economics, IGNOU, New Delhi
12	Dr. Sunil K Mishra Fellow, Institute for Human Development New Delhi	Prof. Narayan Prasad Professor of Economics, IGNOU, New Delhi

Print Production

Mr. Manjit Singh
Section Officer (Pub.)
SOSS, IGNOU, New Delhi

October, 2015

ISBN-978-81-266-

Further information on Indira Gandhi National Open University courses may be obtained from the University's office at Maidan Garhi. New Delhi-110 068.

Printed and published on behalf of the Indira Gandhi National Open University, New Delhi by the Director, School of Social Sciences.

Laser Typeset by : Tessa Media & Computers, C-206, A.F.E.-II, Okhla, New Delhi
Printed at :

BLOCK 3 QUANTITATIVE METHODS-I

Economic theory is basically concerned about economic laws. These laws (or hypotheses) are qualitative statements on the economic behaviour of various agents both at the micro and macro levels. The validity of such qualitative statements, however, depends upon their rigorous empirical verification by means of the available data.

The implicit argument between the economic theory and its empirical verification is that if the stated laws holds good in the real world situation, it should be reflected in the displayed pattern of the relevant data. The identification and the examination of such patterns are conducted by employing various statistical techniques. The primary statistical tool employed for such a purpose is the *Regression Modeling*. Hence, a thorough knowledge of Regression Models is absolutely essential for conducting any serious research in Applied Economics.

An economic law or theory essentially postulates some relationship that exist among the economic variables. Suppose the relationship between two variables Q and P (i.e. Quantity and Price) is expressed in the form (say linear form) $Q = \alpha + \beta P$. Such an expression amount to an exact linear relationship which is hard to find in reality. Usually, various non-deterministic factors (called random factors) affect such a relationship. Therefore, in order to allow for the effect of such random factors, a random component is incorporated into the model i.e. $Q = \alpha + \beta P + U$, where U is a random variable. With this, the strict mathematical relationship expressed before becomes statistical in character. Regression technique can now be applied to estimate the two parameters α and β by using the actual data. This is the crux of a *Two Variable Regression Model* explained in **Unit 9**.

Unit 10 explains the multiple regression models by including more than two independent random variables. The same procedure as done in the case of two variable regression models is applied to examine whether the estimated parameters correspond to the postulated relationship. The treatment accorded to the Two Variable Regression Models and the Multiple Regression Models in this course is first limited to the ordinary least square (OLS) method subsequently, another method viz., the maximum likelihood method (MLM) is also explained. The two methods taken together equip a researcher with the basic tools for undertaking empirical research.

Improvement in well being of the poor has been one of the important goals of economic policy and to a significant extent it is determined by the growth and distribution of its income. Distribution patterns have an important bearing on the relationship between average income and poverty levels. Extreme inequalities are economically wasteful. Further, income inequalities also interact with other life-chance inequalities. Hence reducing inequalities has become priority of public policy. **Unit 11** deals with the conceptual and measurement aspects of income inequality.

The complex social and economic issues like child deprivation, food security, human development, human wellbeing etc. are difficult to measure in terms of single variable. They are expression of several indicators. Composite Index is an important statistical techniques to express the single value of several interdependent or independent variables. Hence, **Unit 12** throws light on various methods to construct Composite Index.

UNIT 9 TWO VARIABLE REGRESSION MODELS

Structure

9.0 Objectives

9.1 Introduction

9.2 The Issue of Linearity

9.3 The Non-deterministic Nature of Regression Model

9.4 Population Regression Function

9.5 Sample Regression Function

9.6 Estimation of Sample Regression Function

9.7 Goodness of Fit

9.8 Functional Forms of Regression Model

9.9 Classical Normal Regression Model

9.10 Hypothesis Testing

9.11 Let Us Sum Up

9.12 Exercises

9.13 Key Words

9.14 Some Useful Books

9.15 Answers or Hints to Check Your Progress Exercises

9.16 Answers or Hints to Exercises

9.0 OBJECTIVES

After reading this unit, you will be able to:

- know the issue of linearity in regression model;
- appreciate the probabilistic nature of the regression model;
- distinguish between the population regression model and the sample regression model;
- state the assumptions of the classical regression model;
- estimate the unknown population regression parameters with the help of the sample information;
- explain the concept of goodness of fit;
- use various functional forms in the estimation of the regression model;
- state the role of classical normal regression model; and
- conduct some tests of hypotheses regarding the unknown population regression parameters.

9.1 INTRODUCTION

The empirical research in economics is concerned with the statistical analysis of economic relations. Often these relations are expressed in the form of regression equations involving the dependent variable and the independent variables. The formulation of an economic relation in the form of a regression equation is called a regression model in econometrics. In such kind of a regression model, the variable under study is assumed to be a function of certain explanatory variables. The major purpose of a regression model is the estimation of its parameters. In addition, it is also useful for testing hypotheses and making certain forecasts. However, our concern will be mainly with the estimation of parameters and testing of some hypotheses. At this stage, we shall refrain from discussing the issue of forecasts from a regression model. A regression model that contains one explanatory variable is called a two variable regression model. In this unit, we shall focus on a two variable regression model. It may be mentioned here that we are essentially focusing on what is known as the two variable classical regression model.

9.2 THE ISSUE OF LINEARITY

When we talk about a regression model; at our level, what we have in our mind is a linear regression model. It is important to have a clear idea about the concept of linearity in the context of a regression model.

Suppose, we have a regression equation like,

$$Y = \alpha + \beta X$$

Generally, we call this a linear regression equation. By linearity we often mean a relationship, in which, the dependent variable is a linear function of the independent variable. In this case, the graphical representation of the regression equation is a straight line. However, there can be another interpretation of linearity. Consider the following regression equation,

$$Y = \alpha + \beta X + \gamma X^2$$

Now, this regression equation is linear in parameter, in the sense that the highest power of any parameter (constant) is one. But this equation is non-linear in variable because the highest power of the independent variable is two. In fact, conventionally speaking, it is a quadratic regression equation. In this way, we can have any number of polynomial regression equations with the highest power of the independent variable going up to any positive integer. And all these equations may be linear in parameter.

But consider this regression equation,

$$Y = \alpha + \beta^2 X$$

This regression equation is linear in variable, but non-linear in parameter because the power of β is two.

We should note now that from the point of view of regression analysis, we shall consider only those models which are linear in parameter, no matter whether they are linear in variable or not. After all, the main purpose of a regression model is the estimation of its parameters. These parameters have to be linear for the purpose of their straightforward estimation, say,

by least square method. For example, the parameters of the second regression equation presented above, can be easily estimated by a simple extension of the least square method that we have studied in Unit 8. It should be mentioned here that there are some regression equations that are non-linear in parameters. By applying suitable transformations they can be reduced to linear in parameter regression equations. We may consider the following regression equation

$$Y = \alpha X^{\beta}$$

This regression equation in its present form is non-linear in parameter. However, by applying log transformation, we obtain the following equation

$$\log Y = \log \alpha + \log \beta X$$

It can be clearly seen that the equation has now been transformed into one that is linear in the log of parameters. We can easily estimate log values of these parameters by using the usual least square method and then obtain the estimated values of the original parameters by applying the antilog procedure. However, some times, the regression equations can be of non-linear in parameter type that no transformation can render them to a linear in parameter form. Such equations should be the basis of non-linear or intrinsically non-regression models. There is no direct method available for the estimation of the parameters of this kind of regression models and they are generally estimated by following some standard iteration procedure. Any discussion on such iteration procedure, however, is beyond the scope of the present unit.

9.3 THE NON-DETERMINISTIC NATURE OF THE REGRESSION MODEL

We have already referred to the statistical nature of the regression equation in the last unit. By the very nature of social science, we cannot expect the relationships that may exist among different variables to be exact or deterministic. There is always some random element involved in them. As a result, for a particular value of the independent variable *X*, the value of the dependent variable *Y* cannot be exactly determined from such a relationship. Let us consider the example of a consumption function. Here, for a given level of income, when we try to identify the corresponding level of consumption, in all probability, we shall not be able to obtain a definite value of consumption; instead, a host of values will be available. And this will be the case with all possible levels of income. The diagram below shows this.

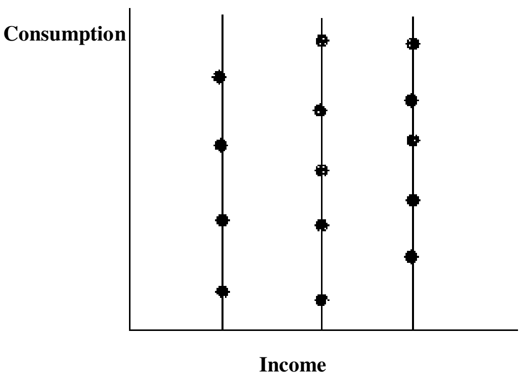


Fig. 9.1: Bi-Variate Population

Thus, the dependent variable *Y* tends to be probabilistic or stochastic in nature. In fact, for each value of the independent variable *X*, there will be a distribution of the values of the dependent variable *Y*. Accordingly; we specify the regression model by incorporating a random or stochastic variable. As mentioned in the previous unit, it should be clear that in our formulation the dependent variable is stochastic but the independent or explanatory variables are non-stochastic in nature. We should point out here that at an advanced treatment of the regression model, even the explanatory variables are considered to be stochastic in nature. It is the element of randomness either in the dependent variable alone or both in the dependent and independent variables, that make the regression model non-deterministic in nature.

9.4 POPULATION REGRESSION FUNCTION

The first step in the regression model is the conceptualization of a population regression function. Let us assume that there is a bi-variate population of (*X, Y*) of size *N*. Suppose, in this population, *Y* is the dependent variable and *X* is the independent variable. Let *X* and *Y* be related to each in a linear fashion in the following way:

$$Y = \alpha + \beta X + U$$

The above equation is called the population regression function. Here, α and β are the unknown parameters. It is clear from our discussion on linearity that this function is linear in variable as well as in parameter.

In this population regression function, the variable *U* deserves some attention. We have already mentioned that in social Science, there cannot be an exact relationship and it is at the most statistical in nature. In our formulation, in fact, *Y* is stochastic or random, but *X* is non-stochastic or deterministic in nature. Consequently, a random variable like *U* is introduced in the population regression function to incorporate this element of randomness of a statistical relationship. The random variable *U* is also called the disturbance term. It is a sort of catch – all variables that represent all kinds of indeterminacies of an inexact relationship. Thus it may represent, for example,

a) *Inherent randomness in human behaviour*: The unpredictability of human psyche is also reflected in the human behaviour. As a result, even after taking into account of all possible factors, a regression equation cannot fully explain the value of the dependent variable for the given values of all possible explanatory variables.

b) *Effect of omitted variables*: Sometimes for the sake of parsimony, all the explanatory variables are not included in a regression model. The disturbance term can be taken as a representative variable of all such omitted variables.

c) *Effect of measurement error*: Sometimes, it is not possible to measure the values of the dependent variable accurately. Consequently, the disturbance term is introduced in the regression model to represent the combined effect of all such possible sources of measurement error.

d) *Error in the formulation*: Often, the functional form of the regression model proves to be far from a correct depiction of the underlying relationship between the dependent variable and the independent variables.

We incorporate the disturbance term to correct the distortions that may arise from a wrong specification of the regression model.

We should note that apart from the above-mentioned factors, there might be other unidentifiable factors that might also be influencing the dependent variable in an unknown manner. We introduce the disturbance term for incorporating the effect of all such unknown random factors. It should be clear that for some of the reasons mentioned above the disturbance term may assume a positive value and for some other factor it may tend to be negative. Consequently, for all practical purposes, the net or the mean effect of the random disturbance can be taken to be zero.

Difference between the Disturbance Term and the Intercept

There is a difference between the interpretation of the disturbance term and that of the intercept term α in our regression equation. As we have noted, that the disturbance represents the random effects of the known and unknown variables that have not been included in the regression model. The intercept term, on the other hand, stands for all such variables, that are known to have some definite non-random effects on the dependent variable. For example, in the demand function, if we just formulate a two variable regression equation between the quantity demanded and the price instead of a multiple regression equation, then we are consciously not including variables like prices of other commodities and the income of the consumer. And all these variables affect the quantity demanded in a known fashion. So, the intercept term here represents the average effect of all such known factors that are not explicitly included in the model.

It should be clear now that it is the disturbance term that makes the relationship statistical in nature and renders the entire procedure so rich in content.

Population Regression Line

The main objective of the regression analysis is to obtain the value of the dependent variable for a given value of the independent variable. This can be attempted by running a regression on the population regression function $Y = \alpha + \beta X + U$, i.e., by fitting a regression line through the population cluster shown in Figure 9.1. Obviously, the value of Y that we shall obtain from such a population regression line will be an average value. In other words, we shall get the equation for the population regression line by applying the expectation to the population regression function, $Y = \alpha + \beta X + U$. Thus, the population regression equation will be given by

$$E(Y / X) = \alpha + \beta X$$

It may be mentioned here that while applying the expectation operator, E (U) can be taken as zero because, as we have discussed above, the mean effect of U tends to be zero.

Sometimes we call the above expression, the population regression line and loosely write it as

$$Y = \alpha + \beta X$$

In this expression, however, we should be clear that Y stands for the conditional mean of Y for a given X.

Assumptions of the Classical Regression Model

As we have mentioned earlier, the population regression model $Y = \alpha + \beta X + U$ is unknown in the sense that that the numerical values of its parameters are not known. As a result, we are not in a position to find the value taken by *Y* on an average for a given value of *X*. This means that we have to estimate these parameters from sample information. We already know that a very simple and popular method of estimation is the least square procedure. However, we should ensure that these least square estimates faithfully represent the unknown population parameters, otherwise, the whole purpose is defeated. This is possible only if the disturbance term *U* satisfies some assumptions. These assumptions along with two other that we have already mentioned about, are known as the assumptions of the classical regression model. The assumptions are:

1) The disturbance term *U* has a zero mean for all the values of *X*, i.e., *E* (*U*) = 0.
2) Variance of *U* is constant for all the values of X, i.e.,$V(X) = \sigma^2$. It may be mentioned here that this assumption is known as the assumption of homoscedasticity in the econometric literature.
3) The disturbance terms for two different values of *X* are independent i.e., $\text{cov}(U_i, U_j) = 0, \text{for } i \neq j$.
4) *X* is non-stochastic.
5) The model is linear in parameter.

The assumption of non-stochasticity of *X* leads to a corollary. It is, *X* and *U* are independent i.e., $\text{cov}(X, U) = 0$.

9.5 SAMPLE REGRESSION FUNCTION

To put the whole discussion in the proper perspective; we have an unknown bi-variate population. The only information that we have is some randomly selected values of *Y* from this population that correspond to some fixed values of *X*. Suppose, in this way we have *n* pairs of values of *X* and *Y*. So, we can say that we have a random sample of size *n*. Our purpose is to estimate the parameters of this unknown population from our sample information so that we can have a reasonable estimate of an average value of *Y* for a given value of *X* that is valid for the entire population. In other words, our objective is to estimate the population regression function from the sample observations. We can proceed in that direction by discussing the concept of a sample regression function. The form of the sample regression function is quite similar to that of the population regression function. It can be presented as

$$\hat{Y} = \hat{\alpha} + \hat{\beta} X + \hat{U}$$

In the above expression, $\hat{Y}, \hat{\alpha}, \hat{\beta}$ and $\hat{U}$ should be interpreted as the sample estimators for their respective unknown population counterparts. Thus, we are hypothesizing that corresponding to the linear population regression function $Y = \alpha + \beta X + U$, there is a linear sample regression function $\hat{Y} = \hat{\alpha} + \hat{\beta} X + \hat{U}$. Our purpose is now to estimate the population regression function from the sample regression function. It means that we have to estimate $\hat{\alpha}$ and $\hat{\beta}$ from the given sample and take them as the estimates for the unknown population parameters. It should be clear that due to sampling fluctuations, we might obtain different values of $\hat{\alpha}$ and $\hat{\beta}$ from different samples. What we have to ensure is that on the average, they represent the population parameters. Thus, the entire issue now boils down to the estimation of $\hat{\alpha}$ and $\hat{\beta}$. This is what is known as the estimation of the sample regression function and we are going to discuss it in the next section.

9.6 ESTIMATION OF SAMPLE REGRESSION FUNCTION

The commonly used procedure is the least square method that we discussed in the previous unit. To recapitulate, we have a sample of observations represented by a cluster of points in the sample space and we have to pass a regression line through this cluster in such a fashion that the sum of the squares of the deviations of the observed values of *Y* from their estimated values from the regression line for different values of *X* is minimum. We are reproducing the two relevant diagrams from Unit 8 for the sake of convenience.

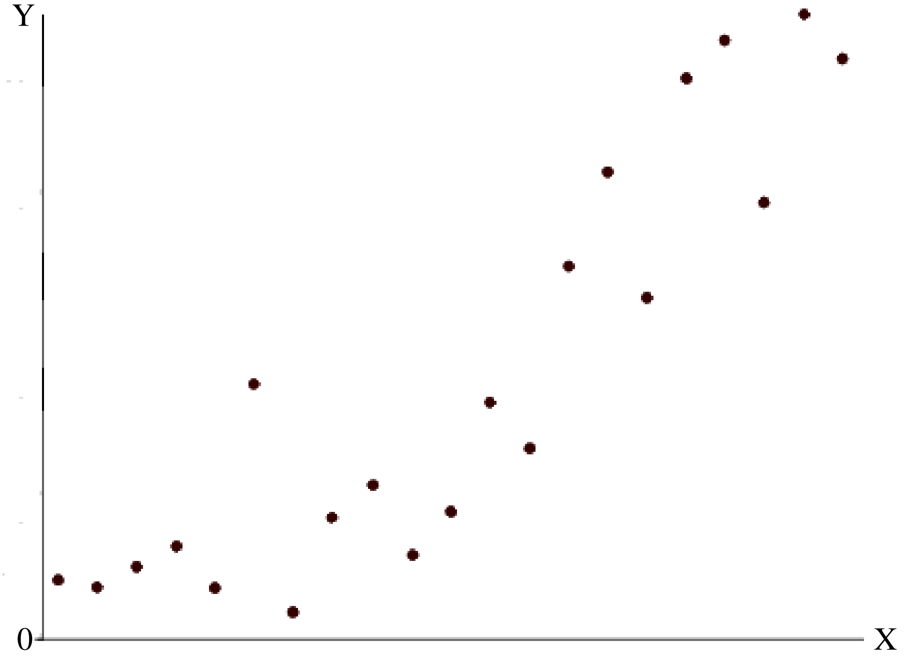


Fig. 9.2: Sample Scatter-Plot

Thus, we have a scatter of sample values as shown in the above diagram and we have to pass a regression line through it. We can summarise the least square procedure for obtaining such a line now. Putting mathematically the procedure amounts to Minimize

$$\sum u^2 = \sum \left(Y - \hat{Y}\right)^2 = \sum (Y - \hat{\alpha} - \hat{\beta}X)^2 \text{ with respect to } \hat{\alpha} \text{ and } \hat{\beta}.$$

Following the usual minimization procedure, we obtain two normal equations given by

$$\sum Y = n\hat{\alpha} + \hat{\beta}\sum X$$

and

$$\sum XY = \hat{\alpha}\sum X + \hat{\beta}\sum X^2 .$$

Solving the two normal equations simultaneously, we obtain

$$\hat{\beta} = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - \left(\sum X\right)^2} = \frac{\sum \left(X - \overline{X}\right)\left(Y - \overline{Y}\right)}{\sum \left(X - \overline{X}\right)^2}$$

and

$$\hat{\alpha} = \frac{\sum X^2 \sum Y - \sum X \sum XY}{n\sum X^2 - \left(\sum X\right)^2} = \overline{Y} - \hat{\beta}\overline{X}\ .$$

Let us now derive an important result for the estimated slope coefficient $\hat{\beta}$.

Writing lower case letters for deviations

$$\hat{\beta} = \frac{\sum xy}{\sum x^2} = \frac{\sum x\left(Y - \overline{Y}\right)}{\sum x^2} = \frac{\sum xY - \overline{Y}\sum \left(X - \overline{X}\right)}{\sum x^2}$$

But $\sum \left(X - \overline{X}\right) = 0$

Thus $\hat{\beta} = \dfrac{\sum xY}{\sum x^2}$

Now putting $k = \dfrac{x}{\sum x^2}$, as x's are given.

We have

$\hat{\beta} = \sum kY$

Thus, $\hat{\beta}$ is a linear function of the observed values of Y. This is an important result, which we shall use later in the issue of hypothesis testing.

The following diagram shows a regression line, with the values of $\hat{\alpha}$ and $\hat{\beta}$ given by the above expressions fitted into our sample scatter.

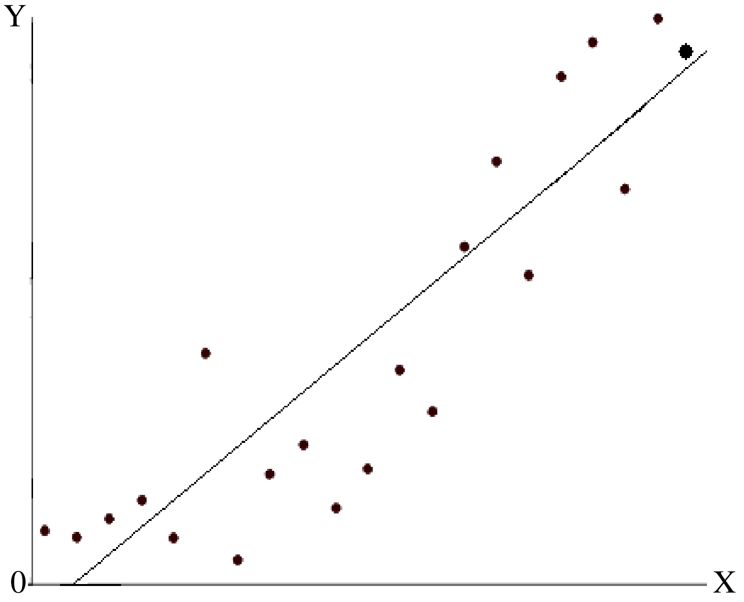


Fig. 9.3: Sample Scatter-Plot with the Sample Regression Line

It may be noted here that the sample regression line has the equation, $Y = \hat{\alpha} + \hat{\beta}X$ with the values of $\hat{\alpha}$ and $\hat{\beta}$ obtained from the above-mentioned least square method. It can be clearly seen that the values of $\hat{\alpha}$ and $\hat{\beta}$ can be calculated in terms of sample observations only and no other information that is not known is required for this purpose. As we can see, both $\hat{\alpha}$ and $\hat{\beta}$ are linear functions of *X* and *Y*. Consequently, they are called linear estimators.

Gauss-Markov Theorem

The least square estimators $\hat{\alpha}$ and $\hat{\beta}$ can be taken as the estimators of the unknown population parameters α and β because they satisfy certain desirable properties. We can state them without proofs below.

Under the assumptions of the classical regression model as mentioned earlier,

1) Least square estimators are linear. As we have already seen from the expressions of $\hat{\alpha}$ and $\hat{\beta}$, they are linear functions of the variables.

2) Least square estimators are unbiased i.e., $E(\hat{\alpha}) = \alpha$ and $E(\hat{\beta}) = \beta$. This means, if we consider different values of the least square estimates obtained from a number of random samples for a given population on the average, they will be equal to the unknown population parameters. They will not be systematically overestimating or underestimating the population parameters.

3) Among all the linear unbiased estimators, least square estimators have the minimum variance. In this sense, they are termed as the efficient estimators.

All these properties of the least square estimators lead to what is known as the Gauss-Markov Theorem. The theorem states:

> **Under the assumptions of the classical linear regression model, among all the linear unbiased estimators, the least square estimators have the minimum variance. In other words, the least square estimators are the best linear unbiased estimators or in short, BLUE.**

Standard Error of the Regression Estimate

We have already stated that due to sampling fluctuations, we should expect to obtain different values of the least square estimates $\hat{\alpha}$ and $\hat{\beta}$ from sample to sample. Consequently, if we measure the standard deviations of these values of the two least square estimates from their respective expectation or mean, we shall get an idea about the extent to which these estimates are affected by the sampling fluctuations. In other words, the standard deviations of the least square estimates can be taken as a measure of the precision of these estimates. **The standard deviations of the least square estimates are known as the standard errors of the estimates.** It should be clear that the standard errors are the standard deviations of the sampling distributions of the least square estimates. The standard deviations or standard errors are obtained by taking the positive square root of the variances of $\hat{\alpha}$ and $\hat{\beta}$. The expressions for both the variance and standard of the least square estimators are given below.

$$\operatorname{var}(\hat{\alpha}) = \frac{\sum X^2}{n\sum\left(X - \bar{X}^2\right)}\sigma^2$$

$$se(\hat{\alpha}) = \sqrt{\frac{\sum X^2}{n\sum\left(X - \bar{X}^2\right)}\sigma^2}$$

$$\operatorname{var}(\hat{\beta}) = \frac{\sigma^2}{\sum\left(X - \bar{X}\right)^2}$$

$$se(\hat{\beta}) = \frac{\sigma}{\sqrt{\sum\left(X - \bar{X}\right)^2}}$$

It should be recollected that c is the standard deviation of the error term U in our population regression model and we assume that it is a constant. However, since the population regression model is unknown, it follows that c is also unknown. As a result, for the calculation of variance and standard error of the least square estimates $\hat{a}$ and $\hat{\beta}$, we need to estimate c . It can be proved that an unbiased estimator of c is $\sqrt{\frac{\sum \hat{U}^2}{n-2}}$. Here, $n - 2$ is what is known as the degrees of freedom in statistics. This concept you must have studied in your compulsory course in statistics. Now, $\hat{U} = Y - \hat{Y}$ is the sample regression error term and accordingly we can calculate it. Thus, by replacing c by its unbiased estimator we can compute the standard errors of both $\hat{a}$ and $\hat{\beta}$**.** It should be noted here that we can write the unbiased estimator of c as

$$\hat{\sigma} = \sqrt{\frac{\sum \hat{U}^2}{n-2}} = \sqrt{\frac{\sum\left(\hat{U} - \bar{\hat{U}}\right)^2}{n-2}}$$

This is, because, $\bar{\hat{U}}$ is the mean of the sample regression errors. Now, while calculating it, the positive errors will tend to cancel the negative errors and as a result, it will be reduced to zero. From above expression of the estimator of c , it can easily be interpreted as the standard deviation of the sample observations about the estimated sample regression line. The sample estimator $\hat{\sigma} = \sqrt{\frac{\sum \hat{U}^2}{n-2}}$ is known as the standard error of estimate or the standard error of the regression.

9.7 GOODNESS OF FIT

In the earlier section, we have discussed the procedure for fitting the sample regression line and its purpose. Once such a regression line is fitted, we may be interested in examining how good has been the fit in the sense how faithfully it can describe the unknown population regression line. This is known as the issue of goodness of fit. In this matter, the regression error term or residual $\hat{U}$

plays an important role. Small quantities of residuals imply that a large proportion of variation in the dependent variable has been explained by the regression equation and consequently, the fit is good. Similarly, large quantities of residuals obviously point to a poor fit. At this stage, what we are interested in is to obtain a quantitative measure of the goodness of fit that is free of any unit for the purpose of comparability. In the previous unit, we have referred to the coefficient of determination, which is the square of the correlation coefficient. It can be shown that this coefficient of determination acts as a measure of goodness of fit. We may consider it now. The variation in the dependent variable Y about its mean can be conceptualized as

$$\operatorname{var}(Y) = \sum \left(Y - \overline{Y}\right)^2$$

We can decompose it into two components. The first is the variation explained by the regression. The second is the portion that remains unexplained by the regression. Thus we can write

$$\left(Y - \overline{Y}\right) = \left(Y - \hat{Y}\right) + \left(\hat{Y} - \overline{Y}\right)$$

Squaring and applying summation on both the sides

$$\sum \left(Y - \overline{Y}\right)^2 = \sum \left(\hat{Y} - \overline{Y}\right)^2 + \sum \left(Y - \hat{Y}\right)^2 + 2\sum \left(\hat{Y} - \overline{Y}\right)\left(Y - \hat{Y}\right)$$

Now let us try to find the value of $\sum \left(\hat{Y} - \overline{Y}\right)\left(Y - \hat{Y}\right)$. We have the regression equation

$$Y = \hat{\alpha} + \hat{\beta}X + \hat{U}$$

or $\overline{Y} = \hat{\alpha} + \hat{\beta}\overline{X} + \overline{\hat{U}}$

Subtracting the second equation from the first equation

$Y - \overline{Y} = \hat{\beta}\left(X - \overline{X}\right) + \hat{U}$, because $\overline{\hat{U}} = \dfrac{\sum \hat{U}}{n} = 0$, as $\sum \hat{U} = 0$.

