UNIT 8: FORECASTING WITH THE LINEAR REGRESSION EQUATION

UNIT STRUCTURE

8.1 Learning Objectives

8.2 Introduction

8.3 Properties of a Good Forecast or Prediction

8.4 Point Prediction

8.5 Interval Prediction

8.6 Let Us Sum Up

8.7 Further Readings

8.8 Answers to Check Your Progress

8.9 Model Questions

8.1 LEARNING OBJECTIVES

After going through this unit, you will be able to

- explain the concepts related to prediction or forecast;
- know about point and interval prediction;
- make to prediction using linear regression model;
- evaluate accuracy of the predictions.

8.2 INTRODUCTION

After the estimation of the model parameters, a common use of regression modeling is for prediction of the dependent variable. We can make a distinction between prediction and forecasting most easily based on the difference between cross section and time-series modeling. Prediction involves using the regression model to predict values of the dependent variable, either within the sample or for observations outside the sample. The same set of results will apply to cross sections, panels, and time series. Forecasting, while largely the same exercise, explicitly gives a role to time and often involves lagged dependent variables and disturbances that are correlated with their past values. This exercise usually involves predicting

future outcomes. In this module, however, we will be using both these terms interchangeably for the sake of simplicity.

The prediction of values consists of two steps. In the first step, the regression coefficients are estimated on the basis of given observations. In the second step, these estimators are then used to construct the predictor which provides the prediction of values of study variables. Prediction can be attempted in two alternative ways: either as a point prediction or as an interval prediction, in just the same way as we can give a point estimate or an interval estimate of a parameter β. In point prediction, we predict a single value in each period, while interval prediction involves prediction of an interval within which the value of the variable will most probably lie.

8.3 PROPERTIES OF A GOOD FORECAST OR PREDICTION

A good forecast or prediction is the one which yields a prediction error with minimum variance. To prove this, let us consider the following two-variable linear regression model:

$$Y_i=\beta_0+\beta_1X_i+U_i \qquad (i=1,2,3,\ldots\ldots,n)$$

Where Ui are identically and independently distributed following $N(0,\sigma^2)$. The parameters β_0 and β_1 are estimated using the ordinary least squares estimation and the fitted model is

$$\hat{Y}_i=b_0+b_1X_i$$

Here, we make prediction for the value of Y for the unit 'p' that lies outside the sample of unit 'p', i.e., X_p is either known or projected. Then the point prediction $(\hat{Y}_p)$ is given by:

$$\hat{Y}_p=b_0+b_1X_p$$

This is called a point prediction as we obtain one specific value of Y as the predicted value. Again, the actual value or observed value of Y at the predicted period 'p' will be:

$$Y_p=\beta_0+\beta_1X_p+U_p$$

Therefore, the prediction error $\hat{U}_p$ is given by

$$\hat{U}_p = \hat{Y}_p - Y_p$$

$$\hat{U}_p = (b_0 + b_1 X_p) - (\beta_0 + \beta_1 X_p + U_p)$$

$$\hat{U}_p = (b_0 - \beta_0) + (b_1 - \beta_1) X_p - U_p$$

We would like the forecast error to be small, implying that our forecast is close to the value we are predicting. Taking the expected value of $\hat{U}_p$, we find

$$E(\hat{U}_p) = E(b_0 - \beta_0) + E[(b_1 - \beta_1) X_p] - E(U_p)$$

$$E(\hat{U}_p) = 0$$

which implies that the mean of prediction error is zero. Therefore, $\hat{Y}_p$ is an unbiased predictor of Y_p. However, unbiasedness does not necessarily imply that a particular forecast will be close to the actual value. The probability of a small forecast error also depends on the variance of the forecast error. Therefore, the variance of the forecast error is

$$Var(\hat{U}_p) = E[\hat{U}_p - E(\hat{U}_p)]^2$$

$$Var(\hat{U}_p) = \sigma^2 \left[1 + \frac{1}{n} + \frac{(X_p - \bar{X})^2}{\Sigma (X_i - \bar{X})^2}\right]$$

Notice that some of the elements of this expression appear in the formulas for the variances of the least squares estimators and affect the precision of prediction in the same way that they affect the precision of estimation. We would prefer that the variance of the forecast error be small, which would increase the probability that the prediction $\hat{Y}_p$ is close to the value Y_p we are trying to predict. Note that the variance of the prediction error is smaller when

a) the overall uncertainty in the model is smaller, as measured by the variance of the random errors σ^2;
b) the sample size n is larger;
c) the variation in the explanatory variable is larger;
d) the value of $(X_p - \bar{X})^2$ is small.

The term $(X_p - \bar{X})^2$ measures how far X_p is from the mean of the X-values. The further the forecasted point X_p is from the center of the sample data, the greater is the degree of uncertainty. Intuitively, this means that we are able to do a better job predicting in the region where we have more sample information, and we will have less accurate predictions when we

try to predict outside the limits of our data. Again, note that the prediction variance is composed of three parts. The second and third become progressively smaller as we accumulate more data (i.e., as n increases). But, the first term, ó2 is constant, which implies that no matter how much data we have, we can never predict perfectly.

In practice, we can estimate the prediction variance by using s2 in place of ó2 as follows

$$\text{Var}(\hat{U}_p) = s^2\left[1 + \frac{1}{n} + \frac{(X_p - \bar{X})^2}{\Sigma(X_i - \bar{X})^2}\right]$$

The square root of this estimated variance is the standard error of the forecast, i.e.,

$$se(\hat{U}_p) = \sqrt{\text{Var}(\hat{U}p)}$$

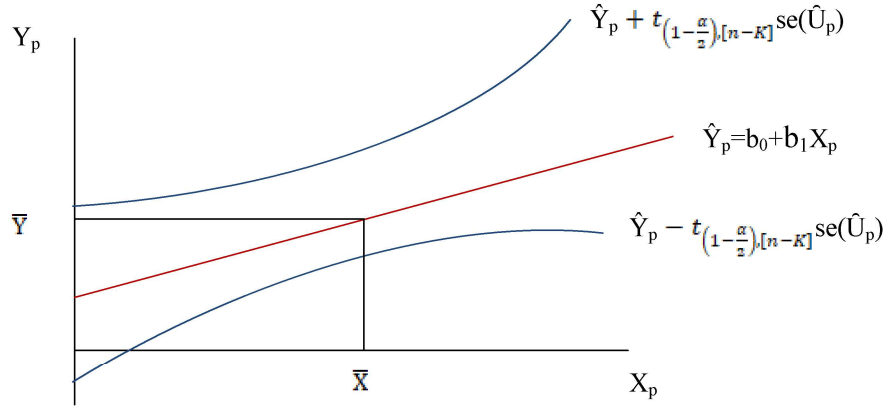
A prediction interval for Yp would then be formed using

$$\text{Prediction interval} = \hat{Y}_p \pm t_{\left(1-\frac{\alpha}{2}\right),[n-K]}\, se(\hat{U}_p)$$

where $t_{\left(1-\frac{\alpha}{2}\right),[n-K]}$ is the appropriate critical value for 100(1 -α) percent level of significance from the t table for n–K degrees of freedom.

From earlier, we know that the further Xp is from the sample mean, the larger the variance of the prediction error will be, and the less reliable the prediction is likely to be. In other words, our predictions for values of Xp close to the sample mean are more reliable than our predictions for values of Xp far from the sample mean. This fact shows up in the size of the prediction intervals. The relationship between point and interval predictions for different values of Xp is illustrated in Figure 1

Figure 8.1: Point And Interval Prediction



A point prediction is given by the fitted least squares line $\hat{Y}_p=b_0+b_1X_p$. The prediction interval takes the form of two bands around the fitted least squares line. Because the forecast variance increases the further X_p is from the sample mean $\bar{X}$, the confidence bands are the narrowest when $X_p=\bar{X}$, and they increase in width as (Xp- $\bar{X}$) increases.



CHECK YOUR PROGRESS

Q 1: What is Point prediction? (Answer in about 20 words)

.....

.....

Q 2: How can we measure predictive accuracy of a forecast? (Answer in about 30 words)

.....

.....

Q 3: State the conditions under which variance of the prediction error is smaller. (Answer in about 60 words)

.....

.....

.....

8.4 POINT PREDICTION

Typically, point prediction tries to predict or forecast a single number for the dependent variable (Y) conditional upon the values of explanatory variable (X). Since the point prediction only reports a single number, the goal is to compute the value that is closest to the eventual outcome. In other words, the point prediction is intended to represent the outcome that is “most likely” to occur. The point prediction represents the expected value of the outcome and, accordingly, is a number that weights different possible outcomes by how likely they are to occur.

The point prediction is based on the assumption that the fitted model still holds in the prediction period. When a new value of Y is observed, it is possible to test this stability assumption. An appealing point prediction is

obtained by inserting the given X values into the regression equation:

$$\hat{Y}_p=\beta_0+\beta_1 X_p$$

That is, the predicted value $\hat{Y}_p$ is given by the point on the least squares fitted line where $X= X_p$. The least squares estimators β_0 and β_1 are random variables—their values vary from one sample to another. It follows that the least squares predictor $\hat{Y}_p$ must also be random. To evaluate how well this predictor performs, we define the forecast error, which is analogous to the least squares residual.

The point predictions are also well suited to computing and communicating revisions in forecasts. For example, if economic growth slows down, the point forecast will likely be revised downward relative to the previous forecast. The amount by which the forecast is reduced conveys a sense of the expected impact of the negative news on the future, uncertain outcome.

Point predictions can be reported across different forecast horizons. For example, suppose the point predictions of a company's revenue growth are as follows:

- 15% revenue growth in the next (first) fiscal year;
- 10% revenue growth in the following (second) fiscal year; and
- 7% revenue growth in the subsequent (third) fiscal year.

In this example, the point predictions signal an expected slowdown in the growth of the company's revenues. Accordingly, point predictions reported across multiple horizons can be used to map out a likely future trajectory that is being forecasted.

While point predictions are easily communicated and represent a reasonable starting point for a forecasting analysis, they also have some shortcomings. Most importantly, point predictions do not convey a sense of how much uncertainty surrounds a predicted outcome. Moreover, point predictions do not indicate if and how the level of uncertainty varies over time.

8.5 INTERVAL PREDICTION

An interval prediction (or prediction interval) gives a range of values within which we expect Yp to lie with a specific probability. For example, assuming that the forecast errors are normally distributed, a prediction interval of 95 percent for the forecast is

$$\text{Prediction interval} = \hat{Y}_p \pm 1.96\text{se}(\hat{U}_p)$$

where $\text{se}(\hat{U}_p)$ is an estimate of the standard deviation of the forecast distribution. More generally, a prediction interval can be written as

$$\text{Prediction interval} = \hat{Y}_p \pm t_{\left(1-\frac{\alpha}{2}\right),[n-K]}\text{se}(\hat{U}_p)$$

where the value of $t_{\left(1-\frac{\alpha}{2}\right),[n-K]}$ depends on the coverage probability.

The value of prediction intervals is that they express the uncertainty in the forecasts. If we only produce point forecasts, there is no way of telling how accurate the forecasts are. However, if we also produce prediction intervals, then it is clear how much uncertainty is associated with each forecast. For this reason, point forecasts can be of almost no value without the accompanying prediction intervals.

As we know, we can use a fitted regression model to generate predictions based on the specific predictor values. However, predictions are not as simple as a single predicted value (i.e., a point prediction). The predicted value is actually the mean response value. Like any mean, there is variability around that mean. Prediction intervals account for the variability around the mean response inherent in any prediction. Like confidence intervals, prediction intervals have a confidence level and can be a two-sided range. Unlike confidence intervals, prediction intervals predict the spread for individual observations rather than the mean.

A common feature of prediction intervals is that they increase in length as the forecast horizon increases. The further ahead we forecast the more uncertainty is associated with the forecast and thus the wider the prediction intervals. For example, the 50% interval prediction for the next year may be (1%, 5%), whereas the 50% interval prediction for following year or two-year ahead may be (–1%, 7%). The extra width of the two-year ahead interval prediction compared to the one-year ahead interval prediction

conveys a sense of how quickly uncertainty increases as we forecast further into the future.



CHECK YOUR PROGRESS

Q 4: Write out the difference between point and interval prediction. (Answer in about 150 words)

.....

.....

.....

.....

.....

Q 5: Differentiate between confidence interval and prediction interval. (Answer in about 60 words)

.....

.....

.....

.....



8.6 LET US SUM UP

- Prediction involves using the regression model to predict values of the dependent variable, either within the sample or for observations outside the sample.
- Prediction can be attempted in two alternative ways: either as a point prediction or as an interval prediction.
- Point prediction tries to predict or forecast a single number for the dependent variable (Y) conditional upon the values of explanatory variable (X). An interval prediction gives a range of values within which we expect Yp to lie with a specific probability.
- A good forecast or prediction is the one which yields a prediction error with minimum variance.
- To increase the probability that the prediction $\hat{Y}_p$ is close to the actual value Y_p, the variance of the forecast error should be small.

- The further the forecasted point Xp is from the center of the sample data, the greater is the degree of uncertainty.
- The point prediction is based on the assumption that the fitted model still holds in the prediction period.
- Point predictions do not convey a sense of how much uncertainty surrounds a predicted outcome.
- The value of prediction intervals is that they express the uncertainty in the forecasts.
- A common feature of prediction intervals is that they increase in length as the forecast horizon increases.



8.7 FURTHER READING

1) Greene, W. H. (2018). *Econometric analysis*. New York: Pearson Education.
2) Hill, R. C., Griffiths, W. E., & Lim, G. C. (2018). *Principles of econometrics*. John Wiley & Sons.
3) Hyndman, R. and Athanasopoulos, G. (2013). *Forecasting: Principles and Practice*.
4) Johnston, J. and Dinardo, J. (1997). *Econometric Methods*. New York: McGraw-Hill.



8.8 ANSWERS TO CHECK YOUR PROGRESS

Ans to Q No 1: The term prediction of value of the study variable involves predicting the values of the dependent variable, either within the sample or for observations outside the sample.

Ans to Q No 2: We know that the further Xp is from the sample mean, the larger the variance of the prediction error will be, and the less reliable the prediction is likely to be. In other words, our predictions for values of Xp close to the sample mean are more reliable than our predictions for values of Xp far from the sample mean. Therefore,

the distance between predicted value and sample mean helps in measuring the predictive accuracy of a forecast.

Ans to Q No 3: The variance of the prediction error is smaller when

(a) the overall uncertainty in the model is smaller, as measured by the variance of the random errors σ^2;

(b) the sample size n is larger;

(c) the variation in the explanatory variable is larger;

(d) the value of $(X_p - \bar{X})^2$ is small.

Ans to Q No 4: Difference between point and interval prediction are as follows:

(a) When a particular value of independent variable is used to estimate a single value for the dependent variable, it is known as point prediction. Here, prediction is made about the value of dependent variable for a particular future value of independent variable. On the other hand, when an attempt is made to make some reasonable statement about the unknown dependent variable at a certain level of significance, then it is called interval prediction. Here, there are two limits under which predicted value of the dependent variable will lie.

(b) In point prediction, the variance of prediction error increases in a non-linear way as the gap between the values of independent variable for prediction unit (X_p) and the sample mean $(\bar{X})$ increases. On the other hand, prediction error in interval prediction is assumed to be normally distributed.

Ans to Q No 5: Difference between confidence interval and prediction interval are as follows:

- Unlike confidence intervals, prediction intervals predict the spread for individual observations rather than the mean.
- The prediction interval is always wider than the confidence interval of the prediction because of the added uncertainty involved in predicting a single response versus the mean response.



8.9 MODEL QUESTIONS

Short Questions (Answer each question in about 150 words)

Q 1: Elaborate the concept of point prediction using a linear regression model.

Q 2: Explain the relationship between point and interval prediction using a suitable diagram.

Long Questions (Answer each question in about 300-500 words)

Q 1: Illustrate the properties of a good forecast or prediction.

*** ***** ***

UNIT 9 : SPECIFICATION ERRORS AND THEIR CONSEQUENCES

UNIT STRUCTURE

9.1 Learning Objectives

9.2 Introduction

9.3 Types of Specification Errors

9.4 Consequences of Specification Errors

9.4.1 Consequences of Omission of a Relevant Variable

9.4.2 Consequences of Inclusion of an Irrelevant Variable

9.5 Let Us Sum Up

9.6 Further Readings

9.7 Answers To Check Your Prgress

9.7 Model Questions

9.1 LEARNING OBJECTIVES

After going through this unit, you will be able to-

- understand what is specification error in a regression model
- know about different types of model specification errors
- explain the consequences of specification errors on the process of estimation of the parameters of a regression model and on the confidence interval and hypothesis testing procedures

9.2 INTRODUCTION

One of the assumptions of the linear regression model is that the model is correctly specified. With this assumption, model parameters are estimated and a few diagnostic statistics like R^2, *t, F* and Durbin-Watson's *d* are also checked. If these diagnostic statistics are found to be satisfactory, then the model is considered as a good model.

In case, one or more diagnostic statistic is not satisfactory, then some other advanced methods of estimation are used. If the diagnostic statistics still remain unsatisfactory, then the problem is considered to

persist due to specification error or bias of the applied model. Sometimes, one or more relevant variables may not be included in the model or some irrelevant variables may be included due to which the functional form of the applied model may be incorrect. All these situations lead to specification errors in the econometric model.

9.3 TYPES OF SPECIFICATION ERRORS

Let us consider that the following model is correct for a particular economic phenomenon:

$$Y_t = \alpha_1 + \alpha_2 X_{1t} + \alpha_3 X_{2t} + \alpha_4 X_{3t} + u_t \quad \text{.....(1)}$$

where, Y_t is the dependent variable, X_s are the explanatory variables, αs are the model parameters and u_t is the random error term.

In certain situations, we may use a model other than the correct model as in equation (1) and may confront with specification errors. Such model specification errors may take the following types:

(i) Omission of a Relevant Variable: Suppose that a researcher has used the following model instead of the correct model as specified in equation (1):

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + u_t \text{ .....(2)}$$

The researcher has made a specification error in this model in the form of omission of a relevant variable X_{3t}. Here, the error term V_t is given by:

$$v_t = u_t + \alpha_4 X_{3t}$$

(ii) Inclusion of Unnecessary or Irrelevant Variable: Sometimes, researchers include some unnecessary variables in their models which lead to another kind of specification error in the form of inclusion of unnecessary or irrelevant variables. Suppose, in a situation when the model in equation (1) is true but the researcher has used the following model:

$$Y_t = \gamma_1 + \gamma_2 X_{1t} + \gamma_3 X_{2t} + \gamma_4 X_{3t} + \gamma_5 X_{4t} + \eta_t \text{ .....(3)}$$

Compared to the correct model in equation (1), the model in equation (3) has a specification error of inclusion of an irrelevant variable X_{4t}. Here, the error term η_t is given by:

$$\eta_t = u_t - \gamma_5 X_{4t}$$

$= u_t$ (as $\gamma_5 = 0$ in the true model)

(iii) Wrong Functional Form: Specification error also occurs if the functional form of the applied model is not correct. For example, the researcher may use the following model:

$$logY_t = \delta_1 + \delta_2 X_{1t} + \delta_3 X_{2t} + \delta_4 X_{3t} + \pi_t \quad \text{.....(4)}$$

The functional form in the model in equation (4) is wrong in relation to the correct model in equation (1). In equation (1), the dependent variable Y is linear but it is log-linear in equation (4). Hence, Equation (4) cannot be considered as the correct functional form of the true model.

(iv) Errors of Measurement: Errors of measurement also lead to specification errors in regression models. Suppose, a researcher uses the following model:

$$Y_t^* = \alpha_1^* + \alpha_2^* X_{1t}^* + \alpha_3^* X_{2t}^* + \alpha_4^* X_{3t}^* + u_t^* \quad \text{.....(5)}$$

Here, instead of Y_t and X_t s, the researcher has used proxy variables Y_t^* and X_t^* s which may lead to some additional errors of measurement. In this case,

$$Y_t^* = Y_t + \varepsilon_t \quad \text{and} \quad X_t^* = X_t + \omega_t$$

where, ε_t and ω_t are the errors of measurement.

It is seen that once a correct model is specified, we can check whether we have committed one or more specification errors in the model we are actually applying.

