UNIT 10 MULTICOLLINEARITY*

Structure

10.0 OBJECTIVES

After going through this unit, you should be able to

- explain the concept of multicollinearity in a regression model;
- comprehend the difference between the near and perfect multicollinearity;
- describe the consequences of multicollinearity;
- [1]explain how multicollinearity can be detected; and
- describe the remedial measures of multicollinearity; and
- explain the concept of ridge regression.

10.1 INTRODUCTION

The classical linear regression model assumes that there is no perfect multicollinearity. Multicollinearity means the presence of high correlation between two or more explanatory variables in a multiple regression model.

* Dr. Pooja Sharma, Assistant Professor, Daulat Ram College, University of Delhi

Absence of multicollinearity implies that there is no exact linear relationship among the explanatory variables. The assumption of no perfect multicollinearity is very crucial to a regression model since the presence of perfect multicollinearity has serious consequences on the regression model. We will discuss about the consequences, detection methods, and remedial measures for multicollinearity in this Unit.

10.2 TYPES OF MULTICOLLINEARITY

Multicollinearity could be of two types: (i) perfect multicollinearity, and (ii) imperfect multicollinearity. Remember that the division is according to the degree or extent of relationship between the explanatory variables. The distinction is made because of the nature of the problem they pose. We describe both types of multicollinearity below.

10.2.1 Perfect Multicollinearity

In the case of perfect multicollinearity, the explanatory variables are perfectly correlated with each other. It implies the coefficient of correlation between the explanatory variables is 1. For instance, suppose want to derive the demand curve for a good Y. We assume that quantity demanded (Y) is a function of price (X_2) and income (X_3). In symbols,

$Y = f(X_2, X_3)$ where X_2 is price of good Y and X_3 is the weekly consumer income.

Let us consider the following regression model (population regression function):

$$Y_i = A_1 + A_2X_{2i} + A_3X_{3i} + u_i \quad \ldots (10.1)$$

In the above equation, suppose

A_2 is < 0. This implies that prices are inversely related do demand.

$A_3 > 0$. This indicates that as income increases, demand for the good increases.

Suppose there is a perfect relationship between X_2 and X_3 such that

$$X_{3i} = 300 - 2X_{2i} \quad \ldots (10.2)$$

In the above case, if we regress X_3 on X_2 we obtain the coefficient of determination $R^2 = 1$.

If we substitute the value of X_3 from equation (10.2), we obtain

$$\begin{aligned} Y_i &= A_1 + A_2X_{2i} + A_3(300 - 2X_{2i}) + u_i \\ &= A_1 + A_2X_{2i} + 300A_3 - 2A_3X_{2i} + u_i \\ &= (A_1 + 300A_3) + (A_2 - 2A_3)X_{2i} + u_i \quad \ldots (10.3) \end{aligned}$$

Let $C_1 = (A_1 + 300A_3)$ and $C_2 = (A_2 - 2A_3)$. Then equation (10.3) can be written as:

$$Y_i = C_1 + C_2X_{2i} + u_i \quad \text{....(10.4)}$$

Thus if we estimate the regression model given at (10.4), we obtain estimators for C_1 and C_2. We do not obtain unique estimators for A_1, A_2 and A_3.

As a result, in the case of perfect linear relationship or perfect multicollinearity among explanatory variables, we cannot obtain unique estimators of all the parameters. Since we cannot obtain their unique estimates, we cannot draw any statistical inferences (hypothesis testing) about them. Thus, in case of perfect multicollinearity, estimation and hypothesis testing of individual regression coefficients in a multiple regression are not possible.

10.2.2 Near or Imperfect Multicollinearity

In the previous section, the presence of perfect multicollinearity indicated that we do not get unique estimators for all the parameters in the model. In practice, we do not encounter perfect multicollinearity. We usually encounter near or very high multicollinearity. In this case the explanatory variables are approximately linearity related.

High collinearity refers to the case of "near" or "imperfect" multicollinearity. Thus, when we refer to the problem of multicollinearity we usually mean "imperfect multicollinearity''

Let us consider the same demand function of good Y. In this case we however assume that there is imperfect multicollinearity between the explanatory variables (in order to distinguish it from the earlier case, we have changed the parameter notations). The following is the population regression function:

$$Y_i = B_1 + B_2X_{2i} + B_3X_{3i} + u_i \quad \text{....(10.5)}$$

Equation (10.5) refers to the case when two or more explanatory variables are not exactly linear. For the above regression model, we may obtain an estimated regression equation as follows:

Equation (10.5):	$\hat{Y}_i = 145.37$	$-$	$2.7975X_{2i}$	$-$ $0.3191X_{3i}$
Standard Error:	(120.06)		(0.8122)	(0.4003)
t-ratio:	(1.2107)		(–3.4444)	(–0.7971)

$R^2 = 0.97778$ ... (10.6)

Since the explanatory variables are not exactly related, we can find estimates for the parameters. In this case, regression can be estimated unlike the first case of prefect multicollinearity. It does not mean that there is no problem with our estimators if there is imperfect multicollinearity. We discuss the consequences of multicollinearity in the next section.

Check Your Progress 1

1) What is meant by perfect multicollinearity?

.....

.....

.....

.....

.....

2) What do you understand by imperfect multicollinearity?

.....

.....

.....

.....

.....

3) Explain why it is not possible to estimate a multiple regression model in the presence of perfect multicollinearity.

.....

.....

.....

.....

.....

10.3 CONSEQUENCES OF MULTICOLLINEARITY

We know from Unit 4 that the ordinary least squares (OLS) estimators are the Best Linear Unbiased Estimators (BLUE). It implies they have the minimum variance in the class of all linear unbiased estimators. In the case of imperfect multicollinearity, the OLS estimators still remain BLUE. Then what is the problem? In the presence of multicollinearity, there is an increase in the variance and standard error of the coefficients. As a result, very few estimators are statistically significant.

Some more consequences of multicollinearity are given below.

(a) The explanatory variables may not be linearly related in the population (i.e., in the population regression function), but they could be related in a particular sample. Thus multicollinearity is a sample problem.

(b) Near or high multicollinearity results in large variances and standard errors of OLS estimators. As a result, it becomes difficult to estimate true value of the estimator.

(c) Multicollinearity results in wider confidence intervals. The standard errors associated with the partial slope coefficients are higher. Therefore, it results in wider confidence intervals.

$$P_r\left[b_2 - t_{\alpha/2}SE(b_2) \leq \beta_2 \leq b_2 + t_{\alpha/2}SE(b_2)\right] = 1 - \alpha \quad \ldots.(10.7)$$

Since the values of standard errors have increased the interval reflected in expression in (10.7) has widened.

(d) Insignificant t ratios: As pointed out above, standard errors of the estimators increase due to multicollinearity. The t-ratio is given as $= \frac{b_2}{SE(b_2)}$. Therefore, the t-ratio is very small. Thus we tend to accept (or do not reject) the null hypothesis and tend to conclude that the variable has no effect on the dependent variable.

(e) A high R^2 and few significant t-ratios: In equation (10.6) we notice that the R^2 is very high, about 98% or 0.98. The t-ratios of both the explanatory variables are not statistically significant. Only the price variable slope coefficient has significant t-value. However, using F-test while testing overall significance $H_0: R^2 = 0$, we reject the null hypotheses. Thus there is some discrepancy between the results of the F-test and the t-test.

(f) The OLS estimators are mainly partial slope coefficients and their standard errors become very sensitive to small changes in the data. If there is a small change in data, the regression results change substantially.

(g) Wrong signs of regression coefficients: It is a very prominent impact of the presence of multicollinearity. In the case of the example given at equation (10.6) we find that the coefficient of the variable income is negative. The income variable has a 'wrong' sign as economic theory suggests that income effect is positive unless the commodity concerned is an inferior good.

10.4 DETECTION OF MULTICOLLINEARITY

In the previous section we pointed out the consequences of multicollinearity. Now let us discuss how multicollinearity can be detected.

(h) High R^2 and Few Significant t-ratios

This is the classic symptom of multicollinearity. If R^2 is high (greater than 0.8), the null hypothesis that the partial slope coefficients are jointly or simultaneously equal to zero $\left[H_0 : \beta_2 = \beta_3 = 0\right]$ is rejected in most cases (on the basis of F-test). But the individual t-tests will reflect that none or very few partial slope coefficients are statistically different from zero. This suggests very few slope coefficients are statistically significant.

(ii) High Pair-wise Correlations among Explanatory Variables

Due to high correlation among the independent variables, the estimated regression coefficients have high standard errors. But this is not necessarily true as demonstrated below. Even low correlation among the independent variables can lead to the problem of multicollinearity.

Let r_{23}, r_{24} and r_{34} represent the pair-wise correlation coefficients between X_2 and X_3 and X_4 respectively. Suppose $r_{23} = 0.90$, reflecting high collinearity between X_2 and X_3. Let us consider partial correlation coefficient $r_{23.4}$ that indicates correlation between X_2 and X_3 (while keeping the influence of X_4 constant). Suppose we find that $r_{23.4} = 0.43$. It indicates that partial correlation between X_2 and X_3 is low reflecting the absence of high collinearity. Therefore, pair-wise correlation coefficient when replaced by partial correlation coefficients does not indicate the presence of multicollinearity. Suppose the true population regression is given by equation (10.8)

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_2 X_{3i} + \beta_4 X_{4i} + u_i \quad \ldots (10.8)$$

Suppose the explanatory variables are perfectly correlated with each other as shown in equation (10.9) below

$$X_{4i} = \lambda_2 X_{2i} + \lambda_3 X_{3i} \quad \ldots (10.9)$$

X_4 is an exact linear combination of X_2 and X_3

If we estimate the coefficient of determination by regressing X_4 on X_2 and X_3, we find that

$$R^2_{4.23} = \frac{r^2_{42} + r^2_{43} - 2r_{42}\, r_{43}\, r_{23}}{1 - r^2_{23}} \quad \ldots (10.10)$$

Suppose, $r_{42} = 0.5, r_{43} = 0.5, r_{23} = -0.5$. If we substitute these values in equation (10.10), we find that $R^2_{4.23} = 1$. An implication of the above is that all the correlation coefficients (among explanatory variables) are not very high but still there is perfect multicollinearity.