Writing lower case letters for the deviations of the variables from their mean, we have the sample regression equation in the form

$$y = \hat{\beta}x + \hat{u}$$

Now

$$\sum\left(\hat{Y} - \overline{Y}\right)\left(Y - \hat{Y}\right)$$
$$= \sum\left(\hat{Y} - \overline{Y}\right)\hat{U}$$
$$= \sum \hat{Y}\hat{U} - \overline{Y}\sum \hat{U}$$
$$= \sum \hat{Y}\hat{U}, \text{ because } \sum \hat{U} = 0$$
$$= \sum\left(\hat{\alpha} + \hat{\beta}\mathrm{X}\right)\hat{U}$$
$$= \hat{\alpha}\sum \hat{U} + \hat{\beta}\sum X\hat{U} \quad \because \ \sum \hat{U} = 0$$
$$= \hat{\beta}\sum X\hat{U}$$

We have

$x = X - \bar{X}$

$\Rightarrow \sum x = \sum\left(X - \bar{X}\right)$

or $\sum x\hat{U} = \sum\left(X - \bar{X}\right)\hat{U} = \sum X\hat{U} - \bar{X}\sum\hat{U} = \sum X\hat{U}$

Now

$\sum x\hat{U} = \sum x\left(y - \hat{\beta}x\right) = \sum xy - \hat{\beta}\sum x^2$

We know, $\hat{\beta} = \dfrac{\sum\left(X - \bar{X}\right)\left(Y - \bar{Y}\right)}{\sum\left(X - \bar{X}\right)} = \dfrac{\sum xy}{\sum x^2}$

Putting the value of $\hat{\beta}$ in the above expression for $\sum x\hat{U}$,

$\sum x\hat{U} = \sum xy - \dfrac{\sum xy}{\sum x^2}\sum x^2 = \sum xy - \sum xy = 0$

Hence,

$\sum \left(\hat{Y} - \bar{Y}\right)\left(Y - \hat{Y}\right) = 0$

Thus,

$\sum \left(Y - \bar{Y}\right)^2 = \sum \left(\hat{Y} - \bar{Y}\right)^2 + \sum \left(Y - \hat{Y}\right)^2$

If $\sum \left(Y - \bar{Y}\right)^2$ is defined as the total sum of squares (TSS), $\sum \left(\hat{Y} - \bar{Y}\right)^2$ as the explained sum of squares (ESS) and $\sum \left(Y - \hat{Y}\right)^2$ as the residual sum of squares (RSS),

We have

TSS = ESS + RSS

Dividing both the sides by TSS,

$\dfrac{\text{TSS}}{\text{TSS}} = \dfrac{\text{ESS}}{\text{TSS}} + \dfrac{\text{RSS}}{\text{TSS}}$

Defining the ratio of ESS to TSS as R^2, we have

$\text{R}^2 = 1 - \dfrac{\text{RSS}}{\text{TSS}} = \dfrac{\text{ESS}}{\text{TSS}}$

It should be clear now that R^2 or coefficient determination can be interpreted as the proportion of total variation in *Y* explained by the regression of *Y* on *X*.

Let us now consider a numerical example to fix the concepts and ideas that we have discussed so far.

Example 9.1

The hypothetical figures for the total labour force and the number of employed out of that for the period 1991-2000 are given below in Table 9.1. Run a regression of the number of employed on the total labour force.

Table 9.1: Actual Employment and Labour Force

Year	Employed (million)	Labour Force (million)
1991	100	120
1992	125	140
1993	140	165
1994	160	185
1995	175	200
1996	195	210
1997	230	250
1998	245	255
1999	270	305
2000	295	320

Let Y be the dependent variable employed and X be the independent variable labour force. The detailed calculation for working out the regression is shown below:

Year	Y	X	XY	X^2	$y = Y - \bar{Y}$	$x = X - \bar{X}$	x^2	y^2	xy
1991	100	120	12000	14400	-93.5	-95	9025	8742.25	8882.5
1992	125	140	17500	19600	-68.5	-75	5625	4692.25	5137.5
1993	140	165	23100	27225	-53.5	-50	2500	2862.25	2675
1994	160	185	29600	34225	-33.5	-30	900	1122.25	1005
1995	175	200	35000	40000	-18.5	-15	225	342.25	277.5
1996	195	210	40950	44100	1.5	-5	25	2.25	-7.5
1997	230	250	57500	62500	36.5	35	1225	1332.25	1277.5
1998	245	255	62475	65025	51.5	40	1600	2652.25	2060
1999	270	305	82350	93025	76.5	90	8100	5852.25	6885
2000	295	320	94400	102400	101.5	105	11025	10302.25	10657.5
Sum	1935	2150	454875	502500			40250	37902.5	38850
Mean	193.5	215							

$$\hat{\beta} = \frac{\sum xy}{\sum x^2} = \frac{38850}{40250} = 0.965217 \text{ and } \hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} = 193.5 - \left(0.965217 \times 215\right) = -14.217$$

Year	$\hat{Y} = \hat{\alpha} - \hat{\beta}X$	$\hat{U} = Y - \hat{Y}$	$\hat{U}^2$	$\hat{Y} - \bar{Y}$	$\left(\hat{Y} - \bar{Y}\right)^2$
1991	101.8043	-1.80434	3.255643	-91.6957	8408.094
1992	121.1087	3.89132	15.14237	-72.3913	5240.503
1993	145.2391	-5.23911	27.44822	-48.2609	2329.114
1994	164.5434	-4.54344	20.64289	-28.9566	838.4821
1995	179.0217	-4.0217	16.17407	-14.4783	209.6212
1996	188.6739	6.32613	40.01992	-4.82613	23.29153
1997	227.2826	2.71745	7.384535	33.78255	1141.261
1998	232.1086	12.89137	166.1873	38.60864	1490.627
1999	280.3695	-10.3695	107.5262	86.86949	7546.307
2000	294.8477	0.15226	0.023183	101.3477	10271.36
Sum		0.00045	403.8043		37498.67

In the next stage we compute the following:

$$V\left(\hat{\beta}\right) = \frac{\sigma^2}{\sum x^2},\ V\left(\hat{\alpha}\right) = \frac{\sum X^2}{n\sum x^2}\sigma^2 \text{ and } \mathrm{R}^2 = \frac{\mathrm{ESS}}{\mathrm{TSS}} = \frac{\sum\left(\hat{Y} - \bar{Y}\right)^2}{\sum\left(Y - \bar{Y}\right)^2}.$$

But the population variance σ^2 is unknown. So we have to use an unbiased sample estimator σ^2 for the same. Such an unbiased estimator is given by

$\hat{\sigma}^2 = \dfrac{\sum \hat{U}^2}{n-2}$, where $n - 2$ is the degrees of feedom. Here, $n - 2 = 10 - 2 = 8$.

Therefore, $\hat{\sigma}^2 = \dfrac{\sum \hat{U}^2}{n-2} = \dfrac{403.8043}{8} = 50.475537.$

$$\text{Thus, } V\left(\hat{\beta}\right) = \frac{\hat{\sigma}^2}{\sum x^2} = \frac{50.475537}{40250} = 0.001254$$

$$s.e.\left(\hat{\beta}\right) = \sqrt{V\left(\hat{\beta}\right)} = \sqrt{0.001254} = 0.0354118$$

$$V\left(\hat{\alpha}\right) = \frac{\sum X^2}{n\sum x^2}\hat{\sigma}^2 = \frac{502500}{10 \times 40250} \times 50.475537 = 63.016042$$

$$s.e.\left(\hat{\alpha}\right) = \sqrt{V\left(\hat{\alpha}\right)} = \sqrt{63.016042} = 7.9382644$$

$$\mathrm{R}^2 = \frac{\mathrm{ESS}}{\mathrm{TSS}} = \frac{37498.67}{37902.5} = 0.989345$$

Finally, we present a summary of all the important results concerning the above-mentioned question.

$\hat{\beta} = 0.965217 \qquad V\left(\hat{\beta}\right) = 0.001254 \qquad \text{and} \qquad s.e.\left(\hat{\beta}\right) = 0.0354118$

$\hat{\alpha} = -14.0217 \qquad V\left(\hat{\alpha}\right) = 63.016042 \qquad \text{and} \qquad s.e.\left(\hat{\alpha}\right) = 7.9382644$

$\hat{\sigma}^2 = 50.475537$

$\text{R}^2 = 0.989345$ Degrees of Freedom $(d.f.) = 8$

The estimated regression line is given by

$\hat{Y} = -\ 14.0217 + 0.965217X$

It may be mentioned here that it is this kind of a summary of the results that is generally used for the further analysis of the regression model. Let us now interpret some of the results from the above summary. The regression line estimates the average employment (*Y*) for a given level of labour force (*X*). The slope coefficient $\hat{\beta} = 0.965217$ estimates the rate of change of employment with respect to labour force. For example, if 100 more persons start looking for jobs, about 97 of them actually get employed. The intercept $\hat{\alpha} = -\ 14.0217$ can be interpreted as the average combined effect of all those variables that might also affect employment but have been omitted for the purpose of the above-mentioned regression. The coefficient of regression $\text{R}^2 = 0.989345$ indicates that about 99 per cent of the variation in employment can be explained by a variation in the labour force, which is indeed high and indicates a good fit to the given sample.

Check Your Progress 1

1) Discuss the meaning of linearity in the regression model.

……………………………………………………………………………………

……………………………………………………………………………………

……………………………………………………………………………………

……………………………………………………………………………………

……………………………………………………………………………………

2) How is the regression model non-deterministic in nature?

……………………………………………………………………………………

……………………………………………………………………………………

……………………………………………………………………………………

……………………………………………………………………………………

……………………………………………………………………………………

……………………………………………………………………………………

……………………………………………………………………………………

3) Discuss the assumptions of the classical regression model.

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

4) State the Gauss-Markov Theorem.

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

5) Explain the concept of goodness of fit.

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

9.8 FUNCTIONAL FORMS OF REGRESSION MODEL

We have already mentioned that our focus is on a linear regression model. The issue of linearity has also been considered in Section 9.2. It is in fact linear in parameter regression model that is relevant for us. However, linear in parameter regression models may also have different functional forms. We shall now briefly discuss four such regression models.

1) **Linear Model**

This functional form is the most common one and we have already discussed it. Its equation is given by

$$Y = \alpha + \beta X + U$$

As we know, it is both linear in parameter and linear in variable. This model can be estimated by the ordinary least square (OLS) method. The least square estimators $\hat{\alpha}$ and $\hat{\beta}$ are the unbiased estimators of the unknown population parameters α and β.

Here the regression coefficient β measures the rate of change of Y per unit change of X.

Example 9.2

In this example we shall consider the issue of fitting a Keynesian consumption function to the Indian data. Table 9.2 presents the GDP at factor cost and final consumption expenditure (FCE) figures for the Indian economy during the period 1980-2001 at 1993-94 prices.

Table 9.2: GDP and Final Consumption Expenditure (Rs. Crore)

Year	GDP	FCE	Year	GDP	FCE
1980	401128	390671	1991	701863	620806
1981	425073	407728	1992	737792	637307
1982	438079	415924	1993	781345	666950
1983	471742	446396	1994	838031	695825
1984	492077	461675	1995	899563	740109
1985	513990	485090	1996	970082	794093
1986	536257	504854	1997	1016595	826834
1987	556778	525558	1998	1082747	888513
1988	615098	557527	1999	1148367	952489
1989	656331	584970	2000	1198592	972932
1990	692871	610169	2001	1267945	1027254

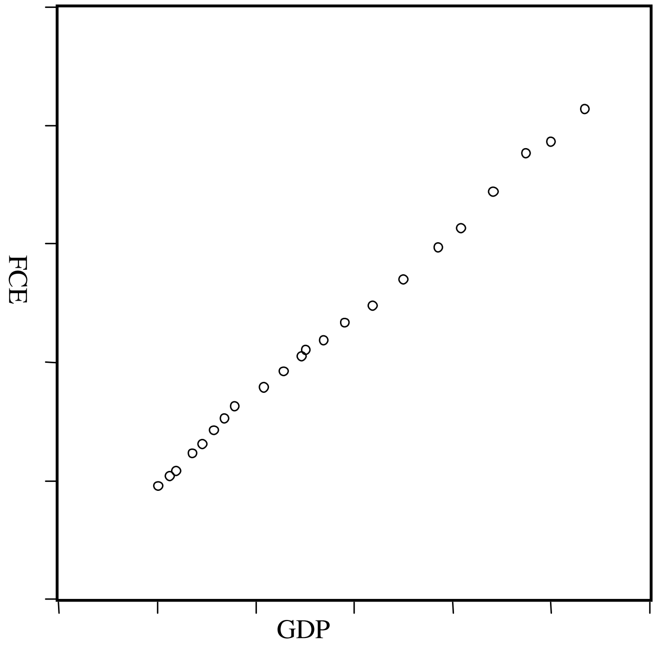


Fig. 9.3: Scatter of Final Consumption Expenditure against GDP

The above scatter of final consumption expenditure against GDP been obtained from Table 9.2. The scatter makes it clear that there seems to be a linear relationship between the two variables. Consequently, we have run a linear regression between the variables by using the data of Table 9.2. The results are presented below:

$$FCE = 108206.4 + 0.719674\ GDP$$

$$s.e.(\hat{\alpha}) = 6233.203 \qquad s.e.(\hat{\beta}) = 0.007865$$

$$t\ (\hat{\alpha}) = 17.35968 \qquad t\ (\hat{\beta}) = 91.50314$$

$$\mathrm{R}^2 = 0.997617 \qquad d.f. = 20$$

A high R^2 of more than 0.99 implies a tight fit. The estimated $\hat{\beta} = 0.719674$ indicates a marginal propensity to consume of about 72 percent and this can be quite expected in the Indian economy. We shall explain the significance of degrees of freedom and *t* values later in the issue of tests of hypotheses. However, we shall continue to present these two statistics in our subsequent examples also.

2) **Log-linear Model**

This model is also known as log-log, double-log or constant elasticity model. Its original form is given by

$Y = \alpha X^{\beta} e^{U}$, Where, *e* is the base of natural log (*e* is approximately equal to 2.718). In the original form, the model is non linear in nature. However, by applying the log transformation, the model is transformed to

$$\ln Y = \ln \alpha + \beta \ln X + U$$

If we put $\ln Y = y,\ \ln \alpha = a,\ \beta = b$ and $\ln X = x$, the model reduces to

$$y = a + bx + U$$

From the above transformation, it is very clear that the model now becomes linear in parameter and linear in log of the variables; and thus, can be estimated by the ordinary least square method. Here, *a* will be estimating $\ln \alpha$ and then by taking the antilog of *a*, we shall obtain the estimate for α . The regression coefficient $b = \beta$ deserves our special attention. It measures a change in log *Y* per unit of a change in log *X*. In the language of calculus, $\beta = \frac{d \ln Y}{d \ln X}$. Now, a change in the log of some variable implies a proportional change in it. Thus, β is the ratio of a proportional change in *Y* to a proportional change in *X*. in other words, β measures the elasticity of *Y* with respect to *X*. This is the rationale for the log linear regression model being termed as the constant elasticity model.

Example 9.3

This is an example from Gujrati (2003). The table below presents the quarterly data on total personal consumption expenditure (PCExp) and expenditure on durables (ExpDur) for the U.S. economy measured in 1992 billions of dollar for the period 1993:1-1998:3.

Table 9.3: Personal Consumption Expenditure and Expenditure on Durables in U.S.A.

Year	PCExp	ExpDur
1993:1	4286.8	504
1993:2	4322.8	519.3
1993:3	4366.6	529.9
1993:4	4398	542.1
1994:1	4439.4	550.7
1994:2	4472.2	558.8
1994:3	4498.2	561.7
1994:4	4534.1	576.6
1995:1	4555.3	575.2
1995:2	4593.6	583.5
1995:3	4623.4	595.3
1995:4	4650	602.4
1996:1	4692.1	611
1996:2	4746.6	629.5
1996:3	4768.3	626.5
1996:4	4802.6	637.5
1997:1	4853.4	656.3
1997:2	4872.7	653.8
1997:3	4947	679.6
1997:4	4981	648.8
1998:1	5055.1	710.3
1998:2	5130.2	729.4
1998:3	5181.8	733.7

Source: *Damodar Gujrati (2003), Basic Econometrics*

We have plotted the log of expenditure on durables against the log of total personal consumption expenditure in Figure 9.4 below.

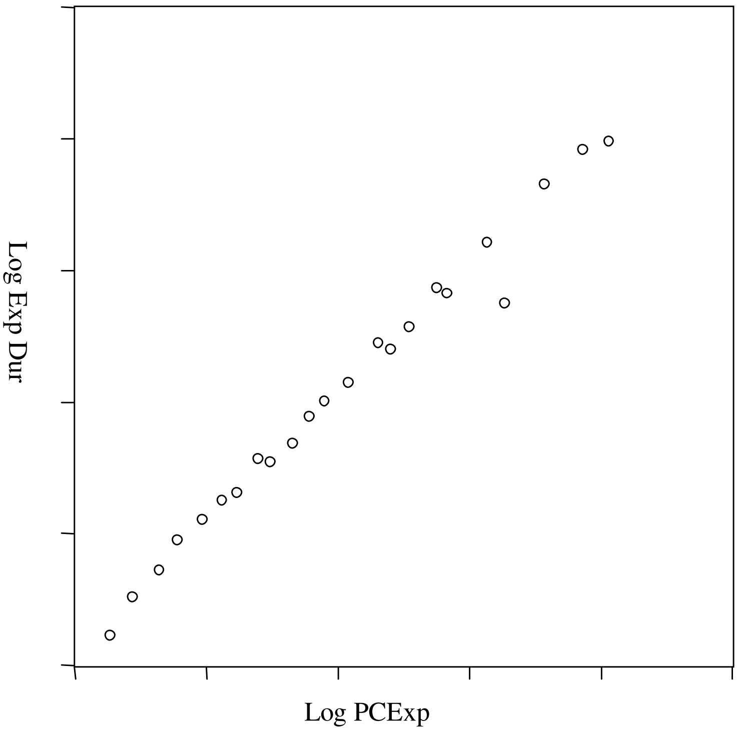


Fig. 9.4: Scatter of Log of Exp on Durables against Log of Personal Con Exp

The scatter above is clearly indicative of a linear relationship between the log of the two variables. As a result, we fit in a log linear model to the data on total personal consumption expenditure and the expenditure on consumer durables. We present some of the results from the fit below.

$$\ln ExpDur = -9.697098 + 1.905633 \ \ln PCExp$$

$$s.e.(\ln \hat{\alpha}) = 0.434127 \qquad s.e.(\hat{\beta}) = 0.051370$$

$$t \ (\ln \hat{\alpha}) = -22.33702 \qquad t \ (\hat{\beta}) = 37.09622$$

$$\mathrm{R}^2 = 0.984969 \qquad d.f. = 21$$

We have an R^2 of about 0.99. This is indicative of a good fit. In this fit, the slope coefficient is an estimate of the elasticity of expenditure on durables with respect to personal consumption expenditure. Thus, $\hat{\beta} = 1.905633$ indicates that the expenditure on durables is highly elastic to total personal consumption expenditure.

3) **Semi-Log Model**

One of the common forms of semi-log regression models is the so-called growth rate model. The model is expressed as

$$\log Y = \alpha + \beta t + U$$

In this model the dependent variable is expressed in log, whereas, the independent variable is the time period, and it is measured in absolute term. Since, log operator is applied only to the left hand side of the equation, the model is called the semi-log model. In this model, the slope coefficient β is a measure of the proportional change in the dependent variable for a unit change in the time period. Thus it measures the growth rate of the dependent variable.

Example 9.4

In this example, we are going to verify the so-called 'Hindu Rate of Growth' of about 3.5 per cent that Indian economy consistently witnessed in 1960s and 1970s.

In Table 9.4, we present India's GDP at 1993-94 prices for the period 1960-1980.

Table 9.4: India's GDP during 1970-1980

Year	GDP	Year	GDP
1960	206103	1971	299269
1961	212499	1972	298316
1962	216994	1973	311894
1963	227980	1974	315514
1964	245270	1975	343924
1965	236306	1976	348223
1966	238710	1977	374235
1967	258137	1978	394828
1968	264873	1979	374291
1969	282134	1980	401128
1970	296278		

Figure 9.5 presents the scatter of the log of GDP against the time period denoted by the letter *t*.

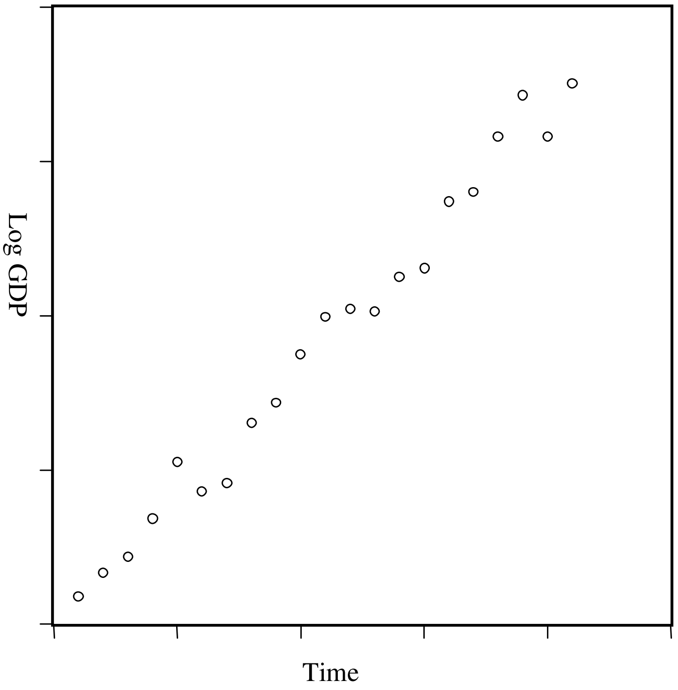


Fig. 9.5: Scatter of India's GDP against Time

The scatter above suggests a linear relationship between the log of GDP and the time period. Therefore, we run a regression of log GDP on time period *t*. we present the result below.

$$\log GDP = 12.19473 + 0.033729\,t$$

$$s.e.(\hat{\alpha}) = 0.012517 \qquad s.e.\left(\hat{\beta}\right) = 0.000997$$

$$t\,(\hat{\alpha}) = 974.2658 \qquad t\left(\hat{\beta}\right) = 33.83607$$

$$\mathrm{R}^2 = 0.983675 \qquad d.f. = 19$$

The value of R^2 is quite high. The estimated slope coefficient of 0.034 indicates a rate of growth of 3.4 per cent during the period 1960-1980. This adequately supports the phenomenon of 'Hindu rate of Growth' in the 60s and the 70s.

4) **Reciprocal Model**

In the reciprocal model, the dependent variable is regressed on the reciprocal of the independent variable. Its functional from is

$$Y = \alpha + \beta \frac{1}{X} + U$$

Although the model is non-linear in variable (because the power of X is –1), it is linear in parameter. Hence, it can be estimated by the OLS method. The model has an important characteristic. As the independent variable X increases indefinitely, the term $\beta \frac{1}{X}$ approaches zero, β being a constant. Consequently, the dependent variable Y approaches a limiting value or an asymptotic value equal to the intercept α . An application of this model can be in the relationship between per capita GNP and the child mortality. As per capita GNP increases, the infant mortality rate is expected to fall but we cannot expect it to fall independently if per capita GNP continues to increase indefinitely. In such a situation, in all probability, the mortality rate will tend to a limiting value. Gujrati (1993) presents an example of applying the reciprocal model to study the relationship between the per capita GNP and the infant mortality rate by using the cross-section data for 64 countries. His results are presented below.

$$Y = 81.79436 + 27237.17 \frac{1}{X}$$

$$s.e.(\hat{\alpha}) = 10.8321 \qquad s.e.\left(\hat{\beta}\right) = 3759.999$$

$$t\,(\hat{\alpha}) = 7.5511 \qquad t\left(\hat{\beta}\right) = 7.2535$$

$$\mathrm{R}^2 = 0.4590$$

where, Y is the child mortality rate (the number of deaths of children under the age of 5 per year per 1000 live births) and X is the per capita GNP at constant prices. The above-mentioned results show that if per capita GNP increases indefinitely, the infant mortality approaches a limiting rate of about 82 deaths per 1000 children born.

Another application of the reciprocal model can be in the estimation of the Phillips curve. This curve proposes an inverse relationship between the change in unemployment rate and that in inflation rate. With a growth in unemployment rate, the growth in inflation rate is expected to fall. In fact Phillips curve postulates that if the unemployment rate increases beyond its so-called natural rate, the growth in the inflation rate should become negative. However, with an indefinite increase in the unemployment rate, the inflation growth cannot be expected to fall indefinitely. It should stabilize at some negative limiting value. Thus, Phillips Curve can be an appropriate case for the use of the reciprocal model.

9.9 CLASSICAL NORMAL REGRESSION MODEL

We have already seen that a two variable classical linear regression model can be presented as

$$Y = \alpha + \beta X + U$$

In this model, *U* is the population disturbance term. We can estimate the unknown parameters of this regression model from the sample information by using the least square or, as it is sometimes called, ordinary least square method. The least square estimates possess some desirable properties if the population disturbance *U* satisfies some five assumptions that we have discussed in Section 9.3. These five assumptions are adequate for the estimation purposes. But the scope of a regression model is just not restricted to the estimation of the parameters. An important purpose of the sample estimates is to test some hypotheses about the unknown population regression parameters with their help. And we can do this if we make some assumption about the distribution of *U*. This we do by making the assumption of normality for *U*. The assumption essentially means that the population regression disturbance term follows normal distribution with mean zero, a constant variance equal to c^2 and a zero covariance. In fact *U* has a conditional distribution, in the sense, that, for each of the given values of the non-stochastic independent variable *X*, we might have a distribution of different values of *U*. This will be clear if we reproduce the diagram of Figure 9.1 with the unknown population regression line fitted into it.

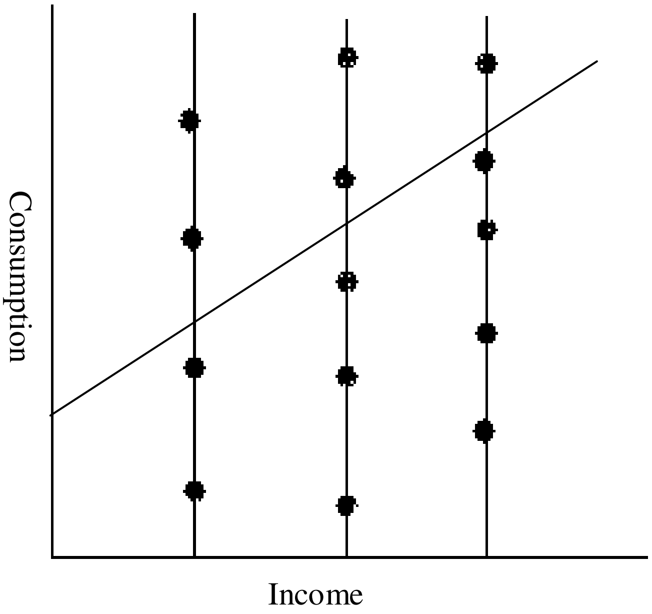


Fig. 9.6: Bi-Variate Population with the Unknown Regression Line

In Figure 9.6, we can see that for a given observed value of income, we can have different observed values of consumption represented by the bold dots on each of the vertical lines. The vertical distance between an observed value of consumption and the corresponding estimated value from the regression line for a given income level measures the value of the disturbance term. Thus there are many possible values of U for each of the given income levels. As a result, there are conditional distributions of U for different levels of income.

The significance of the normality assumption is that these conditional distributions of U should all be independently normally distributed with the same mean equal to zero and, the same variance equal to c^2. Mathematically speaking,

$$E(U_i) = 0 \qquad \text{for all } i\text{s}$$
$$V(U_i) = \sigma^2 \qquad \text{for all } i\text{s}$$
$$Cov(U_i U_j) = 0 \quad \text{for } i \neq j$$

In other words, the population disturbance variable should have independent and identical normal distribution.

If we make this additional normality assumption along with the earlier-mentioned five assumptions about the disturbance term U, then the regression model is called the classical normal regression model.

It may be mentioned here that the normality assumption is important not only for hypothesis testing but also for the fact that under this assumption we can use an alternative procedure for estimating the population regression parameters. This procedure is known as the method of maximum likelihood estimation. Although for regression parameters the method gives the same estimates as the least square estimates, the approach followed is quite different. In fact this procedure has some stronger theoretical characteristics than the least square method. However, we are not discussing this procedure here because it is beyond our scope.

In the next section, we shall discuss the issue of hypothesis testing and see how the normality assumption plays a crucial role there.

9.10 HYPOTHESIS TESTING

Considering the two variable regression model

$$Y = \alpha + \beta X + U$$

We might be interested in examining whether the unknown parameter α or β assumes a particular value or not. This is what is known as hypothesis testing in statistics. We can conduct such a test from the sample estimate of α or β as the case might be. Although we may test some hypothesis about the intercept α, our main concern in regression model is the slope coefficient β. We have the estimated sample regression function

$$\hat{Y} = \hat{\alpha} + \hat{\beta} X + \hat{U}$$

Here, $\hat{\beta}$ is the sample estimate of β and we shall use it for estimating the unknown β. As mentioned earlier that $\hat{\beta}$ might vary from sample to sample collected from the same population. As a result, $\hat{\beta}$ is a random variable with some probability distribution. Now, for the purpose of hypothesis testing, we need to know the form of this distribution of $\hat{\beta}$.

It is here that the assumption of normality helps us. Let us consider it. From Section 9.6, we know that $\hat{\beta} = \sum kY$ and thus, it is a linear function of the observed values of *Y*. But, $Y = \alpha + \beta X + U$. Consequently, we can see, $\hat{\beta} = \sum k(\alpha + \beta X + U)$. Now, we already know that the *k*s, the *X*s and the parameters are all given. As a result, finally, $\hat{\beta}$ becomes a linear function of the random disturbance variable *U*. Thus, $\hat{\beta}$ has the same distribution as *U*. Therefore, the assumption of normality *U*, implies that $\hat{\beta}$ is distributed as normal.