The question here is that why do researchers commit specification errors. The probable answers may be:

- The correct model may not be known leading to a model mis-specification error.
- Data on some of the relevant variables may not be available. In this case, one or more relevant variables may be omitted from the applied model.
- The precise functional form of the model may not be known.



CHECK YOUR PROGRESS

Q 3: Name two diagnostic statistics to check correctness of a regression model. (Answer in about 20 words)

.....

.....

Q 2: What should a researcher do if one or more diagnostic statistics is not found satisfactory for his applied regression model? (Answer in about 30 words)

.....

.....

.....

Q 3: What does it indicate if some of the diagnostic statistics remain unsatisfactory even after applying advanced econometric methods of estimation?
(Answer in about 40 words)

.....

.....

.....

.....

Q 4: In a situation when the following model is correct for a particular event:

$$Y_t = \beta_1 + \beta_2 X_{1t} + \beta_3 X_{2t} + v_t$$

But a researcher has applied the model:

$Y_t = \alpha_1 + \alpha_2 X_{1t} + \alpha_3 X_{2t} + \alpha_4 X_{3t} + u_t$.

What type of specification error has he made?
(Answer in about 20 words)

.....

.....

9.4 CONSEQUENCES OF SPECIFICATION ERRORS

Although model specification error may arise out of various reasons,

here we only discuss the consequences of specification errors committed in the following two ways:

- Omission of a relevant variable
- Inclusion of an irrelevant variable

Before studying the consequences of specification errors, let us consider the following three variable linear regression model and its Ordinary Least Square (OLS) estimators:

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + u_t \quad \text{.....(6)}$$

Here, we have one dependent variable Y_t and two explanatory variables X_2 and X_3, u is the stochastic disturbance term and t denotes the t-th observation (t=1,2,….., n). β_1 is the intercept term representing the mean effect on Y of all the variables excluded from the model. β_2 and β_3 are the two partial regression coefficients.

Estimation of the model is based on the following assumptions:

- Mean effect of U_t is zero, i.e., $E(U_t|X_{2t}, X_{3t}) = 0$ for each t.
- U_ts are not serially correlated, i.e., cov $(U_t, U_s) = 0$ for all $t \neq s$.
- U_ts are homoscedastic, i.e., var $(U_t) = \sigma^2$.
- Covariance between U_ts and explanatory variables are zero, i.e.,

  $$\text{cov}\,(u_t, X_{2t}) = \text{cov}\,(u_t, X_{3t}) = 0.$$
- The model is correctly specified, i.e., there is no specification error.
- There is absence of exact linear correlation between the X variables.

Based on the above assumptions, OLS estimators of the parameters of equation (6) are given by:

$$\hat{\beta}_2 = \frac{(\Sigma y_t x_{2t})(\Sigma x_{3t}^2) - (\Sigma y_t x_{3t})(\Sigma x_{2t} x_{3t})}{(\Sigma x_{2t}^2)(\Sigma x_{3t}^2) - (\Sigma x_{2t} x_{3t})^2} \quad \text{.....(7)}$$

$$\hat{\beta}_3 = \frac{(\Sigma y_t x_{3t})(\Sigma x_{2t}^2) - (\Sigma y_t x_{2t})(\Sigma x_{2t} x_{3t})}{(\Sigma x_{2t}^2)(\Sigma x_{3t}^2) - (\Sigma x_{2t} x_{3t})^2} \quad \text{.....(8)}$$

$$\hat{\beta}_1 = \hat{Y} - \hat{\beta}_2 \bar{X}_2 - \hat{\beta}_3 \bar{X}_3 \quad \text{.....(9)}$$

$$var\,(\hat{\beta}_2) = \frac{\sigma^2}{\Sigma X_{2t}^2 (1 - r_{23}^2)} \quad \text{.....(10)}$$

$$var\,(\hat{\beta}_3) = \frac{\sigma^2}{\Sigma X_{3t}^2 (1 - r_{23}^2)} \quad \text{.....(11)}$$

Where, $x = (X - \bar{X})$ and $y = (Y - \bar{Y})$.

Let us now discuss the consequences of the above mentioned specification errors one by one.



CHECK YOUR PROGRESS

Q 5: In the following three variable regression model:

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + u_t$$

What does β_1 represent? (Answer in about 20 words)

.....

.....

Q 6: What does the assumption of homoscedasticity mean in a linear regression model? (Answer in about 30 words)

.....

.....

9.4.1 Consequences of Omission of a Relevant Variable

We now restate the correct regression model in equation (6):

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + u_t$$

Suppose that the researcher has taken the following model:

$$Y_t = \alpha_1 + \alpha_2 X_{2t} + v_t$$

Here, he has omitted the variable X_{3t} which was included in the correct model. Thus the researcher has made a specification error by omitting a relevant variable X_{3t}.

Consequences:

1. If the omitted variable X_{3t} is correlated with X_{2t}, the included variable, then the estimation of α_1 and α_2 will not only be biased but also inconsistent. Therefore,

   $$E(\hat{\alpha}_1) \neq \beta_1 \quad \text{and} \quad E(\hat{\alpha}_2) \neq \beta_2$$

   This bias will persist even if the sample size is increased sufficiently.
2. If X_{3t} and X_{2t} are not correlated, estimator of α_1 will still be biased although estimator of α_2 will now be unbiased.
3. Variance of the disturbance term, i.e. var (U_t) will be estimated

incorrectly.

4. The estimated variance of α_2 will be a biased estimate of the true variance of β_2.
5. Due to the above consequences, the confidence interval of the parameters and the usual hypothesis testing procedure may become misleading. It results in incorrect conclusions on the statistical significance of the estimated parameters.

9.4.2 Consequences of Inclusion of an Irrelevant Variable

Let us now consider the correct model to be:

$$Y_t = \beta_1 + \beta_2 X_{2t} + u_t \quad \text{.....(12)}$$

In a situation when equation (12) is correct, suppose our researcher has fitted a model including an irrelevant variable as:

$$Y_t = \alpha_1 + \alpha_2 X_{2t} + \alpha_3 X_{3t} + v_t \quad \text{.....(13)}$$

Thus the researcher has made a specification error in the form of inclusion of an irrelevant or unnecessary variable X_{3t}.

Consequences: In this case, the researcher may have to face the following consequences:

1. The OLS estimators of the model parameters will be unbiased and consistent. Symbolically,

   $E(\hat{\alpha}_1) = \beta_1$, $E(\hat{\alpha}_2) = \beta_2$ and $E(\hat{\alpha}_3) = \beta_3$.

   These conditions remain true with increase in the sample size also.
2. The variance of the disturbance term will be correctly estimated.
3. The derivation of the usual confidence interval and the process of hypothesis testing would not be affected.
4. But the estimated parameters would be generally inefficient. In other words, the variances of the estimated αs would be larger than the variances of the βs.

Thus inclusion of irrelevant variables in regression models results in larger variances of the estimated parameters making them less precise.

Let us now compare the consequences of these two types of specification errors, viz, omission of a relevant variable and inclusion of an irrelevant variable. In the case of omission of a relevant variable, the estimated parameters become not only biased but also they become inconsistent. Moreover, the variances of the error terms are incorrectly estimated and the conventional confidence interval and hypothesis testing procedures become invalid. But inclusion of irrelevant variables in regression models does not affect the unbiasedness and consistency properties of the estimated parameters. Again, error variances are estimated correctly and usual confidence interval and hypothesis testing procedures remain valid. However, the only drawback with this specification bias is that the estimated parameters become inefficient compared to the parameters of the correct model.

It seems from the above that inclusion of an irrelevant variable is less harmful than exclusion of a relevant variable in regression models. However, inclusion of irrelevant variables should also be carefully checked as it has the following consequences:

- Reduction in efficiency of the estimated parameters.
- Emergence of the problem of multicollinearity. (If the included irrelevant variables have strong relationship with the other explanatory variables)
- Loss of degrees of freedom.



CHECK YOUR PROGRESS

Q 7: If a researcher omits a relevant variable from his regression model and if the omitted variable is correlated with the included variables, what would be its impact on the estimators of the model parameters? (Answer in about 20 words)

.....

.....

Q 8: If a researcher omits a relevant variable from his regression

model, what would be its impact on the usual hypothesis testing procedure? (Answer in about 20 words)

.....

.....

Q 9: If a researcher includes an unnecessary variable in his regression model, what would be its impact on the efficiency of the estimated parameters? (Answer in about 30 words)

.....

.....

.....

Q 10: If a researcher includes an unnecessary variable in his regression model, what would be its impact on the usual hypothesis testing procedure? (Answer in about 20 words)

.....

.....



9.5 LET US SUM UP

- In the presence of specification errors in the regression models, diagnostic statistics remain unsatisfactory.
- Model specification error may be committed in the form of omission of a relevant variable, inclusion of irrelevant variable(s), using a wrong functional form of the model and/or error of measurement.
- A researcher may commit specification error due to lack of knowledge on the correct model, unavailability of data on some of the relevant variables and/or inadequate knowledge on the precise functional form of the model.
- In case of specification error in the form of omission of a relevant variable, the estimators of the model parameters become not only biased but also inconsistent, the error variances are incorrectly estimated and the usual confidence interval and hypothesis testing procedures become misleading.
- In case of specification error in the form of inclusion of an irrelevant

variable, the estimators of the model parameters remain unbiased and consistent, the error variances are correctly estimated and the usual confidence interval and hypothesis testing procedures remain valid. However, the estimators of the parameters generally become inefficient.



9.6 FURTHER READINGS

1) Gujarati, D. N. (2015). *Basic Econometrics*, McGraw-Hill International Editions

2) Dougherty, C. (2008). *Introduction to Econometrics*, Third edition, Oxford University Press



9.7 ANSWERS TO CHECK YOUR PRGRESS

Ans to Q No 1: Two diagnostic statistics to check correctness of a regression model are R^2 and t-statistic.

Ans to Q No 2: If one or more diagnostic statistics like R^2, t, F and Durbin-Watson's d statistic is not found satisfactory for the applied regression model, the researcher should apply some other advanced econometric methods of estimation.

Ans to Q No 3: If some of the diagnostic statistics remain unsatisfactory even after applying advanced econometric methods of estimation, it indicates to the presence of specification error in the applied regression model.

Ans to Q No 4: In this case, the researcher has made a specification error in the form of inclusion of an irrelevant variable X_{3t}.

Ans to Q No 5: In the given regression model, β_1 represents the mean effect on Y of all the variables excluded from the model.

Ans to Q No 6: The assumption of homoscedasticity in a linear regression model means constant variance of the random disturbance term.

Ans to Q No 7: The estimators of the model parameters in this case

become not only biased but also inconsistent.

Ans to Q No 8: The usual hypothesis testing procedure in this case becomes misleading.

Ans to Q No 9: The efficiency of the estimated parameters in this case would be reduced.

Ans to Q No 10: The usual hypothesis testing procedure in this case would not be affected.



9.7 MODEL QUESTIONS

Short Questions (Answer each question in about 150 words)

Q 1: What is specification error? How would a researcher be sure that his regression model has specification error?

Q 2: What are the different types of specification errors?

Q 3: Discuss why efficiency of the estimated parameters is reduced when the regression model has specification error in terms of inclusion of irrelevant variables.

Long Questions (Answer each question in about 300-500 words)

Q 1: Discuss the consequences of omission of a relevant variable from regression models.

Q 2: Discuss the consequences of inclusion of an irrelevant variable in regression models.

Q 3: Compare the effects of omission of a relevant variable and inclusion of an irrelevant variable in regression models. Justify which one is more serious.

*** ***** ***

UNIT 10 : PROBLEM OF MULTICOLLINEARITY

UNIT STRUCTURE

10.1 Learning Objectives

10.2 Introduction

10.3 The Multicollinearity Problem

10.4 Consequences of multicollinearity

10.5 Methods of detecting multicollinearity

10.6 Remedies to the problem of multicollinearity

10.7 Let Us Sum Up

10.8 Further Readings

10.9 Answers to Check Your Progress

10.10 Model Questions

10.1 LEARNING OBJECTIVES

After going through this unit, you will be able to know-

- About the problem of multicollinearity
- About the concepts of perfect and near multicollinearity
- About consequences of the occurrence of the problem of multicollinearity
- About the ways to detect whether multicollinearity is present in a regression model and
- The remedies to the problem of multicollinearity

10.2 INTRODUCTION

Multicollinearity refers to the problem wherein the explanatory variables of a linear regression model are not independent of each other. As per the standard assumptions of a multiple regression model the explanatory variables are not linearly correlated to each other. The explanatory variables if uncorrelated, it is termed as orthogonality and it indicates absence of multicollinearity. Earlier the concept of multicollinearity meant the existence of a perfect correlation between the explanatory

variables. However, the concept now includes the existence of a high degree of correlation between the same and not only the extreme case of perfect multicollinearity. The present chapter discusses the two concepts and analyses the consequences of both in the regression model.

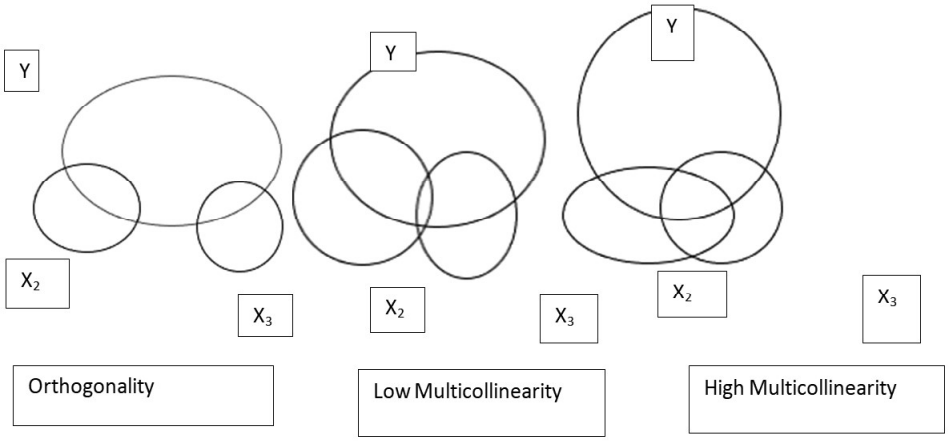
10.3 THE MULTICOLLINEARITY PROBLEM

The term multicollinearity was earlier conceived as a situation where the independent variables in a multiple regression model are perfectly linearly correlated. This creates a problem where the estimates of the regression model become indeterminate. More specifically, this is a problem of perfect or exact multicollinearity. Here, some or all of the explanatory variables are perfectly linearly related. However, the situations of perfect multicollinearity and orthogonality are rare and often we come across situations where there is some high degree of correlation among the explanatory variables owing to the dependence between economic variables. This is the case of near or imperfect multicollinearity. Here, there is a high degree of linear correlation between the explanatory variables. This is not a case of violation of the classical linear regression assumptions. Though the regression coefficients can be estimated in this case but they cannot be estimated with great precision. The following Venn diagram illustrates the problem of multicollinearity. The closed figures represent the variations in Y (dependent variable), X_2 and X_3 (independent variables) of a three variable regression model.

The degree of overlap between the figures explains the degree of multicollinearity.

The degree of overlap between the figures explains the degree of multicollinearity.

Figure 10.1



There are different sources of multicollinearity. Multicollinearitymay arise due to model specifications like use of lagged explanatory variables which may be highly correlated, for example past and present incomes. It may be due to constraints in the population and, the sample being drawn from such part of the population which results in linear correlation between the variables. The method of data collection employed may cause multicollinearity to arise if data is collected over only a range of the variables where they are linearly dependent. Sometimes the variables by their inherent nature are also linearly correlated like income and wealth. Again, sometimes due to too many variables and a correspondingly smaller number of observations multicollinearity may arise.



CHECK YOUR PROGRESS

Q 1: Define multicollinearity. (answer in about 50 words)

.....

.....

.....

.....

.....

Q 2: Differentiate between perfect and near multicollinearity. (Answer in about 50 words)

.....

.....

.....

.....

.....

10.4 CONSEQUENCES OF MULTICOLLINEARITY

The consequences of multicollinearity must be illustrated in two dimensions, viz. the consequences of perfect multicollinearity and that of high multicollinearity. In the case of perfect multicollinearity, the regression coefficients remain indeterminate. In case of perfect multicollinearity in a multiple regression model, we cannot find the inverse of the matrix X'X and thus the estimation of the regression coefficients remains impossible. In this context, we may also consider the three variable linear regression model. The consequences in the same is illustrated as below.

$$Y_i = a + bX_{1i} + cX_{2i} + \mu_i, \rightarrow A$$

where μ_i is the random term and a, b, c are the population parameters. We may have a three variable regression model of consumption expenditure (Y, the dependent variable) as a function of income (X_1) and family size (X_2).

Now, $\hat{y}_{iOLS} = \hat{a} + \hat{b}\, x_{1i} + \hat{c}\, x_{2i}$

$$\hat{b}_{OLS} = \frac{(\Sigma x_{1i} y_i)(\Sigma x_{2i}^2) - (\Sigma x_{2i} y_i)(\Sigma x_{1i} x_{2i})}{(\Sigma x_{1i}^2)(\Sigma x_{2i}^2) - (\Sigma x_{1i} x_{2i})^2}$$

$$\hat{c}_{OLS} = \frac{(\Sigma x_{2i} y_i)(\Sigma x_{1i}^2) - (\Sigma x_{1i} y_i)(\Sigma x_{1i} x_{2i})}{(\Sigma x_{1i}^2)(\Sigma x_{2i}^2) - (\Sigma x_{1i} x_{2i})^2}$$

Suppose, there exists perfect correlation between X_1 and X_2 in our model (A) such that

$$X_{2i} = a + kX_{1i}$$

or $x_{2i} = kx_{1i}$ (in devation form)

where, k = any non-zero constant.

putting $x_{2i} = kx_{1i}$ in $\hat{b}$ and $\hat{c}$ we get

$$\hat{b}_{OLS} = \frac{(\Sigma x_{1i} y_i)(\Sigma k^2 x_{1i}^2) - (\Sigma k x_{1i} y_i (\Sigma x_{1i} k x_{1i})}{\Sigma x_{1i}^2 (\Sigma k^2 x_{1i}^2) - (\Sigma x_{1i} k x_{1i})^2}$$

$$= \frac{k^2 (\Sigma x_{1i} y_i)(\Sigma x_{1i}^2) - k^2 (\Sigma x_{1i} y_i)(\Sigma x_{1i}^2)}{k^2 (\Sigma x_{1i}^2)^2 - k^2 (\Sigma x_{1i}^2)^2}$$

$$= \frac{0}{0}$$ which is indeterminate

Again, $$\hat{c}_{ois} = \frac{(\Sigma x_{1i}{}^2 y_i)(\Sigma x_{1i}{}^2) - (\Sigma x_{1i}{}^2 y_i)(\Sigma x_{1i} x_{2i})}{(\Sigma x_{1i}{}^2)(\Sigma x_{2i}{}^2) - (\Sigma x_{1i} x_{2i})^2}$$

$$= \frac{(\Sigma k x_{1i} y_i)(\Sigma x_{1i}{}^2) - (\Sigma x_{1i} y_i)(\Sigma x_{1i} k x_{1i})}{(\Sigma x_{1i}{}^2)(\Sigma k^2 x_{1i}{}^2) - (\Sigma x_{1i} k x_{1i})^2}$$

$$= \frac{k(\Sigma x_{1i} y_i)(\Sigma x_{1i}{}^2) - k(\Sigma x_{1i} y_i)(\Sigma x_{1i}{}^2)}{k^2 (\Sigma x_{1i}{}^2)^2 - k^2 (\Sigma x_{1i}{}^2)^2}$$

$$= \frac{0}{0}$$ which is again indeterminate.

Further, we may note that in our model (A) if we put instead of x_2, kx_1 then we get

$$y_i = a + bx_{1i} + c\,(kx_{2i}) + \mu_i$$
$$= a + (b + ck)\, x_{1i} + \mu_i$$
$$= a + d\, x_{1i} + \mu_i$$

It is now possible to find the estinate $\hat{d}$. But due to the presence of perfect multicollinearity the seperate effects of X_1 and X_2 on Y i.e. $\hat{b}$ and $\hat{c}$ cannot be estimated.