(iii) Subsidiary or Auxiliary Regressions

Suppose one explanatory variable is regressed on each of the remaining variables and the corresponding R^2 is computed. Each of these regressions is referred to as subsidiary or auxiliary regression. For example, in a regression model with seven explanatory variables, we regress X_1 on X_2, X_3, X_4, X_5, X_6 and X_7 and find out the R^2_1. Similarly, we can regress X_2 on X_1, X_3, X_4, X_5, X_6 and X_7 and find out the R^2_2. By examining the auxiliary regression models we can find out the possibility

of multicollinearity. We take the rule of thumb that multicollinearity may be troublesome if R_i^2 obtained from auxiliary regression is greater than overall R^2 of the regression model.

A limitation of this method is that we have to compute R_i^2 several times, which is cumbersome and time consuming.

(iv) Variance Inflation Factor (VIF)

Another indicator of multicollinearity is the variance inflation factor (VIF). The R_i^2 obtained from auxiliary regressions may not be a reliable indicator of collinearity. In VIF method we modify the formula of variance of the estimators as follows; (b_2) and (b_3)

$$\text{var}\,(b_2) = \frac{\sigma^2}{\sum x_{2i}^2(1-R_2^2)} = \frac{\sigma^2}{\sum X_{2i}^2}.\left(\frac{1}{1-R_2^2}\right) \quad \text{... (10.11)}$$

In equation (10.11), you should note that R_2^2 is the auxiliary regression discussed earlier.

Compare the variance of b_2 given in equation (10.11) with the usual formula for variance of an estimator given in Unit 4. We find that

$$var(b_2) = \frac{\sigma^2}{\sum x_{2i}^2}\ VIF \quad \text{... (10.12)}$$

where VIF $= \left(\frac{1}{1-R_2^2}\right)$

Similarly, $var(b_3) = \frac{\sigma^2}{\sum x_{3i}^2}(VIF)$

Note that as R_i^2 increases the VIF also increases. This inflates the variance and hence standard errors of b_2 and b_3

If $R_i^2 = 0$, $VIF = 1 \Rightarrow V(b_2) = \frac{\sigma^2}{\sum x_{2i}^2}$ and $V(b_3) = \frac{\sigma^2}{\sum x_{3i}^2}$

Therefore, there is no collinearity.

On the other hand,

if $R_i^2 = 1$, $VIF = \infty \Rightarrow V(b_2) \to \infty, V(b_3) \to \infty$

If R_i^2 is high, however $V(b_2)$ tends to ∞.

Note that $var(b_2)$ depends not only on R_i^2, but also on σ^2 and $\sum x_{2i}^2$. It is possible that R_i^2 is high (say, 0.91) but $var(b_2)$ could be lower due to low σ^2 or high $\sum x_{2i}^2$. Thus $V(b_2)$ is still lower resulting in high *t* value. Thus R_i^2 obtained from auxiliary regression is only a superficial indicator of multicollinearity.

Check Your Progress 2

1) Bring out four important consequences of multicollinearity.

.....

.....

.....

.....

.....

2) Explain how multicollinearity can be detected using partial correlations.

.....

.....

.....

.....

.....

3) Describe the method of detection of multicollinearity using the variance inflation factor (VIF).

.....

.....

.....

.....

.....

10.5 REMEDIAL MEASURES OF MULTICOLLINEARITY

Multicollinearity may not necessarily be an "evil'' if the goal of the study is to forecast the mean value of the dependent variable. If the collinearity between the explanatory variables is expected to continue in future, then the population regression function can be used to predict the relationship between the dependent variable Y and other collinear explanatory variables.

However, if in some other sample, the degree of collinearity between the two variables is not that strong the forecast based on the given Regression is of little use.

On the other hand, if the objective of the study is not only prediction but also reliable estimations of the individual parameters of the chosen model then serious collinearity may be bad, since multicollinearity results in large standard errors of estimators and therefore widens confidence interval. Thus, resulting in accepting null hypotheses in most cases. If the objective of the study is to estimate a group

of coefficients (i.e., sum or difference of two coefficients) then this is possible even in presence of multicollinearity. In such a case multicollinearity may not be a problem.

$$Y_i = C_1 + C_2 X_{2i} + u_i \qquad \ldots(10.13)$$

$$C_1 = A_1 + 300A_3, \qquad C_2 = A_2 - 2A_3$$

Running the above regression in equation (10.2), as presented in earlier section 10.2, one can easily estimate C_2 by using OLS method, although neither A_2 nor A_3 can be estimated individually. There can be situation when in spite of inflated S.E., the individual coefficients happened to be numerically significant since the true value itself is so large even or estimate on the downside still shows up a significant test.

Certain remedies prescribed for reducing the severity of collinearity problem which can be listed as OLS estimators can still retain BLUE property despite of near collinearity. Further, one or more regression coefficients can e individually statistically significant or some of them with wrong signs.

10.5.1 Dropping a Variable from the Model

The simplest solution may be to drop one or more of the collinear variables. However, dropping a variable from the model may lead to model specification error. In other words, when we estimate the model without the excluded variable, the estimated parameters of the reduced model may turn out to be biased. Therefore, the best practical advice is not to drop a variable from a model that is theoretically sound. A variable which has *t* value of its coefficient greater than 1, then than variable should not be dropped as it will result in a decrease in $\bar{R}^2$.

10.5.2 Acquiring Additional Data or New Sample

Acquiring additional data implies increasing the sample size. This is likely to reduce the severity of the multicollinearity problem. As we know from equation (10.11),

$$\text{var}\,(b_2) = \frac{\sigma^2}{\sum x_{2i}^2 (1 - R_2^2)}$$

Given σ^2 and R_2^2, if the sample size of X_2 increases, there is an increase in $\sum x_{2i}^2$. It will lead a decrease in var (b_2) and its standard error.

10.5.3 Re-Specification of the Model

It is possible that some important variables are omitted from the model. The functional form of the model may also be incorrect. Therefore, there is a need of looking into the specification of the model. Many times, taking log form of a model leads to solving the problem of multicollinearity.

10.5.4 Prior Information about Certain Parameters

Estimated values of certain parameters are available in existing studies. These values can be used as prior information. These values give us some tentative idea on the plausible value of the parameters.

10.5.5 Transformation of Variables

Transformation of the variables would minimize the problem of collinearity.

10.5.6 Ridge Regression

The ridge regressions are another method of resolving the problem of multicollinearity. In the ridge regression, the first step is to standardize the variables both dependent and independent by subtracting the respective means and dividing by their standard deviations. This mainly implies that the main regression is run by transforming both dependent and explanatory variables into the standardized values.

It is observed that in the presence of multicollinearity, the value of variance inflation factor is substantially high. This is mainly due to a high value of coefficient of determination. The ridge regression is applied when the regression equations are in the form of matrix involving large number of explanatory variables.

The ridge regression proceeds by adding a small value, k, to the diagonal elements of the correlation matrix. The reason that the diagonal of ones in the correlation matrix could be considered as a ridge, this is the reason such regression is referred as ridge regression.

10.5.7 Other Remedial Measures

There are several other Remedies suggested such as combining time series and cross-sectional data, factor or principal component analysis and ridge regressions.

Polynomial Regression Models

Let us consider total cost of production (TC) as a function of output as well as marginal cost (MC) and Average Cost (AC)

$$Y_i = \beta_1 + \beta_2 X_1 + \beta_3 X_i^2 + \beta_4 X_i^3 \qquad \text{…….(10.12)}$$

The cost function is defined as Cubic function for cost as a third-degree polynomial of variable X. This model in equation (10.12) is linear in parameters β^s, therefore satisfy assumption of CLRM of linear Regression Model and can be estimated by usual OLS method. However, one needs to worry about problem of collinearity since it is not linear in variables and at the same time X^2 and X^3 are non-linear function of X and do not violate the assumptions of no perfect collinearity i.e., no perfect linear relationship between variables. The estimated results are presented in equation (10.13).

$$\hat{Y}_i = 141.7667 + 63.4776X_i - 12.9615X_i^2 + 0.9396\ X_i^3 \qquad \text{….(10.13)}$$

Se (6.3753) (4.7786) (0.9857) (0.0591)

$$R^2 = 0.9983$$

$$AC = \frac{RC}{X_i} = \frac{141.7667}{X_i} + 63.4776 - 12.96X_i + (0.9396)X_i^2$$

$$AC_i = 63.4776 - 12.9615X_i + 141.7667X_i + 0.9396X_i^2$$

$$MC = \frac{\partial TC}{\partial X_i} = 63.4776 - 2X(12.9615)X_i + 3 \times 0.9396X_i^2$$

If the cost curves are U-shaped Average Marginal cost curves then the theory suggests that the coefficient should satisfy following

1) β_1, β_2 and $\beta_4 > 0$
2) $\beta_3 < 0$
3) $\beta_3^2 < 3\beta_2\beta_4$

Check Your Progress 3

1) Define two significant methods to rectify the problem of multicollinearity?

.....

.....

.....

.....

.....

2) Describe the method of ridge regression.

.....

.....

.....

.....

.....

10.6 LET US SUM UP

This unit presents a clear understanding of the concept of multicollinearity in the regression model. The unit also presents a clear distinction of near and perfect multicollinearity. The unit familiarizes the consequences of presence of multicollinearity in regression model. The method of detection of multicollinearity has been highlighted in the unit. Finally various techniques that provide remedial measures including the concept of ridge regression have been explained in the unit.