We have already seen that $E(\hat{\beta}) = \beta$ and $se(\hat{\beta}) = \frac{c}{\sqrt{\sum (X - \bar{X})^2}}$. It follows then that

$$\hat{\beta} \sim N\left(\beta, \frac{\sigma}{\sqrt{\sum (X - \bar{X})^2}}\right)$$

You must have learnt from your compulsory course in quantitative methods, that now a standard normal variable can be formed with mean equal to zero and standard deviation equal to one and it can be used for testing of hypotheses regarding β . But the problem here is that the population standard deviation c is unknown and we have to use an estimate for it. However, we know that an unbiased estimate of c is given by $\hat{\sigma}^2 = \frac{\sum \hat{U}^2}{n - 2}$, where, $\hat{U}$ is the sample regression error term and is therefore, computable, and *n* is the size of the sample. Again you must be knowing that when we standardize $\hat{\beta}$ by subtracting its mean from it and divide it by its estimated standard deviation, it no longer follows normal distribution and it in fact follows a *student-t* distribution with *n-2* degrees of freedom. This *student-t* distribution has a mean equal to zero and a standard deviation equal to one. Thus, it is this standard *student-t* distribution that is used for testing of hypothesis about β . We have a table for such a standard *student-t* distribution for different degrees of freedom. And like the standard normal table, this table is put into use for such tests of hypotheses.

Let us see now how we can conduct some test of hypothesis regarding the unknown population regression coefficient by considering some examples that we have already discussed in Section 9.7. Consider Example 9.2. In this example, we have considered the issue of the consumption function in the Indian economy by regressing final consumption expenditure on GDP. To begin with, we might be interested in examining whether the relationship is significant or not. Thus, our focus is on the regression coefficient β and by the nature of the inquiry, we have to conduct a two-tailed test. We can proceed by setting our null hypothesis and the alternative hypothesis in the following manner:

Null hypothesis $\quad H_0 : \beta = 0$
Alternative hypothesis $\quad H_1 : \beta \neq 0$

The acceptance of the null hypothesis implies the rejection of the alternative hypothesis and this in turn implies that on the basis of the sample information there does not seem to be any significant relationship between GDP and the final consumption expenditure. On the other hand, the rejection of the null hypothesis implies the acceptance of the alternative hypothesis and this in turn implies that on the basis of the sample information there seems to exist significant relationship between GDP and the final consumption expenditure. The next step is to construct our test statistic, which is conventionally denoted by *t*. Thus

$$t = \left| \frac{\hat{\beta} - E(\hat{\beta})}{s.e(\hat{\beta})} \right|$$

Under the null hypothesis,

$$t = \left| \frac{\hat{\beta} - 0}{\sqrt{\sum \frac{\hat{U}^2}{n-2}}} \right| = \left| \frac{\hat{\beta}}{\sqrt{\sum \frac{\hat{U}^2}{n-2}}} \right|$$

And this is distributed as a *student-t* with *n-2* degrees of freedom. From the results of the Example 9.2 we can easily compute the value of this *t* statistic since, both the values of the estimated regression coefficient (0.719676) and the standard error of the estimated regression coefficient (0.007865) are mentioned there. In fact, even the computed value of the *t* statistic (91.50314) is given. In addition, the degree of freedom is also given (20). Normally, any test of hypothesis is conducted either at 1 per cent level of significance or at 5% level of significance. From the student-t distribution table we can find out the critical values of the test statistic t for both 1 per cent level of significance and at 5% level of significance at a degree of freedom of 20 for two-tailed test. The values are 2.845 and 2.086 respectively. Thus, the computed value of *t* far exceeds both the critical values. Thus, on the basis of the sample information we cannot accept the null hypothesis. This amounts to accepting the alternative hypothesis. Consequently, personal consumption expenditure does seem to be dependent upon GNP in India during the sample period 1980-2001.

Once the relationship between GDP and personal consumption expenditure has been established some further test can be conducted about the likely value of the marginal propensity to consume out of GNP. For example we can conduct a test regarding whether the marginal propensity to consume is 80 per cent or not. Thus we have

Null hypothesis $H_0 : \beta = 0.80$

Alternative hypothesis $H_1 : \beta \neq 0.80$

Thus our *t* statistic in this case is given by

$$t = \left| \frac{\hat{\beta} - E(\hat{\beta})}{s.e(\hat{\beta})} \right| = \left| \frac{0.719676 - 0.80}{0.007865} \right| = \left| \frac{-0.080324}{0.007865} \right| = 10.212841.$$

Now this value of the test statistic also exceeds the critical values of 2.845 and 2.086 for a degree of freedom of 20 at 1 per cent and 5 per cent levels of significance respectively. Thus, on the basis of the sample information the difference between estimated value of β and its mean is so much that even in 1 out 100 cases or 5 out of 100 cases we do not expect to obtain such a difference. Thus on the basis of the sample information, we are not in a position to accept the null hypothesis. Hence, in all probability during the sample period of 1980-2001, India's marginal propensity to consume out GDP has not been as high as 80 per cent.

In the above example, we considered two-tailed tests of hypotheses. This however, does not rule out the possibility of conducting one-tiled tests. It all depends upon the type of inquiry that we intend conducting.

Thus, a discussion on the test of hypotheses finally marks the end of our discussion on two variable regression model.

Check Your Progress 2

1) When will you use a log linear regression model?

……………………………………………………………………………………

……………………………………………………………………………………

……………………………………………………………………………………

……………………………………………………………………………………

……………………………………………………………………………………

2) How do you interpret the estimated slope coefficient of a log linear regression model?

……………………………………………………………………………………

……………………………………………………………………………………

……………………………………………………………………………………

……………………………………………………………………………………

……………………………………………………………………………………

3) If you want to estimate India's rate of growth of per capita income during the period 1980-2000, what should be the functional form of your regression model?

……………………………………………………………………………………

……………………………………………………………………………………

……………………………………………………………………………………

……………………………………………………………………………………

……………………………………………………………………………………

……………………………………………………………………………………

4) Explain the concept of a classical normal regression model. What is its use?

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

5) Explain briefly the steps for testing the significance of the regression coefficient.

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

9.11 LET US SUM UP

Our focus has been exclusively on the two variable regression model. The purpose of this unit has been to specify what is known as the classical regression model in the literature for the population. It is essentially a linear in parameter regression model. This regression model is stochastic in nature in the sense that here the dependent variable is taken as a random variable as against the independent variable being considered as non-random in nature. The element of stochasticity is introduced by including a random error or disturbance term in the model. Our major concern has been to estimate the unknown parameters of the population regression model from the known sample information and tests some hypotheses about these parameters. In this regression model we usually make five assumptions about the disturbance term in order to effectively estimate the parameters by implementing the ordinary least square procedure. The estimates thus obtained satisfy some desirable properties as enunciated in the Gauss-Markov Theorem. In this connection, it may be mentioned here that if some of these assumptions are not fulfilled, some problems regarding the sample estimates of these unknown parameters arise. These problems and their solutions have not been discussed in this unit. We have also considered different functional forms of the regression model that might be relevant to use in different situations. For the purpose of hypothesis testing, we have to extend the assumptions of the classical regression model to include one more assumption about the probability distribution of the disturbance term. The resultant regression model is then known as the classical normal regression model. We have considered the use of this model for the tests of hypotheses.

9.12 EXERCISES

1) The following table presents the data for broad money supply M_3 (Rs. Crore) for India during 1980-2002. Fit in an appropriate regression model and estimate the rate of growth of money supply in India during this period.

Table 9.5: India's Broad Money Supply 1980-2002

Year	M3
1980	50966
1981	59793
1982	68515
1983	80577
1984	95295
1985	111096
1986	130653
1987	153207
1988	179687
1989	213856
1990	249493
1991	292403
1992	344238
1993	399048
1994	478196
1995	552953
1996	642631
1997	752028
1998	901294
1999	1056025
2000	1224092
2001	1420025
2002	1647976

2) Consider the following hypothetical data. Fit in a log linear regression model and interpret the estimated regression coefficient. Show all the calculations.

Dependent Variable (Y)	Independent Variable (X)
189.8	173.3
172.1	165.4
159.1	158.2
135.6	141.7
132.0	141.6
141.8	148.0
153.9	154.4
171.5	163.5
183.0	172.0
173.2	161.5
188.5	168.6
205.5	176.5
236.0	192.4
257.8	205.1
277.5	210.1
291.1	208.8
284.5	202.1
274.0	213.4
279.9	223.6
297.6	228.2
297.7	221.3
328.9	228.8
351.4	239.0
360.4	241.7
378.9	245.2

Adapted from Maddala (2001)

3) In Example 9.4, the sample estimate of India's rate of growth during the period 1960-80 has been obtained to be about 3.4 per cent per year. Test the hypothesis that during the same period the rate of growth has in fact been 4 per cent per year.

9.13 KEY WORDS

Assumptions of the Classical Regression Model:

1) The disturbance term U has a zero mean for all the values of X, i.e., $E\ (U) = 0$.

2) Variance of U is constant for all the values of X, i.e.,$V(X) = \sigma^2$.

3) The disturbance terms for two different values of X are independent i.e., $\text{cov}(U_i, U_j) = 0, \text{ for } i \neq j$.

4) X is non-stochastic.

5) The model is linear.

Classical Normal Regression Model : A linear regression model, in which, in addition to the five assumptions of the classical regression model, one more assumption of the error term being normally distributed is made.

Classical Regression Model : The conventional regression model whose parameters are estimated by the ordinary least square procedure.

Functional Forms of Linear Regression Model : Different kinds of linear regression models.

Gauss-Markov Theorem : Under the assumptions of the classical regression model, among all the linear unbiased estimators, the least square estimators have the minimum variance. In other words, the least square estimators are the best linear unbiased estimators or BLUE.

Goodness of Fit : The ratio of the explained sum of squares to the total sum of squares. Also known as the coefficient of determination.

Hypothesis Testing : Conducting tests regarding hypotheses made about the unknown parameters of the population regression model with the help of the estimated sample regression function.

Linear Regression Model : A regression model that is essentially linear in parameter.

Non-Deterministic Regression Model : A regression model in which the dependent variable is random or probabilistic but the independent variable is non-random.

Parameters : Unknown intercept and the slope coefficient of a population regression model.

Population Regression Function : A linear regression model with a stochastic error term for a given population.

Sample Regression Function : The sample regression function that is fitted into the sample data for estimating the population regression model.

Standard Errors of Estimate : Standard deviations of the estimated sample regression intercept and slope coefficient.

Two Variable Regression Model : A regression model with one explanatory variable.

9.14 SOME USEFUL BOOKS

Gujrati, Damodar N. (2003); *Basic Econometrics*, McGraw-Hill, New York, U.S.A.

Maddala, G.S. (2002): *Introduction to Econometrics*, John Wiley & Sons (Asia), Singapore.

Pidyck, Robert S. & Rubinfeld, Daniel L. (1991); *Econometric Models & Economic Forecasts*, McGraw-Hill, New York, U.S.A.

9.15 ANSWERS OR HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) See Section 9.2
2) See Section 9.3
3) See Section 9.4
4) See Section 9.6
5) See Section 9.7

Check Your Progress 2

1) See Section 9.8
2) Measures the elasticity of the dependent variable with respect to the independent variable.
3) See Section 9.8
4) See Section 9.9
5) See Section 9.10

9.16 ANSWERS OR HINTS TO EXERCISES

1) 15.9 per cent
2) Do Yourself
3) Do yourself

UNIT 10 MULTIVARIABLE REGRESSION MODELS

Structure

10.0 Objectives

10.1 Introduction

10.2 Regression Model with Two Explanatory Variables

- 10.2.1 Estimation of Parameters
- 10.2.2 Variance and Standard Errors

10.3 Interpretation of Regression Coefficients

- 10.3.1 Goodness of Fit: Multiple Coefficient of Determination: R^2
- 10.3.2 Analysis of Variance (ANOVA)

10.4 Inclusion and Exclusion of Variables

10.5 Generalisation to n-explainatory Variables

10.6 Problem of Multi-co-linearity

10.7 Problem of Hetero-scedasticity

10.8 Problem of Autocorrelation

10.9 Maximum Likelihood Estimations

10.10 Let Us Sum Up

10.11 Key Words

10.12 Answers or Hints for Check Your Progress Exercises

10.0 OBJECTIVES

This Unit aims at equipping the learners with techniques of multiple regressions which come handy in analysing a number of situations where one is required to study impact of a number of independent variables (factors) on some particular dependent variable. After going through this unit, you will be able to:

- have a fairly good comprehension of technique of multiple regression analysis;
- appreciate the need for extending the basic idea of regression as discussed in Unit 8;
- apply those techniques for more realistic model formulation of explaination of economic phenomena; and
- spot and avoid pit falls which may arise (in quantitative analysis).

on accounts of auto-correlation, hetero-scedasticity and multi-co-linearity on the one hand and mis-specification of the model, (in the sense of including irrelevant as well as excluding relevant variables from the model) on the other.

10.1 INTRODUCTION

In Unit 8, you have studied basics of the Classical Linear Regression Model. You regressed dependent variable Y on independent variable X. In other words, you tried to explain changes in Y in terms of the changes in X. You had hypothesised a linear relationship between Y and X of this type:

$$Y = \alpha + \beta X$$

Then, you went on to ‘estimate’ the two constants $\hat{\alpha}$ and $\hat{\beta}$. Once the estimates were ready, you had the estimated model or ‘relationship’ with you given by

$$\hat{Y} = \hat{\alpha} + \hat{\beta}\ x$$

In the present unit we are going to extend this type of analysis further to make it more ‘realistic’ and ‘comprehensive’. We will do it in a number of steps: first of all, we shall introduce one more explainatory variable and re-examine’ the model. Next step will be to generalise the model to n-explainatory variables. At the next stage, we try to interpret the partial regression co-efficients. After that, we will try to examine how large should the number ‘n’ be – that is how many explainatory variables must be included in the model and what should be the statistical touch stone for arriving at such a decision. Then, we examine the conditions, or ‘assumptions’, which make these extensions and generalisations possible. We also examine the possible effects of violation of one or more assumptions. We shall, in particular, pay attention to problems of **multi-co-linearity**, **Hetero-scedasticity** and **auto-correlation**. Then, we take a look at another class of estimates, the Maximum Likelihood Estimators. Finally, we give a brief exposition to some uses of multiple regression analysis in economics.

10.2 REGRESSION MODEL WITH TWO EXPLANATORY VARIABLES

At the outset, please note one small change in notation: in section 10.1 – the introduction to the present unit, we have used the specification of model as

$$Y = \alpha + \beta X$$

However, now onwards, we will write the model in slightly different form:

$Y = \beta_0 + \beta_1 X_1$ This form helps us in presentation in the sense that as and when, we add more explanatory variables X_2, X_3, etc. we simply indicate their coefficients with respective β_1, β_2, β_3, etc.

So the model that we consider in the present section has two explanatory variables X_1, and X_2. The non-stochastic specification will be:

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \qquad (10.1)$$

The same model in stochastic form will have ‘error terms’ included:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \mu \qquad (10.2)$$

$$= E(Y) + \mu \qquad (10.3)$$

We can use subscripts‘t’ with Y_0, X_1, X_2 and μ to denote ‘t^{th}’ observation. Thus, the above equations can be written as

$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t}$

$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \mu_t$ and

$Y_t = E(Y_t) + \mu_t$ respectively

In the relationships above, the component ($\beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t}$) is systematic or deterministic component, which is equal to mean value E(Y_t) — or a point on the regression line.

The component μ_t is non-systematic random component determined by factors other than X_1 and X_2.

The model specified by equations 10.1 and 10.2 is a linear regression model which is linear in parameters.

10.2.1 Estimation of Parameters: The Ordinary Least Squares Approach

The Ordinary Least Squares Approach

We collect a sample of observations on Y_1, X1 and X_2 and write down sample regression function

$$Y_t = b_0 + b_1 X_{1t} + b_2 X_{2t} + e_t \qquad (10.4)$$

Note that b_0, b_1, b_2 have replaced the corresponding population parameters β_1, β_2, β_3 here. The random component μ_t replaced by e_t or the sample error term.

So, simply speaking:

b_0 = the estimator for β_0

b_1 = the estimator for β_1

b_2 = the estimator for β_2

We know from the Unit 8 that the ***principle of ordinary least squares*** selects those values for unknowns (b_0, b_1, b_2) such that residual sum of squares (RSS) = $\sum e_t^2$ is minimum. We can develop this idea by rewriting equation 10.4 as

$$e_t = Y_t - b_0 - b_1 X_{1t} - b_2 X_{2t} \qquad (10.5)$$

Square on both sides and sum up to get RSS:

$$\Sigma e_t^2 = \Sigma \left[Y_t - b_0 - b_1 X_{1t} - b_2 X_{2t} \right]^2 \qquad (10.6)$$

Differentiating the (10.6) w.r.t b_0, b_1, b_2 and equating to zero gives us the three normal equations:

$$\overline{Y}_t = b_0 + b_1 \overline{X}_{1t} + b_2 \overline{X}_{2t} \qquad (10.7)$$

$$\Sigma Y_t X_{1t} = b_0 \Sigma X_{1t} + b_1 \Sigma X_{1t}^2 + b_2 \Sigma X_{1t} X_{2t} \qquad (10.8)$$

$$\Sigma Y_t X_{2t} = b_0 \Sigma X_{2t} + b_1 \Sigma X_{1t} X_{2t} + b_2 \Sigma X_{2t}^2 \qquad (10.9)$$

These three equations give us the following expressions for b$_0$, b$_1$, and b$_2$ respectively:

$$b_0 = \overline{Y} - b_1\overline{X} \; - b_2\overline{X_2} \qquad (10.10)$$

$$b_1 = \frac{\left(\Sigma\, y_t x_{1t}\right)\left(\Sigma\, x_{2t}^2\right) - \left(\Sigma\, y_t x_{2t}\right)\left(\Sigma\, x_{1t}x_{2t}\right)}{\left(\Sigma\, x_{1t}^2\right)\left(\Sigma\, x_{2t}^2\right) - \left(\Sigma\, x_{1t}x_{2t}\right)^2} \qquad (10.11)$$

and

$$b_2 = \frac{\left(\Sigma\, y_t x_{2t}\right)\left(\Sigma\, x_{1t}^2\right) - \left(\Sigma\, y_t x_{1t}\right)\left(\Sigma\, x_{1t}x_{2t}\right)}{\left(\Sigma\, x_{1t}^2\right)\left(\Sigma\, x_{2t}^2\right) - \left(\Sigma\, x_{1t}x_{2t}\right)^2} \qquad (10.12)$$

The lower case letters denote, as usual, the deviations from the respective means. Thus $y_t = (y_t - \overline{Y}), x_{1t} = (X_{1t} - \overline{X})$ and $x_{2t} = (X_{2t} - \overline{X}_2)$..

10.2.2 Variance and Standard Errors

The variances of OLS estimators given in 10.10, 10.11 and 10.12 above are given in terms of means and deviations of X$_1$ and X$_2$ and the variance of population error terms μ$_t$. Thus,

$$Var(\,b_0\,) = \left(\frac{1}{n} + \frac{\overline{X}_1^2\, \Sigma\, x_{2t}^2 + \overline{X}_2^2\, \Sigma\, x_{1t}^2 - 2\,\overline{X}_1\,\overline{X}_2\, \Sigma\, x_{1t}x_{2t}}{\Sigma\, x_{1t}^2\, \Sigma\, x_{2t}^2 - (\,\Sigma\, x_{1t}x_{2t}\,)^2} \right) \sigma^2 \qquad (10.13)$$

Standard Error of $b_0 = \sqrt{Var(\,b_0\,)}$ (10.14)

$$Var(b_1) = \frac{\Sigma\, x_{2t}^2}{\Sigma\, x_{1t}^2\, \Sigma\, x_{2t}^2 - (\Sigma\, x_{1t}x_{2t})^2} .\sigma^2 \qquad (10.15)$$

$$SE(b_1) = \sqrt{Var(b_1)} \qquad (10.16)$$

and

$$Var(b_2) = \frac{\Sigma\, x_{1t}^2}{\Sigma\, x_{1t}^2\, \Sigma\, x_{2t}^2 - (\Sigma\, x_{1t}x_{2t})^2} .\sigma^2 \qquad (10.17)$$

$$SE(b_2) = \sqrt{Var(b_2)}$$

(10.18)

When σ^2 is unknown and its t unbiased OLS estimator is obtained, we find that

$$\hat{\sigma}^2 = \frac{\Sigma\, e_i^2}{n-3} \qquad (10.19)$$

The denominator of 10.19 shows the degrees of freedom. In a sample of size 'n' we exhaust 3 degrees of freedom in estimating b$_0$, b$_1$, and b$_2$. Note that this reasoning is quiet general, in the sense that if out of a sample of size 'n', we estimate ‘k’ parameters, remaining degrees of freedom will be n — k. We call $\hat{\sigma} = \sqrt{\hat{\sigma}^2}$, the standard error of estimate/regression.

The residual sum of squares in 10.19 is $\Sigma e_t^2 = \Sigma(Y_t - \hat{Y}_t)^2$. This expression is equivalent to the following.

$$\Sigma e_t^2 = \Sigma y_t^2 - b_1 \Sigma y_1 x_{1t} - b_2 \Sigma y_t x_{2t} \qquad \text{(10.20}$$

10.3 INTERPRETATION OF REGRESSION COEFFICIENTS

Mathematically, b$_1$ and b$_2$ represent the partial slopes of regression plan with respect to X$_1$ and X$_2$ respectively. In other words, b$_1$ shows the rate of change in Y as X$_1$ alone undergoes a unit change, keeping all other thing, constant. Similarly, b$_2$ represents rate of change of Y as X$_2$ alone changes by a unit while other things are held constant.

Note: *This interpretation is quite similar to what you have been doing in economic theory. Recall the law of demand: quantity demanded of a commodity varied inversely with its price, ‘ceteris paribus’ that is when other things were kept ‘constant’. What were those other things? Those were prices of complements and substitutes, income of the consumer, and tastes/likings/ disliking etc.*

10.3.1 Goodness of Fit: Multiple Coefficient of Determination:R^2

Recall that is case of a single independent variable, r^2 measured goodness of fit of the fitted sample regression line. It gave you the percentage total variation in dependent variable Y which has been explained by the single explanatory variable X. Correspondingly, when we have two explanatory variables X$_1$ and X$_2$, we would like to know the proportion of total variation in $Y = \Sigma y_t^2$ explained by X$_1$ and X$_2$ jointly. This information is conveyed by R^2 — the multiple coefficient of determination. Thus,

$$R^2 = \frac{ESS}{TSS}$$

When ESS is explained sum of squares and TSS is total sum of squares. You are familiar with the relationship between TSS, ESS and RSS from Unit 8.

You know that $ESS = b_1 \Sigma\ y_t x_{1t} + b_2 \Sigma\ y_t x_{2t}$

and $RSS = \Sigma y_t^2 - b_1 \Sigma y_t x_{1t} - b_2 \Sigma y_t x_{2t}$

Therefore R^2 can be computed as

$$R^2 = \frac{b_1 \Sigma\ y_t x_{1t} + b_2 \Sigma\ y_t x_{2t}}{\Sigma y_t^2} \qquad (10.21)$$

You can get the quantities in the above formula from your computation sheet for the normal equations. The R^2 lies between 0 and 1. Closer it is to one;

better is the fit – implying estimated regression line is capable of explaining greater proportions of variation in Y. The positive square root of R^2 is called coefficient of multiple correlation.

10.3.2 Analysis of Variance (ANOVA)

We know the relationship:

TSS = ESS +RSS

This is equivalent to saying:

$$\underset{(TSS)}{\Sigma y_t^2} = \underset{(ESS)}{b_1 \Sigma y_t x_{1t} + b_2 \Sigma y_t x_{2t}} + \underset{(RSS)}{\Sigma e_t^2}$$

A study of the components of TSS is called analysis of variance in the context of regression. You know that every sum of squares has some degrees of freedom (d*f*) associated with it. In our above 2-explainatory variable case, the degrees of freedom will be

TSS = n–1

RSS = n–3

ESS = 2

We can put this information in form of a table:

Table 10.1: ANOVA Table for 2-Explainatory Variable Regression

Source of Variation	Sum of Squares (SS) *df*	Mean S.S = $\frac{\acute{S}S}{\text{d}f}$
Due to regression (ESS)	$b_1 \sum y_t x_t + b_2 \sum y_t x_{2t}$ 2	$\frac{ESS}{2}$
Due to residuals	$\sum e_t^2$ n-3	$\frac{\sum e_t^2}{n - 3}$
TSS	$\sum y_1^2$ n-1	

One may be interested in testing a null hypothesis H_0: $B_1 = B_2 = 0$. In such a case we find that

$\frac{ESS / df}{RSS / df}$ that is ratio of the variance explained by X_1 and X_2 to unexplained variance follows F distribution with 2 and n-3 degrees of freedom. In general, if regression equation estimates 'k' parameters including intercept, than F ratio has (k-1) d*f* in numerator and (n-k) d*f* in the denominator.

***Interpretation*:** Larger the variance explained by fitted regression line, larger the numerator will be in relation to denominator. Thus a larger F value is evidence **against** the 'truthfulness' of H_0: $B_1 = B_2 = 0$. That is, F values larger than one will indicate that hypothesis of both the variables X_1 and X_2 having no effect on Y cannot sustain.

We can also express F values in terms of R^2.

$$F = \frac{R^2/(k-1)}{(1-R^2)/(n-k)}$$

Check Your Progress 1

1) What is multiple regression? Explain with an example.

………………………………………………………………………………

………………………………………………………………………………

………………………………………………………………………………

………………………………………………………………………………

………………………………………………………………………………

2) State the assumptions of classical linear multiple regression model.

………………………………………………………………………………

………………………………………………………………………………

………………………………………………………………………………

………………………………………………………………………………

………………………………………………………………………………

3) What do you mean by linearity of regression model? Explain?

………………………………………………………………………………

………………………………………………………………………………

………………………………………………………………………………

………………………………………………………………………………

………………………………………………………………………………

4) How do you interpret coefficients of multiple regression model?

………………………………………………………………………………

………………………………………………………………………………

………………………………………………………………………………

………………………………………………………………………………

………………………………………………………………………………

………………………………………………………………………………

10.4 INCLUSION AND EXCLUSION OF EXPLANATORY VARIABLES

The Adjusted $R^2 : \overline{R^2}$

It has been noted that as we add more and more explanatory variables X_s , the explained sum of squares keeps on rising. Thus, R^2 goes on increasing as we increase X_s, the explained sum of squares keeps on rising. Thus, R^2 goes on increasing as we increase X_s. But, notice that adding each additional variable 'eats up' one degree of freedom and our definition of R^2 makes no allowance for loss of degrees of freedom. Hence, the thinking that you can improve the goodness of fit by suitably increasing the number of variables may not be justified — We know that TSS always has (n-1) degrees of freedom. Therefore, comparing two regression models with same dependent variable but differencing number of independent variables will not be justified. Therefore, we must adjust our measure of goodness of fit for degrees of freedom. This measure is called adjusted R^2, denoted by $\overline{R^2}$. It can be derived from R^2 as follows:

$$\overline{R}^2 = 1 - (1 - R^2)\frac{(n-1)}{(n-k)} \qquad (10.22)$$

Therefore, it is recommended that, one must include new variables only if (upon inclusion) $\overline{R^2}$ increases and not otherwise. A general guide is provided by 't' statistic, if absolute value of the co-efficient of added variable is greater than one, retain it (Note that 't' value is calculated under the hypothesis that population value of that co-efficient is zero).

10.5 GENERALISATION TO N-EXPLANATORY VARIABLES

In general, our regression model may have a large number of independent variables. Each of those variables can, on priority grounds, be expected to have some influence over the 'dependent' or 'explained' variable. Consider a very simple example. What can be determinants of demand for potatoes in a vegetables market? One obvious choice will be the price of potatoes. What else can affect the quantity demanded? Could it be availability of vegetables which can be paired off with potatoes? In that case, prior of a large number of vegetables which are cooked along with potatoes will become 'relevant explanatory variables'. You cannot ignore income of the community that patronizes the particular market. Needless to say, the dietary preferences of members of the households can also affect the demand and so on. In the next Unit, we shall discuss techniques which help us restrict the analysis to a selected 'few variables, though theoretic considerations may find a huge number of them to be 'useful' and 'powerful' determinants. In fact, in economic theory, we usually append the phrase **Ceteris paribus**, with many a statements. This phrase means keeping all other things constant. That means, we may focus on impact of only a few selected variables on the dependent variable while assuming that all other variables remain 'unchanged' during the period of analysis. This assumption may not hold so, we have to juggle the need to include more and more variables in our model with the 'gains' in explainatory power of the model. We have developed, in previous section (10.4) a working touchstone for inclusion of more variables in terms of improvement in $\overline{R}^2$ and have tried to give it 'practical' shape in form of magnitude of 't' values of the relevant slope parameters.

With these considerations in mind we can generalise the linear regression model as follows:

We hypothesize that in population, the dependent variable Y depends upon n explainatory variables, X_1, X_2, ………..X_n. We also assume that the relationship is linear in parameters. Three more assumptions are made and they have very significant bearing on the analysis. These are:

a) Absence of Multi-co-linearity;

b) Absence of Hetero-scedasticity; and

c) Absence of Autocorrelation

We will discuss the complications, which arise because of violations of these assumptions in section 10.6, 10.7 and 10.8 respectively. So our Classical Linear General Regression Model is :

$$Y_t = \beta_0 + \beta_1 X_{1t} + \cdots\cdots\cdots\cdots\cdots + \beta_n X_{nt}$$

or

where it is understood that Y and each X will have a large numbers of values t=1…………N, forming (n+1) 'tuples' (Y_{11} X_{11}, X_{21}, ……………..X_{n1}).