Moreover, we may recall

$$\text{var}(\hat{b}_{OLS}) = \frac{\sigma_\mu{}^2 \Sigma x_{2i}{}^2}{\Sigma x_{1i}{}^2 \Sigma x_{2i}{}^2 - (\Sigma x_{1i} x_{2i})^2}$$

Putting $x_{2i} = kx_i$

$$\text{var}(\hat{b}_{OLS}) = \frac{k^2 \sigma_\mu{}^2 \Sigma x_{1i}{}^2}{k^2(\Sigma x_{1i}{}^2) - k^2(\Sigma x_{1i}{}^2)^2} = \alpha$$

Similarly $$\text{var}(\hat{c}_{OLS}) = \frac{\sigma_\mu{}^2 \Sigma x_{1i}{}^2}{(\Sigma x_{1i}{}^2 \Sigma x_{2i}{}^2 - (\Sigma x_{1i} x_{2i})^2}$$

$$= \frac{\sigma_\mu{}^2 \Sigma x_{1i}{}^2}{k^2(\Sigma x_{1i}{}^2)^2 - k^2(\Sigma x_{1i}{}^2)^2} = \alpha$$

On the other hand, in high multicollinearity estimation of the regression coefficients are possible. High multicollinearity being no violation of the standard OLS assumptions, the OLS estimators continue to have the BLUE properties. However, they have large variances and covariances. Higher the correlation higher is the variances and covariances of the estimators. The confidence intervals of the coefficients will be very wide in the presence of high multicollinearity. The probability of accepting a false hypothesis increase. Since the estimated standard error increases in the presence of high multicollinearity so the 't' values are smaller. Accordingly, in testing the null hypothesis that the regression coefficients have no effect on the dependent variable(b=c=0), it is hard to reject the null hypothesis. However, the R^2 value is very high in high multicollinearity and the F value are significant such that on the basis of the F test one can reject the null hypothesis that regression coefficients have no effect on the dependent variable(b=c=0). TheOLS estimates in the presence of high multicollinearity become sensitive to changes in data.



CHECK YOUR PROGRESS

Q 3: What problem is associated with perfect multicollinearity? (Answer in about 50 words)

.....

.....

.....

.....

.....

10.5 METHODS OF DETECTING MULTICOLLINEARITY

The following section deals with the methods of identifying the presence of multicollinearity in any regression model.

- **R^2 values and t values:** The presence of multicollinearity is often suggested as in the case the R^2 value, the coefficient of determination is high. The F-test will in the case reject the null hypothesis that the regression coefficients will be 0. However, in contradiction the t values are smaller and will show otherwise.
- **Correlation among the explanatory variables:** If one conducts a correlation between each pair of explanatory variables and there exists a high correlation between any two such variables it suggests the existence of multicollinearity. But this method is not compatible in regression models with more than two explanatory variables where one explanatory variable is a linear combination of two or more explanatory variables. The limitation of the simple pair wise correlation method as suggested has further led to the recommendation that one examines the partial correlation coefficients.
- **Auxiliary Regression:** As the name suggest, in the method each explanatory variable is regressed to the other X's.Suppose there are k independent variables then there will be a regression equation for each of the explanatory variables. Each of these equations must be fitted using OLS. These are called the auxiliary regressions. For each regression the R^2value, the coefficient of determination is computed.

Now, $\{R^2/(k-2\}/\{(1-R^2)/(n-k+1)\}$ follows F distribution with k-2 and n-k+1 degrees of freedom. Here, k is the no. of explanatory variables and n the sample size. Now, a significant F value will mean that the particular X is collinear with the other X's. In the case we may drop the collinear variable to avoid multicollinearity. It is however suggested, that, as a general rule if any R^2 obtained from any of the auxiliary regressions is greater than the overall R^2, that is of the given regression model, then multicollinearity may be a troublesome issue.

- **Tolerance and Variance Inflation Factor:** The variance inflation factor or the measure of tolerance is often suggested as a method of detecting multicollinearity. The variance of any regression coefficient$\hat{a}_j$ in any k-variable multiple regression model, can be expressed as $Var(\beta_j) = \sigma^2/ \sum x_j^2 . (1/(1-R_j^2))$ .Here, R_j^2 is the coefficient of determination of the auxiliary regression of X_j^2 on the remaining X's and β_j isthe coefficient of X_j. This $(1/(1-R_j^2))$ is the variance inflation factor which increases as R_j^2 increases, i.e. as the collinearity of Xj with other X's increases and hence R_j^2 increases. $(1-R_j^2)$ is the measure of tolerance. It increases to unity if X_j is not correlated to the other regressors but decreases to 0 if there is perfect correlation.
- **Eigen values and Condition index:** Multicollinearity can be detected on the basis of Eigen values, condition number and condition indices. However, the discussion on the detection of multicollinearity on the basis of eigen values, condition number or condition indices are beyond the scope of the present chapter.

10.6 REMEDIES TO THE PROBLEM OF MULTICOLLINEARITY

Multicollinearity of mild degree, as suggested by econometricians, may be tolerated in a regression model. In the case of existence of high multicollinearity, the procedures, discussed in the following section, may be adopted, though they may have their own limitations.

- **Using information of prior information or research:** The method suggests that there may exist a priori information from economic theory

or any research about the coefficients of variables or any relation among the parameters of the regression model. This information may then be used in estimating the parameters. For example, in any regression model of the form

$Y=a + bx^1+ cx^2$……(i)

It may happen that

b=0.4 c ……(ii)

This additional information if put in the model we can estimate b and then estimate c from (ii).

Likewise, in the Cobb-Doulas production function of the form, $Q=AL^aK^b$, a+b=1 (where, a and b are the parameters to be estimated). Thus, we have, $Q=AL^{1-b}K^b$. we can now estimate b despite correlation between L (labour) and K (capital).

- **Transformation of variables:** One way to eliminate the problem of multicollinearity is to transform the trouble causing correlated variables from the use of underlying theory to eliminate or reduce the severe multicollinearity problem. Moreover, pooling time series data and cross section data or using cross section data instead of the other if the population is known, can eliminate the problem in cases where multicollinearity occurs in case of time series data.
- **Dropping of collinear variables:** A simple solution to the multicollinearity problem is that we may drop collinear variables from the regression model. In such a case, the multicollinearity problem may be eradicated. But this solution may entail specification bias in the regression model. If the variables all needs to be included then specification bias may arise when one of the colinear variable is removed. Thus, the estimators will be biased.
- **Increasing the sample size:**Multicollinearity is often cited to be a problem arising out of small sample size. Thus, a remedy suggested is to increase the sample size. With a large sample the variance of the estimated coefficients will reduce and we will have more accurate estimates.However, the problems associated with getting additional data including time and resources will persist.

- **Pooling cross sectional and time series data:** Sometimes, multicollinearity problem is regression involving time series data may be removed by way of combining time series and cross-sectional data. Suppose in the three variable regression model, the variables X_2 and X_3 are correlated in time series data. We may than estimate the coefficient of either of the variables from cross section data provided the other variable do not vary in the cross-section data. Thereafter we may use the estimated coefficient in the earlier regression mode to estimate the coefficient of the remaining variable.
- **Adopting other methods of estimation:** Sometimes multivariate statistical techniques of estimation like factor analysis, principal component analysis or ridge regression are used instead of OLS method to overcome the problem of multicollinearity.



CHECK YOUR PROGRESS

Q 4: Give solutions to the problem of multicollinearity. (answer in about 50 words)

.....

.....

.....

.....

.....



10.8 FURTHER READINGS

1) Gujrati, D.N. (1995): *Basic Econometrics*, Tata McGraw Hill, New Delhi.
2) Koutsoyiannis, A. (1977): *Theory of Econometrics*, the MacMillan Press Ltd. , London.
3) Wooldridge, J. F. (2008): *Introductory Econometrics: A Modern Approach*, Third edition, South- Western Cengage Learning India.
4) Johnston, J. (1991): *Econometric Methods*, MacGraw Hill Book Co., London.



10.7 LET US SUM UP

- Multicollinearity is a situation where there exists perfect or a high degree of linear relationship between the explanatory variables.
- In the case of perfect multicollinearity, the regression coefficients are indeterminate.
- When multicollinearity is not perfect but there is near perfect multicollinearity the standard errors are large and the population parameters cannot be estimated with precision.
- R^2 value if high but the t test suggest that the regression coefficients are not significant then this indicates the presence of multicollinearity.
- One may run auxiliary regressions of each explanatory variable on the other X's and calculate the corresponding R^2 values. For any high R_i^2 it will indicate that the corresponding X_i^2 is highly correlated with the other X's.
- The various ways to get rid of the problem of multicollinearity includes using a priori information, increasing the sample size, pooling cross section and time series data, omitting redundant variables, etc.



10.9 ANSWERS TO CHECK YOUR PROGRESS

Ans Q No 1: Multicollinearity refers to a situation where the independent variables are linearly correlated to each other. If there exists perfect correlation among the independent variables then it is a case of perfect multicollinearity. If there exist not perfect but a high degree of correlation between the variables then it is called near or imperfect multicollinearity.

Ans Q No 2: Perfector exact multicollinearity refers to perfect linear correlation among the independent variables. Near or imperfect multicollinearity is when there is a high but not perfect linear correlation among the variables. Perfect multicollinearity results in

that the regression coefficients are indeterminate. However, in near multicollinearity though the coefficients are determinate they cannot be estimated with great precision.

Ans Q No 3: Perfect multicollinearity, a case where there exists linear correlation among the independent variables, the regression coefficients remain indeterminate. In case of perfect multicollinearity in a multiple regression model, the inverse of the matrix X'X does not exist.Thus, the estimation of the regression coefficients remains impossible.

Ans Q No 4: Multicollinearity may be removed by increasing the sample size. Redundant variables may be dropped or the coefficients may be estimated through use of a priori information or through pooling of cross section and time series data. Collinear variables may be transformed with underlying theory.



10.10 MODEL QUESTIONS

Short Questions (Answer each question in about 150 words)

Q 1: What is multicollinearity?

Q 2: Differentiate between multicollinearity and orthogonality?

Q 3: What problems arises in near perfect multicollinearity?

Q 4: How do we correct perfect multicollinearity?

Q 5: How can we detect near perfect multicollinearity?

Long Questions (Answer each question in about 300-500 words)

Q 1: What is multicollinearity? Examine the problems associated in the presence of multicollinearity?

Q 2: Discuss the sources multicollinearity.

Q 3: Describe what consequences may arise in the presence of multicollinearity.

Q 4: Explain the methods of detecting multicollinearity.

Q 5: What are the solutions to the problem of multicollinearity?

*** ***** ***

UNIT 11 : QUALITATIVE FACTORS AS REGRESSORS

UNIT STRUCTURE

11.1 Learning Objectives

11.2 Introduction

11.3 Nature of qualitative variable

11.4 Use of Dummy Variables for capturing the effect of qualitative factors

11.5 Interpretation of coefficients of dummy variables

11.6 Dummy Variable Trap

11.7 Let Us Sum Up

11.8 Further Readings

11.9 Answer to Check Your Progress

11.10 Model Questions

11.1 LEARNING OBJECTIVES

After going through this unit, you will be able to know-

- to get introduced about the qualitative factors and its inclusion in regression model
- to know about the use of dummy variable
- to know about the interpretation of coefficient of dummy variable
- to know the problem of dummy variable trap and its solution

11.2 INTRODUCTION

In regression analysis, quite often, we need to introduce qualitative factors either as explanatory variable or as dependent variable. As qualitative variables are not numeric, it is not possible to introduce them directly in regression analysis. We incorporate the effects of qualitative variables in regression analysis using the variable known as Dummy Variable. For capturing the temporal effect and spatial effect, we need to use dummy

variable. As for example, if we are interested in comparing the sale of particular goods between war time and post-war time, a time dummy variable is to be used in the regression model. Similarly, comparison of sale of the goods across different locations, say states of India, requires use of location dummy variable in the regression model. However, we must be careful in introducing dummy variable in regression model as there may be the problem of Dummy Variable Trap if the number of dummy variable is different from the required number. This chapter discusses the nature of qualitative variable, use of dummy variable and the interpretation of its coefficient and the problem of dummy variable trap.

11.3 NATURE OF QUALITATIVE VARIABLE

Broadly, the variables can be quantitative or qualitative. Quantitative variables are numeric. Age, income, GDP, national income etc are some example of quantitative variable. As quantitative variables have value, they can be added, subtracted, multiplied etc. while these are not possible in case of qualitative variables as they are not numeric. For instance, caste, race, gender, season etc are qualitative variables. Qualitative variables are also known as categorical variable. There is no natural ordering in case of qualitative variable. Further, ordering in numerical scale is not possible for qualitative variables, and such variables are assigned nominal scale.

11.4 USE OF DUMMY VARIABLES FOR CAPTURING THE EFFECT OF QUALITATIVE FACTORS

Many events are affected by different factors. For instance, the wage earned by labour in an industry may be affected by working hours and gender. If we want to examine the impact of working hours and gender on the wage using regression analysis, working hours can be directly included in the regression model as it is quantitative. However, we can't include gender directly in regression model as it is qualitative in nature. For that dummy variable is to be introduced in the model. Thus,

dummy variables are the tools for capturing the effect of a qualitative variable in the regression analysis. Dummy variables are binary in nature and have two possible options-'0' and '1'. A dummy variable takes '1' for a particular aspect of the qualitative factor and '0' for the rest of the aspect(s).

Suppose

Y=f(X, Gender),

Where,

Y is the wage

X is the hours of working

Let

D = 1 for female

= 0 for male

Now the regression equation is given by

$Y_i = \alpha + \gamma D_i + \beta X_i + u_i$ (1)

11.5 INTERPRETATION OF COEFFICIENTS OF DUMMY VARIABLES

The interpretation of the coefficient of a dummy variable is different from the interpretation of the coefficient of quantitative variable. For instance, in case of equation (1),

$\frac{\delta Y_i}{\delta X_i} = \beta$ but

$\frac{\delta Y_i}{\delta D_i} \neq \gamma$

When,

$D = 1,\ Y_i = (\alpha + \gamma) + \beta X_i + u_i$ and

$D = 0,\ Y_i = \alpha + \beta X_i + u_i$

Therefore, γ is the difference of female intercept over male intercept. Thus, the coefficient of dummy variable, is the defferential effect of that category of the qualitative factor for which D = 1 over the category for which D = 0 (reference category).

In general, if there are 'g' categories of a qualitative factor, then we need to use 'g-1' number of dummy variables.

Let

Y = f(X, seasons)

Where,

Y is sale of ice cream and

X is price of ice cream

Let there are four seasons and

S_2 = 1 for spring

= 0, otherwise

S_3 = 1 for summer

= 0, otherwise

S_4 = 1 for autumn

= 0, otherwise

The excluded season, say, winter is known as the reference/base category.

Now

$$Y_i = \alpha + \gamma_2 S_{2i} + \gamma_3 S_{3i} + \gamma_4 S_{4i} + \beta X_i + u_i \quad (2)$$

If,

$S_{2i} = 1$, $Y_i = (\alpha + \gamma_2) + \beta X_i + u_i$

$S_{3i} = 1$, $Y_i = (\alpha + \gamma_3) + \beta X_i + u_i$

$S_{4i} = 1$, $Y_i = (\alpha + \gamma_4) + \beta X_i + u_i$ and

When,

$$S_{2i} = S_{3i} = S_{4i} = 0,\ Y_i = \alpha + \beta X_i + u_i \quad (3)$$

Therefore,

γ_2 is the differential effect of spring.

γ_3 is the differential effect of summer and

γ_4 is the differential effect of autumn.

Thus, the coefficient of a dummy variable is the differential effect of the category for which dummy variable takes value '1' over the category for which all the dummies take value '0'.



CHECK YOUR PROGRESS

Q 1: Discuss the nature of a qualitative variable? (Answer in about 20 words)

.....

.....

Q 2: What is dummy variable? (Answer in about 30 words)

.....

.....

.....

11.6 DUMMY VARIABLE TRAP

The number of dummy variable should be 'number of categories of the qualitative variable – 1', ie, we should use 'g-1' number of dummy variables if there are 'g' categories of the qualitative variable. If the number of dummy variable is equal to the number of categories of the qualitative variable, it is known as 'Dummy Variable Trap (DVT)'. In such situation, we cannot use Ordinary Least Square (OLS) to estimate the parameters the problem of multicollinearty is arised there. This can be explained as below-

Let

Y = f(X, Gender) and

D_1 = 1 for female

= 0 for male

D_2 = 1 for male

= 0 for female

Therefore,

$$Y_i = \alpha + \gamma_1 D_{1i} + \gamma_2 D_{2i} + \beta X_i + u_i \qquad (4)$$

Here, the matrix of explanatory variables is

$$\begin{bmatrix} 1 & 1 & 0 & X_1 \\ 1 & 1 & 0 & X_2 \\ 1 & 0 & 1 & X_3 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & X_n \end{bmatrix}$$

Where,

$D_{1i} + D_{2i} = 1$

Therefore, first three columns are perfectly linearly dependent revealing the case of perfect multicollinearty. Hence, OLS can't be used in such situation to get estimate of parameters as the matrix of explanatory variable is singular here which violates the assumption of full column rank of the matrix.

In the present case, we have two options-

- Use number of dummy which is equal to 'number of categories -1, that is, one dummy variable in present case.
- It is possible to get unbiased estimate of parameters by subtracting the constant part from the model.

For instance, subtracting the constant part of equation (4), we get

$$Y_i = \gamma_1 D_{1i} + \gamma_2 D_{2i} + \beta X_i + u_i \quad (5)$$

If

$D_{1i} = 1$, $Y_i = \gamma_1 + \beta X_i + u_i$

$D_{2i} = 1$, $Y_i = \gamma_2 + \beta X_i + u_i$

And γ_1 and γ_2 are the intercept for female and male respectively.

However, in case of two qualitative factors, even if we subtract the constant part, the problem will resurface.

Suppose,

Y=f(X, Gender, Cast),

Where,

Y is the efficiency of teacher and

X is the years of teaching

Let

$G_1 = 1$ for female

$= 0$ for male

$G_2 = 1$ for male

$= 0$ for female

$C_1 = 1$ for General

$= 0$ for Others

$G_2 = 1$ for Others

= 0 for General

Therefore

$$Y_i = \gamma_1 G_{1i} + \gamma_2 G_{2i} + \delta_1 C_{1i} + \delta_2 C_{2i} + \beta X_i + u_i \quad (6)$$

Here,

$$G_{1i} + G_{2i} = 1 = C_{1i} + C_{2i}$$

Or $G_{1i} + G_{2i} - C_{1i} - C_{2i} = 0$

Thus there is a perfect linear dependence among the first four columns. So, use of OLS will not yield the estimate of parameters.



CHECK YOUR PROGRESS

Q 3: How many dummy variables should be used? (Answer in about 30 words)

.....

.....

.....

Q 4: Discuss the consequence of the dummy variable trap? (Answer in about 40 words)

.....

.....

.....

.....



11.7 LET US SUM UP

- We use dummy variable as a tools to capture the effect of qualitative variables in the regression analysis. Unlike the coefficient of a quantitative variable, the coefficient of a dummy variable indicates the differential effect of the category for which it takes one over the category for which it (all the dummy variables if there are more than two categories) takes zero. The number of dummy variable to be used for a qualitative factor should be equal to the 'number of categories of the factor -1'; otherwise we have to confront with the problem of dummy variable trap.



11.8 FURTHER READINGS

1) Dougherty, Christopher (2008), *'Introduction to Econometrics'*, Third Edition, Oxford University Press.

2) Gujarati, D N (2003), *'Basic Econometrics'*, International edition, McGraw-Hill International Book Company

3) Johnston, J, *'Econometric Methods'*, Second Edition, McGraw-Hill International Book Company

4) Johnston, J (1988), *'Econometric Methods'*, Third Edition, McGraw-Hill International Book Company



11.9 ANSWER TO CHECK YOUR PROGRESS

Ans to Q No 1: Qualitative variable are not numeric. Such variables do not have natural ordering. Qualitative variable are assigned nominal scale.

Ans to Q No 2: Dummy variable are tools used in regression model to capture the effect of qualitative variables. Dummy variables are categorical having only two options – 1 and 0. Thus, dummy variables are binary in nature.

Ans to Q No 3: The number of dummy to be used should be equal to the 'number of categories – 1'. Thus, if there are 'n' categories of the variable, we should use 'n-1' number of dummy variables.