10.7 ANSWERS/ HINTS TO CHECK YOUR PORGRESS EXERCISES

Check Your Progress 1

1) The case of perfect multicollinearity mainly reflects the situation when the explanatory variables and perfectly correlated with each other implying the coefficient of correlation between the explanatory variables is 1.

2) This refers to the case when two or more explanatory variables are not exactly linear this reinforces the fact that collinearity can be high but not perfect. "High collinearity" refers to the case of "near" or imperfect" or high multicollinearity. Presence of multicollinearity implies "imperfect multicollinearity''

3) In the case of perfect multicollinearity it is not possible to obtain estimators for the parameters of the regression model. See Section 10.2 for details.

Check Your Progress 2

1) (i) In case of imperfect multicollinearity, some of the estimators are statistically not significant. But OLS estimates still retain their BLUE property that is, Best Linear Unbiased Estimators. Therefore, imperfect multicollinearity does not violate any of the assumptions, OLS estimators retain BLUE property. Being BLUE with minimum variance does not imply that the numerical value of variance will be small.

(ii) The R^2 value is very high but very few estimators are significant (t-ratios low). The example mentioned in earlier section where the demand function of good Y we computed using the earnings of individuals, reflects the situation where R^2 is quite high about 98% or 0.98 but only price variable slope coefficient has significant t-value. However, using F-test while testing overall significance $H_0 : R^2 = 0$, we reject the hypotheses that both prices and earnings have no effect on the demand of Y.

(iii) The ordinary least square OLS estimators mainly partial slope coefficients and their standard errors become very sensitive to small changes in the data, i.e. they then to be rentable. A small charge of data, the regression results change quite substantially as in case example of near or imperfect multicollinearity mentioned above, the standard errors go down and t-ratios have increased in absolute values.

(iv) Wrong signs of regression coefficients. It is a very prominent impact of presence of multicollinearity. In case of example where earnings of individuals were used in deriving demand curve of good Y, the earning

variable has the 'wrong' sign for the economic theory since the income effect usually positive unless it is case of inferior good.

2) Examining partial correlations: In case of three explanatory variables X_2, X_3 and X_4 very high or perfect multicollinearity between X_4 and X_2, X_3.

Subsidiary or auxiliary regressions: When one explanatory variables X is regressed on each of the remaining X variable and the corresponding R^2 is computed. Each of these regressions is referred as subsidiary or auxiliary regression. A regression Y on X_2, X_3, X_4, X_5, X_6 and X_7 with six explanatory variables. If R^2 comes out to be very high but few significant t-ratios or very few X coefficients are individually statistically significant then the purpose is to identify the source of the multicollinearity or existent of perfect or near perfect linear combination of other X^s.

For this we Regress X_2 on remaining X^s and obtain R_2^2 or also written as $R^2_{2.34567}$

Regress X_3 on remaining X^s, and obtain R_3^2 coefficient of determination also written as $R^2_{3.24567}$ each R_i^2 obtained will lie between 0 and 1. By testing the null hypothesis $H_0 : R_i^2 = 0$ by applying F-test. Let r_{23}, r_{24} and r_{34} represent pairwise correlation between X_2 and X_3, X_2 and X_4, X_3 and X_4 respectively suppose $r_{23} = 0.90$, reflecting high collinearity between X_2 and X_3. Considering partial correlations coefficient $r_{23.4}$ that indicators correlations coefficient between X_2 and X_3, Adding the influence of X_4 constant. If $r_{23.4} = 0.43$. Thus, partial correlation between X_2 and X_3 is low reflecting no high collinearity or low degree of collinearity. Therefore, pairwise correlation when replaced by partial correlation coefficients does not provide indicator of presence of multicollinearity.

3) Variance Inflation Factor (VIF): R^2 obtained variables auxiliary regression may not be completely reliable and is not reliable indicator of collinearity. In this method we modify the formula of var (b_2) and (b_3)

$$\text{var}(b_2) = \frac{\sigma^2}{\sum x_{2i}^2 \left(1 - R_2^2\right)}$$

$$= \frac{\sigma^2}{\sum X_{2i}^2} \cdot \left(\frac{1}{1 - R_2^2}\right)$$

$$\text{VIF} = \left(\frac{1}{1 - R_2^2}\right) \qquad \therefore V(b_2) = \frac{\sigma^2}{\sum x_{2i}^2} \text{V.I.F.}$$

Similarly, $V(b_3) = \frac{\sigma^2}{\sum x_{3i}^2}(VIF)$

VIF is variance inflation factor. As R^2 increases VIF $\frac{1}{1-R^2}$ increased thus inflating the variance and hence standard errors of b_2 and b_3

If $R^2 = 0$, $VIF = 1 \Rightarrow V(b_2) = \frac{\sigma^2}{\sum x_{2i}^2}$ and $V(b_3) = \frac{\sigma^2}{\sum x_{3i}^2}$

$\Rightarrow$ No collinearity

If $R^2 = 1$, $VIF = \infty \Rightarrow V(b_2) \to \infty, V(b_3) \to \infty$

If R^2 is high, however $var(b_2) \to \infty, var(b_3)$ does not only depend on R^2 (auxiliary coefficient of determination) or VIF. If also depends on σ^2 and $\sum x_{2i}^2$ it is possible that R_i^2 is high 0.91 but $var(b_2)$ could be lower due to low σ^2 or high $\sum x_{2i}^2$ thus $V(b_2)$be still lower resulting in high t value not showing any low t end thus defeating the indicator of multicollinearity. Thus R^2 obtained from and binary regression is only a surface indicator of multicollinearity.

Check Your Progress 3

1) (i) Dropping a variable from the Model: The simplest solution might seem to be to drop one or more of the collinear variables. However, dropping a variable from the model may lead to model specification error in either words, where we estimate the model without that variable, the estimated parameters of reduced model may turn out to be biased. Therefore, the best practical advice is not to drop or variable from an economically variable model first because the collinearity problem is serious. A variable which has t value of its coefficient greater than 1, then than variable should not be dipped as it will result in decrease in adjusted $\overline{R}^2$

(ii) Acquiring Additional Data or new sample: Acquiring additional data implies increasing the sample size can reduce the severity of collinearity problem.

$$V(b_2) = \frac{\sigma^2}{\sum x_{3i}^2 (1 - R_2^2)}$$

Given σ^2 and R^2, if the sample size of X_3 increases $\Rightarrow \sum x_{3i}^2$ will increase as a result $V(b_3)$will tend to decrease and standard error b_3 will also.

2) In ridge regression we first standardise all the variables in the model. Go through Sub-Section 10.5.6 for details.

UNIT 11 HETEROSCDASTICTY*

Structure

11.0 OBJECTIVES

After going through this unit, you should be able to

- explain the concept of heteroscedasticity in a regression model;
- identify the consequences of heteroscedasticity in the regression model;
- explain the methods of detection of heteroscedasticity;
- describe the remedial measures for resolving heteroscedasticity;
- show how the use of deflators can help in overcoming the consequences of heteroscedasticity; and
- identify the correct functional form of regression model so that heteroscedasticity is avoided.

* Dr. Pooja Sharma, Assistant Professor, Daulat Ram College, University of Delhi

11.1 INTRODUCTION

A crucial assumption of the Classical Linear Regression Model (CLRM) is that the error term u_i in population regression function (PRF) is homoscedastic. It means that u_i has the same variance σ^2 throughout the population. An alternative scenario arises where the variance of u_i is σ_i^2. In other words, the error variance varies from one observation to another. Such cases are referred to as cases of heteroscedasticity.

11.2 HETEROSCEDASTICITY: DEFINITION

Let us first make a distinction between homoscedasticity and heteroscedasticity. This will help us in understanding the concept of heteroscedasticity better.

11.2.1 Homoscedasticity

Consider a 2-variable regression model, where the dependent variable *Y* is personal savings and the explanatory variable *X* is personal disposable income (or after-tax income).

As personal disposal income (PDI) increases, the mean or average level of savings also increases but the variances of savings around its mean value remains the same at all the levels of PDI. Such a case depicts the case of homoscedasticity or equal variance as shown in Fig. 11.1. In such cases, we have:

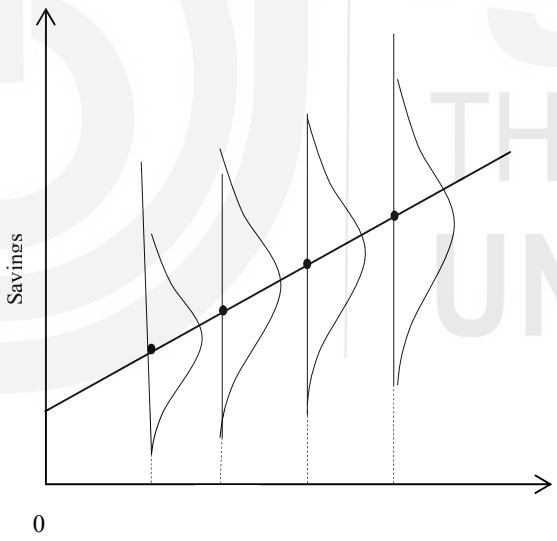

Fig.11.1: Case of Homoscedasticity

$$E(u_i^2) = \sigma^2 \quad \ldots (11.1)$$

We can alternatively express equation (11.1) as a case where:

$$V(u_i) = \sigma^2 \quad \ldots (11.2)$$

In Fig. 11.1, we see a case of homoscedasticity where the variance of the error term is a constant value, σ^2. This is expressed in the form of an equation as in

(11.2). Since the expected value of the error term is zero, the expression $V(u_i) = \sigma^2$ can also be written as $E(u_i^2) = \sigma^2$ as in equation (11.1).