We can simply write

$$Y = \beta_1 X_1 + \beta_2 X_2 \ldots\ldots\ldots\ldots\ldots\ldots \beta_n X_n + \mu \qquad (10.23)$$

in Matrix equation form, we get

$$Y = X\beta + U \qquad (10.24)$$

Where $Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}, \beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_n \end{bmatrix}$ and $U = \begin{bmatrix} U_1 \\ \vdots \\ U_n \end{bmatrix}$

also $X = \begin{bmatrix} X_{11}, X_{21}, X_{31}, \cdots\cdots\cdots\cdots\cdots X_{k1} \\ X_{12}, X_{22}, X_{32}, \cdots\cdots\cdots\cdots\cdots X_{k2} \\ \vdots \\ \vdots \\ X_{1n}, X_{2n}, X_{3n}, \cdots\cdots\cdots\cdots\cdots X_{kn} \end{bmatrix}$

We assume that (1) expected values of error terms are equal to zero; that is $E(u_i) = 0$ for all 'i'. In matrix notation

$$E(u) = \begin{bmatrix} \mathrm{E}(u_{\mathrm{i}}) \\ \vdots \\ E(u_n) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = 0$$

2) The error terms are not correlated with one another and they all have same variance for σ ^{2} all sets values of the variables X. That is,

$$E(u_i u_j) = 0; \ \ \forall i \neq j \text{ and}$$
$$E(u_1^2) = \sigma^2 \forall i$$

in matrix notation:

$$E\left[UU^{'}\right] = \begin{bmatrix} E(u_1^2), E(u_1u_2) \cdots\cdots\cdots\cdots\cdots E(u_1u_n) \\ E(u_1u_2), E(u_2^2) \cdots\cdots\cdots\cdots\cdots E(u_2u_n) \\ \vdots \\ E(u_1u_n), E(u_2u_n) \cdots\cdots\cdots\cdots E(u_n^2) \end{bmatrix} = \begin{bmatrix} \sigma^2 & 0 & 0 & 0 \\ 0 & \sigma^2 & \sigma & 0 \\ 0 & 0 & \sigma^2 & 0 \\ 0 & 0 & 0 & \sigma^2 \end{bmatrix} = \sigma^2 I_n$$

3) Number X_{1i}……………..X_k are all real numbers and are free of random errors or disturbances.

4) The matrix X has been linearly independent columns. It implies that number of observations exceeds number of co-efficient to be estimated. It also implies there exists no exact linear relationships between any of the X variables.

Note: Assumptions that $E(U_iU_j) = 0$ means that error terms are not correlated. The implication of **diagonal** terms in matrix E(UU$'$) being all equal to c^2 is that all error terms have same variance, c^2. This is also called assumption of homo-scedasticity. The last assumption implies absence of multi-co-linearity. We can write the regression relation for the sample as:

e = Y – Xb

where *e, Y, X* and *b* are appropriate matrices.

Sum of squared residuals will be

$$\phi = \sum e_1^2 = \sum (Y_i - b_1 X_{1i} \cdots\cdots\cdots + b_k X_{ki})^2 \quad (10.25)$$
$$= e'e = [Y - Xb]'[Y - Xb]$$
$$= Y^1Y - 2b'X'Y + b'X'Xb$$

Note: b' X$'$Y is scalar and is therefore equal to its transpose Y$'$Xb.

by equating 1st order partials of ϕ w.r.t each b_i, to zero, we get *k* normal equations. This set of equations in matrix form is:

$$\frac{\partial\phi}{\partial b} = -2X'Y + 2X'Xb = 0 \qquad (10.26)$$

$$X'Xb = X'Y \qquad (10.27)$$

when X has rank equal to *k,* the normal equation 10.27 will have a unique solution and least squares estimator *b* is equal to:

$$b = [X'X]^{-1}[X'Y] \qquad (10.28)$$

We have assumed b to be estimator for β and thus E(b) = β , therefore we can rewrite 10.28 as

$$b = [X'X]^{-1} X''[X\beta + u]$$
$$= [X'X]^{-1} X'X\beta + [X'X]^{-1} X'U$$
$$= \beta + \left[X^1X\right]^1 X^1U$$
$$\therefore E(b) = E(\beta) + E\left[\left[X^1X\right]^{-1} X^1E(\mu)\right] = E(B) + [X'X].X'E(\mu)$$
$$= \beta$$

Variance of $b = \sigma^2 (X'X)^{-1}$

Notes

1) In this course our objective is simply to introduce the concepts. Those who plan to pursue the concepts at much more rigorous level can study our course on Econometric Methods (MEC) – included as optional course of M.A.(Economics) Programme.

2) The other ideas regarding coefficient of determination R^2 and adjusted R^2 remain the same as they were developed for two explainatory variable case.

Now we can safely turn to discussions about non-satisfaction or violation of assumptions.

10.6 THE PROBLEM OF MULTI-CO-LINEARITY

Many a times our X variables may be found to have some other linear relationships among themselves. This vitiates our classical regression model.

Let us illustrate it with help of our 2 explainatory variable model.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + U_1$$

Let us give specific names to variables, say, X$_1$ is price of commodity Y and X$_2$ is family income. We expect β$_1$ to be negative and β$_2$ to be positive. Now we go one step further. Let Y be demand for milk, X$_1$ be price of milk and let the family wise demand for milk is being estimated for a family, which produces and sells milk! Clearly, larger the value of X$_1$ higher the magnitude of X$_2$ will be.

In such situations, the estimation of co-efficient will not be possible. Recall, we wanted variables X in our matrix equations to be linearly independent. If that conditions is not satisfied X matrix becomes singular, that is its determinant tends to vanish. Thus, there will be **no solution** to the normal equation **10.26 (or 10.27).**

However, if co-linearity not perfect, we can still get OLS estimates and they remain best linear unbiased estimates (BLUE) – though one or more partial regression co-efficient may turn out to be individually insignificant.

Not only this, the OLS estimates still retain property of minimum variance. Further, it is found that multi co-linearity is essentially a sample regression problem. The X variables may not be linearly related in population but some of our suppositions while drawing a sample may create a situation of multiple linear relations.

The practical consequences of multi-co-linearity. Gujarali, (D.N.) has listed the following consequences of multiplicity of linear relationships:

1) Large variances /SEs of OLS estimates
2) Wider confidence intervals
3) Insignificant 't' ratios for β parameters
4) A high R^2 despite few significant t values
5) Instability of OLS estimators: The estimators and their standard errors (SEs) become very sensitive to small changes in data.
6) Sometimes, even signs of some of the regressions may turn out to be theoretically unacceptable like a rise in income having negative impact on demand for milk.
7) When many regressions have insignificant coefficients, their individual contributions to the explained sum of squares cannot be assessed properly.

The multi-co-linearity can be detected by:

1) high R^2 but few significant 't' ratios,

2) high pair wise correlation between explainatory variables. One can try partial correlations, subsidiary or auxiliary regressions as well. But each such technique increases burden of calculations.

Check Your Progress 2

1) When do you decide to drop a variable from the regression equation? Why?

………………………………………………………………………………

………………………………………………………………………………

………………………………………………………………………………

………………………………………………………………………………

2) When do you include more variable(s) into your model? Why?

………………………………………………………………………………

………………………………………………………………………………

………………………………………………………………………………

………………………………………………………………………………

3) "Inclusion of more variables always increases R^2 the goodness of fit. So to make a regression model 'good' what we need to do is simply increase the number of explanatory variables". Do you agree/disagree with this statement? Give reasons.

……………………………………………………………………………

……………………………………………………………………………

……………………………………………………………………………

……………………………………………………………………………

……………………………………………………………………………

4) What is multi-co-linearity? What are its consequences?

……………………………………………………………………………

……………………………………………………………………………

……………………………………………………………………………

……………………………………………………………………………

10.7 PROBLEM OF HETERO-SCEDASTICITY

The Classical Linear Regression Model has a significant underlying assumption in homo scedasticity, that is, all the error terms are identically distributed with mean equal to zero and standard deviation equal to σ (or variance equal to σ2). σ2 What happens when this second part of assumption regarding distribution of variance does not hold? As a result, in symbolic terms $E(u_i)^2 = \sigma_i^2$, that is, if the expectation of squared errors is no longer equal to σ^2 — each error term has its own σ^2, or variance varies from observation to observation.

It has been observed that usually time series data does not suffer from this problem of hetero scedasticity but in cross-section data, the problem may assume serious dimensions. **The consequences of hetero scedasticity:** If the assumption of homoscedasticity does not hold, we observe the following impact on OLS estimators.

1) They are still linear

2) They are still unbiased

3) But they no longer have minimum variance – that is we cannot call them BLUE: the Best Linear Unbiased Estimators. In fact, this point is relevant both for small as well as large samples.

4) Reason for this problem hinted at in (3) above is that generally, OLS estimators have some bias built into their formulae. We try to rectify that making use of degrees of freedom.

   For instance $\hat{c}^2$, (the estimator for true population σ^2) given by $\sum e_1^2 / df$ no longer remains unbiased. And this very $\hat{c}^2$ enters into calculation of standard errors of OLS estimates.

5) Since, estimates of standard errors are themselves no longer reliable, we may end up drawing wrong conclusions using conventional reasoning based on procedures for testing the hypothesis.

How to detect Hetero scedasticity

In applied regression analysis, plotting the residual terms can give us important clues about whether or not one or more assumptions underlying our regression model hold. The pattern exhibited by e_i^2 plotted against the concerned variable can provide important clue. If no pattern is detected – homoscedasticity holds or hetero scedasticity is absent. On the other hand, if errors form a pattern with variable — expanding, increasing linear or changing is some non-linear manner thus hetero scedasticity is certainly present.

Some tests have been designed to detect presence of Hetero scedasticity, using various statistical techniques. Prominent ones are: Park Test, Glejser Test, Whites General Test, Spearman's Rank correlation Test, Goldfield - Quadnt Test etc. But in this unit, the limitation of space does not permit us to go into their details. We are forced to refer the learners again the course on Econometric Method for details in this regard.

How to tackle the hetero scedasticity?

Our ability to tackle the problem will depend upon the assumptions. We can really make about error variance. Thus, the following situations may emerge

i) When σ_1^2 is known

Here the CLRM

$Y_i = \beta_0 + \beta_1 X_1 + u_i$ can be transformed, dividing each value by corresponding σ_1 thus,

$$\frac{Y_i}{\sigma_1} = \beta_0\left(\frac{1}{\sigma_1}\right) + \beta_1\left(\frac{X_i}{\sigma_i}\right) + \frac{U_i}{\sigma_i}$$

This effectively transforms error terms to $U_i = \frac{U_i}{\sigma_1}$ which is homo-scedastic and therefore, the OLS estimators will be free of disability caused by hetero scedasticity. The estimates of β_0 and β_1 in this situation are called **Weighted Least Squares Estimators (WLSEs)**.

ii) **When σ^2 is unknown:** we make some further assumptions about error variance:

iia) Error variance proportional to the X_i s. Here, the Square Root transformation is enough. We divide on both sides by $\sqrt{X_i}$. Thus, our regression line looks like:

$$\frac{Y_i}{\sqrt{X_i}} = \frac{\beta_0}{\sqrt{X_i}} + \beta_1 \frac{X_i}{\sqrt{X_i}} + \frac{U_i}{\sqrt{X_i}}$$

$$= \beta_0 \frac{1}{\sqrt{X_i}} + \beta_1 \sqrt{X_i} + U_i$$

Here $\text{U}_i = \frac{U_i}{\sqrt{X_i}}$ and this is sufficient to address the problem.

iib) Error Variance proportional to $\text{X}_1{}^2$. Here, instead of division by $\sqrt{X_i}$, we divide by X_i on both the sides and estimate

$$\frac{Y_i}{X_i} = \beta_0 \frac{1}{X_i} + \beta_1 + \frac{U_i}{X_i}$$

$$= \beta_0 \frac{1}{X_i} + \beta_1 + U_i$$

The error term will be $U_i = \frac{U_i}{X_i}$ and this will be free of hetero scedasticity, facilitating use of CLS techniques.

iii) ***Respecification of Model*:** Assigning a different functional form to the model, in place of speculating about the nature of variance may be found expedient. We can estimate this model:

$$\ln Y_i = \beta_0 + \beta_1 \ln X_i + U_i$$

This loglinear model is usually adequate to address our concerns.

10.8 PROBLEM OF AUTOCORRELATION

The classical regression model also assumes that disturbance terms U_is do not have any serial correlation. But, in many situations this assumption may not hold. The consequences of the presence of serial or auto correlation are similar to those of hetero scedasticity: the OLS are no longer BLUE.

Symbolically **no** autocorrelation means $E(U_i\ U_j)= 0$ when $i \neq j$. Autocorrelation can arise in economic data on account of many factors:

i) there may be various reasons which cause cyclical up and down swings in economic time series. These tendencies continue till something happens which reverse them. This is called **inertia**.

ii) Misspecification of model in the form of under specification can be a cause of autocorrelation: you have fewer X variables in the model, leaving out rather large systematic components to be clubbed with errors.

iii) Cob-web phenomenon is another factor which may create this problem in certain types of economic time series (especially agricultural output etc.).

iv) ***Polishing of data*** – like adding monthly data to make quarterly or quarterly to **make** half yearly series etc. can also be responsible for the autocorrelation – as the averaging involved dampens the fluctuations of the original data.

The consequences of autocorrelation are not different from those of hetero scedasticity listed in 10.7 above. Here too OLS estimators are biased or are not BLUE, *t* & F tests are no longer reliable. Therefore, computed value of R^2 is not reliable estimate of true goodness of fit.

There are many tests for detecting autocorrelation – varying from visual inspection of error plots, the Runs Test, Swed-Eisenhart critical runs test. But most commonly used in Durbin-Watson *d* test defined as

$$d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}$$

However, again, we are holding back information on practical detections and avoidance of problem of autocorrelation for the reasons of limitation of space here.

10.9 THE MAXIMUM LIKELIHOOD ESTIMATIONS

Sometimes another class of estimators is used in place of OLS. This class of estimators is called ***maximum likelihood estimators*** (MLEs). The MLEs possess some stronger theoretical properties. But it also requires a stronger assumption about distribution of error terms. Moreover, when errors follow normal distribution, we find that OLS and MLE methods give identical estimates of β parameters, both in simple and in multiple regressions. However, MLE estimate of σ^2 is biased. Hence, if one uses assumption of normal distribution of U_is and persists with OLS, one does not miss out on any thing that may be advantageous in MLE.

The Method: Simple Regression Illustration

In the model $Y_1 = \beta_0 + \beta_1 X_1 + U_i$, the Y$_1$ is normally and independently distributed with means $\beta_0 + \beta_1 X_i$ and variance σ^2. Therefore, the joint probability density function of

Y$_1$, Y$_2$, Y$_3$……………………Y$_n$ with above mean and variance will be

$$f\left(Y_1 \ldots\ldots\ldots\ldots Y_n / \beta_0 + \beta_1 X_i, \sigma^2\right) \qquad \text{ML - 1}$$

But given the independence of Y$_s$, this function can be written as product of ‘*n*’ individual density functions, or

$$f\left(Y_1 \ldots\ldots\ldots\ldots Y_n / \beta_0 + \beta_1 X_i + U_i, \sigma^2\right) \qquad \text{ML-2}$$

$$= \pi\left[Y_i / \beta_0 + \beta_1 X_i + U_i, \sigma^2\right]$$

$$\text{where} \quad f(Y_i) = \frac{1}{\sigma\sqrt{2\pi}} . e^{\left[-Y_2 \frac{(Y_i - \beta_0 - \beta_1 X_i)^2}{\sigma^2}\right]} \qquad \text{ML-3}$$

Putting this value for each Y$_i$, in ML-2. We get the likelihood function (LF):

$$LF(\beta_0,\beta_1,\sigma^2) = \frac{1}{\sigma^n\left(\sqrt{2\pi}\right)^n}.e^{\left[-\frac{1}{2}\sum\frac{(Y_i-\beta_0-\beta_1-U_i)^2}{\sigma^2}\right]} \qquad \text{ML - 4}$$

The method of maximum likelihood estimation is nothing but maximization of the (LF) given in ML-4 above. We can differentiate logarithms of (LF) with respect to β$_0$, β$_1$ and σ^2 and equate the respective partials to zero to get the requisite estimation relations. So as

$$\ln(LF) = n\Big/ n\sigma^2 - \frac{n}{2}\ln(2\pi) - \frac{1}{2}\sum\frac{(Y_i-\beta_0-\beta_1X_1-U_i)^2}{\sigma^2} \qquad \text{ML - 5}$$

therefore

$$\frac{\partial\ln(\mathrm{LF})}{\partial\beta_0} = \frac{-1}{\sigma^2}\sum(Y_1-\beta_0-\beta_1X_i)(-1)=0 \qquad \text{ML - 6}$$

$$\frac{\partial\ln(\mathrm{LF})}{\partial\beta_1} = \frac{-1}{\sigma^2}\sum(Y_1-\beta_0-\beta_1X_1)(-X_i)=0 \qquad \text{ML - 7}$$

$$\frac{\partial\ln(\mathrm{LF})}{\partial c^2} = \frac{-n}{2c^2} + \frac{1}{2c^4}\sum(Y_1-\beta_0-\beta_1X_i)^2=0 \qquad \text{ML - 8}$$

Simplification of ML-6 and ML-7 give us

$$\Sigma Y_i = n\beta_0 + \beta_1\,\Sigma X_i \qquad \text{ML-8a}$$

$$\Sigma X_iY_i = \beta_0\,\Sigma X_i + \beta_1\,\Sigma X_i^2 \qquad \text{ML-9}$$

Which are same as OLS normal equations.

Substituting values of ML (=OLS) estimates obtained by simultaneously solving ML-8 and ML-9 into ML-7 gives us

$$\sigma^2 = \frac{1}{n}\sum(Y_i-\beta_0-\beta_1X_i)^2$$

$$= \frac{1}{n}\sum U_i^2 \qquad \text{ML - 10}$$

Expectation of $\tilde{\sigma}^2, E(\tilde{\sigma}^2) = \frac{1}{n}E(\Sigma U_i^2)$

$$= \left(\frac{n-2}{n}\right)\sigma^2$$

$$\sigma^2 - \frac{2}{n}\sigma^2$$

or $\tilde{c}^2$ is biased downwards.

Multiple Regression: An example: The following regression were run on SPSS. Edited summary of results is as follows:

India’s Imports, Exports and Foreign Investment Inflows

All figures are in Millions of US dollars

Year	var00001	var00002	var00003
1991-1992	19411.00	17865.00	133.00
1992-1993	21882.00	18537.00	559.00
1993-1994	23306.00	22238.00	4153.00
1994-1995	28654.00	26330.00	5138.00
1995-1996	36678.00	31797.00	4892.00
1996-1997	39133.00	33470.00	6133.00
1997-1998	41484.00	35006.00	5385.00
1998-1999	42389.00	33218.00	2401.00
1999-2000	49671.00	36822.00	5181.00
2000-2001	50536.00	44560.00	5862.00
2001-2002	51413.00	43827.00	6693.00
2002-2003	61412.00	52719.00	4555.00

var00001=Imports

var00002=Exports

var00003=Foreign Investment Inflow

Regression of Imports on Exports and Foreign Investment Inflow

R	R^2	$\bar{R}^2$	Standard error of Estimate
0.983	0.966	0.958	2726.228

Coefficients

	B	Standard Error	t
Constant	-1655.737	2658.578	-0.623
var00002	1.263	0.104	12.192
var00003	-0.288	0.522	-0.552

Coefficients Correlation

Correlation		Var00003	Var00002
	Var00003	1.0000	-0.670
	Var00002	-0.670	1.0000
Covariance	Var00003	0.272	-3.625E-02
	Var00002	-3.625E-02	1.073E-02

ANOVA

Sum of Sqrs.	*df*	Mean sqrs.	F
Regression1.9E+09	2	9.4E+08	127.097
Residual 6.7E+07	9	7432320	
Total 2.0E+09	11		

Regression of Foreign Investment Inflow on Exports and Imports

R	R^2	$\overline{R}^{-2}$	Standard error of Estimate
0.684	0.468	0.349	1712.334

Coefficients

	B	Standard Error	t
Constant	-321.545	1702.075	-0.189
var00001	-0.114	0.206	-0.552
var00002	0.272	0.257	1.060

Coefficients Correlation

Correlation		**Var00002**	**Var00001**
	Var00002	1.0000	-0.982
	Var00001	-0.982	1.0000
Covariance	Var00002	6.590E-02	-5.92E-02
	Var00001	-5.192E-02	4.24E-02

ANOVA

	Sum of Sqrs.	*df*	Mean sqrs.	F
Regression	2.3E+07	2	1.2E+07	3.952
Residual	2.6E+07	9	2932089	
Total	5.0E+07	11		

The above regressions are based on certain 'a prior' expectations. We expect that imports into India during the period 1991-92 to 2002-03 depend upon exports from India and foreign investment inflows into the country. The idea is that exports pay for imports and foreign investment inflow 'facilitates' the country to import more. Our results confirm our theoretic expectations. Regression on imports on exports and investment inflows shows that R^2 = 0.966 and $\overline{R}^2 = 0.958$. This shows that model has high explanatory power. It explains over 95 per cent of the variation. Typical computer output gives information on coefficients, correlation and co-variances, co-linearity diagnostics, residual analysis of variance etc. We have run another regression too – the results of which are reported above. This is regression of foreign investment inflows on exports and imports. Theoretic expectation was that investment inflows are determined by magnitudes of exports and imports. However, this model gives rather disappointing results. $R^2 = 0.468$ and $\overline{R^2} = 0.349$ only. In other words, our model explains less than 35 per cent of the variations. It is not desirable to persist with it. The same is reflected in low values of β_1 & β_2 parameters and their high standard errors.

Check Your Progress 3

1) What is hetero-scedasticity? What are its consequences?

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

2) What is auto-correlation? When does it arise? What are its consequences?

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

3) What are maximum likelihood estimations? Why do we persist with least squares estimators most of the time.

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

10.10 LET US SUM UP

We began with extension of the simple linear regression model to incorporate one more explanatory variable. Our next step was to interpret the coefficients of partial regression. Afterwards, we tried to design statistical touch stone for inclusion of more variables into the model. This also gave us some guidelines to drop the 'undesirable' variable as well. We then moved on to more tedious and mathematically more demanding extension to 'n' variable model. We then attempted to analyse the effects of multi-co-linearity, hetero scedasticity and auto-correlation respectively. We also made comments on techniques of identification of these problems and some strategies to get rid of the problems as well. Further, we discussed the class of estimators called maximum likelihood estimators (MLEs). We made comparison between MLEs and Least square estimators as well. Finally, we ran some regressions on SPSS package between India's imports, exports and foreign investment inflows to illustrate some of the points, which arose in course of our discussions. However, we must again draw the attention of learners to the fact that limitations of space did not permit us to go in for an exhaustive discussion of the concepts touched upon. Those who want to have detailed knowledge about the issues and concepts involved in multiple regression models are advised to go through our optional course MECE-001 Econometric Methods.

10.11 KEY WORDS

Multiple Regressions : Regression of one dependent variable on more than one independent variables.

Partial regression Coefficients : Coefficients of individual predictors in multiple regressions.

Coefficient of multiple determination — R^2 : Ratio of variation explained or accounted for by the regression model to the total variation.

Adjusted coefficient of multiple determination — $\overline{R}^2$: It is coefficient of determination adjusted for the loss of degrees of freedom. It helps us against increasing $\overline{R}^2$ by incorporating unnecessary explanatory variables.

Multi-co-linearity : Existence of linear relationships between explanatory variables in addition to one between dependent variable and them.

Hetero scedasticity : Differences between variances of the error terms. This problem is more serious in cross-sectional data.

Auto-co-relation : Correlations between the errors and their previous values. This problem is encountered very often when dealing with time series data.

10.12 ANSWERS OR HINTS TO CHECK YOUR PROGRESS EXERCISES

1) Read Section 10.2

2) Read Section 10.2

3) Read Section 10.2

4) Read Section 10.3

5) Read Section 10.4

6) Read Section 10.4

7) Read Section 10.4

8) Read Section 10.6

9) Read Section 10.7

10) Read Section 10.8

11) Read Section 10.9

UNIT 11 MEASURES OF INEQUALITY

Structure

11.0 Objectives

11.1 Introduction

11.2 Positive Measures

- 11.2.1 Relative Range
- 11.2.2 Relative Inter-Quartile Range
- 11.2.3 Relative Standard Variation
- 11.2.4 Standard Deviation of Logarithms
- 11.2.5 Champernowne Index
- 11.2.6 Hirschman-Herfindahl Indices
- 11.2.7 Kolm's Index

11.3 Gini Index

- 11.3.1 Gini as a Measure of Dispersion
- 11.3.2 Simple Computational Device

11.4 Lorenz Curve

- 11.4.1 Geometrical Definition
- 11.4.2 Properties of Lorenz Curve
- 11.4.3 A Measure Based on Area
- 11.4.4 A Measure Based on Length

11.5 Normative Measures

- 11.5.1 Dalton Index
- 11.5.2 Atkinson Index
- 11.5.3 Sen Index
- 11.5.4 Theil Entropy Index

11.6 Let Us Sum Up

11.7 Key Words

11.8 Exercises

11.9 Some Useful Books

11.10 Answers or Hints to Check Your Progress Exercises

11.0 OBJECTIVES

After going through this Unit, you will be able to:

- explain the various positive measures of inequality;
- discuss the computational device for construction of Gini index; and
- describe the normative measures of inequality propounded by Dalton, Atiknson, Sen and Theil.

11.1 INTRODUCTION

Improvement in well-being of the poor has been one of the important goals of economic policy and to a significant extent it is determined by the growth and distribution of its income. Distribution patterns have an important bearing on the relationship between average income and poverty levels. Extreme inequalities are economically wasteful. Further, income inequalities also interact with other life-chance inequalities. Hence reducing inequalities has become priority of public policy.

It therefore, becomes pertinent to measure income inequalities. Various measures have been developed over a period of time to study the level of inequalities in different situations. Broadly these measures can be put under two categories (i) positive measures, and (ii) normative measures. The measures which capture the inequality of income without value judgment about social well-being are known as positive measures. Range, quartile range, standard deviation, Gini ratio, Lorenz curve etc. are positive measures of inequality. On the other hand, the measures that essentially involve value judgement about social welfare are called normative measures. The index propounded by Dalton, Atkinson, Sen, Theil and Kakwani are normative measures. We shall discuss all these measures one by one in this unit. Gini Coefficient of Inequality and Lorenz Curve will receive particular attention, as they are very popular in literature.

11.2 POSITIVE MEASURES

If all values in a distribution are not equal, which means that there is dispersion in the distribution, hence, there exists inequality in the distribution. If a measure is developed to capture this non-equality in values without giving explicit consideration to its consequences with respect to social well-being or economic significance in a particular context, the measure is known as **positive**. It means that the measure is bothering about the fact whether it is measuring inequality of lengths of iron nails or incomes of wage earners in a village. Nevertheless, many of them are standard statistical measures and their social implications and/or social consequences can still be studied. Some of these measures can be arrived from normative approach as well.

Let us consider an income distribution x_i i=1,2,…,N over N persons and with mean income μ. Let the relative share of total income with person i be designated as q_i, which is naturally given by $x_i/N\mu$. The cumulative share of total income with the persons not having more than x_i income can be given as Q_i. However, when we have an income having frequency more than one, the proportion of people with income x_i can be denoted by p_i and the cumulative proportion of people with income no less than x_i as P_i. In this case, obviously the relative share of total income would be given by $f_i x_i / N\mu$ where f_i denotes the frequency of occurrence of income x_i and $N= \ f_i$. It will not be necessary to use two different subscripts for distinguishing the two cases for the purpose as the context would make it clear whether subscript i stands for a person with income x_i or for the group of persons with income x_i.

We have implicitly assumed that the data are arranged in the increasing (non-decreasing) order by magnitude of income so that symbolic representation is easy. The same could be accomplished by decreasing (non-increasing) order.

We intend to cover important measures along with their variants in this Unit. We shall also, albeit briefly, discuss their properties and weaknesses.

To recapitulate, we recount what we said above in a formal way

$$q_{i} = \frac{x_i}{N\mu} or \frac{f_i x_i}{N\mu}$$

$$Q_i = \sum_{i-1}^{i} q_{i,}$$

$$p_i = \frac{1}{N} or \frac{f_i}{N}$$

$$P_i = \sum_{i-1}^{i} p_{i,}$$

11.2.1 Relative Range

A measure of relative dispersion can be taken a measure of inequality. It is defined as the relative range by

$$RR_1 = \frac{Max_i x_i \text{ - } Min_i x_i}{\mu} \qquad \text{(RR.1)}$$

that is, the relative difference between the highest income and the lowest income. If income is equally distributed, then $RR_1 = 0$ and if one person received all the income, then RR_1 is maximum. If one wants to make the index lie in the interval between 0 and 1, one can define it as

$$RR_2 = \frac{Max_i x_i \text{ - } Min_i x_i}{N\mu} \qquad \text{(RR.2)}$$

which means it is the gap between the maximum share and the minimum share. That is,

$$RR_2 = Max_i q_i \text{ - } Min_i q_i \qquad \text{(RR.3)}$$

Though Cowell has suggested division of range by Min_i $x_{i,}$, which does not serve, in our view, any purpose. Two other normalization or standardization procedures that make it unit-free and contain it in (0,1) interval are suggested below:

$$RR_3 = \frac{Max_i x_i \text{ - } Min_i \overline{x}_i}{Max_i x_i} \qquad \text{(RR.4)}$$

and

$$RR_4 = \frac{Max_i x_i \text{ - } Min_i x_i}{Max_i x_i + Min_i x_i} \qquad \text{(RR.5)}$$

The basic weaknesses of these range-based measures are that they are not based on all values and therefore they do not reflect the change in inequality if there is any transfer of income between two non-extreme recipients.