Ans to Q No 4: Dummy variable trap refers to the situation when we use dummies equal to the number of categories of the variable. In such situation, linear dependency among columns in the matrix of explanatory variables arises. As a result, estimates of parameters are not possible to get under OLS.



11.10 MODEL QUESTIONS

Short Questions (Answer each question in about 150 words)

Q 1: Distinguished quantitative and qualitative variables

Q 2: Discuss the use of dummy variable

Long Questions (Answer each question in about 300-500 words)

Q 1: Explain the interpretation of coefficient of dummy variable.

Q 2: What is 'Dummy Variable Trap'? What happens to OLS estimation in the presence of the problem of dummy variable trap?

*** ***** ***

UNIT 12 : TIME SERIES ANALYSIS

UNIT SRUCTCTURE

12.1 Learning Objectives

12.2 Introduction

12.3 Components of Time Series

12.4 Estimation of Trend

- 12.4.1 Free Hand Curve Method
- 12.4.2 Semi-Average Method
- 12.4.3 Moving Average Method
- 12.4.4 Least Square Method

12.5 Let Us Sum UP

12.6 Further Readings

12.7 Answer to Check Your Progress

12.8 Model Questions

12.1 LEARNING OBJECTIVES

After going through this unit, you will be able to-

- Know components of a time series
- Know the estimation of trend
- Moving average method of estimation of trend
- Curve fitting method of estimation of trend

12.2 INTRODUCTION

A time series is an arrangement of statistical data in accordance with the time of occurance.

There are basically four methods which are generally used for the study and estimation of trend component in a time series such as Free Hand Curve Method, Semi-Average Method, Moving Average Method, and Least Square Method.

Method of moving averages is a very simple and flexible mood of measuring trend.

Curve fitting method is another important method of estimation of trend.

A time series is an arrangement of statistical data in a chronological order, i.e., in accordance with its time of occurrence. It reflects the dynamic pace of movements of a phenomenon over a period of time. Time Series has many applications and it gives us the trend and nature of series with the passage of time. Accordingly, time series have an important and significant place in Business and Economics, and basically most of the statistical techniques and methods have been developed by various economists for the analysis of time series data. However, these techniques can also be applied for the study of behaviour of any phenomenon collected chronologically over a period of time in any discipline relating to natural and social sciences, though not directly related to economics or business. Such data are of particular interest in economics, business and commerce, where the values of a variable are observed chronologically be days, weeks, months, quarters, or years.

Mathematically, a time series is defined by the functional relationship,

Y= f(t)

Where y is the value of the phenomenon (or variable under consideration at time t) e.g;

- The population of a country or a place in different years (t)
- The number of births and deaths(y) in different months (t) of the year
- The sales(y) of a departmental store in different months (t) of the year.
- The temperature (y) of a palace on different days (t) of the week

As t refers to time so if the values of a phenomenon or variable at times $t_1, t_2 \ldots\ldots\ldots\ldots t_n$ are $y_1, y_2 \ldots\ldots\ldots\ldots y_n$ respectively, then the series which is given as follows constitutes a time series:

t : $t_1 t_2 t_3 \quad \ldots t_n$

y: $y_1 y_2 y_3 \quad \ldots y_n$

The values of y fluctuate from time to time and these fluctuations are the results of the joint action of multiple forces and events that occur

with the passage of time. The forces can be changes in tastes and habits of people, increase in population development of new techniques resulting in lower cost of production changes in weather conditions, etc. Thus, values of the variable undergo change with the passage of time. The objective of time series analysis is to isolate and measure the effects of various components. Time series analysis helps to predict the future tendency of the variable in study.

- **Use and Significance of Time Series:** There are various areas where we can apply time series analysis for measurement. Some of the applications of time series as follow:
  - To study the past behaviour of data for formulating the policies for future.
  - With the help of time series, anticipation of changes going to occur in the future becomes possible because studies about past proves beneficial for forecasting about future.
  - One can compare different time series together for understanding the cause and effect relationship.
  - It is useful in studying the present variations in different economic variables.

12.3 COMPONENTS OF TIME SERIES

The value of the phenomenon is not only affected by a single factor but due to the cumulative effect of a multiplicity of factors pulling it up and down. However, if the various forces were in a state of equilibrium, then the time series will remain constant. A time series may be broadly classified into the following four categories, commonly known as the components of a time series, some or all of which are present in a given time series) in varying degrees

- Secular Trend or Trend (T)
- Seasonal Variations (S)
- Cyclical Variations (C).
- Random or Irregular Variations (I)

The value (y) of a phenomenon observed at any time (n) is the

net effect of the interaction of above components. We shall explain these components briefly in the following sections.

- **Secular Trend (T):** The general tendency of the time series data to grow or decline or stagnate during a long period of time is called the secular trend. This phenomenon is usually observed in most of the series relating to Economics and Business, e.g. an upward tendency is usually observed in time series relating to population, production and sales of products, prices, incomes, money in circulation, etc., while a downward tendency is noticed in the time series relating to deaths, epidemics, etc. This trend studies movement over a long time period. It should be clearly understood that trend is the general, smooth, long-term average tendency.Such tendencies are the result of the forces which are more or less constant for a long time or which change very gradually and continuously over a long period of time such as the change in the population, tastes, habits and customs of the people in a society, and so on. For example, the effect of population increase over a long period of time on the expansion of various sectors like agriculture, industry, education, textiles, etc., is a continuous but gradual process. The term "long period of time is a relative term and cannot be defined exactly. It would very much depend on the nature of the data. In certain phenomenon, a period as small as few hours may be sufficiently long while in others even a period as long as 3 or 4 years may not be sufficient. For example, to have an idea about the production of a particular product (agricultural or industrial production), an increase over the past 20 or 30 months will not reflect a secular change for which we must have data for 7-8 years. Quite often, time series exhibit secular trend due to population growth, technological reforms, capital formation, improvements in business organisations etc. Secular Trend is usually of two types:
  - **Linear Trend:** When long term rise or fall in a time series takes place by a constant amount then that is called a linear trend. This is also known as straight line trend. Symbolically this can be represented as: $Y = a + bX$

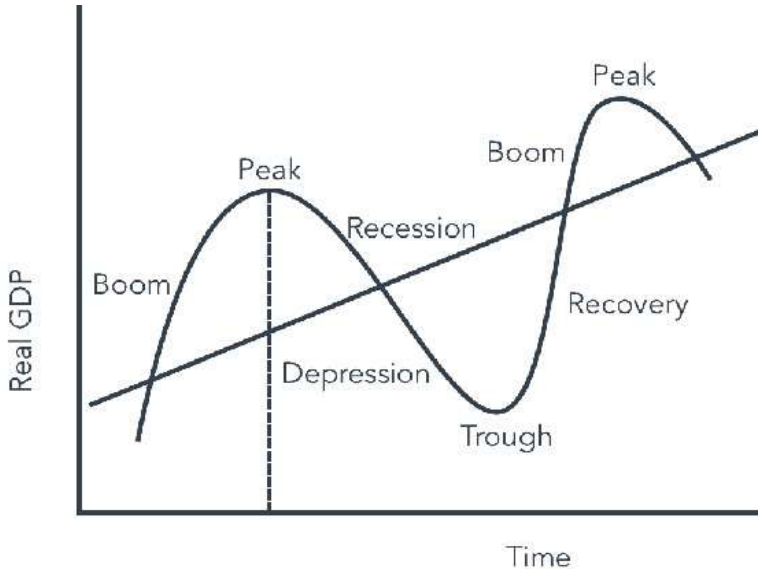
- **Parabolic Trend:** A trend in a time series data is said to be parabolic when long term rise or fall in a time series is not taking place at a definite rate. It has many form but most prominent of them is the second degree parabolic or quadratic trend. Symbolically this can be represented as: $Y = a + bX + cX^2$

- **Seasonal Variations (S):** These variations refer to periodic variations in time series which operate in a regular and periodic manner over a span of less than a year, i.e., during a period of 12 months and have the same or almost same pattern year after year. Thus, scasonal variations in a time series will be there if the data are recorded quarterly every three months, monthly, weekly, daily, hourly, and so on. These variations are of regular type and they are easy to be anticipated. Most of the variations taking place in economic and business world are of this type. For example, Prices of foodgrains tend to be low at the time of harvesting and high at sowing time. During winters, there is rise in the price of woollen clothes and in summer its price falls. Thus, most of the economic time series are influenced by seasonal swings, e.g. prices, production and consumption of commodities; sales and profits in a departmental store: bank clearings and bank deposits etc, are all affected by seasonal variations. The seasonal variations may be attributed to the following two causes: The main objective of the measurement of seasonal variations is to isolate them from the trend and study their effects A study of the seasonal patterns is extremely useful to businessmen, producers, sales managers, etc. in planning future operations and in formulation of policy decisions regarding purchase production, inventory control, personnel requirements selling and advertising programmes. In the absence of any knowledge of seasonal variations, a seasonal upswing may be mistaken indicator of better business conditions while a seasonal slump may be mis-interpreted as deteriorating business conditions Thus to understand the behaviour of the phenomenon in a time series properly, the time series data

must be adjusted for seasonal variations. This is done by isolating them from trend and other components on dividing the given time series values by the seasonal variations Model.

- **Cyclical Variations (C):** The oscillatory movements in a time series with period of oscillation greater than one year are termed as cyclical variations. These variations in a time series are due to ups and downs recurring after a period greater than one year. Cyclical variations arise because of trade cycles occur in the economy. "Business cycles have been described as successive waves of expansion and contraction that occur at about the same time in many economic activities"- Burns and Mitchel. Theoscillatory movements in any business activity are the outcome of the so-called "Business Cycles which are the four phased cycles comprising prosperity (boom), recession, depression and recovery from time to time. These booms and depressions in any business activity follow each other with steady regularity and the complete cycle from the peak of one boom to the peak of next boom usually lasts from 7 to 9 years (Figure 1). Most of the economic and business series, e.g., series relating to production, prices, wages, investments, etc. are affected by cyclical upswings and downswings

Figure 12.1: Phases of Business Cycles



The study of cyclical variations is of great significance to the business executives in formulation of policies aimed at stabilizing the level of business activity in the macroeconomic environment. The awarenessabout cyclic component enables a businessman to have an idea about the periodicity of the different phases of business cycle viz. boom, recession, depression and recovery.Figure 1 makes it clear that business activities tend to be at their height during prosperity, then recession or the fall starts and gradually, there comes the lowest limit of decline in trade activities i.e. depression. According to an estimate, trade cycle takes complete turn from 2 to 10 years and then it repeats itself and thus sequence goes on.

- **Random or Irregular Variations (I):** There is inherent in every time series another factor called random or irregular variations. These fluctuations are purely random and are the result of such unforeseen and unpredictable forces which operate in absolutely irregular manner. Such variations do not exhibit any definite pattern and there is no regular period or time of their occurrence, hence they are named irregular variations. Irregular variations refer to those short term variations in time series which occur irregularly due to certain accidental causes, which are known as random causes. Irregular variations take place due to accidental causes like floods, famines, wars, earthquakes, and epidemics etc., which behave in a very erratic and unpredictable manner. Normally they are short-term variations but sometimes their effect is so intense that they may give rise to new cyclical or other movements. These fluctuation don't happen in definite order or in a systematic way. Because of their absolutely random character, it is not possible to isolate such variations and study them exclusively nor can we forecast or estimate them precisely. The best that can be done about such variations is to obtain rough estimates (from past experience) and accordingly make provisions for such abnormalities during normal times in business.

- **Analysis or Decomposition of a Time Series:**
  - The time series analysis consists of identifying and determining the various forces or influences whose interaction produces the variations in the time series.
  - Isolating, studying, analysing and measuring them independently, i.e., by holding other things variations in the time series constant.

The time series analysis is of great importance not only to businessman or an economist but also to people working in various disciplines in natural, social and physical sciences. Some of its uses are enumerated below:

- It enables us to study the past behaviour of the phenomenon under consideration, i.e. to determine the type and nature of the variations in the data
- The segregation and study of the various components is of paramount importance to a businessman in the planning of future operations and in the formulation of executive and policy decision
- It helps to compare the actual current performance or accomplishments with the expected ones con the basis of the past performances) and analyse the causes of such variations if any.
- It enables us to predict or estimate or forecast the behaviour of the phenomenon in future which is (w) It helps us to compare the changes in the values of different phenomena different times which are very essential for business planning.

- **Mathematical Models for Time Series:** Time series is composed of four components viz. Trend (T), Seasonal variations (S), Cyclical Variations (C), and Irregular Variations (I). There is always some sort of relationship in these four time series components. In order to study the influence of these four components on time series, they are measured separately, which is called as Analysis or Decomposition of time-series. The following are the two models commonly used for the decomposition components of a time series.

- **Additive Model or Decomposition by Additive Hypothesis:** According to the additive model, this has been assumed that the time series is the sum of the four components which can be expressed as

$$O = T + S + C + I$$

This model treats all the constituents as residuals on the basis of which, by deducting trend from the original data, short term fluctuations can be determined. Similarly cyclical variations and irregular variations can be determined by deducting seasonal variations from short term variations. On the basis of additive model, the analysis of various components is as follows:

$$O - T = S + C + I$$
$$O - T - S = C + I$$
$$O - T - S - C = I$$

The additive model assumes that all the four components of the time series operate independently of each other so that none of these components has any effect on the remaining three. This implies that the trend, however, fast or slow it may be, has no effect on the seasonal and cyclical components nor do seasonal swings have any impact on cyclical variations and conversely. However, this assumption is not true in most of the economic and business time series where the four components of the time series are not independent of each other.

- **Multiplicative Model or Decomposition by Multiplicative Hypothesis:** Most of the economic and business time series are characterised by the following classical multiplicative model, which symbolically can be illustrated as:

$$O = T * S * C * I = TSCI \qquad \text{(Equation 1)}$$

This model assumes that the four components of the time series are due to different causes but they are not necessarily independent and they can affect each other. Whatever the

component is to be separated, works as a divisor with respect to original data (O). Analysis of different components on the basis of multiplicative model can be expressed in the following forms:

$$\frac{O}{T} = SCI$$

$$\frac{O}{T * S} = CI$$

$$\frac{O}{T * S * C} = I$$

If we take logarithm on both sides, using Equation 1, we get

$$logO = logT + logS + logC + logI$$

Which is nothing but the additive model fitted to the logarithms of the given time series values.

Most of3 the time series relating to economic and business phenomena conform to the multiplicative model. In practice, additive model is rarely used. Thus, by isolating trend values from the given time series i.e. by dividing the given time series values by the trend values or subtracting trend values from the given time series values, we can study the short-term and irregular movements

- **Use of Trend Analysis:** The study of the data over a long period of time enables us to have a general idea about the pattern of the behaviour of the phenomenon under consideration. This helps in business forecasting and planning future operations. For example, if the time series data for a particular phenomenon exhibits a trend in a particular direction, then under the assumption that the same pattern will continue in the near future, an assumption which is quite reasonable unless there are some fundamental and drastic changes in the forces affecting the phenomenon—we can forecast the values of the phenomenon for future also. The accuracy of the trend curve or trend equation or the estimates obtained from them will depend

on the reliability of the type of trend fitted to the given data.Trend analysis enables us to compare two or more time series over different periods of time and draw important conclusions about them.



CHECK YOUR PROGRESS

Q 1: Define Time Series.

.....

.....

Q 2: What are the methods of measuring time series?

.....

.....

Q 3: What are the different components of time series?

.....

.....

12.4 ESTIMATION OF TREND

The following are the four methods which are generally used for the study and measurement of trend component in a time series.

- Free Hand Curve Method
- Semi-Average Method
- Moving Average Method
- Least Square Method

12.4.1 Free Hand Curve Fitting Method

This is the simplest and the most flexible method of estimating the secular trend. In this method first of all, the original data of the time-series is plotted on a graph paper. Thereafter, taking care of the fluctuations of data, a smooth curve is drawn which passes through the mid points of the fluctuations of time-series. In fact, this curve is called as freehand trend curve. The smoothing of the curve eliminates the other components, viz, seasonal, cyclical and random variations.

In order to obtain proper trend line or curve, the following points may be followed:

(i) It should be smooth

(ii) The number of points above the trend curve/line should be more or less equal to the number of points below it.

(iii) The sum of the vertical deviations of the given points above the trend line should be approximately equal to the sum of vertical deviations of the points below the trend line so that the total positive deviations are more or less balanced against total negative deviations

(iv) The sum of the squares of the vertical deviations of the given points from the trend line/curve isminimum.

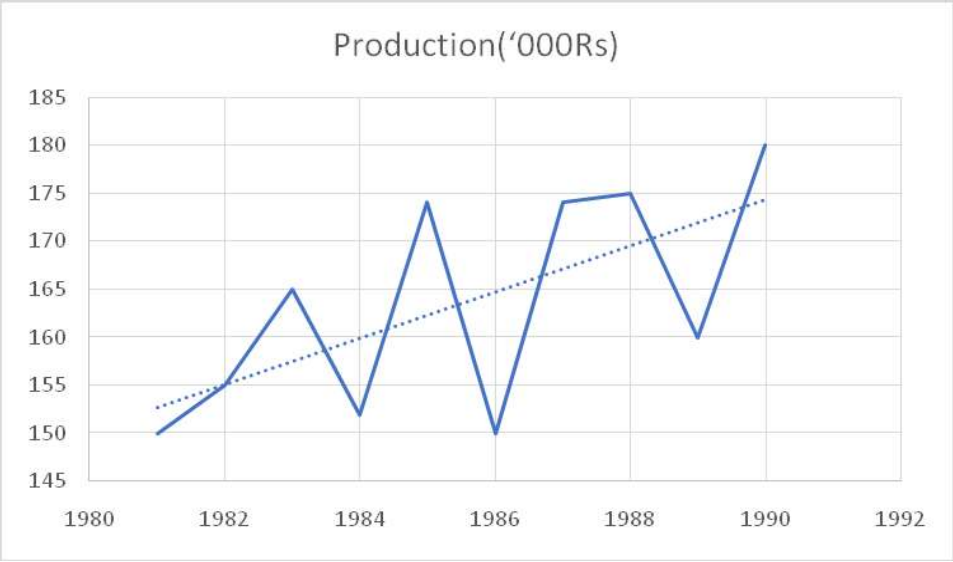
The procedure of this method can be illustrated by the following example:

Example 1: Fit a trend line to the following data by the free hand curve method:

Year	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990
Production (in Qts)	150	155	165	152	174	150	174	175	160	180

Solution:

Figure 12.2: Fitting of Trend Line by Free Hand Curve



Merits and Demerits of Freehand Curve Method:

- **Merits:**
  - It is very simple and time-saving method
  - It does not require any mathematical calculations
  - It is a very flexible method in the sense that it can be used to describe all types of linear trend as well as non-linear
  - This method is also used for forecasting about future.
- **Demerits:**
  - The strongest objection to this method is that it is highly subjective in nature. The trend want to obtained will very much depend on the personal bias and judgement of the investigator handling data and consequently different persons will obtain different trend curves for the same set of data. Thus bias may affect the findings.
  - Long run movement (trend) obtained from this method is not definite. That's why a number of curves can be fitted from the same original data.