11.2.2 Heteroscedasticity

As PDI increases, the average level of savings increases. However, the variance of savings does not remain the same at all the levels of PDI. This is the case of heteroscedasticity or unequal variance. In other words, high-income people, on average, save more than low-income people, but at the same time, there is more variability in their savings. This can be graphically represented as in Fig. 11.2. We now therefore have:

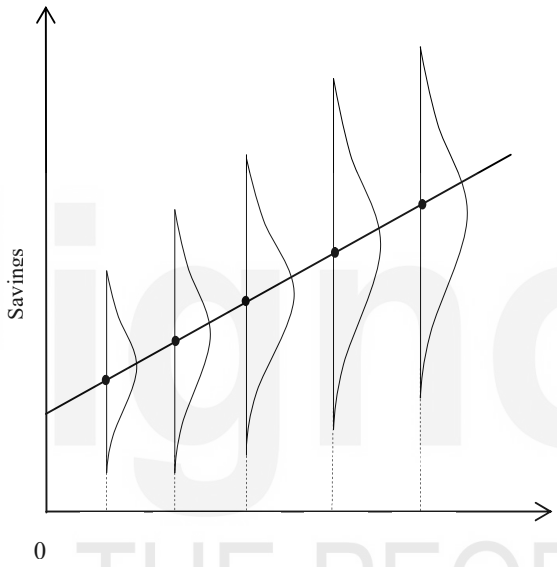

Fig. 11.2: Case of Heteroscedasticity

$$E(u_i^2) = \sigma_i^2 \quad or \quad V(u_i) = \sigma_i^2 \qquad \ldots (11.3)$$

The case of heteroscedasticity reflected in Fig.11.2 indicates that the error variance is not constant. It rather changes with every observation, like

$$V(u_i) = \sigma_i^2.$$

It is observed that heteroscedasticity is usually found in cross-sectional data and not so much in time series data. The reason for its occurrence more in cross-sectional data is mainly because, in the case of cross-sectional data, the members of population are like individuals, firms, industries, geographical division, state or countries. The data in such cases is collected at a point in time. Hence, the members of the population may be of different sizes: small, medium or large. This is referred to as the scale effect. In other words, due to what is called in economics as the 'scale effect', in cross sectional data we find cases of heteroscedasticity more commonly.

In the case of time series, on the other hand, the data of similar variables vary over a period of time. For instance, GDP (gross domestic product) or savings or unemployment varies over a period (like 1960 to 2008).

Check Your Progress 1

1) What is meant by heteroscedasticity?

.....

.....

.....

.....

.....

2) Is the problem of heteroscedasticity related to data? Comment.

.....

.....

.....

.....

.....

11.3 CONSEQUENCES OF HETEROSCEDASTICITY

To avoid the problem of heteroscedasticity, we have made one of the assumptions in the classical linear regression model that the error term is homoscedastic. However, in many regression models and actual data, the disturbance variance varies across observations. Consequently, the model suffers from specific impacts due to heteroscedastic error term.

The following are the characteristics of the OLS model in the presence of heteroscedasticity.

(i) The OLS estimators are linear function of the variables. The regression equation is also linear in its parameters.

(ii) The ordinary least squares (OLS) estimators are unbiased. This means the expected value of estimated parameters is equal to the true population parameters.

(iii) The OLS estimators though unbiased, are no longer with minimum variance, i.e., they are no longer efficient. In fact, even in large samples, the OLS estimators are not efficient. Therefore, the OLS estimators are not BLUE both in small as well as asymptotically large samples.

(iv) In light of the above, the usual formula for estimating variances of OLS estimator is biased, i.e., they are either upward biased (positive bias) or downward biased (negative bias). Note that when the OLS

overestimates the true variances of estimators, a positive bias is said to occur, and when it underestimates the true variances of estimators, we say that a negative bias occurs.

(v) The estimator of true population variance as given by $\hat{\sigma}^2 = \frac{\sum e_i^2}{df} = \frac{RSS}{df}$ is biased. That is

$$E(\hat{\sigma}^2) \neq \sigma^2 \qquad \ldots (11.4)$$

We know that the degrees of freedom for testing an estimated parameter is (n – k), where k is the number of parameters (or explanatory variables) in the regression model. For example, if there are three explanatory variables, d.f. = (n – 3). In the two variables case, df = (n – 2). Note that we are counting the intercept estimate for this purpose of determining the d.f.

(vi) Equation (11.4) implies that in the presence of heteroscedasticity, the estimated value of error variance is not equal to the true population error variance. In view of this, the usual confidence interval and hypothesis testing based on t and F distributions are unreliable (since, the estimator of the error variance is biased). Therefore, the possibility of making wrong inferences (Type–II error) is very high. As a result, in the presence of heteroscedasticity, the results of the usual hypothesis-testing are not reliable raising the possibility of drawing misleading conclusions.

Check Your Progress 2

1) State any two important consequences of heteroscedasticity.

.....

.....

.....

.....

.....

2) In the presence of heteroscedasticity, the OLS estimator will either overestimate or underestimate the error variance. Justify the statement.

.....

.....

.....

.....

.....

11.4 DETECTION OF HETEROSCEDASTICITY

So far, we have discussed the consequences of heteroscedasticity. Now let us discuss how heteroscedasticity can be detected. There are quite a few methods of detecting heteroscedasticity. Some of these methods are described below.

11.4.1 Graphical Examination of the Residuals

We can begin with examining the residuals obtained from the fitted regression line. The residual plot of squared residuals is an indicator of the existence of heteroscedasticity. Since the error terms u_i are not observable, we examine the residuals, e_i.

A plot of the residuals can give us various types of diagrams as in Fig. 11.3.

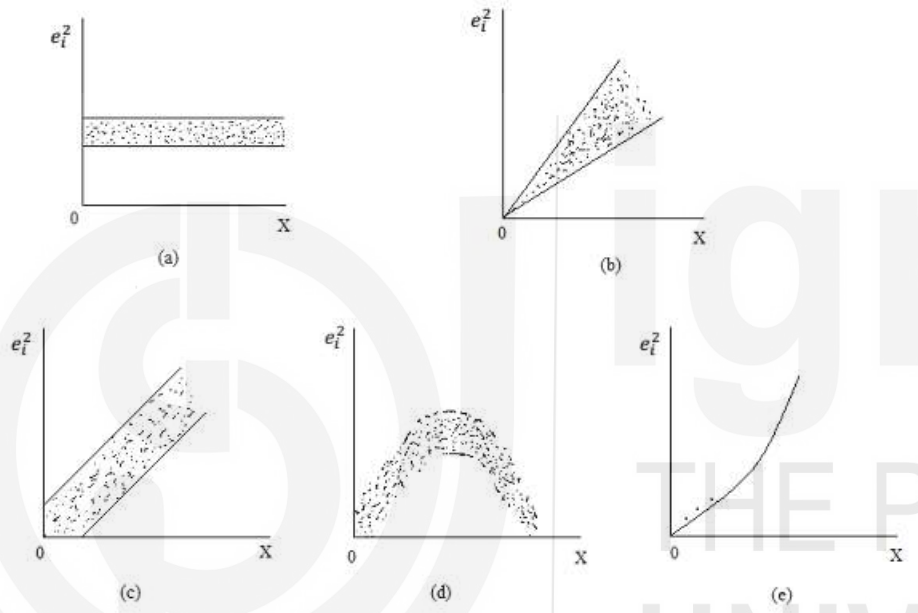

Fig. 11.3: Cases of Homoscedasticity and Heteroscedasticity

In the five situations depicted in Fig. 11.3, we see that Case (a) represents homoscedasticity, i.e., $V(u_i) = \sigma^2$ whereas in the remaining four cases viz., (b), (c), (d) and (e) represent heteroscedasticity, i.e., $V(u_i) = \sigma_i^2$.

11.4.2 Park-Test

If there is heteroscedasticity in a data set, the heteroscedastic variance σ_i^2 may be systematically related to one or more explanatory variables. Therefore, we can regress σ_i^2 on one or more explanatory variables such as

$$\sigma_i^2 = f(X_i)$$

$$\ln\sigma_i^2 = \beta_1 + \beta_2 \ln X_i + v_i \quad \ldots (11.5)$$

In equation (11.5), a non-linear (double-log) regression is run to establish a relationship between the error variance and the explanatory variable with v_i

taken as the residual term. When σ_i^2 are not known, we take the residual term e_i as proxies for u_i. Therefore, we have

$$\ln e_i^2 = \beta_1 + \beta_2 \ln X_i + v_i \quad \ldots (11.6)$$

Now, Park test for detecting heteroscedasticity involves the following steps:

a) Run the original regression in equation (11.5) despite the heteroscedasticity problem.

b) From the regression obtain e_i and square them. Then take the logs of e_i^2.

c) Run the double-log form regression as indicated in equation (11.6) using an explanatory variable in the original model (in the case of more than one explanatory variable). Then run the regression against each *X* variable. In other words, we run the regression against $\hat{Y}_i$, the estimated value of Y_i.

d) Test the null hypothesis $\beta_2 = 0$, i.e., there is no heteroscedasticity.

e) A statistically significant relationship implies that the null hypothesis of no heteroscedasticity is rejected. It suggests the presence of heteroscedasticity which requires remedial measures.

f) If the null hypothesis is not rejected, then it means we accept $\beta_2 = 0$ and the value of β_1, that is, the value of the intercept can be accepted as the common, homoscedastic variance σ^2.

11.4.3 Glejser Test

The Glejser Test is similar to the Park Test. The steps to carry out the Glejser test are as follows:

a) Obtain the residual e_i from the original model.

b) Take absolute value $|e_i|$ of the residuals

c) Regress the absolute values of $|e_i|$ on the *X* variable that is expected to be closely associated with heteroscedastic variance σ_i^2.

d) You can take various functional forms of X_i. Some of the functional forms suggested by Glejser are

$$|e_i| = \beta_1 + \beta_2 X_i + v_i \quad \ldots (11.7)$$

$$|e_i| = \beta_1 + \beta_2 \sqrt{X_i} + v_i \quad \ldots (11.8)$$

$$|e_i| = \beta_1 + \beta_2 \left(\frac{1}{X_i}\right) + v_i \quad \ldots (11.9)$$

The above means that the Glejser test suggests various plausible (linear as well as non-linear) relationships between the residual term and the explanatory variable to investigate the presence of heteroscedasticity.

e) For each of the cases given, test the null hypothesis that there is no heteroscedasticity, i.e., $H_0: \beta_2 = 0$ (no heteroscedasticity).

f) If H_0 is rejected we conclude that there is evidence of heteroscedasticity.