Instead of considering extreme values at either end, which may not be even known, some scholars have toyed with the idea of the ratio between the mean income of the highest fractile (percentile, quintile or decile) and that of the lowest counterpart. They term it as the extreme disparity ratio (EDR). Naturally, this ratio is not contained in the interval (0.1). This ratio is independent of μ as well. The measure will not reflect the transfer of income that does not involve the extreme fractiles.

11.2.2 Relative Inter-Quartile Range

Sometimes, extremism of the relative range is sought to be moderated by restricting the distribution between the 10th and 90th percentile or sometimes to interquartile range. Bowley (1937) suggested relative quartile deviation as the index of inequality:

$$B = \frac{x^{q^3} - x^{q^1}}{x^{q^3} + x^{q^1}} \qquad \text{(B.1)}$$

where x^{q^r} represents the income level which divides the population in r and (4-r) quartiles. B is zero for degenerate distribution where everybody has the same income and unity if the lowest 75 per cent people have no income at all.

Though the extremes are moderated in comparison to the measure of range, it has an obvious weakness that the measure takes into account only 50 per cent of the distribution. Further, a transfer of income between two persons without causing either or both of them cross x^{q1} or x^{q3} would not change the measured level of inequality. Thus, the index suffers from all weaknesses of the earlier proposals except that of extremism. Its highest value reaches when the lowest 75 per cent people do not possess any income.

A variant of this measure is inter-quartile ratio, which can be defined as the 75th percentile (3rd quartile) income minus 25th percentile (1st quartile) income divided by the median (x^{q2}) income.

11.2.3 Relative Standard Variation

The standard deviation divided by the mean can be used as one measure of dispersion. It is:

$$RSD = \frac{c}{\mu} \qquad \text{(RSD.1)}$$

where σ and μ are standard deviation and mean of the distribution.

It can be equivalently defined as the standard deviation of relative incomes. Using definition of c , one can find out that it lies in the interval of 0 and $(N-1)^{1/2}$, not in (0,1). The highest value depends on the size of distribution.

Since the measure uses all values, any transfer of income would be reflected in the measure. However, it should be noted that the measure is equi-sensitive to transfers at all levels. Whether a given amount *d* is transferred between x_j =Rs.400 and x_k =Rs.500, or between x_j =Rs.10,000 and x_k = Rs.10,100, the changed in RSD is exactly the same.

We may finally note that the square of RSD is also quite often used as another measure of inequality, which is known as the coefficient of variance. Quite a few scholars suggest use of variance as a measure of inequality but we have not considered it here primarily because it is not unit-free. We think that an inequality measure must be unit-free.

11.2.4 Standard Deviation of Logarithms

One way of attaching greater importance to transfers at lower end (as required by Sen) is to consider some transformation of incomes. This transformation can easily be attained by considering the logarithms that stagger the income at lower levels.

This measure is defined in either of the following two ways:

$$SDL_1 = \left(\frac{1}{N} \sum_{i=1}^{N} (\log \mu - \log x_i)^2 \right)^{1/2} \qquad \text{(SDL.1)}$$

$$SDL_2 = \left(\frac{1}{N} \sum_{i=1}^{N} (\log \hat{\mu} - \log x_i)^2 \right)^{1/2} \qquad \text{(SDL.2)}$$

where μ and $\hat{\mu}$ are the arithmetic and geometric means respectively. While standard statistical literature prefers use of geometric mean the more common practice in literature on income inequality is one of using arithmetic means.

Cowell (1995) prefers to define these in terms of variance and calls the square of SDL_1 as the logarithm variance (V_1) and the square of SDL_2 as the variance of logarithms (V_2). Name of the second is clear from the expression but that of the first is derived from the fact that (*log x-log* $\bar{x}$) could be written as *log* ($x/\bar{x}$). One can see that V_1 is equal to V_2 plus log ($\hat{\mu}$ /μ).

As these measures are in terms of ratios of incomes, any proportionate change in incomes would leave the magnitude of inequality unchanged when measured by these indices. But, unfortunately, a transfer from a richer person to a poorer person may raise the magnitude of inequality, particularly if the poorer person has income more than 2.72 times the mean of the distribution.

While the lower limit, irrespective of formula, is zero when everybody has the same income, the upper limit depends on the size of distribution and approaches infinity when *N* is large and when everybody except the richest, receives income equal to one unit (as zero is inadmissible in logarithmic transformation.) Further, if we face grouped data, it is convenient to use μ in place of $\hat{\mu}$ and μ_i in place of x_i.

The variance of logarithms is however decomposable. It is a property that is being given emphasis of late. It can be shown that V_2 is the sum of between-group component and within group component, latter being population-weighted sum of within-group V_2's.

11.2.5 Champernowne Index

Champernowne (1973) makes use of the idea of geometric mean. It is a well known fact of an unequal distribution that its geometric mean is smaller than the arithmetic mean. The additive inverse of the ratio of geometric mean to arithmetic mean can duly be considered as an index of inequality. Formally, the index could be written as:

$$CII = 1 - \frac{\hat{\mu}}{\mu} \qquad \text{(CII.1)}$$

where μ and $\hat{\mu}$ are, as stated earlier, arithmetic and geometric means of the income distribution. It is easy to see that its value is bound between 0 and 1.

One can obviously think of another measure where geometric mean is replaced by harmonic means.

These measures are sensitive to transfer to income and change is greater when the transfer takes place at lower end of the distribution. They are sensitive to transfer of income between two persons. One can try it by replacing x_j and x_k by (x_j-*d*) and (x_k+*d*) respectively and finding out the direction of the change. Or, one can use differential calculus.

The trouble with these indices is that they cannot be defined when any of the income is zero.

11.2.6 Hirschman-Herfindahl Indices

These indices were developed in the course of studying the commodity concentration in trade by Hirschman (1945) and in characterizing market monopoly in industry by Herfindahl (1950). Later, they were more used in capturing autonomy and dependence of units in a federation.

If each unit is a class in itself, p_i=1/N, *i*=1,2…,N. Then concentration could be captured through use of q_i's . As the sum of q_i's is always 1, Hirschman devised a measure which would capture the inequality among them. He proposed square root of the sum of squares of shares q_i *i*=1,2…,N. That is,

$$H_1 = \left(N \sum_{i=1}^{N} q_i^2 \right)^{1/2} \qquad \text{(H.1)}$$

which could be generalized as

$$H_1^* = \left(N \sum_{i=1}^{N} q_i^a \right)^{1/a}, a > 1 \qquad \text{(H.2)}$$

Herfindahl devised a very similar measure, which has been more popular than the original (H.1). This is just the sum of share squares:

$$H_2 = \sum_{i=1}^{N} q_i^2 \qquad \text{(H.3)}$$

$$H_2^* = \sum_{i=1}^{N} q_i^a, a \geq 1. \qquad \text{(H.4)}$$

It is clear that, besides inequality among the shares, the value of these measures depends on *N*—the fewness or largeness of the number of units. For *N*=2, it has been suggested that (1/*N*) could be subtracted from (H.3)

$$H_3 = \sum_{i=1}^{N} q_i^2 - \frac{1}{N} \qquad \text{(H.5)}$$

The minimum value of H_3 is zero. But it serves no great purpose. When *N*=2, for q_1=0.99 and q_2=0.01, while H_2=0.98, H_3=0.48. H_2 scores definitely better than H_3 in characterizing the scene of monopoly.

11.2.7 Kolm's Index

Let there be *N* incomes such that *N*= *nm* where *n* is the number of different incomes and each income has *m* recipients. The number of equal pairs with a given income would be *m*(*m*-1)/2 and total number of equal pairs would be *n.m*(*m*-1)/2. Total number of all pairs would obviously be *N*(*N*-1)/2-*nm*(*nm*-1)/2. One can think of an `equality' index in terms of *nm*(*m*-1)/*nm*(*nm*-1)=(*m*-1)/(*N*-1). The inequality index could then be constructed by subtracting it from1: 1-(*m*-1)/(*N*-1)-(*N*-*m*)/(*N*-1)=*m*(*n*-1)/(*nm*-1)=(*nm*-*m*)/(*nm*-1). In case, income x_1 has f_i recipients, the measure is:

$$K = 1 - \frac{\Sigma f_i^2 - N}{N(N-1)} = \frac{N^2 - \Sigma f_i^2}{N(N-1)} \qquad \text{(K.1)}$$

The purpose of developing this curiosum due to Kolm (1996) is just to make one feel that there could be a variety of simple ways to approach the issue of measurement of inequality.

Check Your Progress 1

1) Define relative range measures of inequality. List out relative merits.

………………………………………………………………………………..

………………………………………………………………………………..

………………………………………………………………………………..

………………………………………………………………………………..

………………………………………………………………………………..

………………………………………………………………………………..

2) How relative inter-quartile range is better than relative range?

……………………………………………………………………………..

……………………………………………………………………………..

……………………………………………………………………………..

……………………………………………………………………………..

……………………………………………………………………………..

3) What is the relative mean deviation? If a transfer of income is between two persons both having income lower than the mean, will it change the magnitude of this index?

……………………………………………………………………………..

……………………………………………………………………………..

……………………………………………………………………………..

……………………………………………………………………………..

……………………………………………………………………………..

4) Compare the two versions of standard logarithmic deviations.

……………………………………………………………………………..

……………………………………………………………………………..

……………………………………………………………………………..

……………………………………………………………………………..

……………………………………………………………………………..

5) What is import of Champernowne Index?

……………………………………………………………………………..

……………………………………………………………………………..

……………………………………………………………………………..

……………………………………………………………………………..

……………………………………………………………………………..

……………………………………………………………………………..

6) What is Hirfindahl index? What are its areas of application?

……………………………………………………………………………..

……………………………………………………………………………..

……………………………………………………………………………..

……………………………………………………………………………..

……………………………………………………………………………..

7) What is the message from the Kolm's index? Calculate the Kolm index for a distribution, which frequency 5 for value Rs.5 lakh and frequency 5 with value Rs.10 lakh and therefore total size 10 and the arithmetic mean 7.5.

………………………………………………………………………………..

………………………………………………………………………………..

………………………………………………………………………………..

………………………………………………………………………………..

………………………………………………………………………………..

11.3 GINI INDEX

This coefficient of concentration, as it is usually called, owes to an Italian statistician by the name of Corrado Gini (1912). Modern practice is to call it just Gini. This index in its origin is positive. There are a number of ways in which this coefficient can be expressed. There are also a number of ways in which it can be interpreted. People have also derived it as a measure of inequality under plausible axioms in welfare theoretic framework. The index satisfies many axioms proposed in literature for an index of inequality.

First, we shall discuss those definitions and expressions, which can be derived as a measure of dispersion. Besides giving its expressions for its frequency data for grouped observations, we shall discuss its welfare theoretic interpretations.

11.3.1 Gini as a Measure of Dispersion

Recall that mean deviation and standard deviation, which are measures of dispersion, seek the deviation from arithmetic mean. Also recall that one of the logarithmic measure sought deviation from the geometric mean. However, one may ask why to seek dispersion in terms of deviations from any mean? Why not compare all pairs and seek the differences. In order to consider positive values of differences, we either take modal values of deviations before averaging (mean deviation) or sum the squares of deviations and take root of the mean of the squared differences.

Corrado Gini (1912) proposed to consider all the differences, that is all pairs of values. By contrast, the range measure of dispersion considers only one pair of highest value and lowest value. When x_i and x_j denote *i*th and *j*th incomes respectively and *i, j*=1, 2,…, *N*, we can see that the aggregate of absolute differences is given by

$$\sum_{i=1}^{N}\sum_{j=1}^{N}\left|x_{i-}x_j\right| \qquad \text{(G.1)}$$

and because total number of differences is N^2, the mean of absolute differences can obviously be written as

$$\frac{1}{N^2}\sum_{i=1}^{N}\ \sum_{i=1}^{N}\left|x_{i-}x_j\ \right| \qquad \text{(G.2)}$$

where the differences with the self have also been counted and the difference of x_i with x_j is treated as separate from that of x_j and x_i though numerically they are the same. This is also said to be the case with replacement. Expression (G.2) ranges between 0 and 2μ.

In the case of without replacement, the sum is obviously to be divided by $N(N\text{-}1)$ as there are N deviations with the self. It is not difficult to see that the numerical value of the sum remains the same.

In order to make it serve as a measure of inequality, (G.2) can be divided by μ to produce what can be called coefficient of mean difference (CMD):

$$CMD = \frac{1}{N^2\mu}\sum_{i=1}^{N}\sum_{i=1}^{N}\left|x_i - x_j\right| \qquad \text{(G.3)}$$

CMD fulfils the idea of scale independence. However, the expression (G.3) ranges between 0 when everybody has the same income and 2[=2*N*/(*N*-1)] when only one person has all the income. In order to make it satisfy the interval (0,1), we can further divide it by 2. The result is Gini coefficient of concentration or Gini index of inequality:

$$G = \frac{1}{2N^2\mu}\sum_{i=1}^{N}\sum_{j=1}^{N}\left|x_i - x_j\right| \qquad \text{(G.4)}$$

or

$$G = \frac{1}{N^2\mu}\sum_{i=1}^{N}\sum_{x_j \le x_i}^{N}(x_i - x_j) \qquad \text{(G.5)}$$

conceived as an aggregate of only positive differences, though normalized by the number of all differences and the mean income. Kendall and Stuart define this as 'one half of the average value of absolute differences between all pairs of incomes divided by the mean income'.

The index can also be defined in terms of population proportions and income shares. If the income-share of individual *i* is denoted by q_i, that is,

$$q_i = \frac{x_i}{N\mu}, \qquad \text{(G.6)}$$

then the expression (G.4) can also be written as

$$G = \frac{1}{2N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left|q_i - q_j\right| \qquad \text{(G.7)}$$

In the case of a discrete distribution, each individual constitutes (1/*N*)th of the population, that is,

$$p_i = \frac{1}{N}. \qquad \text{(G.8)}$$

Therefore one can also write (G.7) as

$$G = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} |p_j q_i - p_i q_j|. \qquad (G.9)$$

An obvious question is: why not $|p_i q_i - p_j q_j|$ in the expression (G.9)? For understanding this, let us consider the Gini coefficient for the groups.

Let μ_r and μ_s denote mean incomes of r^{th} and s^{th} groups (say, families) respectively and *r,s*=1,2…g. Then, Gini for the groups can be defined as

$$G = \frac{1}{2N^2\mu}\sum_{r=1}^{g}\sum_{s=1}^{Ng} |\mu_r - \mu_s| f_r f_s \qquad (G.10)$$

where f_r and f_s are frequencies of the groups *r* and *s* respectively. This can obviously be written as

$$G = \frac{1}{2}\sum_{r=1}^{g}\sum_{s=1}^{g} \left|\frac{\mu_r}{\mu} - \frac{\mu_s}{\mu}\right| p_r p_s \qquad (G.12)$$

where

$$p_r = \frac{f_r}{N} \text{ and } p_s = \frac{f_s}{N} \qquad (G.12)$$

Now, let us note the share of total income with the group r:

$$q_r = \frac{\mu_r f_r}{N\mu} = p_r \frac{\mu_r}{\mu} \qquad (G.13)$$

Then, the expression (G.12) can be written in either of the two ways (G.14) and (G.15)

$$G = \frac{1}{2}\sum_{r=1}^{g}\sum_{s=1}^{g} \left|\frac{q_r}{p_r} - \frac{q_s}{p_s}\right| \qquad (G.14)$$

or

$$G = \frac{1}{2}\sum_{r=1}^{g}\sum_{s=1}^{g} |p_s q_r - p_r q_s| \qquad (G.15)$$

It is easy to see that $g=N$ and $p_s=p_r=(1/N)$ when f_r and f_s are all equal to1.

In statistics literature we emphasize frequency aspects, in economics literature we find it convenient, expression-wise, to treat each individual with single income though there is no bar for $x_i=x_j$.

11.3.2 Simple Computational Device

Two years after giving his index to terms of relative mean differences, Gini (1914) showed that the index is exactly equal to one minus twice the area under Lorenz curve (to be discussed later). That is,

$$G = 1 - 2\overline{A} \qquad (LG.1)$$

where $\overline{A}$ is the area under the Lorenz curve, as shown in Fig. 11.1

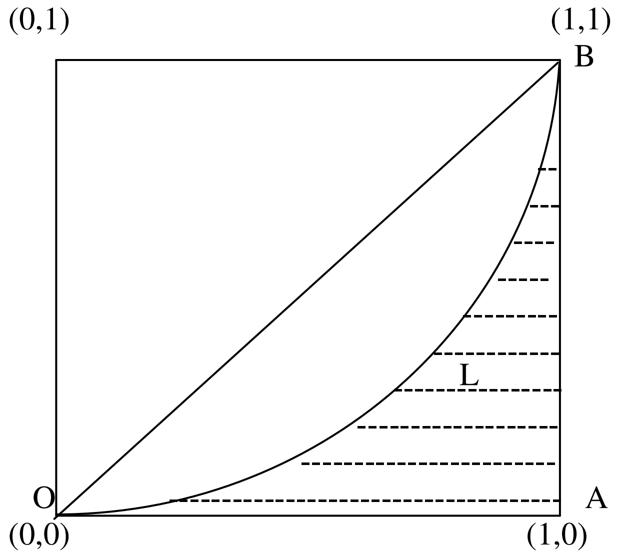


Fig. 11.1

However, normally, we do not specify and estimate a smooth relationship between Q and P. Instead, we obtain the curve by plotting cumulative proportions of people in classes and cumulative shares of their incomes, where classes are arranged according to increasing per capita income values:

$$\mu_1 \leq \mu_2 \leq \mu_3 \leq \ldots. \leq \mu_r \leq \ldots \leq \mu_g . \qquad \text{(LG.4)}$$

strictly speaking, equality sign is useless in this presentation.(We have written it in deference to Theil (1967). Plotting P_r and Q_r, we obtain the Fig. 11.2. We can see that the area below the Lorenz curve can be conceived as consisting of several trapeziums. A trapezium could be seen as consisting of a rectangle and a triangle. Summing the areas of all trapeziums (say g in number), we can get the area $\overline{A}$. Substituting it in (LG.1), we get the following expression for computing Gini coefficient G:

$$G = 1 - \sum_{r=1}^{g} (P_r - P_{r-1})(Q_r + Q_{r-1}) \qquad \text{(LG.5)}$$

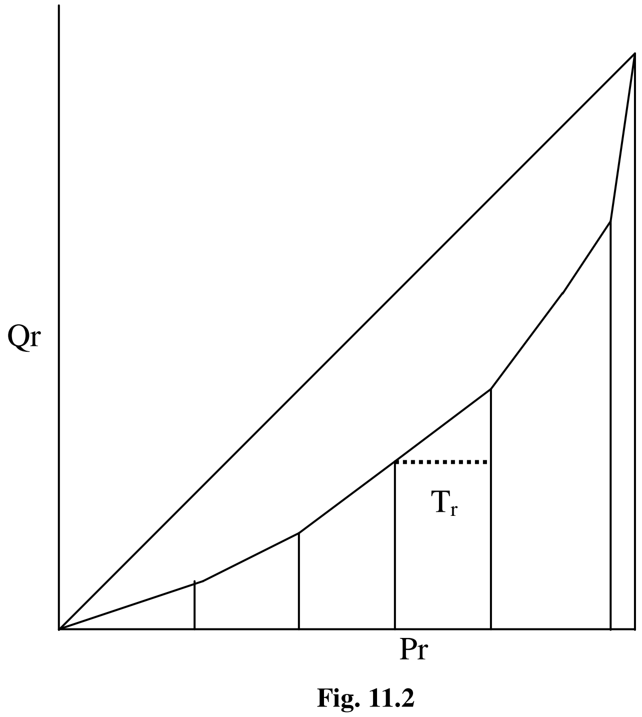


Fig. 11.2

Check Your Progress 2

1) Define Gini ratio.

……………………………………………………………………………

……………………………………………………………………………

……………………………………………………………………………

……………………………………………………………………………

……………………………………………………………………………

2) Show the difference in approach in defining Gini from other measures of dispersion.

……………………………………………………………………………

……………………………………………………………………………

……………………………………………………………………………

……………………………………………………………………………

……………………………………………………………………………

3) Write the expression for computing Gini.

……………………………………………………………………………

……………………………………………………………………………

……………………………………………………………………………

……………………………………………………………………………

……………………………………………………………………………

4) How can you compute Gini ratio?

……………………………………………………………………………

……………………………………………………………………………

……………………………………………………………………………

……………………………………………………………………………

……………………………………………………………………………

11.4 LORENZ CURVE

Lorenz curve is a powerful geometrical device to compare two situations of distribution with regard to their level of inequality. Devised some hundred years ago by Max O. Lorenz (1905) to measure concentration of wealth, it is still very widely used in empirical studies on inequality. The device can be used for comparing inequality of distribution of any measurable entity such as income, wealth (land, capital), consumption, expenditure on an item (say, food or education), etc. The distribution may be over persons or households. But the device can also be used to measure inequality of tax collection or expenditure incurred by states or federal grants received by different states. We can compare pre-tax and post-tax distributions in order to study the efficacy of instrument of tax.

Lorenz (1905) studied a number of methods then in use to gauge the level of, or change in the level of, inequality. Most of these measures used fixed-income classes in data tabulation and made inter-temporal comparison, employed changes in percentage of recipients of class incomes in each of the fixed-income classes or movement of persons from one class to another and so on. Finding them unsatisfactory, he comes to the conclusion that changes in income and changes in population both have to be simultaneously taken into account and in a manner that 'fixed-ness' of income classes gets neutralized.

In fact, this measurement relates to comparison and in most cases, we are in a position to compare but there are situations of non-comparability. However, a few of inequality measures that are capable of numerical representation in terms of a scalar number, and therefore called summary measure, are found to be based on the Lorenz curve.

It may be pointed out that the curve was independently introduced by Gini (1914). It is therefore, quite often referred to as Lorenz-Gini curve as well. We shall, however, stick to more common usage and call it Lorenz Curve.

11.4.1 Geometrical Definition

The Lorenz curve of concentration of incomes is the relationship between the cumulative proportions of recipients, usually plotted on the abscissa, and the corresponding cumulative shares of total income with the recipients, usually plotted on the ordinate. If population proportions and income shares of class *j* are denoted by p_j and q_j and cumulative proportions and shares upto class *i,* by P_i and Q_i then

$$P_i = \sum_{j=1}^{i} p_j, \qquad 1 \geq p_j \geq 0 \qquad \text{(GD.1)}$$

and

$$Q_i = \sum_{j=1}^{i} q_j, \qquad 1 \geq q_j \geq 0 \qquad \text{(GD.2)}$$

The relationship between P_i and Q_i is given by the curve

$$Q_i = L(P_i), \qquad 1 \geq P_i \geq 0, \quad 1 \geq Q_i \geq 0 \qquad \text{(GD.3)}$$

and the point on the curve by $(P_{i\,,}\ Q_i\,)$. Naturally, the first point is (0,0) and the last one on the curve, (1,1). It is also clear that $Q_i \leq P_i$ i=1, 2,…, N-1 if there are N classes of incomes. It means no point will make an angel of more than 45° with the abscissa at the origin. Then, one can be sure that the Lorenz curve lies in the lower triangle of Lorenz Box of the unit square. See Fig. 11.3 in which OLB shows the Lorenz curve (Fig. 11.3).

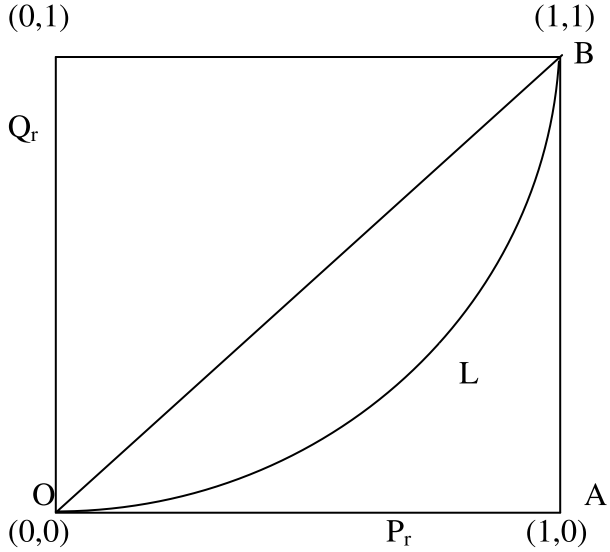


Fig. 11.3

11.4.2 Properties of the Lorenz Curve

Now it is easy to see that the extreme case of perfect equality is given by the diagonal OB which represents P_i=Q_i, i=1, 2,…, N. The other extreme of perfect inequality will be given by a curve OAB. The diagonal OB is often designate as the egalitarian line or line of equality. The other diagonal CA is known as the alternative diagonal and is useful to study the symmetry of the curve. The tringle OAB with sharp kink of 90° at A can be said to be the line of perfect inequality. See Fig. 11.4.

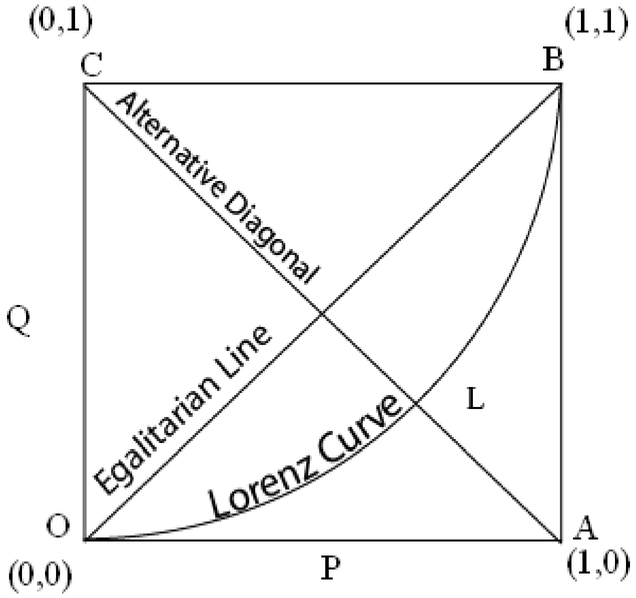


Fig. 11.4

We can finally note the following properties:

i) $1 \geq p_i \geq 0; 1 \geq q_i \geq 0,\ i = 1,2,...,N$

ii) $1 \geq P_i \geq 0; 1 \geq Q_i \geq 0, i = 1,2,...,N - 1$

iii) $P_0 = Q_0 = 0; P_N = Q_N = 1$

iv) $P_i \geq Q_i, i = 1,2,.., N - 1$

By drawing a Lorenz Curve, we can know whether a given distribution is equal or unequal. We do not yet know how much unequal a given distribution is. When we draw two or more Lorenz Curves, we can compare the distributions as regards their levels of inequality. The curve closer to the diagonal of equality has lower level of inequality than the one away from it (Fig. 11.5). But we do not know yet the level of inequality. And even this comparison is possible only when the curves do not intersect (Fig. 11.6).

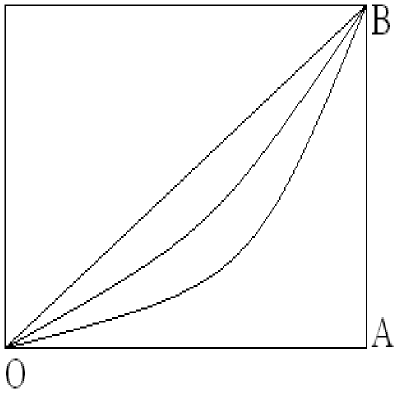


Fig. 11.5

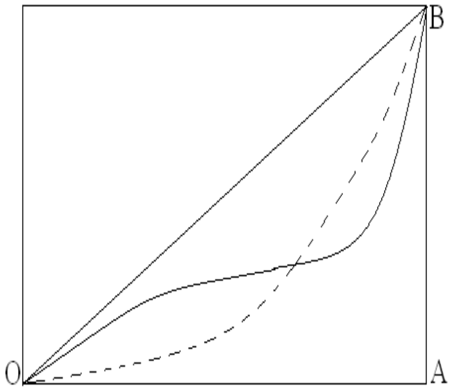


Fig. 11.6

However we can devise some measures, which are based on the Lorenz curve. In case the Lorenz curves intersect, reducing the distributions into single real number is the only option. So we shall discuss only two such proposals.

11.4.3 A Measure Based on Area

We have noted that if Lorenz curve coincides with the diagonal of equality, the inequality is nil and if Lorenz curve coincides with the two sides of the square, the inequality is full.

In the case of non-intersecting Lorenz curves, it is clear that the curve closer to the diagonal of equality will circumscribe smaller area between itself and the diagonal of equality than the one, which is farther. Which is what it should be. We can therefore devise a measure of inequality by dividing the area OLB (Fig. 11.4) by the area of triangle OAB, which is the maximum possible area between the diagonal of equality and Lorenz curve. As the area of OAB is (1/2), the measure turns out to be twice the area between the diagonal of equality and the Lorenz curve. In other words, Lorenz coefficient of concentration (LCC) is:

$$LCC = \frac{Area\,OLB}{\Delta OAB} = 2Area\,OLB$$

Since this turns out to be exactly equal to Gini coefficient, we are not elaborating it any further.

11.4.4 A Measure Based on Length

This is a measure proposed by Kakwani (1980). The length of the Lorenz curve, denoted by ℓ, cannot fall below √2, which is the length of the egalitarian line and cannot exceed 2, which is the sum of the lengths of the two arms of the lower triangle. In order to produce a measure with the minimum value 0 and the maximum value 1, following exercise can be suggested:

	Minimum	Actual	Maximum
Length of the Curve	$\sqrt{2}$	ℓ	2
Length of the Curve - $\sqrt{2}$	0	$\ell-\sqrt{2}$	$2-\sqrt{2}$
$\frac{\text{Length of the Curve }-\sqrt{2}}{\text{Maximum length }-\sqrt{2}}$	0	$\frac{\ell-\sqrt{2}}{2-\sqrt{2}}$	1

So this measure is clearly:

$LK= (\ell-\sqrt{2})/ (2-\sqrt{2})$

In both the cases, one can draw actual graphs and actually measure the area and the length and calculate the indices for level of inequality. Those who wish to carry out a more sophisticated exercise will have to estimate smooth functions.