12.4.2 Semi-Average Method

As compared with graphic method, this method has more objective approach. In this method, the whole time series data is classified into two equal parts. For example, if we are given the time series values for 10 years from 1975 to 1984 then the two equal parts will he the data corresponding to periods 1975 to 1979 and 1980 to 1984. However, in case of odd number of years, the two equal parts are obtained by omitting the value for the middle period. Thus, for example, for the data for 9 years from 1980 to 1988, the two parts will be the data for years 1980 to 1983 and 1985 to 1988, the value for the middle year, viz. 1984 being omitted. Having divided the given series into two equal parts, we next compute the arithmetic mean of time-series values for each

half separately. These means are called semi-averages. Then these semi-averages are plotted as points against the middle point of the respective time periods covered by each part. The line joining these points gives the straight line trend fitting the given data

The procedure of this method can be illustrated by the following example:

Case 1: When the number of years in a series is even

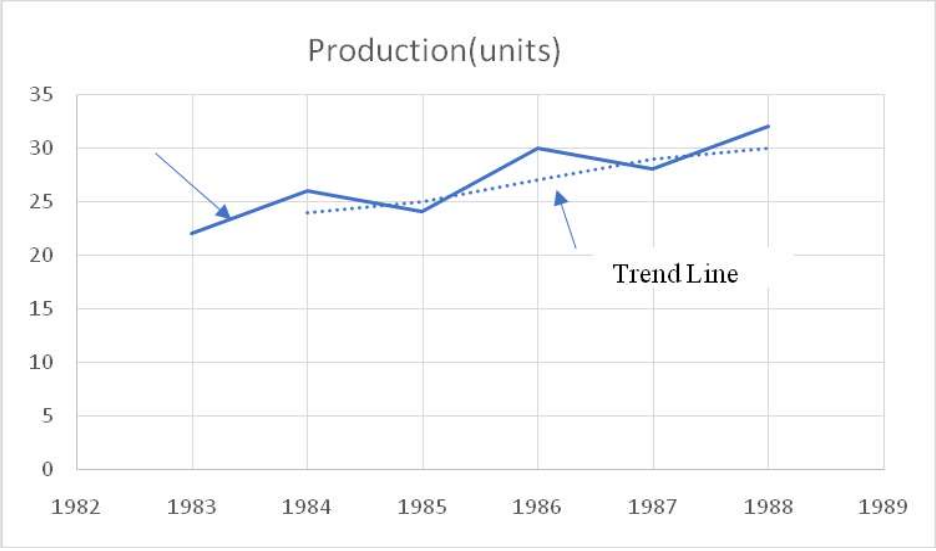
Example 2: Fit a trend line by the method of semi-average to the data given below:

Year	1983	1984	1985	1986	1987	1988
Production (units)	22	26	24	30	28	32

Solution:

Year (X_1)	**Production (units) (X_2)**	**Semi Total**	**Semi-Average**	**Middle Year**
1983	22	72	$\overline{X_1}$= 72÷3=24	1984
1984	26			
1985	24			
1986	30	90	$\overline{X_2}$ = 90÷3=30	1987
1987	28			
1988	32			

Figure 12.3



Case 2: When the number of years in a series is Odd

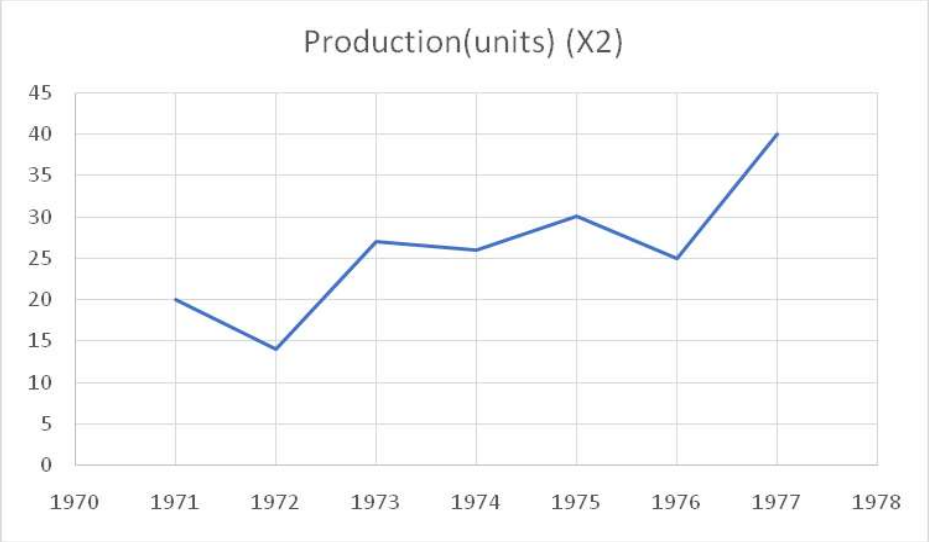
Example 3: Fit a trend line by the method of semi-average to the data given below:

Year	1971	1972	1973	1974	1975	1976	1977
Profit ('000 Rs)	20	14	27	26	30	25	40

Solution:

Year (X_1)	**Production (units) (X_2)**	**Semi Total**	**Semi-Average**	**Middle Year**
1971	20	61	$\overline{X_1}$ = 61÷3=20.33	1972
1972	14			
1973	27			
1974	26	95	$\overline{X_2}$ = 95÷3=31.67	1976
1975	30			
1976	25			
1977	40			

Figure 12.4



The estimated profit for the year 1978 is Rs 35000

Merits and Demerits of Semi-Average Method:

- **Merits:**
  - It's an easy method
  - It's free from any bias
  - Trend values obtained are finite
  - An obvious advantage of this method is its objectivity

in the sense that it does not depend on personal judgement

- **Demerits:**
  - This method is affected by extreme values
  - This method ignores the effect of cyclical fluctuations
  - This method is based on straight line trend assumption which doesn't always holds true.

Exercise:

1: Fit a trend line by method of semi average to the data give below:

Year	1991	1992	1993	1994	1995	1996	1997
Production ('000tonnes)	28	27	24	38	40	30	46

2: Fit a trend line to the following data by semi-average method:

Year	1961	1962	1963	1964	1965	1966
Profits ('000Rs)	80	82	85	70	89	95

12.4.3 Method of Moving Averages

Method of moving averages is a very simple and flexible mood of measuring trend. Under this method, moving averages are calculated. In moving average computations one has to decide which year moving average: 3 year, 4 year, 5 year etc. The period of moving average depends upon the periodicity of data and there is no specific rule for that. The period is determined by plotting the data on the graph paper and noticing the average time interval of successive peaks or troughs. However, it is important to consider while selecting the period of moving average that after how many years most of the fluctuations occur in the data. This method is studied in two different situations:

- Odd Period Moving Average
- Even Period Moving Average
- **Odd Period Moving Average:** When period of the moving average is odd then the following steps are to be taken for the estimation of moving average:

- First of all add up the values corresponding to first 3 years in the time series and put the sum before the middle year (i.e. 2nd year)
- Thereafter, leaving the first year value, add up second, thirst and fourth year values and put the sum in front of middle year (i.e. 3rd year). We need to follow this process till we reach the last value.
- Moving totals thus obtained are to be divided by the period of the moving average and show the trend values of different years.

* Similarly 5 yearly, 7 yearly moving averages can be obtained.

The procedure of this method can be illustrated by the following example:

Example 4: From the following data, calculate the trend values using 3 yearly moving average:

Year	1983	1984	1985	1986	1987	1988	1989
Production	412	438	446	454	470	483	490

Solution:

Year	**Production**	**Three-yearly Moving Totals**	**Three Yearly Moving Average (Trend Values)**
1983	412	-	-
1984	438	→ 412+438+446=1296	1296÷3=432
1985	446	→ 438+446+454= 1338	1338÷3=446
1986	454	→ 446+453+470= 1370	1370÷3=457
1987	470	→ 454+470+483= 1407	1407÷3= 469
1988	483	→ 470+483+490=1443	1443÷3=481
1989	490		

- **Even Period Moving Average:** When moving average period is even, say 4 years, then moving averages have to be centered. The computation procedure of 4 yearly moving average is as follows:
  - First of all, add up first 4 values corresponding to the first 4 years and put the sum in between second and

third year. Thereafter the next total (i.e. from2nd to 5th year total) is to be put in between 3rd and 4th year. We need to follow this process till the last value of the series.

- Now add up the 1st and 2nd, 4 year totals and put them in front of 3rdyear. Similarly add up the 2nd and 3rd, 4 year total and put them in front of the 4th year. We need to follow this process till the last value.
- 8 years' total thus obtained are to be divided by 8. These values are 4 yearly moving averages and show the trend values for different years.

The computation procedure is made clear by the following example:

Example 5: From the following data, calculate the trend values using 4 yearly moving average:

Year	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989
Sale in Crores	7	8	9	11	10	12	8	6	5	10

Solution:

Year	**Sales (in Crores) (2)**	**Four-yearly Moving Totals (3)**	**2 period moving total of 4 yearly moving totals (4)**	**Four Yearly Moving Average (Trend Values) (5)=(4)÷8**
1980	7	-	-	-
1981	8	-	-	-
		→ 7+8+9+11=35		
1982	9	-	→ 35+38=75	75÷8=9.125
		→ 8+9+11+10=38		
1983	11		→ 38+42=80	80÷8=10.00
		→ 9+11+10+12=42		
1984	10		→ 42+41=83	83÷8=10.375
		→ 11+10+12+8=41		
1985	12		→ 41+36=77	77÷8=9.625
		→ 10+12+8+6=36		

1986	8		→ 36+31=67	67÷8=8.375
		→ 12+8+6+5=31		
1987	6		→ 31+29=60	60÷8=7.1
		→ 8+6+5+10=29		
1988	5	-	-	-
1989	10	-	-	-

- **Period of Moving Average:** Sometimes the period of moving average is not given, in that case one has to determine the best suitable period of the moving average. The duration of the cycle can be found out by plotting the original data on a graph paper and reading the time distances between various peaks. The average of these time distances would give the average duration of the cycle and this should be taken as the period of moving average.

The following example would make the things more clear.

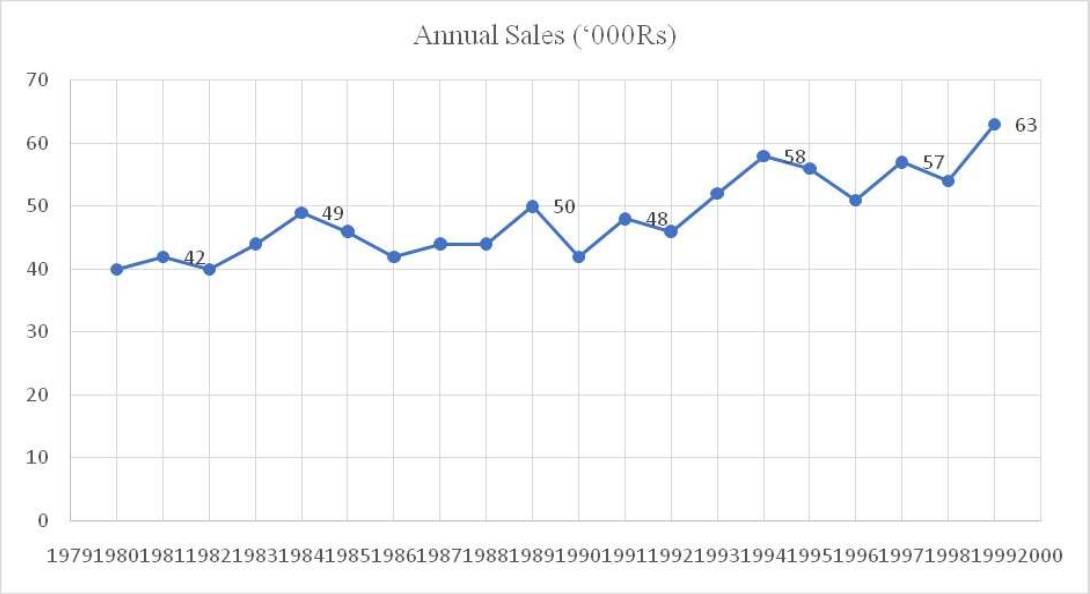
Example 6: Determinethe period of moving average, find trend values by moving average for the following data:

Year	Annual Sales ('000Rs)	Year	Annual Sales ('000Rs)
1980	40	1990	42
1981	42	1991	48
1982	40	1992	46
1983	44	1993	52
1984	49	1994	58
1985	46	1995	56
1986	42	1996	51
1987	44	1997	57
1988	44	1998	54
1989	50	1999	63

Solution:

As discussed, we know that the appropriate period of moving average is the period of the cyclic variation. First of all we need to plot the graph and see whether the data reveal a regular cycle of any fixed period.

Figure 12.5



The graph doesn't reveal a regular cycle of any fixed period. Thus, to determine the appropriate period of moving average, we will examine the graph carefully and we have the peaks at the following points:

Year	1981		1984		1989		1991		1994		1997		1999
Peak Value	42		49		50		48		58		57		63
Period		3		5		2		3		3		2	

Thus, the data under estimation shows 6 cycles with varying periods viz. 3, 5, 2, 3, 3, 2 respectively. The appropriate period of the moving average is given by the arithmetic mean of period of different cycles. Hence the period of moving average is given by:

$$Average\ Cyclical\ period = \frac{3+5+2+3+3+2}{6} = \frac{18}{6} = 3$$

Thus, period of moving average is determined and now the same procedure would be followed to find the trend values by moving averages i.e. 3 yearly moving average in this case (Same as example 4).

- **Measurement of short term fluctuations:** Time series is

a mixture of both trend and short term fluctuations. We can find the short term fluctuations by deducting the trend component from the original data.

Example 7: From the following data, calculate the trend values using 3 yearly moving average:

Year	1983	1984	1985	1986	1987	1988	1989
Production	412	438	446	454	470	483	490

Solution:

Year	**Production (Y)**	**Three-yearly Moving Totals**	**Three Yearly Moving Average (Trend Values=T)**	**Short Term Fluctuations (Y-T)**
1983	412	-	-	-
1984	438	412+438+446=1296	1296 3=432	6
1985	446	438+446+454= 1338	1338 3=446	0
1986	454	446+453+470= 1370	1370 3=457	-3
1987	470	454+470+483= 1407	1407 3= 469	1
1988	483	470+483+490=1443	1443 3=481	-2
1989	490	-	-	-

Thus, deducting trend values (T) from the original series (Y), gives the short term fluctuations.

Merits and Demerits of Moving Average Method

- **Merits:**
  - This method is easy to understand and simple to use.
  - The method is most suitable for cyclical fluctuations.
  - This method has more practical usefulness
- **Demerits:**
  - The limitations of arithmetic mean affect this method adversely.
  - If the periodicity in the series is not clearly visible, then this method should not be used.
  - If the proper period is not ascertained then the results will be misleading and inaccurate.

12.4.4 Least square method

This is the best method of trend fitting in a time series data and is most used in practice. This is a mathematical method and a trend line in this method is fitted or obtained in such a way that following two conditions are fulfilled:

- $\sum(Y - Y_c) = 0$ meaning thereby that the sum of the deviations of the actual values of Y and computed trend values (Y_c) is zero
- $\sum(Y - Y_c)^2$ is the minimum i.e. the sum of the squares of the deviations of the actual and computed trend values from this line is least

Least square method can be used to fit straight line trend or parabolic trend or exponential trend. Trend line fitted under this method is also called as Line of Best Fit.

- **Fitting the straight line trend**

  A straight line trend can be expressed by the following equation:

$$Y = a + bX$$

$$Where, Y = Trend\ Values, X = Unit\ of\ Time,$$

$$a = intercept\ and\ b\ is\ the\ slope\ of\ the\ line$$

For determining two constant , we need to generate the two normal equations and these two normal equations will satisfy the two conditions mentioned above:

$$\sum Y = Na + b\sum X \quad (1)$$

$$\sum XY = a\sum X + b\sum X^2 \quad (2)$$

The values of $a\ and\ b$ are determined by solving the above said two normal equations and finally the calculated values of $a\ and\ b$ are put in $Y = a + bX$ and trend values are computed.

Short- cut Method: If any of the deviations are taken from the middle year (or arithmetic mean of two middle years, the value of $\sum X$ would be zero i.e. $\sum X$=0 and the equations are reduced to the following:

$$\sum Y = Na$$
$$\sum XY = b\sum X^2$$

When further solved, these equations gives the following values of the constants a and b:

$$a = \frac{\sum Y}{N} \text{ and } b = \frac{\sum XY}{\sum X^2}$$

The procedure of this method can be illustrated by the following example:

Example 7: Fit a trend straight line by the method of Least Square (taking 1978 as year of origin) to the following data and also obtain the trend values:

Year	1979	1980	1981	1982	1983	1984
Production (lakh tonnes)	5	7	9	10	12	17

Solution:

Year	**Production (Y)**	**Deviations from 1978 (X)**	**XY**	**X^2**
1979	5	1	5	1
1980	7	2	14	4
1981	9	3	36	9
1982	10	4	40	16
1983	12	5	60	25
1984	17	6	102	36
N=6	$\sum Y$=60	$\sum X$=21	$\sum XY$=248	$\sum X^2$= 91

The straight line trend is defined by the equation

$Y = a + bX$

Two Normal Equations are:

$$\sum Y = Na + b\sum X$$
$$\sum XY = a\sum X + b\sum X^2$$

Substituting the values we get,

$60 = 6a + 21b \qquad (i)$

$248 = 21a + 91b \qquad (ii)$

Solve the two equations $(i) and\ (ii)$

Multiply (i) by 7 and (ii) by 2 and then subtracting, we get

$b = 2.17$

By substituting the value of b in equation (i), we get

$a = 2.40$

Hence the trend equation is: $Y = 2.4 + 2.17X$; origin 1978, X unit=1

Computation of Trend values:

$$For\ 1979, X = 1, Y = 2.4 + 2.17(1) = 4.57$$
$$For\ 1980, X = 2, Y = 2.4 + 2.17(2) = 6.74$$
$$For\ 1981, X = 3, Y = 2.4 + 2.17(3) = 8.91$$
$$For\ 1982, X = 4, Y = 2.4 + 2.17(4) = 11.08$$
$$For\ 1983, X = 5, Y = 2.4 + 2.17(5) = 13.25$$

(b) Fitting of Second Degree Parabolic Trend or Quadratic Trend:

Many a times in economics and business fields, straight line trend may not represent the long term tendency of the time series data. In such cases a second degree parabolic trend or quadratic trend is fitted. The equation of the second degree parabolic trend or the quadratic trend is:

$$Y = a + bX + cX^2$$

Where a is the Y intercept, b is the slope of the curve at the origin and c is the rate of change in slope

Under the method of Least Squares, the values of the constants, $a, b\ and\ c$ are obtained by solving the following three Normal Equations:

$$\sum Y = Na + b\sum X + cX^2$$
$$\sum XY = a\sum X + b\sum X^2 + c\sum X^3$$
$$\sum X^2 Y = a\sum X^2 + b\sum X^3 + c\sum X^4$$

Short- cut Method: If any of the deviations are taken from the middle year (or arithmetic mean of two middle years, the value of $\sum X\ and\ \sum X^3$ would be zero i.e. $\sum X = 0$ and the equations are reduced to the following:

$$\sum Y = Na + cX^2$$
$$\sum XY = b\sum X^2$$
$$\sum X^2 Y = a\sum X^2 + c\sum X^4$$

When further solved, these equations gives the following values of the constants a and b:

$$a = \frac{\sum Y - c\sum X^2}{N};\ b = \frac{\sum XY}{\sum X^2} \text{ and } c = \frac{N\sum X^2 Y - (\sum X^2)(\sum Y)}{N\sum X^4 - (\sum X^2)^2}$$

After finding the values of a, b and c, the trend equation can be fitted to obtain the trend values of the given time series by substituting the respective values of X therein.

(c) Fitting of an Exponential Trend : If the time series is increasing or decreasing by a constant percentage rather than constant absolute amount, the fitting of exponential trend is considered appropriate. Such tendency is found in many economic and business events.

The equation of the exponential trend is

$$Y = ab^x$$

Where a is Y intercept and b the slope of the curve at the origin of X

In the logarithmic form, the above equation is written as under:

$$logY = loga + Xlogb$$

When plotted on a semi-logarithmic graph, the curve gives a straight line (or called logarithmic straight line), However, on an arithmetic scale chart, the curve give a non-linear trend.

The two Normal equations in this case would be as follows:

$$\log\sum Y = N.loga + logb\ \sum X$$

$$\sum(XlogY) = loga\sum X + logb\sum X^2$$

Short-cut Method: If any of the deviations are taken from the middle year (or arithmetic mean of two middle years, the value of $\sum X$ would be zero i.e. $\sum X = 0$ and the equations are reduced to the following:

$$\log\sum Y = N.loga \qquad (i)$$

$$\sum(XlogY) = logb\sum X^2 \qquad (ii)$$

From (i), we get

$$loga = \frac{\sum logY}{N}$$

And to find the value of constant a, we need to take an antilog, thus

$$a = AL\left[\frac{\sum logY}{N}\right]$$

From(ii), we get

$$logb = \frac{\sum X logY}{\sum X^2}$$

And to find the value of constant b, we need to take an antilog, thus

$$b = AL\left[\frac{\sum logY}{\sum X^2}\right]$$

After obtaining the values of $a, and\ b,$ the trend equation can be fitted to obtain the trend values of the given time series by substituting the respective values of X therein.