You should note that the error term v_i can itself be heteroscedastic as well as serially correlated. Thus, in the case of Glesjer test also, we follow the same steps as in the Park Test. The difference between the two tests is in the functional forms to be considered.

11.4.4 White's General Test

Let us consider the following PRF:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i \quad \ldots (11.10)$$

The steps to carry out White's general test for heteroscedasticity are as follows:

a) Estimate the population regression equation (11.10) by OLS and obtain the residuals e_i.

b) Find the square of the residuals e_i^2.

c) Run the following auxiliary regression:

$$e_i^2 = A_1 + A_2 X_{2i} + A_3 X_{3i} + A_4 X_{2i}^2 + A_5 X_{3i}^2 + A_6 X_{2i} X_{3i} + v_i \ldots (11.11)$$

d) Obtain the coefficient of determination R^2 from the auxiliary regression under the null hypothesis that there is no heteroscedasticity (i.e., all the slope coefficient are zero). That is,

$$H_0: A_2 = A_3 \ldots A_6 = 0 \quad \ldots (11.12)$$

The null hypothesis given at equation (11.12) implies that all the partial slope coefficients are simultaneously zero. Note that we do not include the intercept term A_1 in equation (11.12).

e) Test the null hypothesis in equation (11.12) by using the chi-square distribution as follows:

$$nR^2 \sim \chi_{k-1}^2 \quad \ldots (11.13)$$

Equation (11.13) tells us that the product of sample size (n) and the coefficient of determination (R^2) follows χ^2 distribution with degrees of freedom (k–1). Here k is the number of regressors in the auxiliary regression (equation 11.11).

f) If $\chi_{calculated}^2 > \chi_{critical}^2$ we reject the H_0, and conclude that the null hypothesis of homoscedasticity is to be rejected, i.e., there is heteroscedasticity. Alternatively, we can also decide on the basis of the p value (readily given by econometric softwares). If the p value is < 0.05, we reject H_0. If $\chi_{calculated}^2 < \chi_{critical}^2$. On the other hand, if $p > 0.05$ we do not reject the null hypothesis of no heteroscedasticity. This implies the existence of homoscedasticity.

11.4.5 Goldfeld-Quandt Test

The Goldfeld-Quandt (G-Q) test is applicable if heteroscedasticity is related to only one of the explanatory variables. Let us assume that the error variance σ_i^2 is related to one of the explanatory variables (say, X_i) in the regression model.

Suppose σ_i^2 is positively related to X_i as given below.

$$\sigma_i^2 = \sigma^2 X_i^2 \qquad \ldots (11.14)$$

In order to carry out the G-Q test we proceed as follows:

a) Arrange the observations in increasing order of X_i

b) Omit some of the observations (say, C out of the no observations in the sample) in the middle of the series. There is no hard and fast rule for the exact value of C and the choice is quite arbitrary. In practice about one fourth observations are omitted.

c) Run a regression on the first $n_1 = (n - C)/2$ observations. Find out the error sum of squares for this regression, i.e., ESS_1.

d) Run a regression on the last $n_2 = (n - C)/2$ observations. Find out the error sum of squares for this regression, i.e., ESS_2.

e) Take the following null hypothesis:

$$H_0: \sigma_i^2 = \sigma^2 \qquad \ldots (11.15)$$

f) Find out the ratio:

$$\lambda = \frac{RSS_1 / \frac{n_1 - C - 2k}{2}}{RSS_2 / \frac{n_2 - C - 2k}{2}}$$

In case $n_1 = n_2$, the above ratio becomes

$$\lambda = \frac{RSS_1}{RSS_2} \qquad \ldots (11.6)$$

The above ratio (λ) follows F-distribution with degrees of freedom

$$\left(\frac{n_1 - C - 2k}{2}, \quad \frac{n_2 - C - 2k}{2} \right) \qquad \ldots (11.17)$$

g) We compare the value of λ obtained above with the tabulated value of F given at the end of the book. If $\lambda > F_{critical}$ we reject $H_0: \sigma_i^2 = \sigma^2$and conclude that there is heteroscedasticity in error variance. It implies $\sigma_i^2 \neq \sigma^2$. If $\lambda < F_{critical}$ we do not reject H_0. We conclude that there is homoscedasticity in error variance, i.e., $\sigma_i^2 = \sigma^2$.

Check Your Progress 3

1) State the steps in conducting the Park test for detection of heteroscedasticity.

.....

.....

.....

.....

.....

11.5 REMEDIAL MEASURES OF HETEROSCEDASTICITY

Heteroscedasticity means that the OLS estimators are unbiased but no longer efficient; not even in large samples. Therefore, if heteroscedasticity is present, it is important to seek remedial measures. For proceeding with remedial measures, it is important to know if the true error variance σ_i^2 is known or not. In such cases, use of a 'deflator' may help rectify the problem of heteroscedasticity. We will learn about the use of deflators in this section.

11.5.1 Case I: σ_i^2 is Known

If we know σ_i^2, we can use the method of Weighted Least Squares (WLS). We explain the procedure of carrying out WLS below.

Let us consider the two-variable Population Regression Function (PRF).

$$Y_i = \beta_1 + \beta_2 X_i + u_i \qquad \ldots(11.18)$$

Let us assume that u_i has heteroscedastic error variance. Here, since the true variance is known, we can use it to divide the equation (11.18) by σ_i. By dividing both sides of (11.18) by σ_i, we obtain:

$$\frac{Y_i}{\sigma_i} = \beta_1 \left(\frac{1}{\sigma_i}\right) + \beta_2 \left(\frac{X_i}{\sigma_i}\right) + \frac{u_i}{\sigma_i} \qquad \ldots (11.19)$$

Note that the error term gets transformed due to the division by σ_i. Let the new error term be v_i. Squaring the new error term we get:

$$v_i^2 = \frac{u_i^2}{\sigma_i^2} \qquad \ldots(11.20)$$

Since the variance of error term is given by $var(v_i) = E(v_i^2)$, taking the expectation of both sides of the equation (11.20) we get:

$$E(v_i^2) = E\left(\frac{u_i^2}{\sigma_i^2}\right)$$

$$= \left(\frac{1}{\sigma_i^2}\right).E(u_i^2)$$

$$= \frac{\sigma_i^2}{\sigma_i^2} = 1$$

Thus, the transformed error-term v_i is homoscedastic. Therefore, equation (11.19) can be estimated by the usual OLS method. The OLS estimators of β_1 and β_2 thus obtained are called the Weighted Least Squares (WLS) estimators.

11.5.2 Case II: σ_i^2 is Unknown

When the error variance σ_i^2 is not known, we need to make further assumptions to use the WLS method. Here, we consider the following two cases.

(i) **Error variance σ_i^2 is Proportional to X_i**

In this case, we follow what is called as the square root transformation. The proportionality assumption means that:

$E(u_i^2) = \sigma^2 X_i$

Or, $V(u_i) = \sigma^2 X_i$ … (11.21)

Now, the square root transformation requires that we divide both sides of equation (11.18) by $\frac{1}{\sqrt{X_i}}$ to get:

$$\frac{Y_i}{\sqrt{X_i}} = \beta_1 \frac{1}{\sqrt{X_i}} + \beta_2 \frac{X_i}{\sqrt{X_i}} + \frac{u_i}{\sqrt{X_i}}$$

$$= \beta_1 \frac{1}{\sqrt{X_i}} + \beta_2 \sqrt{X_i} + v_i \quad \text{… (11.22)}$$

where $v_i = \frac{u_i}{\sqrt{X_i}}$ … (11.23)

The error term in equation (11.23) is a transformed error term. In order to see whether v_i is devoid of heteroscedasticity, we square both the sides of equation (11.23) to get:

$v_i^2 = \frac{u_i^2}{X_i}$ … (11.24)

Now, the variance of the transformed error term, i.e., equation (11.24) is:

$E(v_i^2) = \frac{E(u_i^2)}{X_i} = \frac{\sigma^2 X_i}{X_i}$ ... (11.25)

$= \sigma^2 \Rightarrow$ homoscedasticity

Thus, when we apply the square root transformation ($v_i = \frac{u_i}{\sqrt{X_i}}$), we could make the error variance to become homoscedastic.

(ii) Error Variance is Proportional to X_i^2

Here, we have:

$E(u_i^2) = \sigma X_i^2$ … (11.27)

$$V(u_i) = \sigma X_i^2$$

Dividing both sides of equation (11.18) by X_i,

$$\frac{Y_i}{X_i} = \beta_1\left(\frac{1}{X_i}\right) + \beta_2 + \left(\frac{u_i}{X_i}\right)$$

$$= \beta_1\left(\frac{1}{X_i}\right) + \beta_2 + v_i \quad \ldots (11.28)$$

Equation (11.28) is the transformed PRF in which the error term is:

$$v_i = \frac{u_i}{X_i}, \quad \ldots (11.29)$$

Squaring both the sides of equation (11.29), we get:

$$v_i^2 = \frac{u_i^2}{X_i^2} \quad \ldots (11.30)$$

The variance of the error term of the transformed equation in (11.30) is homoscedastic because:

$$E(v_i^2) = \frac{E(u_i^2)}{X_i^2} = \frac{\sigma X_i^2}{X_i^2} = \sigma \quad \ldots (11.31)$$

11.5.3 Re-Specification of the Model

Instead of speculating about σ_i^2, sometimes choosing a different functional form can reduce heteroscedasticity. For instance, instead of running the usual regression model, we can estimate the model in its log form.

$$\ln Y_i = \beta_1 + \beta_2 \ln X_i + u_i \quad \ldots (11.32)$$

In many cases transforming original model as above will take care of the problem of heteroscedasticity.