Check Your Progress 3

1) Enumerate the properties of Lorenz curve.

……………………………………………………………………………

……………………………………………………………………………

……………………………………………………………………………

……………………………………………………………………………

2) When will comparison between two Lorenz curve fail to compare inequality in two distribution?

……………………………………………………………………………

……………………………………………………………………………

……………………………………………………………………………

……………………………………………………………………………

3) What is the relationship between Lorenz curve and Gini coefficient.

……………………………………………………………………………

……………………………………………………………………………

……………………………………………………………………………

……………………………………………………………………………

4) What is Kakwani's measure of inequality, which is based on the Lorenz curve.

……………………………………………………………………………

……………………………………………………………………………

……………………………………………………………………………

……………………………………………………………………………

……………………………………………………………………………

11.5 NORMATIVE MEASURES

The measures that essentially involve judgement about values through specification of social welfare function are called normative measures. The arguments of this nature were first advanced by Dalton, pretty eight decades ago in 1920 for constructing what are today called normative measures of inequality.

Reacting to an observation by Pearsons (1909) that 'the statistical problem before the economists in determining upon a measure of inequality in the distribution of wealth is identical with that of the biologist in determining upon a measure of the inequality in the distribution of any physical characteristic', Dalton (1920) pointed out that 'economist is interested, not in distribution as such, but in effects of the distribution upon the distribution (and total amount) of economic welfare which may be derived from income'. The objection to great inequality of income, he further points out, is due to the resulting loss of potential economic welfare that could accrue to people in the absence of it.

Yet, it has to be noted that inequality though defined in terms of economic welfare, has to be measured in terms of income. This idea due to Dalton has been conceded by subsequent contributions. Using the notion of social welfare function in construction gives rise to **normative** measures of inequality.

It may be instructive to remember that the discussion would revolve around three issues:

1) the relationship between income of a person and his welfare;
2) the relationship between personal income-welfare functions; and
3) the relationship between personal welfare and social welfare.

It may be noted that utility is the word mostly used for personal welfare whereas for welfare of society the phrase social utility is rarely used.

There are two major indices in this category: Dalton's index and Atkinson's index. In Atkinson's index a new idea is introduced and that is of equally distributed equivalent income. Actually there are two sub-approaches within normative approach. One is Dalton's and the other is Atkinson's. While in Dalton's approach present social welfare is compared with that could be obtained by equally distributing the total income, in Atkinson's approach the present level of income is compared with that of equally distributed level of income, which generates the present level of social welfare. Sen has generalised the Atkinson's index. Theil's index based on information theory could be suggested here only to sort of complete the unit.

11.5.1 Dalton Index

For each individual, Dalton assumes, marginal economic welfare diminishes as income increases. It means income-welfare function

$$U_i = U_i(x_i), i = 1,2,...N \qquad \text{(D.1)}$$

(where U_i is welfare of person i possessing income x$_{i)}$

is concave, suggesting that $(\partial U_i / x_i) > 0$ but $(\partial^2 U_i / x_i^2) < 0$. Dalton further assumes that economic welfare of different persons is additive. Thus, in his scheme, social welfare is a simple aggregation of personal welfares. In other words, social welfare *W* is given by

$$W = \sum_{i=1}^{N} U_i(x_i) \qquad \text{(D.2)}$$

He further assumes that the relation of income to economic welfare is the same for all members of the community. That is,

$$U_i = U(x_i),\ \text{i=1,2,…,N} \qquad \text{(D.3)}$$

In that case, the relation (D.2) can be expressed as

$$W = \sum_{i=1}^{N} U(x_i) \qquad \text{(D.4)}$$

which makes it clear that whosoever gains in welfare, the addition to the social welfare is the same. For any given level of social welfare, any distribution of welfare among the members of the society is permissible. However, one must remember that the relation of individual income to their welfares is concave. Therefore, transfer of income from A to B will not lead to symmetric change in welfares of those two persons involved in the transaction. The result is some impact on *W* the measure of social welfare.

From Fig. 11.7, we may compare the situation when two individuals, both possessing the same relation, have two different income levels, with that when they have the same (mean) income. We may note that the sum of the welfare of person 1(BB’) and the welfare of person 2(DD’) is less than the twice of CC’ which is the level of welfare enjoyed by both the persons when they have equal income. It is easy to see that the loss suffered by person 2, that is D’E, is overcompensated by the gained by person 1, which is C’F.

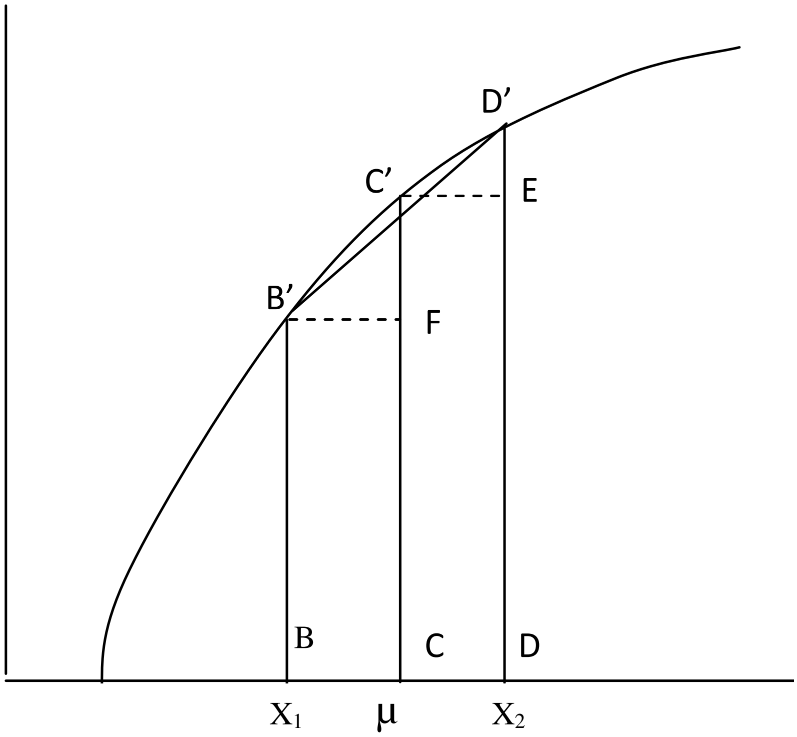


Fig. 11.7

This demonstrates that, under assumptions by Dalton, an equal distribution is preferable to an unequal one for a given amount of total income from the viewpoint of social welfare. In fact, for a given total of income, the economic welfare of the society will be maximum when all incomes are equal. The inequality of any given distribution may therefore be defined as

$$D_1 = \frac{\sum_{i=1}^{N} U(\mu)}{\sum_{i=1}^{N} U(x_i)} = \frac{NU(\mu)}{\sum_{i=1}^{N} U(x_i)} \qquad \text{(D.7)}$$

which is equal to unity for an equal distribution and greater than unity for an unequal one. It may therefore be preferred to define the Dalton's index as

$$D_2 = \frac{NU(\mu)}{\sum_{i=1}^{N} U(x_i)} \qquad \text{(D.8)}$$

which is obviously zero for an equal distribution. How large can it be? It will depend on the values of *U(0)*, *U(μ)* and *U(Nμ)* when *N* and *μ* are given, not necessarily 1. Later writers have therefore preferred to define Dalton's index in the following form, which inverts the arguments of D_2 subtract it from 1:

$$D = 1 - \frac{\sum_{i=1}^{N} U(x_i)}{NU(\mu)} = 1 - \frac{\overline{U}}{U(\mu)} \qquad \text{(D.9)}$$

It looks as if the index is contained in the interval (0,1). However, there are many valid concave functions where it may not hold true. For example, if we have $U(x_i) = \log x_i$, then $D = 1 - \{\log\hat{\mu}/\log\mu\}$. Given the fact that $\hat{\mu} < \mu$, D would turn out to be a negative number for $\mu<1$. And μ could be less than 1 as *x* can be measured in any unit. It would be the same case $U(x_i) = 1/x_i$.

However, in order to obtain numerical magnitude, it is not sufficient to define the index. Dalton (1920) points out that though defined in terms of economic welfare, inequality has to be measured in terms of income. Then, no unique measure of inequality will emerge. It will verily depend on the particular functional relationship assumed. Dalton himself considered two such functions for the purpose of illustration. The first is related to Bernaulli's hypothesis. It holds that proportionate additions to income (in excess of that required for bare subsistence-poverty line) make equal additions to personal welfare, That is,

$$dU_i = \frac{dx_i}{x_i} orU_i = \log x_i + c_i \qquad \text{(D.10)}$$

Under the assumption that every person has the same functional relationship, the Dalton's index can be given as

$$D = 1 - \frac{\log\hat{\mu} + c}{\log\mu + c} \qquad \text{(D.11)}$$

where $\hat{\mu}$ is the geometric mean of personal incomes. The other formulation he discusses is given as

$$dU_i = \frac{dx_i}{x_1^2} or U_i = c - \frac{1}{x_i} \qquad \text{(D.12)}$$

where c is the maximum welfare one can obtain when $x -$ ∝ . Dalton’s index in this case would turn out to be:

$$D = 1 - \frac{c \text{ - } (1/\tilde{\mu})}{C - (1/\mu)} \qquad \text{(D.13)}$$

where $\tilde{\mu}$ is the harmonic mean.

11.5.2 Atkinson Index

Atkinson (1970) objects to Dalton’s measure because D is not invariant with respect to positive linear transformations of personal income-welfare functions. This was pointed out by Dalton himself but he could not resolve it.

Atkinson seeks to redefine the index in such a way that measurement would be invariant with respect to permitted transformations of welfare numbers. Atkinson does it through devising what he calls 'equally distributed equivalent income'. Both the distributions, the original and the new one, are supposed to yield the same level of welfare.

In order to make the concept clear, we put a few artifacts along with the actual distribution. Let us first note that for an actually distributed income vector x_i, i=1,2…,N (call it vector a), there is only one equally distributed income vector with each element equal to μ (call it vector b) but there are a number of equivalently distributed, vectors (call them vectors c). See Chart 1. An equivalent income distribution is one, which has the same level of welfare as that of currently given distribution. However, one of these equivalent distributions (vectors c) is `equal’ as well. This is called equally distributed equivalent income vector, shown as vector (d) in the Chart 1. As μ is the mean level of current distribution, μ^* may be used for designating the level of equally distributed equivalent income.

CHART-I

Vector (a) Actually distributed income vector	$x_1,\ x_2\ .....,\ x_i,\ .....,\ x_N.$
Vector (b) Equally distributed income vector	μ, μ, …., μ, …, μ.
Vector (c) Equivalently distributed income vector	$x_1^*,\ x_2^*,\ \ldots,\ x_1^*,\ \ldots,\ x_N^*$.
Vector (d) Equally distributed equivalent income vector	μ*, μ*, …, μ*, …, μ*

If should be obvious that $W_b \ge W_a = W_c$ *and* $W_c = W_{d.}$ *Then,*$W_a = W_{d.}$ W represents social welfare with respective distributions of income vectors. It is clear that $\mu \ge \mu^*$. μ^* is defined by the additive social welfare function having symmetric individual utility functions such as:

$$U(\mu) = \frac{1}{N}\sum_{i=1}^{N} U(x_i) \qquad \text{(A.1)}$$

or equivalently

$$\mu^* = U^{-1}\left[\frac{1}{N}\sum_{i=1}^{N} U(x_i)\right] \qquad \text{(A.2)}$$

The index due to Atkinson is then defined as the additive inverse of the ratio of equivalent mean income to actual mean income:

$$A = 1 - \frac{\mu^*}{\mu} \tag{A.3}$$

which is said to lie between zero (complete equality) and 1 (complete inequality). We can see that *A* cannot be 1 unless μ^* is zero, which is an impossibility for any distribution with $\mu > 0$. If complete inequality is defined as the situation when only one person grabs all the income, we can see that

$$A = 1 - \frac{\mu_m^*}{\mu} \tag{A.4}$$

where

$$NU(\mu^*) = \sum_{i=1}^{N} U(x_i)$$

and

$$NU(\mu^*_m) = (N - 1)U(0) + U(N\mu).$$

This index is not scale independent unless some restriction is imposed on the relationship *U*. If this requirement has to be met, Atkinson points out, we may have to have the following form

$$U(x_i) = \begin{cases} \alpha + \dfrac{\beta}{1-\varepsilon} x_i^{1-\varepsilon}, & \varepsilon \neq 1 \\ \log_e x_i, & \varepsilon = 1 \end{cases} \tag{A.5}$$

Note that we need $\epsilon \geq 0$ for ensuring concavity and $\epsilon > 0$ for ensuring strict concavity. This is a homothetic function and is linear when $\epsilon = 0$. We may note that ϵ cannot exceed 1 as in that case the varying component assumes inverse relationship. α is usually negative so that U(x_i) is not positive for $x_i = 0$. Otherwise, when $x_i = 0$, Ui = α which means that welfare is positive even when income is zero. This is generally not acceptable. On the contrary, a negative α would be more acceptable. When $\epsilon = 1$, α is infinitely large and negative.

Since ϵ can be zero, Atkinson's requirement is not strict concavity. Sen (1973) has a question. He asks to consider two distributions (0,10) and (5,5) along with

$$U(x_i) = \alpha + \beta(x_i) \tag{A.6}$$

Then, he points out the level of social welfare would be ($2\alpha + 10\beta$) whatever the distribution. μ^* would be 5 in both the cases. μ is of course 5. The measure of inequality *A* is therefore zero. So, both the distributions are ethically equal. This is obviously absurd. Therefore, the relation (A.5) should be defined with the restriction $\epsilon > 0$. We should also note that (A.4) is an iso-elastic marginal utility function.

ϵ is the inequality-aversion parameter and has close resemblance with risk-aversion premium. Atkinson proposed to draw on the parallel formally with the problem of measuring risk. He finds his concept of equally distributed equivalent income very closely resembles with risk-premium or certainty equivalent income as used in the theory of decision-making under uncertainty.

In case we seek to introduce this restrictive personal income-welfare function along with simple aggregation of individual welfares to constitute the social welfare, into the inequality measure A, we will have

$$A = 1 - \left[\frac{1}{N} \sum_{i=1}^{N} \left(\frac{x_i}{\mu} \right)^{1-\varepsilon} \right]^{1/(1-\varepsilon)} \qquad ,\varepsilon \neq 1 \tag{A.7}$$

The question is now narrowed down to choosing ϵ. As ϵ rises, more weight is attached to transfers at the lower end of the distribution and less to that at the top. When ϵ rises, (A.7) assumes the function min$_i$ (x_i), which only takes account of transfers to the very lowest income group (and is therefore not strictly concave). When ϵ = 0,U_i is linear. As a consequence, *A* is always zero. This means *A* has no descriptive content at all. When ϵ→1, *A* turns out to be

$$A = 1 - \prod_{i=1}^{N} \left(\frac{x_i}{\mu} \right)^{1/N} = 1 - \frac{\hat{\mu}}{\mu} \tag{A.8}$$

which is the same as Champernowne index (CII.1). For values of ϵ between 0 and 1, the expressions may not be very neat. Parameter ϵ is often chosen to be ½ or 1/3 or 2/3.

11.5.3 Sen Index

There are people who feel rather strongly that the social valuation of the welfare of individuals should depend crucially on the incomes of their neighbours too. Then, why should society add simply individual welfares? One may also question the assumption of one welfare function for all individuals. If we do so, we should go for broad social welfare function such as

$$W = W(x_i, x_2, ..., x_N) \tag{S.1}$$

which is just symmetric, quasi-concave and increasing in individual income levels. Then, a more general normative measure of inequality can be defined by devising the concept of 'generalized equally distributed equivalent income'. This is obviously the level of per capita income *x** which, if shared by all, would produce the same level of *W* as is generated by the present actual distribution. That is,

$$x^* = x \big| W(x^*, x^*, .., x^*) = W(x_1 x_2, ..., x_N) \tag{S.2}$$

Under the assumption that (S.1) is quasi-concave, $x^* \leq \mu$ for every distribution of income. The index S would then be

$$S = 1 - \frac{x^*}{\mu} \tag{S.3}$$

which is but a generalized version of A. If utilitarian framework is employed, then S and A turn out to be indistinguishable.

These measures, it may be noted, clearly suggest that there exists a redistribution equivalent of growth so far as the concern is about raising the welfare.

11.5.4 Theil Entropy Index

Theil (1967) poses a question: Does information theory supply us with a 'natural' measure of income inequality among N individuals, which is based on income shares? He answers: Yes. Here is a short introduction.

Let us start with income share of individual i:

$$q_i = \frac{x_i}{N_\mu} > o \quad such\, that \sum_{i=1}^{N} q_i = 1 \tag{T.1}$$

When $x_i = \mu$, i=1, 2,…, N, that is, when distribution is equal, we have

$$q_i = \frac{1}{N} \quad i = 1,\ 2, ... N \tag{T.2}$$

We have complete inequality when some x_i=$N\mu$ and $x_j = 0, j \neq i.$ It implies that q_i=1 for some i and $q_j = 0, i \neq j$.

In information theory, one way of defining entropy of probabilities p_i is

$$H = \sum_{i=1}^{N} p_{i.} \log \frac{1}{P_{i}.} \tag{T.2}$$

Replacing probabilities by shares, we have

$$H = \sum_{i=1}^{N} q_{i.} \log \frac{1}{q_{i}.} \tag{T.3}$$

which can be taken as a measure of equality. For the situation of complete equality, we can see that H is equal to $logN$ and for that of complete inequality H is zero. We can therefore define Theil index T as

$$T = \log N \sum_{i=1}^{N} q_{i.} \log \frac{1}{q_{i}.}$$

$$= \sum_{i=1}^{N} q_i \log N - \sum_{i=1}^{N} q_{i.} \log \frac{1}{q_i}$$

$$= \sum_{i=1}^{N} q_i \log N . q_i \tag{T.4}$$

This measure is motivated by the notion of entropy in information theory. But one can see that it can be interpreted in the traditional normative framework with

$$U_i = q_i \log\frac{1}{q_i} \tag{T.5}$$

and

$$W=\sum_{i=1}^{N} U_i(q_i). \tag{T.6}$$

We may note that (T.5) depends on x_i as well as on μ along with *N* and *U* that it is concave with respect to $x_{i.}$

While the lower limit of *T* is zero, its upper limit *log N* increases as the number of individuals increases. To many people, it is objectionable. However Theil (1967) chooses to defend it. When society consists of two crore persons and one grabs all and when society consists of two persons and one grabs all, cannot have the same level of inequality. The former case is equivalent to the situation in which one crore out of two crore people have nothing and the other one crore have equal income. Maximum value for two-person society is log 2, and that for two crore-person society is 7 log 2. Some researchers still insist that the measure should be normalized by dividing it by log N.

Check Your Progress 4

1) What is social welfare function, according to Dalton?

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

2) Discuss Dalton index of inequality.

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

3) Give the logic behind Atkinson index.

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

4) How is Sen index distinct from Atkinson's index?

……………………………………………………………………………………

……………………………………………………………………………………

……………………………………………………………………………………

……………………………………………………………………………………

……………………………………………………………………………………

5) Discuss Theil's entropy index of inequality.

……………………………………………………………………………………

……………………………………………………………………………………

……………………………………………………………………………………

……………………………………………………………………………………

……………………………………………………………………………………

11.6 LET US SUM UP

Owing to adverse impacts of economic inequality both on poverty and on growth, reducing inequality has been a priority of public policy. Various measures of income inequality can be put under two categories: positive measures and normative measures. The positive measures capture the inequality of income without value judgements. These include range quartile range, standard deviation, Gini ratio, etc. Lorenz curve belongs to this category. It measures inequality to the extent of comparing two distributions. The measures, which essentially involve value judgements about social welfare, are normative measures. These include indices propounded by Dalton, Atkinson, Sen, and Theil.

Easy comprehension and easy computation, range of variation and amount of information needed the desirable properties of the measures of economic inequality. In order to judge the efficacy of an inequality index, several axiom have been set up. However, these axioms have been relegated to the Appendix.

11.7 KEY WORDS

Co-efficient of Mean Difference : Mean of all pair-wise differences divided by the mean of differences has been termed as coefficient of mean difference in this text.

Dispersion : The fact that values of a variable are not all the same is known as dispersion. The spread or scattering of the distribution is measured by a measure of dispersion.

Extreme Disparity Ratio : The ratio of the highest value to the lowest value is known as extreme disparity ratio.

Normative Measures of Inequality : Measures of inequality, which are articulated through the explicit incorporation of social welfare function or social welfare considerations, are known as the normative measures of inequality.

Positive Measures of Inequality : Measures of inequality, which are based in statistical properties of distribution, are known as the positive measures of inequality.

Relative Standard Deviation : Standard deviation of a distribution divided by its mean is known as Relative Standard Deviation.

Standard Logarithmic Deviation : Standard deviation of logarithms of values in a distribution is known as Standard Logarithmic Deviation. Though logically the deviations of logarithmic values should be taken from the logarithm of geometric mean but at times they are taken from logarithm of arithmetic mean. Therefore, there are two versions.

Social Welfare Function : An index of social well-being, often articulated as a function of individual utilities or individual incomes or individual consumption baskets, with or without labour disposition, or individual rankings of potential state of affairs.

11.8 EXCERCISES

1) Following the adapted distribution of monthly per capita expenditure (in Rs.) in rural India in the 60th round of the NSS over January to June 2004:

Class	50-225	225-255	255-300	300-340	340-380	380-420	420-470	470-525	525-615	615-755	755-950	950-1200
Avg. Exp.	100	240	280	325	365	405	450	500	580	700	850	1100
Percentage of persons	2.4	2.7	6.4	8.3	9.6	9.6	10.8	10.0	12.2	12.6	6.7	8.8

Calculate as many positive measures as you can.

2) Following is distribution data of operational holdings from agriculture census 1976-77:

	Holding	Marginal	Small	Semi-medium	Medium	Large	All
Definition	Unit	0.0-1.0 Ha	1.0-2.0 Ha	2.0-4.0 Ha	4.0-10.0 Ha	Above 10.0 Ha	
Number	’000	44523	14728	11666	8212	2440	81569
Area	’000 Ha	17509	20905	32428	49628	42673	163343

Draw Lorenz curve and compute Gini ratio.

11.9 SOME USEFUL BOOKS

Chaubey, P. K. (2004), *Inequality: Issues and Indices*, Kanishka Publishers, Distributors, Delhi.

Cowell, Frank A. (1995), *Measuring Inequality*, Prentice Hall/Harvester Wheatsheaf, London;

Sen, A.K. (1997) *On Economic Inequality*, Oxford University Press, Oxford. (with Annexe by James E. Foster.

11.10 ANSWERS OR HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) Range is the difference between the maximum and minimum value. With a view to ensuring that the index of inequality based on this measure of dispersion is unit-free and/or is confined in the interval of (0,1), various ways of normalizations could be considered. See Sub-section 11.2.1.

2) In the relative range only extreme values are considered, which may not be representatives of the distribution. It is like comparing the poorest (who may be a few) with the richest (who may be one). Inter-quartile measures take into account the middlemost distribution with 50 percent recipients. See Sub-section 11.2.2

3) In the mean deviation, all values in a way are considered. The mean of absolute deviations is divided by the mean of the distribution to yield relative mean deviation. See Sub-section 11.2.3. Since the sum of deviations on one side of the mean will not change with the transfer of income contemplated, the magnitude of the index would not change.

4) See Sub-section 11.2.4. In one version the deviations are taken with respect to logarithm of geometric mean and in the other with respect to that of arithmetic mean. With some mathematical manipulation, one can find out that former is smaller than the latter by square of the difference between the logarithms of geometric mean and arithmetic mean.

5) It is a straight application of the fact that geometric mean of a distribution is smaller than its arithmetic mean. Of course, when the values are greater than 1!

6) Hirfindahl index is sum of squares of the shares with each recipient, which of course varies with the number of recipients. Equal distribution of shares between two recipients will yield a value of 0.5 and between three recipients, 0.333. It is therefore more used as a measure of concentration. See Sub-section 11.2.5.

7) The message is that one can try on one's own to devise new methods. See Sub-section 11.2.6. For second part, the answer is 5/9. Try it out.

Check Your Progress 2

1) Gini ratio is one half of the average value of absolute differences between all pairs, including with self, of values divided by the arithmetic mean.

See the definition by Kendall and Stuart in Sub-section 11.3.1. Have a look at expression (g.4) and (G.7).

2) The basic difference lies Gini index from other measures based on dispersion that in articulating all differences are considered while in others either few differences are considered or deviations from arithmetic/ geometric means are considered.

3) See Sub-section 11.3.2.

4) By writing out in a table, columns for class intervals or values, frequencies, class total values, cumulative frequencies, cumulative total values, cumulative proportions and cumulative shares in respect of each class. For using expression (LG.5), consecutive moving differences (or sums) of proportions and consecutive moving sums (or differences) need to be computed in two additional columns. MS-Excel will do well.

Check Your Progress 3

1) Look at the Fig. 11.4 and Sub-section 11.4.2. Try writing the properties in language.

2) When the two Lorenz curve will intersect, it will not be possible to say on balance which distribution is more unequal. In fact, one section in that case will be more unequal and the other section less unequal in distribution 1 in comparison to their counterparts in the distribution 2.

3) The value of Gini coefficient is equal to twice the areas inscribed between line of equality and the Lorenz curve. It is equivalent to saying that G is equal to (1-A) where A is the areas below Lorenz curve in the unit box.

4) Kakwani's measure of inequality is normalized length of the Lorenz curve.

Check Your Progress 4

1) Dalton's social welfare function is a simple aggregation of welfare (utility) functions of the individuals constituting the society. In addition, all individuals are supposed to have the same income utility function.

2) Sub-section 11.5.1. Write Dalton's proposal and its modern version. Also point out that though conceived in terms of utility, Dalton held that the index has to be measured in terms of income only.

3) An inequality measure should not change with linear transformation of personal utility function. Since Dalton's index does not respect this property, Atkinson is not happy. He therefore devises a new artifact called 'equally distributed equivalent income' and suggests a new utility function called iso-elastic marginal utility function.

4) Sen index is different from the Atkinson's index in one respect that he opts for a social welfare function, which has as individual incomes its arguments and is symmetric and quasi-concave. See Sub-section 11.5.3.

5) See Sub-section 11.5.5.

Appendix

AXIOMS OF INEQUALITY MEASURES

For any statistical measure, some of the desirable properties that are described in standard textbooks are (i) simplicity of comprehension, (ii) ease of computation, (iii) range of variation, and (iv) amount of information needed. However, we discuss below only those properties, which are peculiar to the measures of economic inequality.

We are often faced with situations where we have to compare two distributions with the help of an index with regard to their level of inequality. The two distributions may belong to two different countries at a point of time, to a country at two points of time, or to two situations—say, one before tax and the other after tax or before and after interpersonal transfers etc.

People have set up some intuitively appealing properties in order to judge the efficacy of an inequality index. The first set of properties was given as ‘principles’ by Dalton (1920). Today, in literature, they are known as axioms. We propose to discuss some common axioms. It may be pointed out at the outset that these axioms almost ignore the question whether inequality is an issue, which matters more (or less) in an affluent society or in a poor one. The whole discussion will assume that all incomes are positive though we know for sure in case of business failure or crop failure, incomes can well be negative or zero.

1) **Axiom of Scale Independence**

If there are two distributions of equal size such (N) that each element of one distribution is a multiple θ of the corresponding element of the other distribution, i.e.,

$$x_1^2 = \theta x_1^1 \qquad \text{i=1,2,…,N,}$$

then the numerical magnitudes of inequalities of both the distributions should be the same, i.e.,

$$I(x_1^1, x_2^1, ..., x_N^1) = I(x_1^2, x_2^2, ..., x_N^2)$$

where the inequality measure *I* is shown as a function of the distribution

$$(x_1, x_{2,} ...., x_N).$$

Obviously, it also satisfies the idea that the level of inequality should not change when the scale of measurement changes, say, from rupees to paise or bushels to quintals.

It does also mean that equal proportionate additions to all incomes would not change the level of inequality for

$$x_i (1 + \lambda) = \theta x_{\mathrm{i}} \qquad \text{i =1,2,….,N,}$$

The proportionate addition could even be negative. Thus, it is a question of shares in the cake, not the size of the cake. It is very obvious that an inequality measure is defined in terms of shares s_i because a proportionate change in all incomes leaves the shares unchanged.

However, this axiom goes against Dalton's principle of proportionate additions to income, which stated that equal proportionate additions (subtractions) should diminish (increase) the level of inequality. Perhaps, Dalton could not see that, equal proportionate addition is theoretically, equivalent to change in the scale of measurement. It should so happen in the case of a measure of relative dispersion is obvious enough.

The axiom covers the cases of proportionate taxation/subsidies. It may be noted that such additions do not change individual (class) shares of total income and, therefore, the Lorenz curve remains unchanged. All measures based on the Lorenz curve shall therefore satisfy this axiom.

It should not mean that change in the size of cake is immaterial. In the social welfare, size of cake and distribution of cake both matter. It is only in a limited context of measurement of inequality that this property is considered desirable.

Lorenz (1997) mentions an objection raised against this axiom in terms of non-proportionate increase in well-being of different income holders, which means diffusion of well-being when incomes increase but concentration when incomes decrease. Thus, this idea existed much before Dalton mentioned it. One could easily see that this objection incorporates the idea of diminishing marginal utility. The true province of the axiom then is the unit of measurement.