12.5 LET US SUM UP

- A time series is an arrangement of statistical data in accordance with the time of occurrence of an event
- The value (y) of a phenomenon observed at any time is the net effect of the interaction of above components such as secular trend, seasonal variations, cyclical variations and random or irregular variations
- The different methods for estimation of trend are free hand curve method, semi-average method, moving average method and least square method
- There are various areas where we can apply time series analysis for measurement such as to study the past behaviour of data for formulating the policies for future. anticipation of changes going to occur in the future, understanding the cause and effect relationship etc.



12.6 FURTHER READING

1) Freund J E and Ronald E. Walpole (1987), *Mathematical Statistics 4e*: Prentice Hall Englewood Cliffs, NJ

2) Goon, A.M, M.K. Gupta and B, Das Gupta (1987): *Basic Statistics*, The World Press Pvt. Ltd., Calcutta

3) Mantzopoulos Victoria L (1995), *Statistics for the Social Sciences*: Prentice Hall, Englewood Cliffs, NJ.

4) Fleming Michael C. and Joseph G. Nellis (1996), *The Essence of Statistics for Business 2e* :Prentice Hall, Hemel Hempstead

5) Gupta, S P (1978) ,Statistical Methods, Sultan Chand and Sons, Delhi.

6) Gupta S C (1981), Fundamentals of Statistics: Himalaya Publishing House



12.7 ANSWERS TO CHECK YOUR PROGRESS

Ans to Q No 1: The time series can be defined as arrangement of statistical data in accordance with the time of occurrence.

Ans to Q No 2: The four methods which are generally used for the study and estimation of trend component in a time series such as Free Hand Curve Method, Semi-Average Method, Moving Average Method, and Least Square Method.

Ans to Q No 3: The different components of time series are Trend (T), seasonal variations (S), cyclical variations (C), and irregular variations (I).



12.8 MODEL QUESTIONS

Short Questions (Answer each question in about 50 to 100 words)

Q 1: Define time series

Q 2: What is semi-average method?

Q 3: Define secular trend

Q 4: Write two merits of semi-average method

Q 5: What is meant by random variation in time series?

Q 6: What does cyclical variation mean in time series?

Q 7: Write two demerits of moving average method.

Long Questions: (Answer each question in about 300 words)

Q 1: Explain the different components of time series.

Q 2: Discuss the use and significance of time series.

Q 3: Explain different mathematical model for decomposition of time series.

Q 4: Discuss different methods for estimating trend a time series.

Q 5: Explain the merits and demerits of different methods of estimating trend of a time series.

Q 6: Write a short note on mathematical model for time series.

Q 7: Fit a trend line to the following data by the free hand curve method:

Year	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000
Production (in **Qts**)	164	177	235	214	181	162	181	186	211	236

Q 8: Fit a trend line to the following data by the free hand curve method:

Year	1961	1962	1963	1964	1965	1966	1967	1968	1969	1970
Sales ('000Rs)	22	27	25	31	19	27	21	33	17	20

Q 9: Calculate 4 yearly moving average for the following data:

Year	1980	1981	1982	1983	1984	1985	1986	1987
Production ('000tonnes)	13	18	15	21	23	20	24	27

Q 10: Determine the period of moving average for the following data and calculate moving average for that period:

Year	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Values	130	127	124	135	140	132	129	127	145	158	153	146	145	164	170

Q 11: Fit the straight line trend by the method of Least Square to the following data:

Year	1991	1992	1993	1994	1995	1996	1997
Production('000tonnes)	80	90	92	83	94	99	92

Q 12: Fit the straight line trend by the method of Least Square to the following data:

Year	1971	1972	1973	1974	1975	
Values	45	56	78	46	75	

*** ***** ***

UNIT 13 : STOCHASTIC TIME SERIES

UNIT STRUCTURE

13.1 Learning Objectives

13.2 Introduction

13.3 Idea of Stochastic Time Series

13.4 Stationary Time Series

13.5 Non-stationary Time Series

13.6 Integrated Stochastic Process

13.7 Tests of Stationarity

13.8 Lets Sum Up

13.9 Further Readings

13.10 Answers to Check Your Progress

13.11 Model Questions

13.1 LEARNING OBJECTIVES

After going through this unit, you will be able to-

- understand about stochastic processes
- know various types of stochastic processes
- differentiate between stationary and non-stationary time series
- check stationarity of a time series.

13.2 INTRODUCTION

Analysis of economic behaviour is highly dependent upon the type of data used. We have already learned about the basics of time series analysis in the previous chapters. In this we will be discussing particularly about the stochastic time series process and its different types, why it is important in regression analysis? And the various tests to identify and differentiate within a stochastic series (stationary and non-stationary). The rationality behind placing stochastic process in a higher altar of time series is because unlike a deterministic time series, stochastic time series is a bit complicated to handle and it finds its place in most economic

analysis. So we start our learning process by having a brief idea of the stochastic process.

13.3 IDEA OF STOCHASTIC TIME SERIES

When it comes to econometrics, the term *stochastic* can be considered as an academic jargon for randomness. The word *stochastic* comes from the greek word *stokhos* which means to *aim at*. When we aim at a particular object there is a chance of either hitting or missing the target i.e. there is random event of probability attached to it. Therefore when we mention stochastic series we consider the presence of random variables that make up the event. In econometrics these events are in different space and time thus represented by different random variables and time periods. Therefore, a stochastic time process can be defined as a collection of random variables ordered in time. Let us understand this with the help of a simple example. Suppose Y(t) or Y_t is the number of customers in a restaurant at time t. As the time passes new customers will enter the restaurant and customers who have finished eating will leave the restaurant so the value of Y(t) or Y_t changes. But for any given time t, Y(t) or Y_t can take exactly one value of 0,1,2,...; and t can range from $-\infty$ to $+\infty$ which means from infinite past to infinite future.

A stochastic time series can either be continuous or discrete. For a random variable Y, we denote it as Y(t) if it is continuous or Y_t if it is discrete. Real world example of a continuous stochastic time series is the readings from an electrocardiogram. Whereas data sets like that of GDP, investments, expenditures are discrete in nature. From the above example itself we can understand continuous and discrete time series. For a continuous stochastic time series we observe the number of customers continuously at every smallest change in time, and find that as a customer enters and leaves the restaurant the value of Y(t) increases and decreases by one respectively. Here it is important to understand that it is not necessary that customers continuously enter or leave the restaurant, rather a customer can enter or leave the restaurant at any time they want. Which means that Y(t) should not necessarily change at every instant of

time rather Y(t) is allowed to change at any time (smallest continuous unit of time possible) and not just at t = 0,1,2.... Thus, in a continuous stochastic process Y(t) : t ∈ T, has a random number Y(t) recorded at every instant in time. For a discrete stochastic time series we observe the restaurant at unit intervals, say once every hour. Within this period of an hour more than one customer can either enter or leave the restaurant. This leads to greater fluctuations in the values of Y_t. Thus, in a discrete stochastic time series Y_t : t ∈ T, has a random number Y_t recorded at T = 0,1,2....In most cases the economic data sets represents value at discrete points in time, such values are just a realization of the different possible values. The distinction between the stochastic process and its realization is similar to the distinction between population and sample in cross-sectional data. Just as we use sample data to draw inferences about a population, in time series we use the realization to draw inferences about the underlying stochastic process.



CHECK YOUR PROGRESS

Q 1: Define a stochastic process. (Answer in about 20 words)

.....

.....

Q 2: Why most economic data sets are discrete in nature? (Answer in about 40 words)

.....

.....

.....

.....

13.4 STATIONARY TIME SERIES

When we discuss the stochastic process, it is also important to discuss about whether it is stationary or non-stationary. But why stationarity or non-stationarity of a stochastic process matters? To answer this question, we have to first try to understand; what is stationary and non-stationary

stochastic process?

A stationary time series process is one whose probability distributions are stable over time in a sense that if we take any collection of random variables in the sequence and then shift that sequence ahead 'k' time periods, the joint probability distribution must remain unchanged. Which means a random or stochastic process $\{x_t : t= 1, 2, \ldots\}$ is stationary if for every collection of time indices $1 \leq t_1 < t_2 < \ldots < t_m$, the joint distribution of $(x_{t1}, x_{t2}, ..., x_{tm})$ is the same as the joint distribution of $(x_{t1+k}, x_{t2+k}...x_{tm+k})$ for all integers $k \geq 1$. In simple terms, a stationary stochastic process can be defined as one whose realization statistics are the same for any value of time, i.e. there is no systematic change in mean, there is no systematic change in variance, and it contains no strictly periodic variations thus implying that it is time invariant. Such way of describing a stationary process is known as strict stationarity.

Another way of describing a stationary stochastic process is by considering *its mean and variance are constant over time and the value of the covariance between the two time periods depends only on the distance or gap or lag between the two time periods and not the actual time at which the covariance is computed.* Such a stochastic process is known as a weakly stationary, or covariance stationary, or second-order stationary, or wide sense, stochastic process. Weak stationarity can be explained with the help of the following properties for any stochastic time series Y_t:

Mean: $E(Y_t) = \mu$; (13.4.1)

Variance: $var\ (Y_t) = E(Y_t - \mu)^2 = \sigma^2$; (13.4.2)

Covariance: $\gamma_k = E[(Y_t - \mu)(Y_{t+k} - \mu)]$; (13.4.3)

here γ_k, the covariance (or autocovariance) at lag k, is the covariance between the values of Y_t and Y_{t+k}, that is, between two Y values k periods apart. If k = 0, we obtain γ_0, which is simply the variance of Y (= σ^2); if k = 1, γ_1 is the covariance between two adjacent values of Y.

Now if any of these properties is absent in a stochastic time series then it is called a non-stationary stochastic time series. In other words, a non-stationary time series will have a time-varying mean or a time-varying variance or both.

Now usually non-stationary stochastic series is norm in the real world.

Then the question arises why do we even have to consider the stationarity of the series? This brings us to the initial question that we asked at the start of this section that why stationarity or non-stationarity of a stochastic process matters? Well we can understand and describe a non-stationary series but when it comes to forecasting or predicting non-stationary series is of little use. Because if a time series is non-stationary, we can study its behavior only for the time period under consideration as it changes with time. Each set of time series data will therefore be for a particular episode. But a certain level of prediction and forecasting is required for our daily life decisions. Hence there is a need for stationary processes. Still, strict stationarity is only an analysts' sweet dream therefore to bridge the gap between real world phenomena and a hypothetical assumptions, weak stationarity is used. And for use of weak stationarity, we have to understand what a non-stationary time series looks like and what are the different variations of it. Therefore we further discuss this topic in the next section.



CHECK YOUR PROGRESS

Q 3: Define weak stationarity? (Answer in about 30 words)

.....

.....

.....

Q 4: What are the different names of weak stationarity? (Answer in about 20 words)

.....

.....

Q 5: Does all strictly stationary process also follow a weakly stationary process? (Answer in about 20 words)

.....

.....

13.5 NON-STATIONARY TIME SERIES

Non-stationary series as mentioned in the previous section is

time variant, which means the variables in the series changes with time. One of most popular example of non-stationary time series is the random walk model (RWM). A random walk is often compared with a drunkard's walk. After leaving a bar, as the drunkard moves a random distance u_t at time t, and, continuing to walk indefinitely, will eventually drift farther and farther away from the bar. Now exactly predict the drunkard's next step is quite difficult. As the direction and distance of each step may or may not vary. But we know one thing for sure that the each step of the drunkard depends to a certain extent upon his last step. When we compare this to economic data sets such as stock prices we tend to observe a similar random pattern over time where each value depends upon its past value but there is also a certain level of randomness attached to it. Thus such data series are said to follow a random walk model. There are two types of random walk models: (1) simple random walk or random walk without drift and (2) random walk with drift. The random walk model with or without drift indicates the presence or absence of a constant term in the model respectively.

- **Random walk without drift:** A series Yt is said to be a random walk if

$$Y_t = Y_{t-1} + u_t \qquad (13.5.1)$$

In this model we can see that the value of Y at time t is equal to the value of Y in time (t-1) plus a random shock (u_t). Here, u_t is considered to be a white noise term. White noise refers to a special type of stochastic process considered to be purely random as it has zero mean, constant variance, and is serially uncorrelated.

Now, from (13.5.1) we can also write

$$Y_1 = Y_0 + u_1 \qquad (13.5.2)$$

$$Y_2 = Y_1 + u_2 = Y_0 + u_1 + u_2 \qquad (13.5.3)$$

$$Y_3 = Y_2 + u_3 = Y_1 + u_2 + u_3 = Y_0 + u_1 + u_2 + u_3 \qquad (13.5.4)$$

From the above equations, we get the idea that if the process started at some time 0 with a value of Y_0, then we have

$$Y_t = Y_0 + \text{“ } u_t \qquad (13.5.5)$$

Therefore, $E(Y_t) = E(Y_0 + \sum u_t) = Y_0$

(Since, $E(u_t) = 0$)

Likewise, we can show that var (Y_t) = tó ^{2}

From the above expressions we find that for the given series the mean of Y is equal to the original value of Y at the start of the series, which is a constant but as time progresses the variance will increase with time meaning it is time variant and thus non-stationary. It is important to note here that in a random walk model the random shocks (u_t) are persistent. It implies that the impact of a shock in any period does not stop affecting the Y. Suppose for a time period say t_7 if the value of u_7 = 10 and $\neq$ 0. Now for all Y values after t_7 will be 10 units higher until any opposite shock occurs.

- **Random walk with drift:** For a random walk with drift model we just modify the model of random walk without drift (13.5.1) into the following:

$$Y_t = \delta + Y_{t"1} + u_t \qquad (13.5.6)$$

Here, δ is a constant and also known as drift parameter. It leads to Yt drifting upward or downward, depending on δ being positive or negative. Now just like random walk without drift, we find the mean and variance for the random walk with drift model,

$$E(Y_t) = Y_0 + t \cdot \delta \qquad (13.5.7)$$

$$\text{var } (Y_t) = t\sigma^2$$

As we can see, for random walk with drift the mean as well as the variance increases over time, again violating the conditions of (weak) stationarity. In short, RWM, with or without drift, is a non-stationary stochastic process.

- **Unit Root Stochastic Process:** Unit root stochastic process is another example of non-stationary stochastic time series. It is actually quite similar to random walk model. The model can be stated as:

$$Y_t = \rho Y_{t-1} + u_t \quad -1 \leq \rho \leq 1 \qquad (13.5.8)$$

In this process if ρ=1, then it becomes a random walk model without drift as shown in (13.5.1). The process is called unit root because ρ=1 and this leads to non-stationarity as the variance of the model depends on time and is not stationary. Further, if $|\rho| < 1$, then we can show that the process is a stationary process.



CHECK YOUR PROGRESS

Q 6: What are the different random walk models? (Answer in about 20 words)

.....

.....

13.6 INTEGRATED STOCHASTIC PROCESSES

We know that most economic data are non-stationary stochastic time series and they cannot be used for forecasting whereas a stationary series can be used for forecasting. Therefore in most cases, non-stationary series are converted to a stationary series so that we are able to predict future possibilities. Suppose in a random walk model from (13.5.1)

$$Y_t = Y_{t-1} + u_t \qquad (13.6.1)$$

we can write this as $(Y_t - Y_{t-1}) = u_t = \Delta Y_t$ (13.6.2)

here ΔY_t is the first difference of the series Y_t as Δ is the first difference operator . It is easy to show that, while Y_t is nonstationary, but its first difference is stationary. In other words we can show that the first difference of a random walk process is stationary. Thus after such conversion of non-stationary series to its first difference, we can use it for forecasting as it becomes stationary. Such a stationary series obtained by taking first difference of a non-stationary series is known as stationary at first difference or time series integrated of order 1, denoted as I(1). Similarly, if a time series has to be differenced twice (i.e., take the first difference of the first differences) to make it stationary, we call such a time series integrated of order 2. In general, if a (nonstationary) time series has to be differenced n times to make it stationary, that time series is said to be integrated of order n. A time series Y_t integrated of order n is denoted as $Y_t \sim I(n)$. If a time series Y_t is stationary to begin with (i.e., it does not require any differencing), it is said to be integrated of order zero, denoted by $Y_t \sim I(0)$. Thus, we will use the terms “stationary time series” and “time series integrated of order zero” to mean the same thing. Most economic time series are generally I(1); that is, they generally become stationary only after taking their first

differences.

One of the main reasons to know whether a time series is stationary or non-stationary is because sometimes when non-stationary time series are used for regression analysis there is an obvious danger of obtaining a significant regression result even though the variables may not have anything in common or completely unrelated. Such a regression analysis is called a spurious regression. Let us understand this through the following illustration. If Y_t and X_t are two independent non-stationary series such that

$$Y_t = Y_{t-1} + u_t \qquad (13.6.3)$$

$$X_t = X_{t-1} + v_t \qquad (13.6.4)$$

where u_t and v_t are both independent white noise error terms, still after regression analysis impact of X_t on Y_t may show a significant result along with significant t values and high R^2 values. But as we have already mentioned that both these series have nothing in common and nor are they even casually related. It means that the apparent significance of the relationship obtained through regression is false. Thus the question about how we test whether a series is stationary or non-stationary becomes important. This question has been addressed in the next section.

13.7 TESTS OF STATIONARITY

In this chapter so far we have repeatedly hinted upon the importance of a stationary process in time series, so it is almost apparent to have a curiosity about how we would know if a time series is stationary or not. Therefore we discuss the tests for stationarity in this section. Though there are several popular tests for stationarity in literature here we discuss only graphical analysis, autocorrelation function and unit root test in brief.

- **Graphical analysis:** Graphical analysisis an initial intuitive method to identify if a series is stationary or not. If we plot a series over time, the trend of the series (upward or downward) can give us the intuition of its stationarity.

Figure 13.1

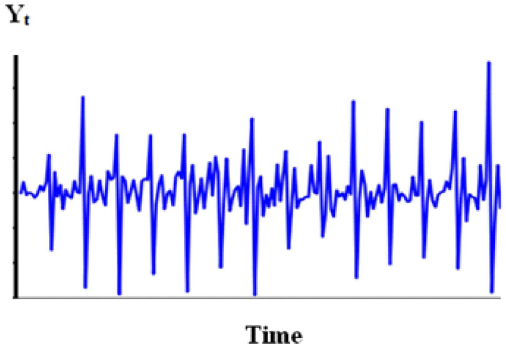
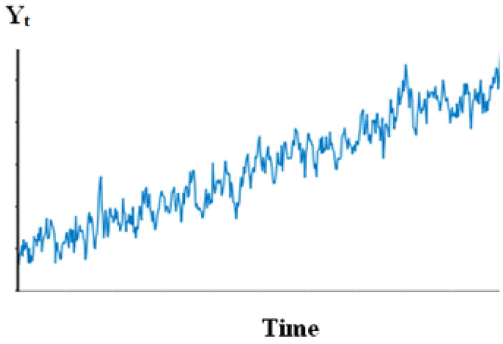


Figure 13.2



In the figure 13.1 and 13.2, we use graphical method to have a notion about the time series of Yt. We observe in figure 13.1 that as time increases Yt fluctuates but over the period remains almost the same, thus it is time invariant and therefore the series is stationary. In figure 13.2 as time increases Yt also increases, thus it is time variant and therefore the series in non-stationary. This is just a simple example of how we can use graphical method to deduce an idea regarding stationarity of a time series.

- **Autocorrelation function:** Autocorrelation function (ACF) is a simple test of stationarity and is also known as a correlogram test. For ACF at lag k, denoted by ρ_k, is defined as

$$\rho_k = \frac{\text{covariance at lag k}}{\text{variance}} = \frac{\gamma_k}{\gamma_0}$$

where covariance at lag k is γ_k, the covariance between the values of Y_t and Y_{t+k}, that is, between two Y values k periods apart. And the variance which is constant is γ_0. If k=0 then covariance becomes γ_0 which is also the variance and thus the value of ρ_k becomes 1. Since both covariance and variance are measured in the same units of measurement, ρ_k is a unitless, or pure number. It lies between "1 and +1, as any correlation coefficient does. If we plot ρ_k against k, the graph we obtain is known as the population correlogram. But as we know that we only have a realization (sample) of the stochastic process, what we compute is known as the sample autocorrelation function. To compute this, we must first compute the sample covariance at lag k, and the sample variance. We then plot the sample autocorrelation values against the lags k to obtain the

sample correlogram. Suppose in the figures below (13.3 and 13.4) we have shown a correlogram up to 30 lags.