We used the word 'deflator' in the beginning of this section. The cases we have considered above basically involve dividing both sides of the original regression model by a known value to transform the variables. Such transformation of variables by division amounts to deflating the original values. The known values used to perform the division act are known as the 'deflators'.

Check Your Progress 4

1) How does the use of deflators work as a solution for the problem of heteroscedasticity?

.....

.....

.....

.....

2) Explain how the usage of deflators serve to tackle the problem of heteroscedasticity when the error variance is proportional to X_i^2.

.....

.....

.....

.....

.....

11.6 LINEAR VERSUS LOG – LINEAR FORMS

The regression model can be run in various functional forms depending upon: (i) the relationship of dependent and independent variable, and (ii) the data. Suppose there is a choice of running two types of regression models: (i) a linear regression model, and (ii) a log-linear model. To help decide in such cases, a test for the selection of the appropriate functional form for regression is proposed by Mackinnon, White and Davidson (MWD). The MWD test is applied as follows:

Let there be two distinct functional forms of a regression like:

Model 1: $Y_i = \beta_1 + \beta_2 X_i + u_i$ …….(11.33)

Model 2: $\ln Y_i = \beta_1 + \beta_2 lnX_i + u_i$ …….(11.34)

In Model 1, the dependent variable is linearly related to one (or more than one) of the *X*s. In Model 2, the relationship between the dependent and independent variable is non-linear. The MWD test involves considering a null and an alternate hypothesis as follows:

H_0: Linear Model, i.e., *Y* is a linear function of regressors (equation (11.33))

H_1: Log- Linear Model, i.e., ln *Y* is a linear function of lnX_i (equation (11.34))

Following are the steps for carrying out the MWD test:

(i) Estimate the linear model and obtain the estimated *Y* values. Let the estimated *Y* values be denoted as Y_f.

(ii) Estimate the log-linear model and obtain the estimated ln*Y* values. Let the estimated values of the log-linear *Y* be denoted as $\ln Y_f$.

(iii) Obtain $Z_1 = (\ln Y_f - Y_f)$

(iv) Regress *Y* on X_s and Z_1 obtained in Step (iii) Reject H_0 if the coefficient of Z_1 is statistically significant by the usual *t*-test.

(v) Obtain $Z_2 = (\text{antilog } \ln Y_f - Y_f)$

(vi) Regress log of Y on the logs of X_s and Z_2. Reject H_1 if the coefficient of Z_2 is statistically significant by the usual t-test.

Suppose the linear model I in equation (11.33) is in fact the correct model. In that case, the constructed variable Z_1 should not be statistically significant in Step (iv). For, in that case the estimated Y values from the linear model and those estimated from the log-linear model (after taking their antilog values for comparative purposes) in equation (11.34) should not be different. The same logic applies to the alternative hypothesis H_1.

Check Your Progress 5

1) Outline the MWD test for choosing the appropriate functional form of the regression model between its linear and log-linear forms.

.....

.....

.....

.....

.....

11.7 LET US SUM UP

In this Unit, we have discussed the concept of heteroscedasticity in regression models. The unit outlines the consequences of the presence of heteroscedasticity and the methods of its detection. Various techniques to provide remedial measures are explained in the unit. The remedial measures involve understanding of the use of deflators. The unit has also explained a method for the choice of selecting the functional form by way of the MWD test.

11.8 ANSWERS/ HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) A crucial assumption of the Classical Linear Regression Model CLRM is that the error term u_i is population regression function (PRF) is homoscedastic, i.e., they have the same variance σ^2. However, if the variance of u_i is σ_i^2 (in other words, it varies from one observation to another), then the situation is referred to as heteroscedasticity.

2) Heteroscedasticity is usually found is cross-sectional data and not in time series data. This is because, in the case of cross-sectional data, the members of population are in the form of individual firms, industries, geographical division, state or countries. The data collected for such units at a point of time from the members of population may be of different sizes: small, medium or large firms. This is referred to as scale effect.

Due to the scale effect, in cross-sectional data, there is a greater chance of coming across heteroscedasticity in the error terms.

Check Your Progress 2

1) The OLS estimators are unbiased but they no longer have minimum variance, i.e., they are no longer efficient. Even in large samples the OLS estimators are not efficient. Therefore, the OLS estimators are not BLUE in small as well as large samples (asymptotically).

   The usual formula for estimating the variances of OLS estimator are biased i.e. there is either upward bias (positive bias) or downward bias (negative bias).

2) The OLS estimator of error variance is a biased estimator. Thus it will either overestimate or underestimate. In fact, the OLS estimator of error variance is inefficient, thereby meaning that it is very high; thus it is always an overestimate.

Check Your Progress 3

1) In the presence of heteroscedasticity, the heteroscedastic variance σ_i^2 may be systematically related to one or more explanatory variables. Therefore, we can regress σ_i^2 on one or more of *X*- variables as:

   $\sigma_i^2 = f(X_i)$ or $\ln\sigma_i^2 = \beta_1 + \beta_2 \ln X_i + v_i$

   where $v_i = new$ residual term. If σ_i^2 are not known, estimated e_i can be used as proxies for u_i. A statistically significant relationship implies that the null hypothesis of no heteroscedasticity is rejected suggesting the presence of heteroscedasticity which requires remedial measures. If null hypothesis is not rejected then it means we accept $\beta_2 = 0$ and value of β_1 can be taken as the common, homoscedastic variance σ^2.

2) Heteroscedasticity means that the OLS estimators are unbiased but estimators are no longer efficient, not even in large samples. This lack of efficiency makes the conventional hypothesis testing of OLS estimators unreliable. For remedial measures, it is important to know whether the true error variance σ_i^2 is known or not. In such cases, use of deflators will help rectify the problem of heteroscedasticity. Various deflators can be used to convert the error variance ti make them homoscedastic.

   When σ_i^2 is known, the method of Weighted Least Squares (WLS) can be considered. In this, the error variance σ_i^2 is used to divide both sides of the equation by σ_i. See Section 11.5 for details.

3) The estimated residuals show a pattern similar to earlier case I, but error variance is not linearly related to *X* but increases proportional to square of X. Hence, $E(u_i^2) = \sigma X_i^2$ and $V(u_i) = \sigma X_i^2$. Dividing both sides by X_i, we get:

   $$\frac{Y_i}{X_i} = \beta_1\left(\frac{1}{X_i}\right) + \beta_2 + \left(\frac{u_i}{X_i}\right)$$

   $$= \beta_1\left(\frac{1}{X_i}\right) + \beta_2 + v_i$$

$v_i = \frac{u_i}{X_i}, v_i^2 = \frac{u_i^2}{X_i^2}$

$E(v_i^2) = \frac{E(u_i^2)}{X_i^2} = \frac{\sigma X_i^2}{X_i^2} = \sigma$

Thus, the transformed equation is homoscedastic.

Check Your Progress 5

1) The test for selection of the appropriate functional form for regression as proposed by Mackinnon, White and Davidson is known as MWD Test. The MWD test is used to choose between the two models. See Section 11.6 for details.

UNIT 12 AUTOCORRELATION *

Structure

12.0 OBJECTIVES

After going through this unit, you should be able to:

- outline the concept of autocorrelation in a regression model;
- describe the consequences of presence of autocorrelation in the regression model;
- explain the methods of detection of autocorrelation;
- discuss the procedure of carrying out the Durbin-Watson test for detection of autocorrelation;
- elucidate the remedial measures for resolving autocorrelation; and
- outline the procedure of dealing with situations where autocorrelation exists in models with a lagged dependent variable.

12.1 INTRODUCTION

In the previous unit, you studied about heteroscedasticity. You saw that heteroscedasticity is a violation of one of the assumptions of the Classical Linear Regression Model (CLRM), viz., homoscedasticity. If the variance of the error term is not constant across all observations, then we have the problem of heteroscedasticity. In this unit, we discuss about the violation of another assumption of the CLRM. Recall that one of the assumptions about the error

* Dr. Pooja Sharma, Assistant Professor, Daulat Ram College, University of Delhi

terms is that the error term of one observation is not correlated with the error term of another observation. If they are correlated, then the situation is said to be one of autocorrelation. This is also called as the problem of serial correlation. This can be present in both cross-section as well as time series data. Let us discuss the concept of autocorrelation in a little more detail.

12.2 CONCEPT OF AUTOCORRELATION

The classical linear regression model (CLRM) assumes that the correlation among various error terms is zero. We know that heteroscedasticity is associated more with cross sectional data. Autocorrelation is usually more associated with time series data. Of course, autocorrelation can be present even in cross-section data. Some authors use the term autocorrelation only for time-series data. They use the term 'serial correlation' for describing autocorrelation in cross-section data. Many authors use the terms autocorrelation and serial correlation as synonyms. They use the term across both cross-section as well as time-series data.

Autocorrelation occurring in cross-sectional data is also sometimes called spatial correlation (correlation in space rather than in time). In CLRM we assume that there is no autocorrelation. This implies:

$$E(u_i, u_j) = 0 \qquad i \neq j \qquad \ldots(12.1)$$

Equation (12.1) means that the stochastic error term associated with one observation is not related to or influenced by the disturbance term associated with any other observation. For instance, the labour strike in one quarter affecting output may not affect the output in the next quarter. This implies there is no autocorrelation in the time series. Similarly, in a cross-section data of family consumption expenditure, the increase in one family's income on consumption expenditure in not expected to affect the consumption expenditure of another family. In the example of output affected due to labour strike above, if $E(u_i, u_j) \neq 0,\ i \neq j$, this implies a situation of autocorrelation. This means the disruption caused by the strike in one quarter is affecting the output in the next quarter. Similarly, increase in consumption expenditure of one family may influence the consumption expenditure of other families in the neighbourhood due to the 'demonstration effect' (cross-sectional data). It is thus more a case of spatial correlation. It is therefore important to analyse the data carefully to bring out what exactly is causing the correlation among the disturbance terms. Let us see more carefully the different situations or cases of autocorrelation as depicted in Fig.12.1. In panels (a) to (d) of Fig. 12.1 we find distinct pattern among u_t . In panel (e) of Fig. 12.1 we do not see any such pattern. Note that since autocorrelation is seen mostly in time series data, we use the subscript '*t*' in place of '*i*' to indicate individual observations. Let us now study the reasons of autocorrelation with some specific examples from economics.