2) **Axiom of Population Size Independence**

The level of inequality remains unaffected if a proportionate number of persons is added to each income level.

This suggests that the magnitude of inequality in the distribution of the cake should depend on the relative number of receivers with different levels of income. If we merge two economies of identical distributions of the same size *N*, then, in the consequent economy of size *2N*, there shall be the same proportion of the merged population for any given income. Such replications will leave the inequality level unchanged. The axiom is also known as the Principle of Population Replication.

This exactly corresponds to Dalton's principle of proportionate additions of persons. Since the Lorenz curve remains unchanged so long as proportions of people in each class remains the same, the measures based on the Lorenz curve would satisfy this axiom.

Let us have however a counter-intuitive example. Let us a have two-person world in which one person is having no income and the other having is having all. Let us replicate the economy. Now there is a four-person world in which two are sharing destitution with zero income but the other two are sharing positive income equally. Earlier there was no equality; now each 50 per cent of population is sharing the income equally. So, some scholars do not accept it.

3) Axiom of Equal Income Addition

If the distribution 2 x_i^2,i=1,2,….,N is obtained by addition of equal amount d (say, through pension) to each element of distribution 1 x_i^1, i= 1,2,…N, i.e.,

$$x_i^2 = x_i^1 + d$$

then inequality level of distribution 2 should be lower than that of distribution 1 i.e.,

$$I(x_1^2, x_2^2, ..., x_N^2) < I(x_1^1, x_2^1, ...x_N^1)$$

naturally, subtractions (say, taxation) of equal amount from each income would reverse the inequality sign. It can be noted that in the former situation, the shares of the poorer persons increase and in the latter, they decrease. This axiom exactly corresponds to Dalton's principle of equal additions to incomes.

Now, we propose to discuss two very important axioms relating to transfer of an income from a person to another when other things remain the same. The former may be called Pigou-Dalton condition and the latter, Sen condition.

4) First Axiom of Income Transfer (Pigou-Dalton Condition)

If an equal transfer from a richer person to a poorer person takes place, then the level of inequality is strictly diminished, provided that the equalizing transfer amount is not more than the difference between two incomes involved. Any number of such transfers taking place between any two consecutive income units will not cause any change in the ranking of income units and therefore such a process of transfers may be called the rank-preserving equalization.

This axiom requires an inequality measure to be sensitive to transfers at all levels of income and, thus, at least a function of all incomes.

This axiom corresponds to Dalton's principle of income transfer. Dalton (1920) argued that an inequality measure must have this minimal property. Since in this context Pigou's contribution (1912) is found significant, Sen (1973) designated this axiom as Pigou-Dalton condition. Following him, a number of contributors in the field have given it the name of 'P-D condition'.

Most of the indices, barring relative range and relative mean deviation, pass this test. This axiom is also known as weak transfer axiom because it suggests the direction but not the magnitude of change in the level of inequality.

5) Second Axiom of Income Transfer (Sen Condition)

If we consider two transfers, one at a time, at different points of scale, then the transfer at lower end of scale should have greater impact than its counterpart at higher end of the scale. According to Sen, (1973), the impact on the index should be greater if the transfer takes place from a person with an income level of, say Rs.1000 to someone with Rs.900 than a similar transfer from a man with Rs.1000100 to someone with Rs.1000000.

We may see many measures do not satisfy any of the two conditions of transfer and some satisfy only the first one. Those that satisfy the second transfer axiom automatically satisfy the first transfer axiom.

6) **Axiom of Symmetry**

If distribution $(x_1^p, x_2^p, ..., x_N^p)$ were a permutation of distribution $(x_1, x_2, ..., x_N)$, then the inequality level of both the distributions would be the same.

This implies that if two persons interchange their income positions, inequality measure does not change. Thus the axiom ensures impartiality between individuals for non-income characteristics: The evaluator does not distinguish between Amar, Akabar and Anthony; nor between Shiela and Peter; or between Mr. Pygmy and Ms. Dwarfy. Further, it means that the inequality depends only on the frequency distribution of incomes.

7) **Axiom of Interval**

The inequality measure should lie in the closed interval of (0,1).

The measure is supposed to assume the value of zero when all incomes are equal, which means when all persons have equal income and the value of unity when only one individual gets all the income (and other have zero incomes, not negative incomes).

Most people tend to agree with the axiom. A few, notably Theil (1967) and Cowell (1995), disagree. They hold that the situation of one person grabbing all the income in a society of 2 persons cannot be described by the same level of inequality as that of one person doing so in a society of 2 crore persons. It would not be easy to assert that in the case of 2-person society the level of inequality is unity when one person has all the income and the other has none. In one case, 50 percent population is having positive income, in the other only 0.00000005 percent. Some people therefore qualify the axiom by saying that when one person gets all the income the measure approaches unity in the limit as the number increases.

When a measure has a finite maximum, it is easy to transform such an index into the one, which has maximum value 1. Most measures, though not all, have zero as their minimum value. But question that Cowell (1995) raises is that there are many ways in which the measure could be transformed so that it lies in the zero-to-one range but each transformation has different cardinal properties.

8) **Axiom of Decomposability**

Suppose population can be sub-divided into several groups and an over-all index of inequality was a function of group-wise indexes and if the population mean can be expressed as weighted average of group means, the population index of inequality can be regarded as decomposable. The groups might be defined as comprising of people in different occupations, residents of different areas, with different religious or educational backgrounds etc.

However we find a lot of overlapping in these groups. This leads sum of the weights to differ from unity.

Not all indices are found to be decomposable. Gini coefficient, a very popular measure is decomposable only if the constituent groups are non-overlapping. Cowell (1995) has conducted a beautiful experiment. First, he computes four inequality measures for two distributions of same size and same mean-each divided into two groups of equal size in a manner that there is no overlapping:

Population A: (60,70,80), (30,30,130)

Population B: (60,60,90), (10,60,120)

Now, it is found that the group means and population means in two distributions are the same and group inequalities in B are higher than their counterparts in A. But when we compute overall inequalities, one of the measures suggests that the magnitude of inequality in B is lower than that in A. And the measure used is Gini, which is very popular among economists. As he says, 'strange but true'. If the component inequality magnitudes are higher and the weights are the same, how could overall measure be lower? It is therefore impossible to express overall inequality (change) as some consistent function of inequality change in the consisted groups.

These are all intuitively appealing axioms. There does remain scope for formulating other axioms. In the literature on poverty measurement one finds a plethora of axioms developed by a number of contributors working in that areas. But we shall be content with these only.

I

UNIT 12 CONSTRUCTION OF COMPOSITE INDEX IN SOCIAL SCIENCES

Structure

12.0 Objectives

12.1 Introduction

12.2 Composite Index: The Concept

12.3 Steps in Constructing Composite Index

12.4 Dealing with Missing Values and Outliers

12.5 Methods to Construct Composite Index

12.5.1 Simple Ranking Method

12.5.2 Indices Method

12.5.3 Mean Standardization Method

12.5.4 Range Equalization Method

12.6 Principal Component Analysis (PCA)

12.6.1 Conducting PCA and Analyses of Results

12.6.2 Use of Output Indicators

12.6.3 Use of Weight

12.6.4 Limitations of Principal Component Analysis

12.7 Merits and Limitations of Composite Index

12.8 Let Us Sum Up

12.9 Exercises

12.10 Some Useful Books/References

12.11 Answer or Hints to Check Your Progress Exercises

12.0 OBJECTIVES

After going through this Unit, you will be able to:

- state the concept of composite index;
- describe the process of constructing the composite index;
- explain the various methods of composite index;
- discuss the merits and limitations of composite index; and
- learn how to interpret the results derived from composite indexes.

12.1 INTRODUCTION

In social sciences research, many a times the complex social and economic issues like child deprivation, food security, human well-being, human development etc. are difficult to measure in terms of single variable. The reason being that such issues have several dimensions and indicators. For example, it is difficult to explain the status of development of a district in terms of a single variable because development is reflected in terms of several indicators. Some of such variables/indicators are quantitative type while others are of qualitative nature. In such situations, Composite index plays an important role to express the single value of several inter-dependent or independent variables. Further, the composite index makes it possible to compare the performance among region/states or districts etc. Hence, composite indexes are being recognized as a useful tool for policy analysis. Composite indexes can also be resorted to make comparison among different regions/sectors where wide range of variables are used.

In this Unit, we shall therefore discuss the concept of composite index, the process of their construction, various methods, their uses, and limitations. The study of the unit will also enable you to learn how to interpret the results. Let us begin to explain the concept of composite index.

12.2 COMPOSITE INDEX: THE CONCEPT

A composite index is an expression of a single score made by combination of different scores to measure a given variable or a group of variables. It expresses quantity or place (position) of multi facet aspects of a concept. The UNDP (2005) has explained that ‘a composite index expresses a quantity or a position on a scale of qualitative multi-faceted aspects….. which is relevant for information of the society’. An index can be a combination of independent indicators, or the average of an accumulation of selected indicators. The index represent specific concept or highlighting specific sector or areas like status of dalit (dalit deprivation index), food situation (food security/insecurity index), status of child (child deprivation/development index), quality of human development (human development index) etc.

The index that we construct is the outcome of some unidirectional variables or indicators. If an index is constructed by taking positive indicators, the higher value of index implies higher development and lower values imply lower development. For example, in case of index related to child development, if the variables are positive directional, the final index can be termed as ‘child development index’. On the other hand, if the variables are negative directional, it is called ‘Child Deprivation Index’. Let us take an example that child mortality is a negatively directional variable whereas percentage of children immunized is a positively directional variable. Such index is very useful in case of qualitative data. An index is more robust than a single indicator or variable.

The choice of indicator is a big challenge for researchers. The major issues in identifying measurable indicators are: whether data are available or not, whether we have reliable data, whether the data are cross-section or time series, and minimization of double counting raised from overlap or redundancy. Again if the variables to be chosen are easy to understand, they are more acceptable to a wider audience.

12.3 STEPS IN CONSTRUCTING COMPOSITE INDEX

The following steps are required to follow in construction of composite index:

Step	Why it is needed
1. **Theoretical framework** A theoretical framework need to be developed because it provides the basis for the selection and combination of variables into a meaningful composite indicator under a fitness-for-purpose principle (involvement of experts and stakeholders is envisaged at this step).	• To get clear understanding and definition of the multidimensional phenomenon to be measured. • To structure the various sub-groups of the phenomenon (if needed). • To compile a list of selection criteria for the underlying variables, e.g., input, output, process.
2. **Data selection** Indicators should be selected on the analytical soundness, measurability, country coverage, and relevance of the indicators to the phenomenon being measured and relationship to each other. The use of proxy variables should be considered when data are scarce (involvement of experts and stakeholders is envisaged at this step).	• To check the quality of the available indicators. • To discuss the strengths and weaknesses of each selected indicator. • To create a summary table on data characteristics, e.g., availability (across country, time), source, type (hard, soft or input, output, process).
3. **Imputation of missing data** Consideration should be given to different approaches for imputing missing values. Extreme values should be examined as they can become unintended benchmarks.	• To estimate missing values. • To provide a measure of the reliability of each imputed value, so as to assess the impact of the imputation on the composite indicator results. • To discuss the presence of outliers in the dataset.
4. **Multivariate analysis** Should be used to study the overall structure of the dataset, assess its suitability, and guid subsequent methodological choices (e.g., weighting, aggregation).	• To check the underlying structure of the data along the two main dimensions, namely individual indicators and countries (by means of suitable multivariate methods, e.g., principal components analysis, cluster analysis). • To identify groups of indicators or groups of countries that are statistically "similar" and provide an interpretation of the results. • To compare the statistically-determined structure of the data set to the theoretical framework and discuss possible differences.
5. **Normalisation** Should be carried out to render the variables comparable	• To select suitable normalization procedure(s) that respect both the theoretical framework and the data properties.

	• To discuss the presence of outliers in the dataset as they may become unintended benchmarks. • To make scale adjustments, if necessary. • To transform highly skewed indicators, if necessary.
6. Weighting and aggregation Should be done along the lines of the underlying theoretical framework.	• To select appropriate weighting and aggregation procedure(s) that respect both the theoretical framework and the data properties. • To discuss whether correlation issues among indicators should be accounted for. • To discuss whether compensability among indicators should be allowed.
7. Uncertainty and sensitivity analysis Should be undertaken to assess the robustness of the composite indicator in terms of e.g., the mechanism for including or excluding an indicator, the normalization scheme, the imputation of missing data, the choice of weights, the aggregation method.	• To consider a multi-modeling approach to build the composite indicator, and if available, alternative conceptual scenarios for the selection of the underlying indicators. • To identify all possible sources of uncertainty in the development of the composite indicator and accompany the composite scores and ranks with uncertainty bounds. • To conduct sensitivity analysis of the inference (assumptions) and determine what sources of uncertainty are more influential in the scores and/pr ranks.
8. Back to the data Is needed to reveal the main drivers for an overall good or bad performance. Transparency is primordial to good analysis and policymaking.	• To profile country performance at the indicator level so as to reveal what is driving the composite indicator results. • To check for correlation and causality (if possible). • To identify if the composite indicator results are overly dominated by few indicators and to explain the relative importance of the sub-components of the composite indicator.
9. Links to other indicators Should be made to correlate the composite indicator (or its	• To correlate the composite indicator with other relevant measures, taking into consideration the results of sensitivity analysis.

dimensions) with existing (simple or composite) indicators as well as to identify linkages through regressions.	• To develop data-driven narratives based on the results.
10. Visualisation of the results Should receive proper attention, given that the visualization can influence (or help to enhance) interpretability.	• To identify a coherent set of presentational tools for the targeted audience. • To select the visualization technique which communicates the most information. • To present the composite indicator results in a clear and accurate manner.

Source: OECD (2008), 'Handbook on Constructing Composite Indicators Methodology and User Guide'

Caution on choosing variable

1) Whenever we choose any variable or a particular dimension for the index, we have to justify the inclusion of the variable into the index. This justification should come from empirical evidences or policy based research studies or from theoretical explanation.

2) If the variable is not unidirectional, the entire variables used should be converted to unidirectional. For example, in construction of the food security index, two variables like 'proportion of agricultural worker to total workers' and per capita value of agricultural output' is used for index. Here the first variable is a negatively directional whereas the second variable is positively directional. In this case we have to convert the entire variables into either positive direction or negative direction. If we want to convert this to positive direction, the first variable which has a negative direction should be deducted from 100. On the other hand, if we want to convert the entire variables into negative direction, we have to work out the reciprocal of per capita value of agricultural output.

12.4 DEALING WITH MISSING VALUES AND OUTLIERS

After selecting the indicators, you have to have a clear idea on missing values for each selected variable. Data can be missing in a random or non random fashion. In such situation, easy solution is to drop cases for which data is missing. However, before dropping the variable, you have to see the number of cases for which values are missing, because exclusion of such households for which data is missing could significantly lower sample size. Further, deleting such household may lead to some biases also. For example, in livelihood study of households, if there is a missing value, we have to find out the frequency of missing value because in many cases, the chances of missing value is high for lower economic class as compared to upper economic class. In such cases if we delete the households having missing value, it may result biases towards upper economic classes of household. In case, inclusion or exclusion of households having missing value put a little impact on the final result, we can delete such cases.

The second solution is to impute the missing value by applying some methodology or logic. But use of imputation is more common in academic indices and datasets. Here you can try to impute the missing value which should be as accurate as possible..For example, in calculating the consumption expenditure of 50 households of which about 5 cases are missing. In this case in imputing we can substitute the average value of consumption expenditure and replace the missing value by the averages. But for a more accurate estimation we can group theses 50 households on the basis of value of assets holding and see to which asset category these 'MPCE missing' households belonged to. Then we can substitute the average consumption expenditure of the respected asset group.

Such approach has been used in a number of more recent indices such as the Corruptions Perceptions Index (CPI) developed by Transparency International and the World Bank's Worldwide Governance Indicators (WGI). Such approaches carry a dual advantage. They allow scores to be estimated for a maximal number of countries, and can use a broader range of indicators to triangulate indices for nebulous constructs.

Many a time,outliers may also disturb the analysis. For example if the income of four persons is 18000, 17000, 18500 and 19000 then the average is 18125. If we include the income of fifth person as 60000 then the average turns out to be 26500. In such matters, you have to drop the cases with high extreme value. Both in case of missing value and outlier you are expected to document and explain the selected imputation procedures and the results in detail.

Check Your Progress 1

1) What do you understand by the term Composite Index?

………………………………………………………………………………..

………………………………………………………………………………..

………………………………………………………………………………..

………………………………………………………………………………..

………………………………………………………………………………..

2) Give some examples of Composite Index.

………………………………………………………………………………..

………………………………………………………………………………..

………………………………………………………………………………..

………………………………………………………………………………..

………………………………………………………………………………..

3) List the steps involved in construction of Composite Index.

………………………………………………………………………………..

………………………………………………………………………………..

………………………………………………………………………………..

………………………………………………………………………………..

………………………………………………………………………………..

4) State the alternative methods to deal with the missing values of the selected variables.

…………………………………………………………………………………..

…………………………………………………………………………………..

…………………………………………………………………………………..

…………………………………………………………………………………..

…………………………………………………………………………………..

…………………………………………………………………………………..

…………………………………………………………………………………..

…………………………………………………………………………………..

12.5 DIFFERENT METHODS OF COMPOSITE INDEX

The various methods are used in construction of composite index. These include: Simple Ranking Method, Indices Method, Mean standardization method, Range Equalization Method, and Principal Component Analysis Method. Let us discuss one by one with simple examples.

12.5.1 Simple Ranking Method

Rank Method is one of the simplest methods to analyse the status of a region/district/state/country. In this method, the first step is to arrange/convert the variable into unidirectional for each district or state. Let us take an example of development status of different districts in Orissa state. First, we have to convert the relevant variables in either positive or negative direction. For example, the variable says 'the proportion of agricultural labour is negatively associated with the development of a region. Hence we have to convert this variable to positive direction by deducting this variable from 100 and change the labeling of variable as 'proportion of other than agricultural labour to total labour'. Higher the value of this variable, higher the development of the district or region. Depending on the value, each variable was ranked in the similar manner. Highest rank (1st) was given to the variables with high value and vice-a-versa. Alternatively, we can do reverse ranking of the variable i.e. highest rank is given to the variables with lowest value. The individual ranks are added to get the total rank value for the district. This has been illustrated in the example given in table 12.1. Our objective is to find out most backward tribal district in Odisha (having more than 50 per cent of tribal population to total population). A total of 11 districts qualify for tribal dominant area having 50 per cent or more proportion of population. We have selected 10 variables for identifying most backward tribal areas. Each of the variables for 11 tribal dominated districts was ranked according to total value of variable. The individual rank of all the variables of all the districts is given in table 12.2. As all the variable are unidirectional, rank 1 is given to district having higher value and vice-a-versa. After doing so we can add together the value of all the variables and find out the average rank for each district (Sum of total rank divided by number of variable i.e. 10). The average value of variable shows the status of district on development indicators.

Table 12.1: Districtwise Development Indicators of Tribal Concentrated Districts in Odisha

District	% of other than agricultural labour to all labour	% of net irrigated area to total sown area	Per capita value of agrictural output	Per capita consumption expenditure	Female Literacy rate	women work force participation rate	% of household getting safe drinking water	Percentage of villages having access to paved road	Percentage of villages having access to Phcs	Average casual wage rate
Actual Value										
Malkangiri	73	31	1304	201	18	67	82	21	16	36
Rayagada	50	23	479	201	18	67	78	34	13	37
Sundargarh	61	19	537	280	43	58	57	38	28	32
Nabarangapur	46	6	839	201	18	66	80	45	28	37
Kandhamal	62	14	410	201	33	68	32	21	19	36
Koraput	55	31	611	201	16	67	67	22	22	40
Mayurbhanj	60	30	572	280	35	62	44	42	21	31
Gajapati	52	25	529	331	24	77	43	29	20	34
Sambalpur	62	34	1075	280	50	62	56	31	13	39
Kendujhar	60	23	537	280	44	46	52	46	15	34
Jharsuguda	68	20	662	280	54	43	63	47	32	39
Average	59	23	687	249	32	62	60	34	21	36

Table 12.2: Districtwise Development Scenario of Tribal Orissa by Simple Ranking Method

District	% of other than agricutral labour to all labour	% of net irrigated area to total sown area	Per capita value of agrictural output	Per capita consumption expenditure	Female Literacy rate	women work force participation rate	% of household getting safe drinking water	Percentage of villages having access to paved road	Percentage of villages having access to primary health centre	Average casual wage rate	Average Rank/indices
Malkangiri	1	2	1	7	8	5	1	10	8	7	5.0
Rayagada	10	7	10	7	9	3	3	6	11	4	7.0
Sundargarh	5	9	7	2	4	9	6	5	3	10	6.0
Nabarangapur	11	11	3	7	10	6	2	3	2	5	6.0
Kandhamal	3	10	11	7	6	2	11	11	7	6	7.4
Koraput	8	3	5	7	11	4	4	9	4	1	5.6
Mayurbhanj	7	4	6	2	5	7	9	4	5	11	6.0
Gajapati	9	5	9	1	7	1	10	8	6	8	6.4
Sambalpur	4	1	2	2	2	8	7	7	10	3	4.6
Kendujhar	6	6	8	2	3	10	8	2	9	9	6.3
Jharsuguda	2	8	4	2	1	11	5	1	1	2	3.7

By this method, the most developed tribal district is Jharsuguda and most backward district is kandhamal.

12.5.2 Indices Method

The indices method is another simple method for calculating the status of development of an area/district or state. Like rank method, here also, at the first stage the variables are converted into one direction. After converting the variable into one (positive or negative) direction, we calculate the index. In this method we have to convert the district figures based on the average figure for the entire 11 districts as 100. Let us take an example of table 12.1. If the proportion of other than agricultural labour to total labour in malkangiri district is 73 and the average of that variable for all the 11 districts is 59, then the indices of that variable for Malkangiri district will be: $\frac{100}{59}x73 = 124$

In the same manner, we can find out the indices for all the variables for 11 districts as is given in Table 12.3. The final index of development for the districts can be obtained by taking the arithmetic mean for all the indicators (given in last column of table 12.3). After working out the average indices value for all the districts, we can rank all the districts. The district with highest indices will be most developed and the district with lowest indices will be least developed.

Table 12.3: Districtwise Development Scenario of Tribal Orissa by Indices Method

District	% of other than agricultural labour to all labour	% of net irrigated area to total sown area	Per capita value of agricultural output	Per capita consumption expenditure	Female Literacy rate	women work force participation rate	% of households getting safe drinking water	Percentage of villages having access to paved road	Percentage of villages having access to primary health centre	Average casual wage rate	Average Rank/indices
Malkangiri	124	133	190	81	57	108	138	62	78	99	107.0
Rayagada	85	99	70	81	57	108	131	98	64	103	89.6
Sundargarh	104	83	78	113	135	93	96	111	135	90	103.6
Nabarangapur	78	25	122	81	56	107	135	132	137	102	97.4
Kandhamal	106	59	60	81	102	109	54	60	92	100	82.3
Koraput	93	133	89	81	49	108	113	65	108	113	95.1
Mayurbhanj	101	129	83	113	109	101	74	123	103	86	102.2
Gajapati	89	106	77	133	76	123	73	86	95	95	95.3
Sambalpur	105	147	157	113	155	100	95	90	65	108	113.3
Kendujhar	102	99	78	113	136	74	88	136	71	94	99.1
Jharsuguda	114	88	96	113	167	69	105	138	153	109	115.3

In our example Kandhamal is the most backward with indices (82.3) and Jharsuguda (115.3) is the most developed district.

Let us make a comparison between the results obtained by rank method and indices method. Results in Table 12.4 shows that Jharsuguda, Sambalpur, and Malkangiri are most developed in both the methods whereas Raygada and Kandhamal are most backward in both the methods. While comparing both rank and indices method, we can also run correlation between the two set of indices. If the correlation is high then we can say that the finding in both methods is almost similar.

Table 12.4: A Comparison of Simple Ranking Method and Indices Method

District	Status of District in Rank Method	District	Status of District in indices Method
Jharsuguda	3.7	Jharsuguda	115
Sambalpur	4.6	Sambalpur	113
Malkangiri	5.0	Malkangiri	107
Koraput	5.6	Sundargarh	104
Sundargarh	6.0	Mayurbhanj	102
Nabarangapur	6.0	Kendujhar	99
Mayurbhanj	6.0	Nabarangapur	97
Kendujhar	6.3	Gajapati	95
Gajapati	6.4	Koraput	95
Rayagada	7.0	Rayagada	90
Kandhamal	7.4	Kandhamal	82

Note: Correlation of status of district in both rank and indices method is 0.927

12.5.3 Mean Standardization Method

Mean standardization Method is another simple method used both as a process of normalization and also a composite index. In this method we normalize the value of each variable and than work out the average of the normalized value for all the variables. The average of normalized value will be the composite index value. The normalization is done by dividing the actual value of variables by their respective means. Let us take an example from Table 12.1. For example the standardized value of the variable '% of other then agricultural labour to all workers' for Malkangiri district can be found out by

Normalized value MS Method=Actual value (73)/Mean value (59) =1.237

In the same process we can find out the normalized value of the entire variable for 11 selected districts. The standardized value of each selected indicator for different district is given in Table 12.5. The composite index value of each district given in the last column of Table 12.5 has been worked out by taking average of all the 10 selected indicators (row vector of the each district).

Table 12.5: Index Value in Mean Standardization Method

District	Percentage of other than agrictural labour to all labour	Percentage of net irrigated area	Per capita value of agricutural output	Per capita consumption expenditure	Female Literacy rate	women work force participation rate	Percentage of household getting safe drinking water	Percentage of villages having access to paved road	Percentage of villages having access to primary health centre	Average casual wage rate	Composite Index
Malkangiri	1.237	1.332	1.899	0.808	0.561	1.079	1.379	0.614	0.775	1.003	1.069
Rayagada	0.847	0.988	0.697	0.808	0.561	1.079	1.312	0.995	0.630	1.030	0.895
Sundargarh	1.034	0.816	0.782	1.126	1.340	0.934	0.959	1.112	1.357	0.891	1.035
Nabarangapur	0.780	0.258	1.222	0.808	0.561	1.063	1.346	1.316	1.357	1.030	0.974
Kandhamal	1.051	0.602	0.597	0.808	1.028	1.095	0.538	0.614	0.921	1.003	0.826
Koraput	0.932	1.332	0.890	0.808	0.499	1.079	1.127	0.644	1.066	1.114	0.949
Mayurbhanj	1.017	1.289	0.833	1.126	1.091	0.999	0.740	1.229	1.018	0.863	1.020
Gajapati	0.881	1.074	0.770	1.331	0.748	1.240	0.723	0.848	0.969	0.947	0.953
Sambalpur	1.051	1.461	1.565	1.126	1.558	0.999	0.942	0.907	0.630	1.086	1.132
Kendujhar	1.017	0.988	0.782	1.126	1.371	0.741	0.875	1.346	0.727	0.947	0.992
Jharsuguda	1.153	0.859	0.964	1.126	1.683	0.693	1.060	1.375	1.551	1.086	1.155

12.5.4 Range Equalization Method

Range Equalization (RE) method otherwise known as max-min approach is adopted by UNDP in computation of Human Development Index. Under this approach, an index is constructed for each variable by applying Range Equalization formula derived by UNDP. This is worked out by subtracting an indicator's minimum value from each observation and then dividing it by its range.

$$\text{RE Index} = \frac{X_i - Min\ X}{Max\ X - Min\ X}$$

where X_i Value of the variable, min X- Minimum value of X in the scaling , max X- Maximum value of X in the scaling. The RE index is also a normalization technique. Without it, a composite index can be biased towards an indicator with very high range. For example the per capita value of agricultural output which is measured in rupees ranges from Rs. 410/- to Rs. 1304/-. On the other hand, the variable proportion of female literacy rate is measured in percentage. Different variables measured in different units sometimes give upward or downward biases. To ensure the index value biasfree, we convert all the variables into equal scaling from 0 to 1.

In RE method, as a first step we undertake scaling exercise. In undertaking the scaling procedure, desirable norms are followed for each indicator. In some cases, scaling of indicators is self-selected, while in others element of value judgment is involved. This scaling exercise is called goal post where we identify the maximum and minimum goal. The goalpost basically visualizes the extent of minimum value and maximum value in the future time period (say next 5 years). The scaling norm that we have adopted is given in Table 12.6. In the table the third and fourth column shows the maximum and minimum district value of selected 10 indicators. The first and second column is the maximum and minimum goal post.

Table 12.6: Construction of Food Security Radar

Variable	Description of Variables	Goalposts		District Value	
		Minimum	Maximum	District Minimum	District Maximum
Oth_agl	Percentage of other than agricultural labour to all labour	30	85	46	73
Irr	Percentage of net irrigated area to net cropped area	2	55	6	34
pcvao	Per capita value of agricultural output	200	2500	410	1304
Mpcce_ia	Inequality adjusted per capita consumption expenditure	150	450	201	331
Lit_f	Female Literacy (adult) rate	10	70	16	54
Wfpr_f	women work force participation rate	30	85	43	77
Hhsdw	Percentage of household having access to safe drinking water	20	90	32	82
Paved_r	Percentage of villages having access to paved road	10	60	21	47
v_phcs	Percentage of villages having access to Primary health centre	10	50	13	32
Wage_c	average Casual wage rate	25	70	31	40

Based on the goalpost, we normalize the value of the variable for all the districts and all the indicators. In our example, the values have been presented in Table 12.7.