Figure 13.3

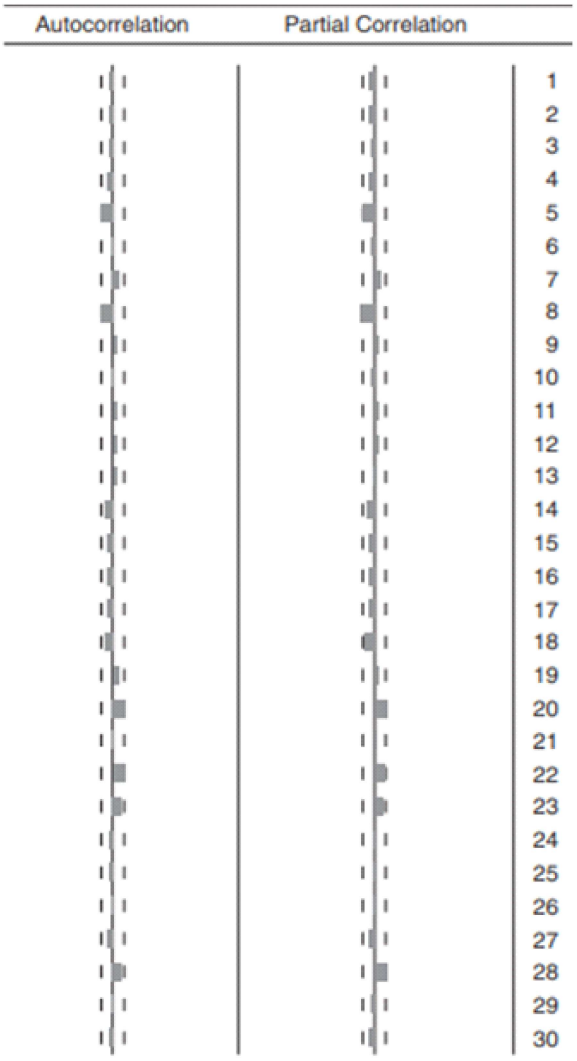
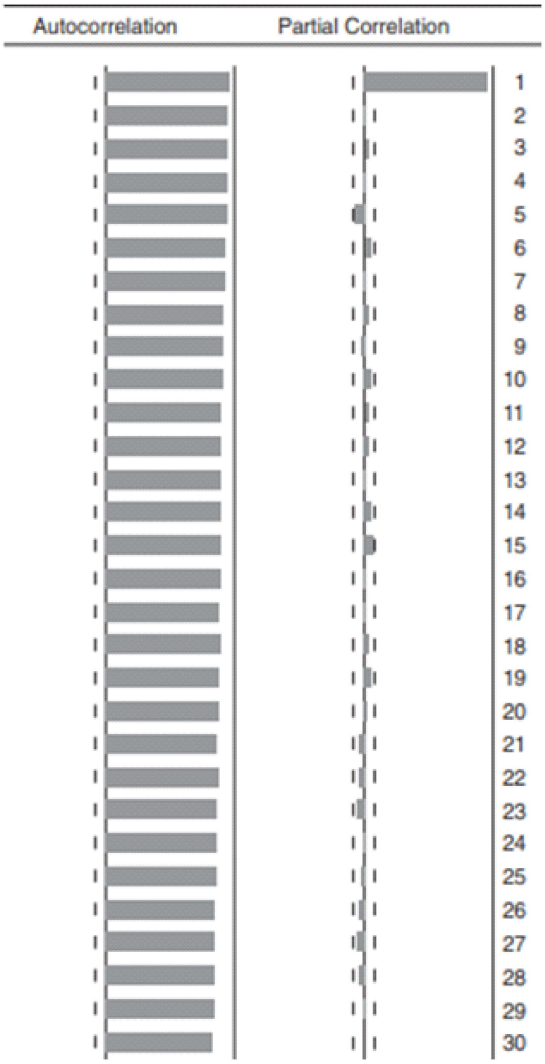


Figure 13.4



In both the figures the solid vertical line represents the zero axis; observations above the line are positive values and those below the line are negative values. In figure 13.3, which is a purely white noise process the autocorrelations at various lags hover around zero. This is the picture of a correlogram of a stationary time series. Thus, if the correlogram of an actual (economic) time series resembles the correlogram of a white noise time series, we can say that time series is probably stationary.

In the next figure 13.4 a correlogram of a random walk series is

shown. The most striking feature of this correlogram is that the autocorrelation coefficients at various lags are very high even up to a lag of 30. Figure 13.4 is the typical correlogram of a nonstationary time series: The autocorrelation coefficient starts at a very high value and declines very slowly toward zero as the lag lengthens.

Here, an interesting question is that how much lag length we should consider. A general rule of thumb is that we need to compute autocorrelation function up to one-third or one-fourth of the length of the time series. That is, if a time series say has 100 time periods than taking a lag length between 25-34 is sufficient. It is also advisable to take large lags and than reduce it by some statistical criterion like Akaike or Schwarz Information Criterion.

- **Unit root test:** We start the unit root test from a unit root process, which is a variation of random walk model.

$$Y_t = \rho Y_{t-1} + u_t \qquad -1 \leq \rho \leq 1 \qquad \text{(from 13.5.8)}$$

Now as mentioned earlier, if $\rho = 1$, that is, in the case of the unit root, it becomes a random walk model without drift, which we know is a non-stationary stochastic process. Therefore, why not simply regress Y_t on its (one period) lagged value Y_{t-1} and find out if the estimated ρ is statistically equal to 1? If it is, then Y_t is non-stationary. This is the general idea behind the unit root test of stationarity. For theoretical reasons, we manipulate (13.5.8) as follows:

Subtract Y_{t-1} from both sides of (13.5.8) to obtain:

$$Y_t - Y_{t-1} = \rho Y_{t-1} - Y_{t-1} + u_t \qquad (13.7.1)$$

$$= (\rho - 1)Y_{t-1} + u_t$$

which can be alternatively written as:

$$\Delta Y_t = \delta Y_{t-1} + u_t \qquad (13.7.2)$$

where $\delta = (\rho - 1)$ and ", as usual, is the first-difference operator. In practice, therefore, instead of estimating (13.5.8), we estimate (13.7.2) and test the (null) hypothesis that $\delta = 0$. If $\delta = 0$, then $\rho = 1$, that is we have a unit root, meaning the time series under consideration is non-stationary. Before we proceed to estimate

(13.7.2), it may be noted that if $\delta = 0$, (13.7.2) will become $\Delta Y_t = \delta Y_{t-1} + u_t$ (13.7.2). Since u_t is a white noise error term, it is stationary, which means that the first differences of a random walk time series are stationary, a point we have already made before.



CHECK YOUR PROGRESS

Q 7: What is spurious regression? (Answer in about 30 words)

.....

.....

Q 8: When does a unit root process become a random walk model? (Answer in about 20 words)

.....

.....



13.8 LET US SUM UP

- Stochastic time series is a collection of random variables ordered in time.
- Stochastic series can be discrete or continuous.
- Stochastic series can also be stationary or non-stationary.
- There are two types of stationarity- strict stationarity and weak stationarity.
- Strict stationarity implies a series whose mean and variance are constant and the joint probability distribution remains the same.
- If the mean and variance of a series are constant over time and the value of the covariance between the two time periods depends only on the distance or gap or lag between the two time periods and not the actual time at which the covariance is computed then a series is known a weak stationary series.
- Non-stationary occurs when either the mean or variance or both are not constant.

- Random walk models are examples of non-stationary time series.
- Unit root is a special case of random walk models.
- There are different tests of stationarity like graphical analysis, autocorrelation function and unit root test.



13.9 FURTHER READINGS

1) Gujarati, D. N., Porter, D. C., & Gunasekar, S. (2012). *Basic Econometrics*. Tata McGraw-Hill Education.

2) Hill, R. C., Griffiths, W. E., & Lim, G. C. (2011). *Principles of Econometrics*. John Wiley & Sons.

3) Wooldridge, J. M. (2015). *Introductory econometrics: A modern approach*. Cengage learning.



13.10 ANSWERS TO CHECK YOUR PROGRESS

Ans to Q No 1: Stochastic process can be defined as a collection of random variables ordered in time.

Ans to Q No 2: Most economic data sets are discrete in nature because it is very difficult to measure a particular variable in a continuous time frame. Capturing data in a continuous time process will incur high cost, require efficient technology along with continuous labour and/or machinery. Therefore until and unless it is absolutely necessary, most data sets are in a discrete form.

Ans to Q No 3: A stochastic process is said to be weakly stationary if its mean and variance are constant over time and the value of the covariance between the two time periods depends only on the distance or gap or lag between the two time periods and not the actual time at which the covariance is computed.

Ans to Q No 4: Weak stationarity is also known as covariance stationary, or second-order stationary, or wide sense stochastic process.

Ans to Q No 5: Yes, all strictly stationary process will also follow a weakly stationary process as in strict stationary process the mean and variance are constant and since the joint distribution of two series are same therefore the covariance will also be the same.

Ans to Q No 6: There are two types of random walk models; a simple random walk or random walk without drift and random walk with drift. Other process like unit root can also follow a random walk model, but with certain conditions.

Ans to Q No 7: When unrelated and non-stationary time series is used for regression analysis and still yields significant results, then such a regression is called spurious regression.

Ans to Q No 8: A unit root process becomes a random walk model when the ρ values equals to 1 or when δ (which is ρ-1) value is equal to 0.



13.11 MODEL QUESTIONS

Short Questions (Answer each question in about 150 words)

Q 1: Differentiate between weak and strict stationarity.

Q 2: Suppose we regress the scores of football matches with agricultural output over three decades, and find a significant positive relationship between the two. What inference can be made out of this result?

Q 3: If ρ value is 0.08, mention whether the unit root process is stationary or not?

Q 4: What do you understand if a time series is integrated of order 1?

Q 5: If GDP of a country over the years increases but the rate of growth is constant, will the mean and variance of GDP and growth rate satisfy the conditions of weak stationarity?

Q 6: Differentiate between continuous and discrete stochastic process.

Q 7: Why the result of a spurious regression is not valid?

Long Questions (Answer each question in about 300-500 words)

Q 1: Write a brief note on the various tests of stationarity?

Q 2: Discuss the random walk models.

Q 3: Write a note on unit root test.

Q 4: Write a note on correlogram test.

*** *****

UNIT 14 : STRUCTURAL SIMULTANEOUS EQUATION MODELS

UNIT STRUCTURE

14.1 Learning Objectives

14.2 Introduction

14.3 Structural and Reduced Forms

14.4 Identification

14.5 Indirect Least Squares

14.6 Two Stage least Squares

14.7 Let's Sum Up

14.8 Further Reading

14.9 Answer to Check your Progress

14.10 Model Questions

14.1 LEARNING OBJECTIVES

After going through this unit, you will be able to-

- describe the structural simultaneous equation
- derive the meaning of simultaneity
- discuss indirect least square
- explain two stage least square

14.2 INTRODUCTION

A model is said to constitute a system of simultaneous equations if all of the relationships involved are needed for determining the value of at least one of the endogenous variables included in the model. In the model whose values are explained by the model and those that contribute to providing such an explanation, the former are called endogenous and the latter predetermined variables.

14.3 STRUCTURAL AND REDUCED FORMS

- **Specification and interpretation**
  - Two or more equations with one variable explained in one equation appears as an explanatory variable in another equation implying that the endogenously variables are simultaneously determined.
  - Variables are also of two types: Endogenous– those determined in the model and Exogenous– those determined outside the model.

  The system is complete if it meets the following condition:
  - Number of equations = Number of endogenous variables
  - The system is linear in parameters.
  - The system can be static or dynamic. In that case, the classification of the variables is as follows: endogenous variables, predetermined variables and lagged endogenous variables. Symbolically

$$EV = f(\ Other\ EV, PDV, SET)$$

$Where; EV = Endogeneous\ Variable\ (determined\ within\ the\ model)$

$PDV = Pre\ Determined\ Variable(\ determined\ outside\ the\ model)$

$SET = Stochastic\ Error\ Term$

- **Reduced Form Equations:**

$EV = f(\ PDV, SET)$

$Where\ PDV = Pre\ Determined\ Variable$

$SET = Stochastic\ Error\ Term$

A Reduced form Equation is one that expresses an endogenous variable solely in terms of the pre-determined variable and the stochastic error term. Basically, we get the Reduced Form Equations from Structural Equations:

Example 1: Consider the relationship:

$$Y_t = \alpha + \beta C_t + U_t \quad (1)$$

$$C_t = Y_t + I_t \quad (2)$$

Putting value of C_t in (1), we get

$$\Rightarrow Y_t = \alpha + \beta(Y_t + I_t) + U_t$$

$$(1-\beta)Y_t = \alpha + \beta I_t + U_t$$

$$Y_t = \frac{\alpha}{1-\beta} + \frac{\beta}{1-\beta} I_t + \frac{U_t}{1-\beta}$$

$$C_t = \frac{\alpha}{1-\beta} + \frac{1}{1-\beta} I_t + \frac{U_t}{1-\beta} \qquad (3)$$

Thus, (3) is the Reduced Form Equation of (2), as it doesn't include other endogenous variable.

- **Problems caused by Endogenous Variables:**

Example 2: Consider the relationship

$$Y_t = \alpha_0 + \alpha_1 X_{1t} + \alpha_2 X_{2t} + U_{1t} \qquad (A)$$

$$X_{2t} = \beta_0 + \beta_1 Y_t + U_{2t} \qquad (B)$$

If we see the Covariance between $X_{2t} and\ U_{1t}$ then

$$Cov(X_{2t}U_{1t}) = E[X_{2t} - E(X_{2t})][U_{1t} - E(U_{1t})]$$

$$Assume\ E(U_{1t}) = 0$$

Thus,

$$Cov(X_{2t}U_{1t}) = E[X_{2t} - E(X_{2t})][U_{1t}] \qquad (C)$$

Putting the value of Y_t in (B), we get

$$X_{2t} = \beta_0 + \beta_1[\alpha_0 + \alpha_1 X_{1t} + \alpha_2 X_{2t} + U_{1t}] + U_{2t}$$

$$= \beta_0 + \beta_1\alpha_0 + \beta_1\alpha_1 X_{1t} + \beta_1\alpha_2 X_{2t} + \beta_1 U_{1t} + U_{2t}$$

$$(1-\beta_1\alpha_2)X_{2t} = \beta_0 + \beta_1\alpha_0 + \beta_1\alpha_1 X_{1t} + \beta_1 U_{1t} + U_{2t}$$

$$X_{2t} = \frac{\beta_0 + \beta_1\alpha_0}{(1-\beta_1\alpha_2)} + \frac{\beta_1\alpha_1 X_{1t}}{(1-\beta_1\alpha_2)} + \frac{\beta_1 U_{1t} + U_{2t}}{(1-\beta_1\alpha_2)}$$

Now, by putting the value of X_{2t} in (C), we get

$$Cov(X_{2t}U_{1t}) = E\left[\frac{\beta_0 + \beta_1\alpha_0}{(1-\beta_1\alpha_2)}(U_{1t}) + \frac{\beta_1\alpha_1 X_{1t}}{(1-\beta_1\alpha_2)}(U_{1t}) + \frac{\beta_1 U_{1t} + U_{2t}}{(1-\beta_1\alpha_2)}(U_{1t}) - U_{1t}E(X_{2t})\right]$$

$$Cov(X_{2t}U_{1t}) = E\left[\frac{\beta_0 + \beta_1\alpha_0}{(1-\beta_1\alpha_2)}(U_{1t}) + \frac{\beta_1\alpha_1 X_{1t}}{(1-\beta_1\alpha_2)}(U_{1t}) + \frac{\beta_1 (U_{1t})^2 + U_{2t}(U_{1t})}{(1-\beta_1\alpha_2)} - U_{1t}E(X_{2t})\right]$$

$$Assume\ E(U_{1t}) = 0$$

$$Cov(X_{2t}U_{1t}) = \frac{\beta_1(\sigma_u)^2 + U_{2t}(U_{1t})}{(1-\beta_1\alpha_2)} \neq 0$$

In SEM the problem is the existence of endogenous variable on RHS as it is correlated. Thus, SEM bias problem is there because of presence of Endogenous variable and we obtain bias results. Here, we can't apply OLS to estimate a structural equation assuming it is identified. The reason here is, the exogenous variable in the equation are necessarily correlated with the error term of the equation and estimated coefficients will be inconsistent.

14.4 IDENTIFICATION PROBLEM

Identification refers to structural parameters, equation by equation. There exist two equivalent approaches to identify the parameters: by study the relation between variables; or by study of the linear combination of the equation of the system. The parameters of a structural equation is said to be identified, if and only if they can be obtained from the reduced form equation in a unique way. A SEM is said to be identified, if and only if the structural parameters of all the equations are identified. If structural parameters of a particular equation are identified, that means the equation is identified, not the whole model. This problem related with whether or not we can derive unique estimates of structural coefficients from reduced form equations.

Example 3: Consider the following set of 4 equations

I	$Q_t^D = \alpha_1 + \alpha_2 P_t + U_{1t}$	(i)
	$Q_t^S = \beta_1 + \beta_2 P_t + U_{2t}$	(ii)
II	$Q_t^D = \alpha_1 + \alpha_2 P_t + \alpha_3 Y_t + U_{1t}$	(iii)
	$Q_t^S = \beta_1 + \beta_2 P_t + U_{2t}$	(iv)
III	$Q_t^D = \alpha_1 + \alpha_2 P_t + \alpha_3 Y_t + U_{1t}$	(v)
	$Q_t^S = \beta_1 + \beta_2 P_t + \beta_3 T_t + U_{2t}$	(vi)
IV	$Q_t^D = \alpha_1 + \alpha_2 P_t + \alpha_3 Y_t + \alpha_4 W_t + U_{1t}$	(vii)
	$Q_t^S = \beta_1 + \beta_2 P_t + \beta_3 T_t + U_{2t}$	(viii)

The identification would be as follows: We need to see the number of excluded variables and total number of endogenous variables and then we need to compare this with the numbers of endogenous variables $(G) - 1$. Thus,

- If the number of excluded variables are less that the value of $G - 1$, then equation is Unidentified.
- If the number of excluded variables are equal to the value of, then equation is Just Identified.
- If the number of excluded variables are more that the value of, then equation is Over Identified.

This is generally known as Rank and Order condition.