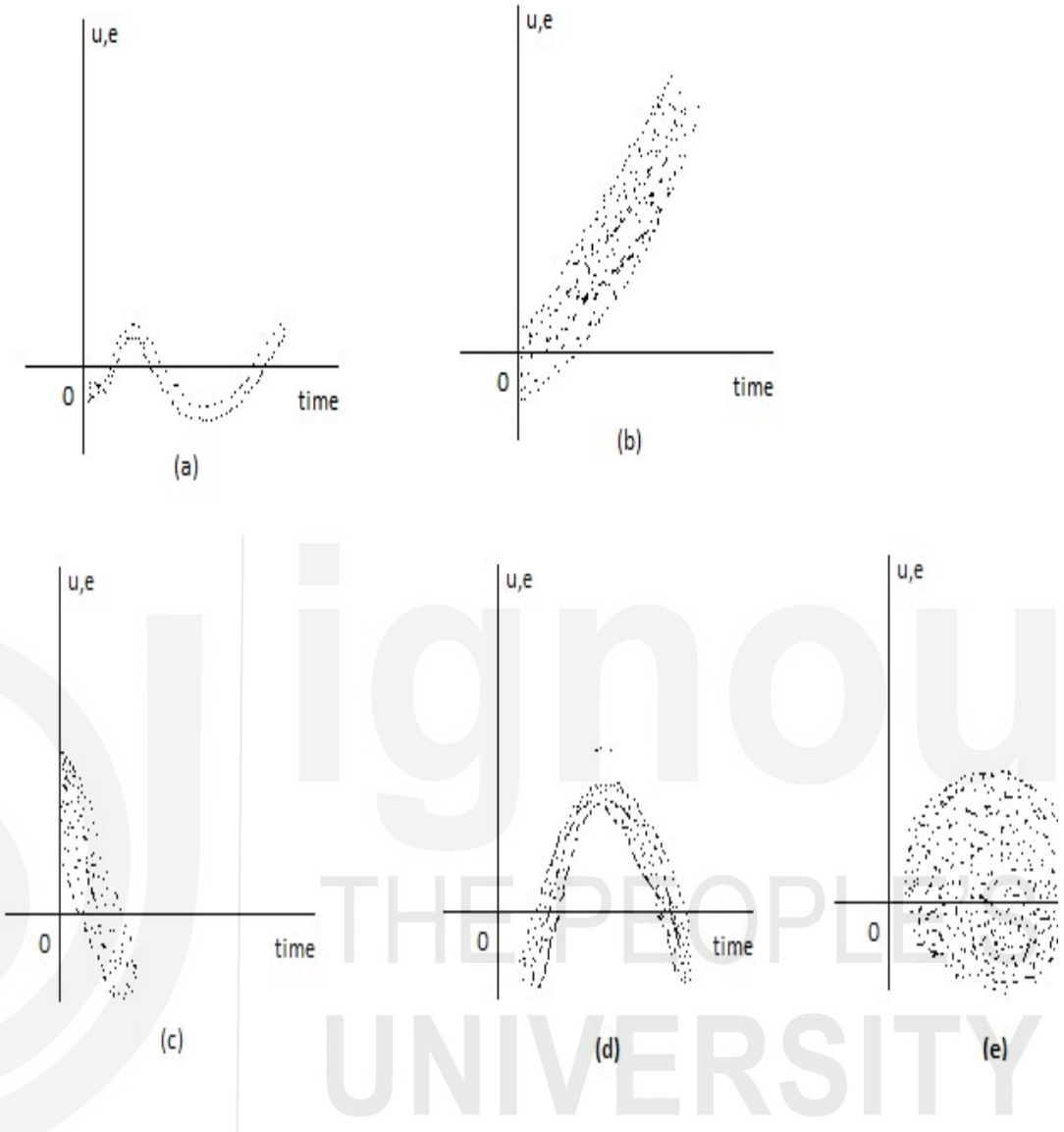

Fig. 12.1: Cases of Autocorrelation

12.3 REASONS FOR AUTOCORRELATION

The various reasons for the presence of autocorrelation can be discussed under the following broad heads.

(a) Inertia or Sluggishness

Most of the economic time series data displays inertia or sluggishness. For instance, gross domestic product (GDP), production, employment, money supply, etc. reflect recurring and self-sustaining fluctuations in economic activity. When an economy is recovering from recession, most of the time series will be moving upwards. This means any subsequent value of a series at one point of time is always greater than its previous time value.

Such a momentum continuous till it slows down due to, say, a factor like increase in taxes or interest or both. Hence, in regressions involving time series data, successive observations would generally be inter-dependent or correlated. Such an uptick effect is termed as 'inertia' which literally means a situation that continues to hold in a similar manner for many successive time periods. We see its opposite effect in periods of recession when most of the economic activity will be suffering, i.e., will be sluggish.

(b) Specification Error in the Model

By an incorrect specification of model, certain important variables that should be included in the model may not be included (i.e. a case of under-specification). If such model-misspecification occurs, the residuals from such an incorrect model will exhibit systematic pattern. If the residuals show a distinct pattern, it gives rise to serial correlation.

(c) The Cobweb Phenomenon

Many agricultural commodities reflect what is called as a 'cobweb phenomenon'. In this, supply reacts to price with a lag of time. This is mainly because supply decisions take time to implement. In other words, there is a gestation period involved. For instance, farmers' decision to plant crop might depend on the prices prevailing in the previous year's supply position or function. This can be written as:

$$S_t = \beta_1 + \beta_2 P_{t-1} + u_t \qquad \ldots (12.2)$$

In (12.2), the error term u_t may not be purely random. This is because, if the farmers over-produce in year t, they are likely to under-produce in year $(t + 1)$ since they want to clear away the unsold stock. This usually leads to a cobweb pattern.

(d) Data Smoothing

Sometimes we need to average the data presented. Considering averages implies 'data smoothing'(see Unit 5 of BECC 109 for an example). We may prefer to convert monthly data into quarterly data by averaging the data over every three months. However, this smoothness, desired in many contexts, may itself lead to a systematic pattern in disturbances, resulting in autocorrelation.

Autocorrelation may be positive or negative depending on the data. Generally, economic data exhibits positive autocorrelation. This is because most of them either move upwards or downwards over time. Such a trend continues at least for some time i.e. some months, or quarters. This means, they are not generally expected to exhibit a sudden upward or downward movement unless there is a reason or a shock.

12.4 CONSEQUENCES OF AUTOCORRELATION

When the assumption of no-autocorrelation is violated, the estimators of the regression model based on sample data suffers from certain consequences. More specifically, the OLS estimators will suffer from the following consequences.

a) The least squares estimators are still linear and unbiased. In other words, the estimated values of parameters continue to be unbiased. However, they are not efficient because they do not have minimum variance. Therefore, the usual OLS estimators are not BLUE (best linear unbiased estimators).

b) The estimated variances of OLS estimators $(b_1 \text{ and } b_2)$ are biased. Hence, the usual formula used to estimate the variances, and their standard errors underestimate the true variances and standard errors. Consequently, the decision of rejecting a parameter on the basis of *t*-values, concluding that a particular coefficient is statistically different from zero, would be an incorrect conclusion. In other words, the usual *t* and *F* tests become unreliable.

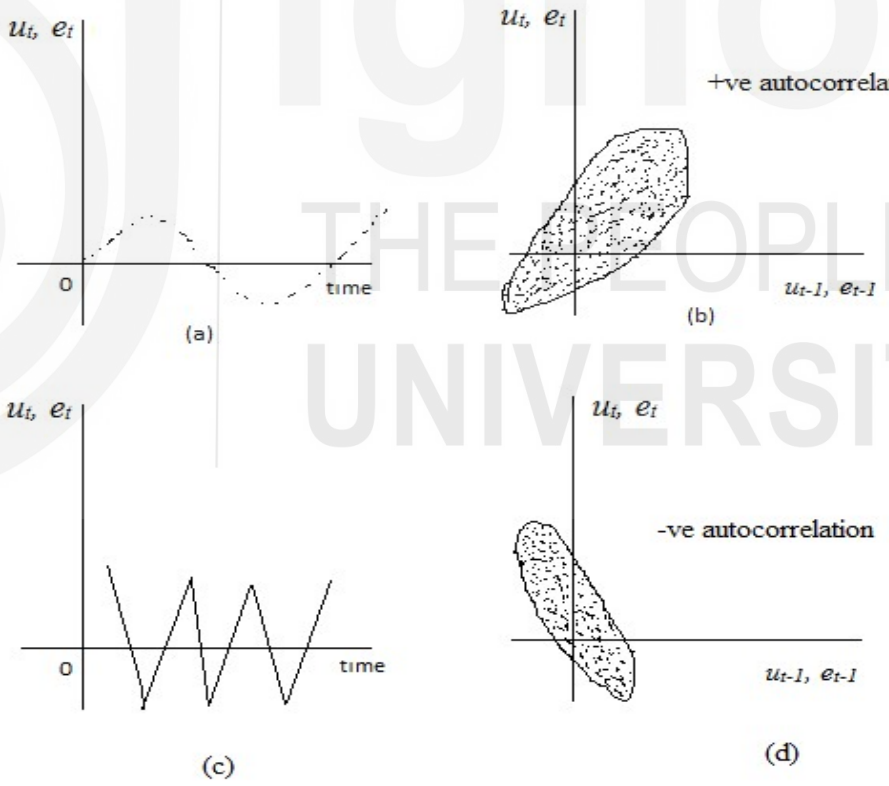

Fig. 12.2: Patterns of the Error Term in Autocorrelation

a) As a direct consequence of the above, the usual formula for estimating the population error variance, viz., $\hat{\sigma}^2 = (\text{RSS}/\text{df})$ yields a biased estimator

of true σ^2. In particular, it underestimates the true σ^2. As a consequence, the computed R^2 becomes an unreliable measure of true R^2.