Table 12.7: Index Value in Range Equalization Method

District	Percentage of other than agrictural labour to all labour	Percentage of net irrigated land to net sown area	Per capita value of agrictural output	Per capita consumption expenditure	Female Literacy rate	women work force participation rate	Percentage of households getting safe drinking water	Percentage of villages having access to paved road	Percentage of villages having access to primary health centre	Average casual wage rate	Composite Index
Malkangiri	0.782	0.547	0.480	0.170	0.133	0.673	0.886	0.220	0.150	0.244	0.429
Rayagada	0.364	0.396	0.121	0.170	0.133	0.673	0.829	0.480	0.075	0.267	0.351
Sundargarh	0.564	0.321	0.147	0.433	0.550	0.509	0.529	0.560	0.450	0.156	0.422
Nabarangapur	0.291	0.075	0.278	0.170	0.133	0.655	0.857	0.700	0.450	0.267	0.388
Kandhamal	0.582	0.226	0.091	0.170	0.383	0.691	0.171	0.220	0.225	0.244	0.300
Koraput	0.455	0.547	0.179	0.170	0.100	0.673	0.671	0.240	0.300	0.333	0.367
Mayurbhanj	0.545	0.528	0.162	0.433	0.417	0.582	0.343	0.640	0.275	0.133	0.406
Gajapati	0.400	0.434	0.143	0.603	0.233	0.855	0.329	0.380	0.250	0.200	0.383
Sambalpur	0.582	0.604	0.380	0.433	0.667	0.582	0.514	0.420	0.075	0.311	0.457
Kendujhar	0.545	0.396	0.147	0.433	0.567	0.291	0.457	0.720	0.125	0.200	0.388
Jharsuguda	0.691	0.340	0.201	0.433	0.733	0.236	0.614	0.740	0.550	0.311	0.485

After calculating the index of each variable, we have averaged them to give each of the five dimensions of food security. The composite food security index is again derived by averaging the five dimensions.

The normalized value for the variable percentage of other than agricultural labour to total worker' for Malkangiri district is found out in the following manner.

$$\frac{(\text{Actual value } (73) - \text{Minimum goalpost } (30))}{(\text{Maximum goalpost } (85) - \text{Minimum goalpost } (30))} = 0.782$$

Likewise we have converted and worked out the normalized value of all the variables and for all the districts shown in Table 12.7.

Comparing MS Method and RE Method

The composite index value worked out by both the RE and MS methods can be compared and analyzed (Table 12.8). We can also examine correlation between two indices. Here the correlation between final index value worked out by both the methods is very high (0.991).

Table 12.8: Comparison between RE and MS Methods

District	Composite Index Value by RE method	Composite Index Value by MS method
Jharsuguda	1.155	0.485
Sambalpur	1.132	0.457
Malkangiri	1.069	0.429
Sundargarh	1.035	0.422
Mayurbhanj	1.020	0.406
Kendujhar	0.992	0.388
Nabarangapur	0.974	0.388
Gajapati	0.953	0.383
Koraput	0.949	0.367
Rayagada	0.895	0.351
Kandhamal	0.826	0.300

Note: Correlation of index value is 0.991

The high degree of the correlation between the values of composite index computed by applying both methods separately indicate the high degree of homogeneity between the two approaches. However, the Range equalization method is preferred because it accounts for wider variations and strong correlations to the PCA composite.

One important point should be kept in mind that the index value as discussed above are based on equal weight. We can also give weights to variables/ components discussed in section 12.5.

Check Your Progress 2

1) State the procedure to workout Composite Index by way of Simple Ranking Method.

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

2) How will you compute the final value of Composite Index by Indices Method?

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

3) What is the Distinction between Range Equalization Method and Mean Standarisation Method?

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

………………………………………………………………………………

12.6 PRINCIPAL COMPONENT ANALYSIS (PCA)

PCA is one of the important composite index method used to analyse the various problems in social sciences. PCA is a mathematical technique that transfers a number of correlated variables into smaller number of uncorrelated variables called the principal components. For our analysis we have an array of variables which may have high correlation with each other. In case where there is high relation between variables, the PCA is a suitable technique where the correlation between different components is low.

Categorical data are not suitable for PCA analysis because in such cases categories converted into quantitative scale has no meaning. To avoid this type of problems, we should recode these variables into binary variables. Let us take an example of social category of household that are Scheduled caste (1), Scheduled tribe (2), Other backward caste (3) and General (4). In PCA analysis such type of variables have no meaning. Such type of categorical variables can be converted into a bianary variable. For example if we want to study the dalit status, variable can be converted to a bianary variable as dalit (code 2) and non-dalit (code 2, 3, and 4). Likewise, if we want to study the impact of reservation for other backward castes we can categorize as OBC (code 3) and other than OBC (code 1, 2 and 4).

The PCA is a data reduction technique. Under this technique, the original data set is transformed into a new set of uncorrelated variables called principal components. PCA reduces the number of variables in a data set to smaller number of dimension/s. In a series of variables if a_{mn} is the weight of m_{th} principal component and nth variable then the matrix will be:

$$\left\{\begin{matrix} PC_1 \\ PC_2 \\ PC_3 \\ . \\ . \\ . \\ PC_n \end{matrix}\right\} = \left\{\begin{matrix} a_{11} & a_{12} & a_{13} & .... & a_{1n} \\ a_{21} & a_{22} & a_{23} & .... & a_{2n} \\ . & & & & \\ . & & & & \\ . & & & & \\ a_{m1} & a_{m2} & a_{m3} & .... & a_{mn} \end{matrix}\right\} \left\{\begin{matrix} X_1 \\ X_2 \\ X_3 \\ \\ \\ \\ X_n \end{matrix}\right\}$$

The matrix can be transformed into equation for principal component i.e.

$$PC_1 = a_{11}x_1 + a_{12}x_2 + ... + a_{1n}x_n$$

$$PC_2 = a_{21}x_1 + a_{22}x_2 + ... + a_{2n}x_n$$

....

$$PC_n = a_{m1}x_1 + a_{m2}x_2 + ... + a_{mn}x_n$$

The components are ordered so that the first component (PC_1) explains the largest possible amount of variation in the original data, subject to the constraint that the sum of the squared weights of the vectors $(a_{11}^2 + a_{12}^2 + \cdots + a_{1n}^2)$ is equal to one.

In the output the eigen value is one of the important determinant of PCA. The eigen value is a number that tell you how much variance there is in the data set. In other words the eigen value is a number telling us how spread out the data is on the line. The eigen vector with the highest eigen value is therefore the principal component.

As the sum of the Eigen values equals the number of variables in the initial data set, the proportion of the total variation in the original data set accounted by each principal component is given by λ_i/n. The second component (PC_2) is completely uncorrelated with the first component, and explains additional but less variation than the first component, subject to the same constraint. Each subsequent component captures additional dimension in the data and adds smaller and smaller proportion of variation in the original variables.

Before conducting PCA, we should run correlation between variables. If the correlation between two variables is very high, we may remove one of those variables. The reason being that the two variables seem to measure same thing. The higher the degree of correlation among variables, the lower will be the number of components required to capture common information. Sometimes two variables may be combined in some ways (taking average). The PCA can be applied either to the original values of variables or to the normalized values of the variables.

In general, normalization can be done by three methods, i.e (i) by deviation of the variables from their respective means (i.e.); (ii) by dividing the actual values by their respective means; (iii) and deviation of value of a variable from the mean which is then divided by standard deviation {i.e. ()/σ}.We are applying here the second method.

Let us try to apply and analyse PCA by using the database given in Table 12.1. We are applying here the second method for normalization that is found out in Table 12.5 column 2 to 10.

Table 12.9: KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.394
Bartlett's Test of Sphericity	Approx. Chi-Square	47.933
	Df	45
	Sig.	.355

When the normalized database is prepared, you have to apply Kaiser-Meyer-Olkin (KMO) technique and Measure of Sampling Adequacy and Bartlett's Test of Sphericity (BTS). In KMO technique, the value varies between 0 and 1 and the value closer to1 is better. However, according to some statisticians, the minimum value should be 0.6. The BTS tests the null hypothesis that the correlation matrix is an identity matrix. Let us remember that identity matrix is a matrice in which the main diagonal elements are 1 and off-diagonal elements are 0.

The above two tests provide the minimum standard before conducting any PCA.

12.6.1 Conducting PCA and Analyzing Results

For the purpose of Principal Component analysis and interpreting its results, SPSS is user friendly software. However, you can try to run the same in STATA also. Let us take our previous example of Table 12.5. In the table, there are 10 standardised variables and we can run our PCA model. Table 12.10 give the first output called 'communalities'

Table 12.10: (Output 1) Communalities: Extraction by Principal Component Analysis.

Variable name	Initial	Extraction
% of other than agricultural labour to all labour	1.000	.721
% of net irrigated area to net sown area	1.000	.748
Per capita value agricultural output	1.000	.710
Monthly per capita consumption expenditure	1.000	.702
Female Literacy rate (adult)	1.000	.876
women work force participation rate	1.000	.858
% of household access to safe drinking water	1.000	.789
% of villages having access to paved road	1.000	.773
% of villages having access to primary health centres	1.000	.572
Average Casual wage rate	1.000	.491

Communalities represent the percentage of variance explained by the extracted components. This explains the proportion of each variable’s variation explained by PCA. The short notation of communalities is ‘h^2’ and defined as sum of squared factor loading. In table 12.10, the first column is the name of variables used in our PCA. The second column is the ‘initials’. The initial value of the communalities in PCA is 1. The second column ‘extraction’ shows the proportion of each variable’s variation captured by PCA. This value ranges between 0 and 1. When the value is near to 1 that means the variable is well represented and the reverse for value is closer to ‘0’.

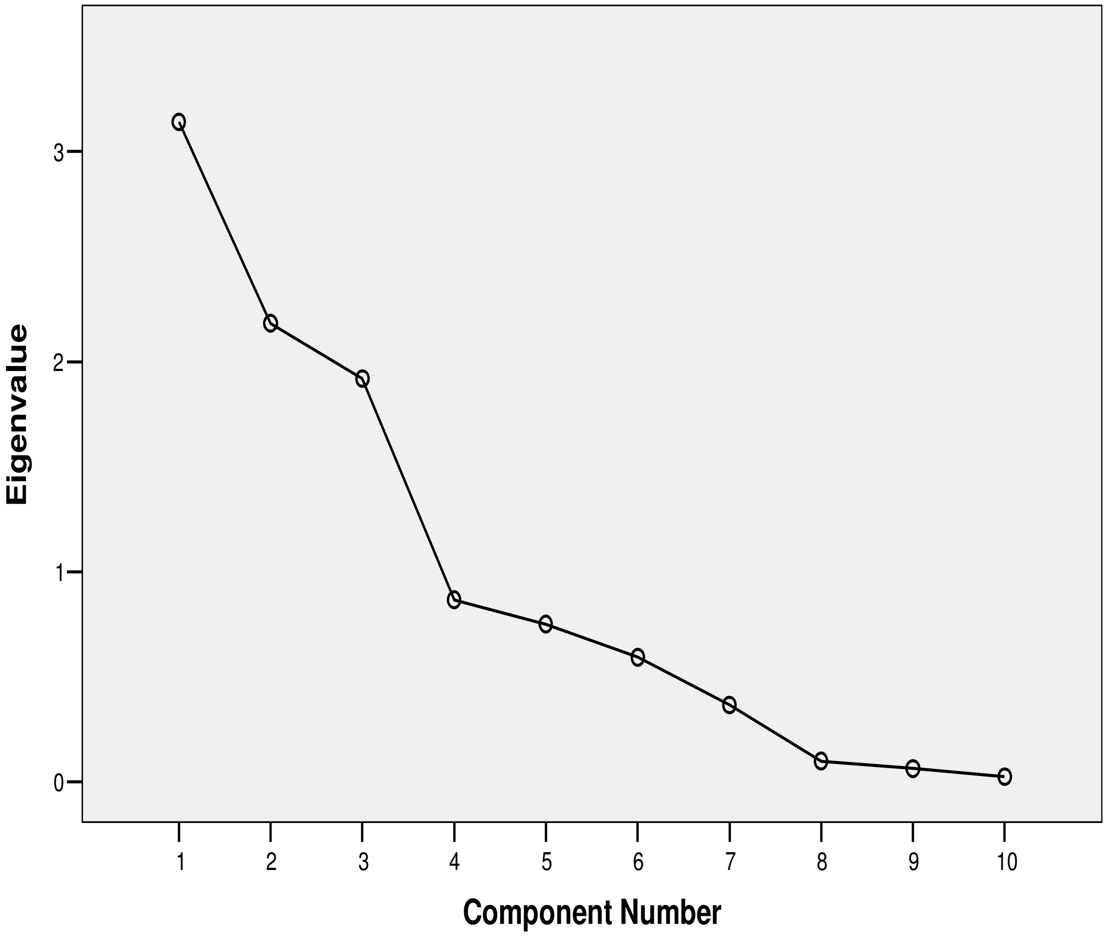
If the communality is low for an item, the reason might be that the item was poorly designed. If the item has very little variance the communality is low. If the different items usually resulting from large positive or negative skewed the communalities is low. We can take an example that if everyone ticks strongly agree, the variation within the variable is low and the communality is low. If the communality is low, we can either remove the item from the analysis to exclude it from any further analyses or we can treat it as a stand alone variable.

Table 12.11: (Output 2) Total Variance Explained: Extraction Method: Principal Component Analysis.

	Initial Eigen values			Extraction Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3.141	31.413	31.413	3.141	31.413	31.413
2	2.182	21.820	53.233	2.182	21.820	53.233
3	1.919	19.195	72.428	1.919	19.195	72.428
4	.866	8.658	81.086			
5	.750	7.502	88.587			
6	.593	5.927	94.514			
7	.365	3.645	98.160			
8	.098	.976	99.135			
9	.062	.625	99.760			
10	.024	.240	100.000			
	10.0					

The Table 12.11 shows ‘total variance explained’. The first column shows the number of components. The number of components should be equal to number of variables used for PCA. In our example, we have used 10 variables and hence the total number of component is 10. The second part of the table reflect initial Eigen value which consists of three columns i.e. Total, percentage of Variance, Cumulative percentage. The initial Eigen values are the variances of principal components. Here as we use the standardized values of variable, the variance becomes equivalent to 1 and the total variance is equal to the total number of variables i.e. ‘10’. The second column ‘total’ contains the Eigen values. Here it can easily be seen that the Eigen values of subsequent components gradually reduce implying that the successive components add less and less variation. The third column shows the percentage variation on the components and the fourth column explains the cumulative addition of percentage variation of cumulative percentage components. In our example the first component explain 31.4 per cent of total variation and the second component explain 21.8 percent of total variation and so on. The third row of fourth column shows that the first three components explain 72 per cent of total variation.

The second part of the table is ‘Extraction Sums of Squared Loadings’ which consists of three columns – Total, percentage of Variance and Cumulative percentage. This has actually reproduced the three rows of figures of the first part of table reflecting the components whose eigen value is greater than 1.



Graph 12.1: Scree plot

The ‘scree plot’ is a graph whose ‘X axis’ represent the component number and ‘Y axis’ represents ‘Eigen values’. In other words scree plot demonstrates the graph of first two column of ‘output 2’ table. From the graph one can select the number of components to be taken for analysis. It can be clearly seen from the above graph that from the fourth component onward the line becomes flat indicating that when the successive addition is less and less, the line of the graph becomes more and more flatter.

Table 12.12: (Output 4) Component Matrix(a)

	Component[a]		
	1	2	3
Percentage of other than agricultural labour to all labour	.202	.807	.174
Percentage of net irrigated area to total sown area.	-.185	.793	-.293
Per capita value of agricultural output	-.403	.596	.438
Monthly per capita consumption expenditure	.717	.217	-.378
Female Literacy rate	.837	.406	.105
women work force participation rate	-.721	-.233	-.533
Percentage of households having access to safe drinking water	-.504	-.041	.730
Percentage of villages having access to paved road	.733	-.291	.389
Percentage of villages having access to primary health centre	.441	-.417	.451
Average Casual wage rate	-.430	.145	.535

Extraction Method: Principal Component Analysis.
a 3 components extracted.

Table 12.12 (output 4) shows Component Matrix of first three components. The first column is the name of variable used in PCA. The values in the output table are the component loading which are the correlation between the variables and components. The values in the tables range from -1 to +1. One can note that the sum of square of all three component matrix of a variable is equal to the extraction value of the communalities of that variable. In this example three component has been extracted as the eigen value for these component is greater than 1.

Table 12.13: (Output 5) Reproduced Correlations

	Percentage of other than agricultural labour to all labour	Percentage of net irrigated area to total sown area	Per capita value of agricutural output	Per capita consumption expenditure	Female Literacy rate	women work force participation rate	Percentage of of households getting safe drinking water	Percentage of villages having access to paved road	Percentage of villages having access to primary health centre	Average casual wage rate
Reproduced Correlation										
Percentage of other than agricutural labour to all labour	0.721	0.554	0.477	0.253	0.512	-0.424	-0.007	-0.022	-0.171	0.120
Percentage of net irrigated to total sown area	0.554	0.748	0.419	0.151	0.137	0.102	-0.152	-0.479	-0.543	0.032
Per capita value of agricutural output	0.477	0.419	0.710	-0.324	-0.049	-0.083	0.500	-0.297	-0.228	0.491
Per capita consumption expenditure	0.253	0.151	-0.324	0.702	0.648	-0.366	-0.645	0.314	0.055	-0.479
Female Literacy rate	0.512	0.137	-0.049	0.648	0.876	-0.754	-0.362	0.536	0.247	-0.246
women work force participation rate	-0.424	0.102	-0.083	-0.366	-0.754	0.858	-0.015	-0.668	-0.460	-0.010
Percentage of households having access to safe drinking water	-0.007	-0.152	0.500	-0.645	-0.362	-0.015	0.789	-0.074	0.123	0.603
Percentage of villages having access to paved road	-0.022	-0.479	-0.297	0.314	0.536	-0.668	-0.074	0.773	0.620	-0.144
Percentage of village access to primary health centre	-0.171	-0.543	-0.228	0.055	0.247	-0.460	0.123	0.620	0.572	-0.005
Average Casual wage rate	0.120	0.032	0.491	-0.479	-0.246	-0.010	0.603	-0.144	-0.005	0.491
Residual(a)										
Percentage of other than agricutural labour to all labour		-0.149	-0.039	-0.161	-0.040	0.002	-0.087	-0.161	0.154	-0.134
Percentage of net irrigated area to total sown area	-0.149		-0.069	0.096	-0.066	-0.030	0.121	0.111	0.016	0.033
Per capita value of agricutural output	-0.039	-0.069		0.139	-0.028	0.145	0.045	0.089	0.046	-0.180
Monthly per capita consumption expenditure	-0.161	0.096	0.139		-0.039	0.119	0.120	0.084	0.067	0.045
Female Literacy rate	-0.040	-0.066	-0.028	-0.039		0.008	-0.082	-0.054	-0.052	0.151
women work force participation rate	0.002	-0.030	0.145	0.119	0.008		0.001	-0.016	0.147	-0.025
Percentage of households having access to safe drinking water	-0.087	0.121	0.045	0.120	-0.082	0.001		0.143	-0.070	-0.173
Percentage of villages having access to paved road	-0.161	0.111	0.089	0.084	-0.054	-0.016	0.143		-0.183	-0.113
Percentage of villages having access to primary health centre	0.154	0.016	0.046	0.067	-0.052	0.147	-0.070	-0.183		-0.007
Average Casual wage rate	-0.134	0.033	-0.180	0.045	0.151	-0.025	-0.173	-0.113	-0.007	

Extraction Method: Principal Component Analysis.

a) Residuals are computed between observed and reproduced correlations. There are 29 (64.0%) no redundant residuals with absolute values greater than 0.05.

b) Reproduced communalities

Table 12.13 (Output 5) has two parts of analysis: reproduced correlation and residuals. The reproduced correlation is the correlation among the extracted components. From this table, we intend to ensure that the correlation between the original variables and reproduced matrix should be as close as possible. The lower part of the table 'residual matrix' shows the difference between original variable and residual matrix. For a good PCA, it was expected that the difference between original variable and extracted matrix should be near to zero. Once the difference is near to zero, it can be said that the extracted components accounted a larger variation in the original correlation matrix. Here we can take an example from the table that the original correlation between irrigation and other than agricultural labour is 0.403, Whereas the extracted correlation between these two variable is 0.554 and the difference between the two correlation given in residual part is 0.403-0.554= -0.149.

Final PCA Index Value

The final index is calculated by the addition of multiplication of normalized value of the variable and the Eigen vector of that variable (first component). In our example the PCA index value for all the 11 selected district can be calculated by using two Table 12.5 and Table 12.12 given above. The final PCA index of first district (Malkangiri) is

PCA index Malkangiri District = a11x1+a12x2+……+a1nxn

= (0.202*1.237) + (-0.185*1.332) + (-0.403*1.899) + (0.717*0.808) + (0.837*0.561) + (-0.721*1.079) + (0.504*1.379) + (0.733*0.614) + (0.441*0.775) + (-.430*1.003)= 0.049

Likewise the PCA Index value of all other districts can be calculated as given in Table 12.14

Table 12.14: PCA Index Value

District	PCA Index
Malkangiri	0.049
Rayagada	1.273
Sundargarh	1.322
Nabarangapur	1.145
Kandhamal	1.641
Koraput	0.517
Mayurbhanj	1.333
Gajapati	1.757
Sambalpur	0.465
Kendujhar	0.290
Jharsuguda	1.551

The table 12.14 reveals that Gajapati is the most developed district followed by Kandhamal district. On the other hand, Malkangiri is most backward district.

12.6.2 Use of Output Indicators

After finding out the Index findings need to be validated. This can be done by comparing the result with some output indicators. Let us illustrate with example discussed in section 12.6. District wise final index as arrived out by PCA method is given in Table 12.14. This finding can be validated be taking an output indicator say infant mortality rate (IMR). We can run a correlation between PCA index and the output indicator 'Infant mortality rate'. This has been provided in Table 12.15 which is 0.893. The table shows that the most backward district has a higher degree of mortality. Hence our final index is validated. If suppose the correlation between these two variable is -0.120, then our index is not validated as the correlation between the two variable is very low.

Table 12.15: PCA Index and Comparing with Output Indicators

District	PCA Index	Infant Morality Rate*
Malkangiri	0.049	120
Kendujhar	0.290	100
Sambalpur	0.465	110
Koraput	0.517	92
Nabarangapur	1.145	89
Rayagada	1.273	100
Sundargarh	1.322	80
Mayurbhanj	1.333	75
Jharsuguda	1.551	65
Kandhamal	1.641	60
Gajapati	1.757	55

Note: Imaginary numbers

12.6.3 Use of Weight

Assigning weight i.e. all the indicators/dimensions are treated equally or differentially is an important issue in construction of composite index.. Sometimes different dimensions are given differential weight but in construction of overall index, an equal weight approach is followed. The human development index assigns differential weights to indicators/dimension but for overall index, an equal weight approach is followed. In construction of human development index, in 2001 by Planning Commission, the overall index was calculated by using equal weight whereas the different sub-indices like Composite indicator on educational attainment, Composite indicator on health attainment differential weights were given. In calculating educational attainment, two variables i.e. 'literacy rate for the age group 7 years and 'adjusted intensity of formal education' were used where the first and second variable were given 1/3 and 2/3 weight respectively. The other 'health attainment was calculated by taking two variables 'life expectancy at age one,' and ‘infant mortality rate’. The first variable has a weight of 2/3 whereas the second variable 1/3. But the overall weight is calculated by taking equal weight for both the variables. Take another example of our range equalization method where we have given equal weight to all variables. In some other cases, the individual variables of sub indices were given equal weights but the overall index was calculated by giving differential weights to sub-indicators.

Let us denote the weight in mathematical notation:

$$I=XW,$$

where ‘X’ is a matrix with m rows and n column, ‘I’ is identity matrix and ‘W’ is weight. Hence the final indices can be arrived at by taking the weighted component/variables. There are basically two ways in assigning weights. According to Munda and Nardo (2005), we can define weight by taking the importance of the particular variable or group of variables. In this process, the weights are arrived at by the past knowledge or observation of individual and the probable effects of that variable in the overall analysis. Let us illustrate with an example. Suppose we want to find out the sanitation and hygiene Index. For this we have selected some variables like ‘proportion of people having access to toilet’, ‘proportion of people having drainage’, ‘proportion of people with wash hand’, ‘proportion of people safely disposed child exgratia’ etc . Here either on the basis of literature or on the basis of our observation from some villages, we found that not having toilet is very important then wash hand before eating. In this case we can put higher weight for having toilet and less weight for washing hand.

In case of literature, weights are derived from on the basis of theoretical or statistical consideration.

12.6.4 Limitations of Principal Component Analysis

The major criticisms against the use of PCA is that the technique is arbitrary in constructing indices. The number of components and the number of variables used are not well defined. This method entirely depends on the first component. If the first component does not explain large variation, this method is not useful. Such possibility depends on nature of data and on relationship of variable.

Alternative methods to PCA that can reduce the dimensionality of the data include: correspondence analysis, multivariate regression or factor analysis. The details about these methods have been provided in Unit 13 and Unit 16 of Block 4.

12.7 MERITS AND LIMITATIONS OF COMPOSITE INDEX

One of the important merits of a composite index is that this can summarize the complex and multidimensional indicators into one indicator which helps the policy makers for implementation of a particular programme or policy decision. A composite index can easily be taken to interpret about the development or backwardness of a particular region or area or sector. The progress of a state in India say, e.g. Bihar can easily be accessed and we can compare the status of Bihar at two or more points of time. The composite index can reduce the number of indicators without reducing the underlying information base. The composite index facilitates communication to general public and promote accountability. The composite index is not free from limitations. It has the risk of misleading policy message if it is poorly constructed or misinterpreted. The calculation may be misleading and faulty if sound statistical or conceptual principle is not applied. The assignment of weight many times creates a debate. One of the principles of assigning weight is on the basis of value judgment. This always being a source of criticism. Again allocation of upper and lower bound (mainly in range equalization method) is a point of criticism.

Check Your Progress 3

1) In the context of Composite Index, what are the uses of Principal Component Analysis (PCA).

   ……………………………………………………………………………………

   ……………………………………………………………………………………

   ……………………………………………………………………………………

   ……………………………………………………………………………………

   ……………………………………………………………………………………

2) What is explained by the term ‘communality’?

   ……………………………………………………………………………………

   ……………………………………………………………………………………

   ……………………………………………………………………………………

   ……………………………………………………………………………………

   ……………………………………………………………………………………

3) How is Final Index Value calculated?

   ……………………………………………………………………………………

   ……………………………………………………………………………………

   ……………………………………………………………………………………

   ……………………………………………………………………………………

4) In construction of Human Development Index, how weights are assigned to the different indicators?

   ……………………………………………………………………………………

   ……………………………………………………………………………………

   ……………………………………………………………………………………

   ……………………………………………………………………………………

5) State the limitations of Composite Index.

…………………………………………………………………………

…………………………………………………………………………

…………………………………………………………………………

…………………………………………………………………………

…………………………………………………………………………

12.8 LET US SUM UP

Composite index is an important analytical technique to analyse the developmental related issues like status of development, food security, human development etc. at the district/region/state level. It is an expression of single score made of different scores measuring the various dimensions and indicators of particular issue. This Unit describes the process to derive composite index by applying the various methods. These methods include simple ranking method, indices method, mean standarisation method, and range equalization method. The main advantage of the rank and indices method is that they are easy to understand and interpret. The RE, MS and PCA methods uses all the variables in reducing the dimensionality of the data. These methods are very much useful in comparing across districts, states, countries or areas such as rural and urban. Again these methods are useful to compare over time by constructing composite index at two points of time by taking same indicators/variables. The principal component analysis enables the students to reduce the indicators that are uncorrelated to explain the variation in the original data set. To run the PCA, SPSS and STATA softwares are preferred.

12.9 EXERCISES

1) Think and collect some gender related variables from the data sources like women literacy, women workforce participation rate in the age group 15-59, mean year schooling, mortality rate, immunization rate etc and find out the gender development index.

2) Explain the process of using Principal Component Analysis (PCA) to reduce the number of variables in a data set to smaller number of dimensions.

3) By using range equalization method find out the food security index from the report prepared by IHD-WFP given in website

   'http://122.180.7.122/displaymorePub.asp?itemid=84&subchkey=11&chname=Publications.'

12.10 SOME USEFUL BOOKS/REFERENCES

A. Saltelli, G. Munda and M. Nardo (2006), 'From Complexity to Multidimensionality: the Role of Composite Indicators for Advocacy of EU Reform', Tijdschrift voor Economie en Management Vol. LI, 3

Downloaded from http://www.econ.kuleuven.be/rebel//jaargangen/2001-2010/2006/TEM%202006-3/TEM_2006-3_03_Saltelli.pdf

Planning Commission (2001), 'National Human Development Report', Government of India downloaded from http://planningcommission.nic.in/reports/genrep/index.php?repts=nhdcont.htm

OECD (2008), 'Handbook on Constructing Composite Indicators Methodology and User Guide', Organization For Economic Co-operation and Development.

12.11 ANSWER OR HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) See Section 12.2
2) See Section 12.2
3) See Section 12.3
4) See Section 12.4

Check Your Progress 2

1) See Sub-section 12.4.1
2) See Sub-section 12.4.2
3) See Sub-section 12.5.4

Check Your Progress 3

1) See the heading 'conducting PCA and analyzing result' under Section 12.6
2) See Sub-section 12.6.1 (Under the head conducting the PCA and analyzing results)
3) See Sub-section 12.6.1 (Under the head final PCA index value)
4) See Sub-section 12.6.3
5) See Section 12.7