Consider the above set of equations, the solution for its identification is as follows:

I	Equation (i) in Set I	Number of Endogenous Variable = 2
		Number of Variables Excluded = 0
		$0 < 1 = G - 1$ $\therefore Equation\ (i)\ is\ Unidentified$
	Equation (ii) in Set I	Number of Endogenous Variable = 2
		Number of Variables Excluded = 0
		$0 < 1 = G - 1$ $\therefore Equation\ (ii) is\ Unidentified$
II	Equation (iii) in Set II	Number of Endogenous Variable = 2
		Number of Variables Excluded = 0
		$0 < 1 = G - 1$ $\therefore Equation\ (iii)\ is\ Unidentified$
	Equation (iv) in Set II	Number of Endogenous Variable = 2
		Number of Variables Excluded = 1
		$1 = 1 = G - 1$ $\therefore Equation\ (iv)\ is\ Just\ Identified$
III	Equation (v) in Set III	Number of Endogenous Variable = 2
		Number of Variables Excluded = 1
		$1 = 1 = G - 1$ $\therefore Equation\ (iii)\ is\ Just\ Indentified$

	Equation (vi) in Set III	Number of Endogenous Variable = 2
		Number of Variables Excluded = 1
		$1 = 1 = G - 1$ $\therefore Equation\ (iii)\ is\ Just\ Identified$
IV	Equation (vii) in Set III	Number of Endogenous Variable = 2
		Number of Variables Excluded = 0
		$1 = 1 = G - 1$ $\therefore Equation\ (vii)\ is\ Just\ Identified$
	Equation (viii) in Set III	Number of Endogenous Variable = 2
		Number of Variables Excluded = 2
		$2 > 1 = G - 1$ $\therefore Equation\ (viii)\ is\ Over\ Identified$

- **Derivation of Rank and Order Condition:**

The system of Simultaneous Equations is as follows:

$$\beta_{11}y_{1t} + \beta_{12}y_{2t} + \dots\dots\dots\dots\dots\beta_{1G}y_{Gt} + \gamma_{11}x_{1t} + \gamma_{12}x_{1t}\dots\dots\dots\dots\gamma_{1k}x_{kt} = U_{1t}$$

$$\beta_{21}y_{1t} + \beta_{22}y_{2t} + \dots\dots\dots\dots\dots\beta_{2G}y_{Gt} + \gamma_{21}x_{1t} + \gamma_{22}x_{1t}\dots\dots\dots\dots\gamma_{2k}x_{kt} = U_{2t}$$

$$\vdots$$

$$\vdots$$

$$\vdots$$

$$\beta_{G1}y_{1t} + \beta_{G2}y_{2t} + \dots\dots\dots\dots\dots\beta_{GG}y_{Gt} + \gamma_{G1}x_{1t} + \gamma_{G2}x_{1t}\dots\dots\dots\dots\gamma_{Gk}x_{kt} = U_{Gt}$$

Let's have the matrix form of generalised equations:

$$BY + \tau X = U \qquad (4)$$

$$Where, B = \begin{bmatrix} \beta_{11} & \cdots & \beta_{1G} \\ \vdots & \ddots & \vdots \\ \beta_{G1} & \cdots & \beta_{GG} \end{bmatrix}; Y = \begin{bmatrix} y_{1t} \\ \vdots \\ y_{Gt} \end{bmatrix}; \tau = \begin{bmatrix} \gamma_{11} & \cdots & \gamma_{1k} \\ \vdots & \ddots & \vdots \\ \gamma_{G1} & \cdots & \gamma_{Gk} \end{bmatrix}; X = \begin{bmatrix} x_{1t} \\ \vdots \\ x_{kt} \end{bmatrix} and\ U = Y = \begin{bmatrix} U_{1t} \\ \vdots \\ U_{Gt} \end{bmatrix}$$

The matrix order is as follows: B_{G*G};Y_{G*1}; τ_{G*k}; X_{k*1} and U_{G*1}

Considering Equation (4),

$$BY + \tau X = U$$

$$BY = -\tau X + U$$

$$Y = -B^{-1}\tau X + B^{-1}U \qquad (5)$$

In Reduced Form, we can write it as,

$$Y = \pi X + V \qquad (6)$$

From (5) and (6), by comparing the coefficients of X

$$\Rightarrow -B^{-1}\tau = \pi$$

Multiply both sides with B,

$$\Rightarrow B\pi = -\tau$$

This is the relation used in determining the identification status of the given SEM equations. We shall now try to use this relation for deriving a general identification criterion for each structural equation.

Its first equation would be,

$$\beta_g \pi = -\tau_g$$

Where: $\beta_g = [\beta_{01} \quad \beta_{02} \quad \ldots \beta_{gG}]$

$$\tau_g = [\tau_{01} \quad \tau_{02} \quad \ldots \tau_{0k}]$$

Let G^{Δ} = *number of endogenous variables which appears in the g[th] equation (i.e., number of the nonzero elements in* β_g.

$G^{\Delta\Delta} = G - G^{\Delta}$

K^* = *number of predetermined variables which appear in the g[th] equation (i.e., number of the nonzero elements in* τ_g

Without a loss of generality, we assume that the elements of β_g and τ_g are arranged in such a way that the nonzero elements appear first, being followed by the zero elements. Then we can partition β_g and τ_g, as

$$\Rightarrow [\beta^{\Delta} \quad 0^{\Delta\Delta}] \begin{bmatrix} \pi^{\Delta *} & \pi^{\Delta **} \\ \pi^{\Delta\Delta *} & \pi^{\Delta\Delta **} \end{bmatrix} = -[\gamma^* \quad 0^{**}]$$

$$\Rightarrow \beta^{\Delta}\pi^{\Delta *} = -\gamma^* \qquad and \qquad \beta^{\Delta}\pi^{\Delta **} = -\,0^{**} \qquad (7)$$

Where the matrix order of the equations is as follows:

$$\beta^{\Delta} \rightarrow 1 * G^{\Delta};\ \pi^{\Delta *} \rightarrow G^{\Delta} * K^*;\ \gamma^* \rightarrow 1 * K^*$$

and

$$\beta^{\Delta} \rightarrow 1 * G^{\Delta};\ \pi^{\Delta **} \rightarrow G^{\Delta} * K^{**};\ 0^{**} \rightarrow 1 * K^{**}$$

In equation (7),

$$Number\ of\ Equations = K^{**}$$

$$Number\ of\ Unknowns = G^{\Delta} - 1$$

$$\therefore K^{**} \geq G^{\Delta} - 1$$

This is known as the order condition for identifiability. This condition, in fact states that a necessary condition for identification of a given structural equation. Therefore, Order Condition is number of variables excluded should be greater than or equal to number of endogenous variables less than 1.Order condition is only a necessary condition: if order condition is not satisfied than equation is not identified. However, if it is satisfied, we have to still check for the rank condition

Rank Condition: For equation to be independent, rank should be equal to $G^{\Delta} - 1$. Thus, a necessary and sufficient condition for identification is that the number of independent equations in (7) is $G^{\Delta} - 1$. This will be the case if and only if the order of the largest nonzero determinant that can be formed from all square sub-matrices of $\pi^{\Delta **}$ is

This is known as rank condition for identifiability. Thus, for the identification problem, rank condition is the necessary condition for the order condition to be satisfied. The order and rank conditions enables us to set up the following general rule for determining the identification status of a structural equation.

$$If\ K^{**} > G^{\Delta} - 1\ and\ rank\ (\pi^{\Delta **}) = G^{\Delta} - 1, we\ have\ over\ identification$$

$$If\ K^{**} = G^{\Delta} - 1\ and\ rank\ (\pi^{\Delta **}) = G^{\Delta} - 1, we\ have\ exact\ identification$$

$$If\ K^{**} < G^{\Delta} - 1\ and\ rank\ (\pi^{\Delta **}) = G^{\Delta} - 1, we\ have\ under\ identified$$

Certain examples of problems of Identification:

Example 4:

$$C_t = \alpha_0 + \alpha_1 Y_t + U_{1t} \qquad (a)$$

$$I_t = \beta_0 + \beta_1 Y_{t-1} + U_{2t} \qquad (b)$$

$$Y_t = C_t + I_t + G_t \qquad (c)$$

Solution:

For each of first two structural equations, we have Number of Endogenous Variables (G) =3 and Number of Equations (K) =2

For Rank and Order condition, we need to get a coefficient matrix of this, which is as follows:

Equations	C_t	I_t	Y_t	Y_{t-1}	G_t	Excluded Variables	Order Condition
a	-1	0	α_1	0	0	3	$> G-1 \quad \therefore Over\ Identified$
b	0	-1	0	β_1	0	3	$> G-1 \quad \therefore Over\ Identified$
c	1	1	-1	0	1		

For Rank condition:

Equation (a): ρ[coefficient of excluded variables from equation]

$$= \rho \begin{bmatrix} -1 & \beta_1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$$

$$= 2$$

$$= G - 1$$

The Rank condition is satisfied; therefore by order condition equation (a) is over identified.

Equation (a): ρ[coefficient of excluded variables from equation]

$$= \rho \begin{bmatrix} -1 & \alpha_1 & 0 \\ 1 & -1 & 1 \end{bmatrix}$$

$$= 2$$

$$= G - 1$$

The Rank condition is satisfied; therefore by order condition equation (b) is over identified.

Both equations are over-identified. So model is over identified.

14.5 INDIRECT LEAST SQUARE

We have two types of methods to Solve Simultaneous Equation Model

(i) Single Equation Method i.e. Indirect Least Square (ILS), 2 Stage Least Square (2SLS), LIML

(ii) System Method i.e. 3 SLS (3 Stage Least Square)

We would be taking up ILS and 2SLS methods

- **Indirect Least Square:**

Consider the following relationship

$$C_t = \alpha + \beta Y_t + U_t \qquad (1)$$

$$Y_t = C_t + Z_t \qquad (2)$$

If we apply OLS then we get,

$$\hat{\beta} = \frac{\sum C_t y_t}{\sum y_t^2} = \frac{\sum(\alpha + \beta Y_t + U_t) y_t}{\sum y_t^2} = \frac{\sum(\alpha + \beta Y_t + \beta \bar{Y} + U_t) y_t}{\sum y_t^2} \quad \text{As} \quad y_t = Y_t - \bar{Y}$$

$$\frac{\sum \alpha y_t + \beta \sum y_t^2 + \beta \bar{Y} \sum y_t + \sum y_t U_t}{\sum y_t^2}$$

As $\sum y_t = 0$;

$$\therefore \hat{\beta} = \beta + \frac{\sum y_t U_t}{\sum y_t^2}$$

$$E(\hat{\beta}) \neq \beta$$

$$\therefore \textit{Estimates are biased}$$

For Large Sample:

$$Y_t = \alpha + \beta Y_t + U_t + Z_t$$

$$(1 - \beta) Y_y = \alpha + U_t + Z_t$$

$$Y_t = \frac{\alpha}{1 - \beta} + \frac{Z_t}{1 - \beta} + \frac{U_t}{1 - \beta}$$

In Reduced Form:

$$y_t = \frac{z_t}{1 - \beta} + \frac{U_t + \bar{U}}{1 - \beta}$$

$$\hat{\beta} = \beta + \frac{\sum u_t y_t}{\sum y_t^2}$$

$$= \beta + \frac{\sum u_t \left(\frac{z_t}{1 - \beta} + \frac{U_t + \bar{U}}{1 - \beta} \right)}{\sum y_t^2}$$

$$= \beta + \frac{\frac{\sum u_t^2}{1 - \beta}}{\sum y_t^2} \qquad As \sum u_t z_t = 0 \text{ and } \sum u_t \bar{u} = 0$$

For Large Sample:

$$Prob \lim_{n\to\infty} \hat{\beta} = \beta + \frac{\lim_{n\to\infty} \frac{1}{1-\beta} \frac{\sum u_t^2}{n}}{\lim_{n\to\infty} \frac{\sum y_t^2}{n}}$$

$$Prob \lim_{n\to\infty} \hat{\beta} = \beta + \frac{1}{1-\beta} \frac{\sigma_u^2}{\sigma_y^2}$$

Thus, even for large sample it is not unbiased. Now to solve this we first find Reduced Form Equations of (1) and (2)

$$Y_t = C_t + Z_t$$

$$= \alpha + \beta Y_t + U_t + Z_t$$

$$(1-\beta)Y_t = \alpha + U_t + Z_t$$

$$Y_t = \frac{\alpha}{1-\beta} + \frac{Z_t}{1-\beta} + \frac{U_t}{1-\beta}$$

$$Y_t = \pi_1 + \pi_2 Z_t + V_{1t} \quad (3)$$

Now,

$$C_t = \alpha + \beta\left[\frac{\alpha}{1-\beta} + \frac{Z_t}{1-\beta} + \frac{U_t}{1-\beta}\right] + U_t$$

$$= \frac{\alpha}{1-\beta} + \frac{\beta Z_t}{1-\beta} + \frac{\beta U_t}{1-\beta} + U_t$$

$$= \pi_3 + \pi_4 Z_t + V_{2t} \quad (4)$$

$$\therefore \pi_1 = \frac{\alpha}{1-\beta};\ \pi_2 = \frac{1}{1-\beta};\ \pi_3 = \frac{\alpha}{1-\beta};\ \pi_4 = \frac{\beta}{1-\beta}$$

$$\hat{\beta} = \frac{\widehat{\pi_4}}{\widehat{\pi_2}}$$

From (3),

$$\widehat{\pi_2} = \frac{\sum y_t z_t}{\sum y_t^2}$$

From (4),

$$\widehat{\pi_4} = \frac{\sum c_t z_t}{\sum z_t^2}$$

$$\Rightarrow \widehat{\beta_{ILS}} = \frac{\widehat{\pi_4}}{\widehat{\pi_2}} = \frac{\sum c_t z_t}{\sum y_t z_t}$$

$$Thus, \widehat{\alpha_{ILS}} = \bar{C} - \beta_{ILS}\bar{Y}$$

- **Properties of ILS Estimator**:

$$\widehat{\beta_{ILS}} = \frac{\sum c_t z_t}{\sum y_t z_t}$$

$$= \frac{\sum \left[\frac{\beta z_t}{1-\beta} + \frac{U_t - \bar{U}}{1-\beta}\right] z_t}{\sum \left[\frac{z_t}{1-\beta} + \frac{U_t - \bar{U}}{1-\beta}\right] z_t}$$

$$= \frac{\sum \beta {z_t}^2 + \sum u_t z_t}{\sum {z_t}^2 + \sum u_t z_t}$$

$$Prob \lim_{n\to\infty} \widehat{\beta_{ILS}} = \frac{Prob \lim_{n\to\infty} \left(\beta \frac{\sum {z_t}^2}{n} + \frac{\sum u_t z_t}{n}\right)}{Prob \lim_{n\to\infty} \left(\frac{\sum {z_t}^2}{n} + \frac{\sum u_t z_t}{n}\right)}$$

$$= \frac{\beta {\sigma_z}^2}{{\sigma_z}^2}$$

$$= \beta \qquad As\ Prob\ \lim_{n\to\infty} \frac{\sum {z_t}^2}{n} = {\sigma_z}^2$$

This means that β is asymptotically unbiased estimator, similarly we can prove that

$$Prob \lim_{n\to\infty} (\beta_{ILS} - \beta)^2 = 0$$

i.e. property of consistency

Thus ILS estimates are biased but consistent.

- **Assumptions for ILS:**
  - Model is just identified
  - Applying OLS means, all assumptions of OLS should be fulfilled by Reduced Form Equations.
    - Error term of Reduced For Equations $\sim iid\ N(0,\sigma^2)$
    - There is no multicollinearity
    - No aggregation errors.

- **Two-Stage Least Square Method (2SLS)**

This method is an extension of indirect least squares method. It is the most important method of the single equation techniques for the estimation of the just identified and over identified cases. It aims at the elimination of simultaneous bias. In 2SLS the endogenous variables is purged by its own estimated value by taking into consideration all the exogenous variables. This method presupposes a full understanding of exogenous variables. The 2SLS method is as follows:

Step 1. To eliminate the possibility of a correlation between the explanatory variables and the error term, regress the explanatory variable on all of the system's predetermined variables.

Step 2. Obtaining the estimates of the explanatory variables and replacing it in the original equation by its estimated value and then using OLS to the modified equation.

For 2 Stage Least Square Estimates; Equations are

$$C_t = \alpha + \beta Y_t + U_t \quad (1)$$

$$Y_t = C_t + Z_t \quad (2)$$

Reduced* form of (2) can be written as

$$Y_t = \pi_1 + \pi_2 Z_t + U_t$$

$$\widehat{\pi_2} = \frac{\sum y_t z_t}{\sum {z_t}^2}$$

(*how to derive reduced form equation, this is already explained in the previous section of ILS)

In deviation form:

$$y_t = \pi_2 z_t + U_t + \bar{U}$$

$$\hat{y}_t = \widehat{\pi_t}\, z_t$$

Now,

$$C_t = \alpha + \beta \hat{Y}_t + W_t$$

$$\hat{\beta}=\frac{\sum c_t\hat{y}_t}{\sum \hat{y}_t^{\,2}}$$

$$\hat{\beta}=\frac{\sum c_t\hat{\pi}_2 z_t}{\sum \hat{\pi}_2^{\,2} z_t^{\,2}}$$

$$\hat{\beta}=\frac{1}{\pi_2}\frac{\sum c_t z_t}{\sum z_t^{\,2}}$$

$$\widehat{\beta_{2SLS}}=\frac{\sum c_t z_t}{\sum z_t^{\,2}\dfrac{\sum y_t z_t}{\sum z_t^{\,2}}}=\frac{\sum c_t z_t}{\sum y_t z_t}$$

- **Assumptions of Two Stage Least Square Method:**

  The 2SLS method involves the application of the classical least squares to two type of functions: (i) to the reduced form equations (ii) to the transformed structural functions. The transformation consisting of the replacement of the endogenous variable by then estimated values obtained from the reduced form equations. The 2SLS is based on the following assumptions:

  - The random variable 'u' of the original structural equations must satisfy the usual stochastic assumptions.
  - The error term of the reduced form equations must satisfy the stochastic assumptions.
  - The specification of the model is to be correct so far as the exogenous variables are concerned.
  - The sample must be large and the number of observations is greater than the number of predetermined variables in the structural model i.e. n>k
  - The explanatory variables are not perfectly multicollinear.

- **Properties of 2SLS**
  - From small sample the estimates are biased
  - In large samples (as n approaches to ∞) the bias tends to zero
  - 2SLS provides only one estimate per parameter, whereas ILS provides multiple estimates of parameters in the over identified equations.
  - The 2SLS method can also be applied to exactly identified equations. But then ILS and 2SLS will give identical estimates

- The method of 2SLS gives consistent estimated under conditions in which the OLS method fails.

- **Most widely used Method**

  2SLS method is extensively used in practice for the following points over the ILS procedure of estimation.

  - When compared to ILS, which provides many parameter estimates when applied to over-identified equations, 2SLS only provides one estimate per parameter.
  - Although 2SLS was specifically created for handling over-identified equations, this method can be used for exactly identified equations too. In this circumstance, ILS and 2SLS produce the same results.
  - There is no straightforward way of estimating standard errors of structural coefficients from the standard errors of reduced form coefficients in the application of ILS. However, in the case of 2SLS estimation, this is simple because the structural coefficients are estimated directly from the second stage (OLS) regressions. However, the estimated standard errors in the second-stage regression need to be modified.



CHECK YOUR PROGRESS

Q 1: State the definition of reduced form equation.

.....

.....

Q 2: When a system of simultaneous equations are said to be complete?

.....

.....

Q 3: Write one Properties of 2SLS.

.....

.....

.....

.....



14.6 LET US SUM UP

- A model is said to constitute a system of simultaneous equations if all of the relationships involved are needed for determining the value of at least one of the endogenous variables included in the model.
- Variables are also of two types: Endogenous– those determined in the model and Exogenous– those determined outside the model
- Identification refers to structural parameters, equation by equation. There exist two equivalent approaches to identify the parameters
- Two stage least square is most important method of the single equation techniques for the estimation of the just identified and over identified cases.
- The 2SLS method involves the application of the classical least squares to two type of functions: (i) to the reduced form equations (ii) to the transformed structural functions.



14.7 FURTHER READING

1) Dougherty, C. (2007), *Introduction to Econometrics*, 3rd Edition., Oxford University Press, New Delhi.
2) Greene, William H. 2005. Econometric Analysis (5th ed.). Upper Saddle River, NJ: Prentice Hall.
3) Gujarati, Damodar. 2003. Basic Econometrics (4th ed.). New York: McGraw-Hill.
4) Johnston, J. 1984. Econometric Methods (3rd ed.). New York: McGraw-Hill. (Note: pages for Johnston, J. and J. DiNardo's Econometric Methods 4th ed. (in parentheses)).
5) Kmenta, Jan. 1986. Elements of Econometrics (2nd ed.). New York: Macmillan
6) Maddala, G.S. (Ed.) (1993), *Econometric Methods and Applications* (2 Vols.) Aldershot U.K.
7) Woolwridge Jefferry (2009): *Introductory Econometrics: A Modern*

Approach, 5th Edition, South-Western Cengage Learnin



14.8 ANSWER TO CHECK YOUR PROGRESS

Ans to Q No 1: A Reduced form Equation is one that expresses an endogenous variable solely in terms of the pre-determined variable and the stochastic error term.

Ans to Q No 2: The system is complete if it meets the following conditions:

- Number of equations = Number of endogenous variables
- The system is linear in parameters.

Ans to Q No 3: The method of 2SLS gives consistent estimated under conditions in which the OLS method fails.



14.9 MODEL QUESTIONS

Q 1: Derive the reduced form of structural equation model

Q 2: Explain the prosperities of 2SLS

Q 3: Explain the method of indirect least square.

Q 4: Describe rank and order condition of simultaneous equation model.

Q 5: Explain the problems caused by Endogenous Variables

*** ***** ***