Fig. 12.2 shows the pattern of error terms under different situations of autocorrelation. Note that since the population error terms (u_t) are not known, we are plotting the sample residuals (e_t).

Check Your Progress 1 [Answer the questions in 50-100 words within the space given]

1) What is meant by autocorrelation in a regression model?

……

……

……

……

……

2) In which type of data the problem of autocorrelation is more common? Why?

……

……

……

……

……

3) State the broad reasons for autocorrelation.

……

……

……

……

……

4) Enumerate the consequences of autocorrelation.

……

……

……

……

……

12.5 DETECTION OF AUTOCORRELATION

There are many methods of detecting the presence of autocorrelation. Let us discuss them now.

12.5.1 Graphical Method

A visual examination of OLS residuals e_t quite often conveys the presence of autocorrelation among the error terms u_t. Such a graphical presentation (Fig. 12.3) is known as the 'time sequence plot'. The first part of this figure does not show any clear pattern in the movement of the error terms. This means there is an absence of autocorrelation. In the lower part of Fig. 12.3, you will notice that the correlation between the two residual terms is first negative and then becomes positive. Therefore, plotting the sample residuals gives us the first indication on the presence or absence of autocorrelation.

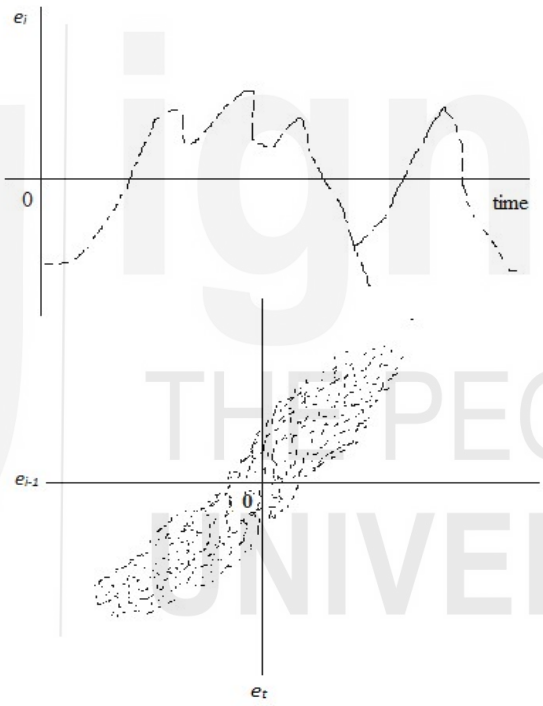

Fig. 12.3: Graphical Method for Detection of Autocorrelation

12.5.2 Durbin-Watson Test

The Durbin-Watson test, or the DW test as it is popularly called, is an analytical method of detecting the presence of autocorrelation. Its statistic is given by:

$$d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} \quad \ldots (12.3)$$

Equation (12.3) defines the *d*-statistic suggested by Durbin-Watson as the ratio of the sum of squared differences in the successive residuals to the residual sum of squares. For computing the *d*-statistic, we take the sample size to be (*n*–1) since

one observation is lost in taking the successive differences. There are certain assumptions underlying the *d*-statistic. These are:

(a) The regression model includes an intercept term. Therefore, this method cannot be used to determine autocorrelation in regression models without the intercept term (i.e. regression equation which passes through the origin).

(b) The *X* variables are non-stochastic, i.e., their values are fixed in repeated samples.

(c) The error term evolves as follows :

$$u_t = \rho u_{t-1} + v_t, \qquad -1 \leq \rho \leq 1 \qquad \ldots (12.4)$$

Equation (12.4) states that the value of error term at time period *t* is dependent on the value of the error term in time-period (*t*–1) and a purely random term v_t. The extent of dependence on past value is measured by ρ which lies between –1 and 1.

The regression model given in equation (12.4) is referred to as the first-order auto-regression scheme. It is denoted by *AR(1)*. The usage of the term 'autoregressive' implies that the error term u_t is regressed on its own lagged value of one period, i.e., u_{t-1}. It is therefore called the first-order autoregressive scheme. If we include 2 lagged values (i.e., u_{t-1} and u_{t-2}) then we have the *AR(2)* scheme. Likewise, when we extend the number of lagged values to '*p*', we have the *AR(p)* scheme.

(d) The regression model does not contain any lagged value of the dependent variable as one of the explanatory variables. In other words, the test is not applicable to models like:

$$Y_t = \beta_1 + \beta_2 X_t + \beta_3 Y_{t-1} + u_t \qquad \ldots (12.5)$$

where Y_{t-1} is the one-period lagged-value of the dependent variable *Y*. Models of the above type are known as auto-regressive (AR) models. For such cases, the *d*-statistic cannot be used.

We can estimate ρ from equation (12.4) as follows:

$$\hat{\rho} = \frac{\sum_{t=2}^{n} e_t e_{t-1}}{\sum_{t=1}^{n} e_t^2}$$

[Recall that the estimator of b_2 in the two variable regression model is $b_2 = \frac{\Sigma x_i y_i}{\Sigma x_i^2}$.

We apply the same logic to derive $\hat{\rho}$ above]

We can expand equation (12.3) to obtain

$$d = \frac{\sum e_t^2 + \sum e_{t-1}^2 - 2\sum e_t e_{t-1}}{\sum e_t^2}$$

The above can be approximated to

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

We can take an approximate value of d as:

$$d \approx 2(1 - \hat{\rho}) \quad \ldots (12.6)$$

where the symbol $\approx$ denotes 'approximately'. In equation (12.6), $\hat{\rho}$ is an estimator of the first order autocorrelation scheme. Table 12.1 presents the value of the d-statistic for different values of $\hat{\rho}$.

From Table 12.1 we find that $0 \leq d \leq 4$. The Durbin-Watson statistic thus provides a lower limit d_L and an upper limit d_U. The computed value of d is therefore a value between 0 and 4. From such a value, we can infer on the nature of autocorrelation as follows:

a) If d is closer to 0, there is evidence of positive autocorrelation.

b) If d is closer to 2, there is evidence of no autocorrelation.

c) If d is closer to 4, there is evidence of negative autocorrelation.

Table 12.1: Value of *d*-Statistic according to $\hat{\rho}$

Value of $\hat{\rho}$	Implication	Value of *d*-statistic
$\hat{\rho} = -1$	Perfect negative autocorrelation	4
$\hat{\rho} = 0$	No autocorrelation	2
$\hat{\rho} = 1$	Perfect positive autocorrelation	0

The steps in applying the DW test are therefore the following:

1. Run the OLS regression and obtain the residuals e_t.
2. Compute d as:

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

3. Find out the critical Table values d_L and d_U for given sample size and given number of explanatory variables.

Follow the decision rule, as depicted in Fig. 12.4.

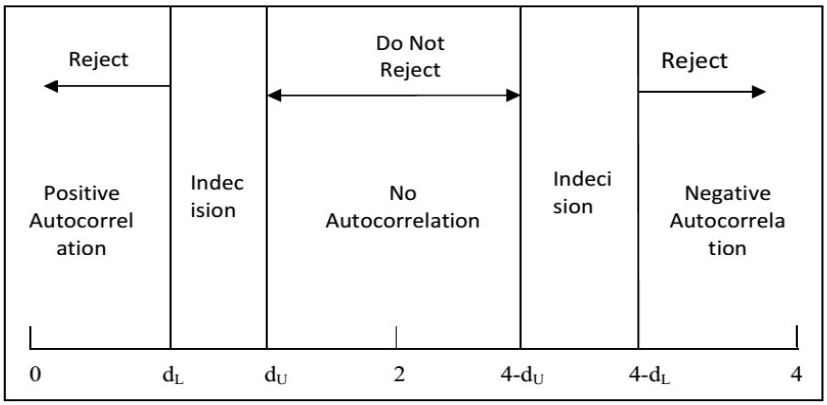

Fig. 12.4: Range of Values of Durbin-Watson Statistic

One drawback of the *d*-test is that it has two zones of indecision viz. $d_L < d < d_U$ and ($4 - d_U < d < 4 - d_L$).

12.5.3 The Breusch-Godfrey (BG) Test

To avoid the pitfalls of the Durbin Watson *d*-test, Breusch and Godfrey have proposed a test criterion for autocorrelation that is general in nature. This is in the sense that:

(a) It can handle non-stochastic regressors as well as the lagged values of Y_t ;

(b) It can deal with higher-order autoregressive schemes such as AR(2), AR(3) … etc.

(c) It can also handle simple or higher order moving averages.

The BG-Test is also referred to as the LM (Lagrange Multiplier) Test (see Unit 8). Let us now consider a two-variable regression model to see how the BG test works.

$$Y_t = \beta_1 + \beta_2 X_t + u_t \quad \ldots (12.7)$$

where u_t follows a P^{th} order auto regressive scheme AR(P) like:

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \ldots + \rho_\rho u_{t-p} + v_t \quad \ldots (12.8)$$

where v_t is the white noise or the stochastic error term. We wish to test:

$$H_0: \rho_1 = \rho_2 = \ldots \rho_p = 0 \quad \ldots (12.9)$$

The null hypothesis says that there is no autocorrelation of any order. Now, the BG test involves the following steps:

(a) Estimate the model $Y_t = \beta_1 + \beta_2 X_t + u_t$ by OLS method and obtain the residuals e_t.

(b) Regress the residuals e_t on the *p*-lagged values of estimated residuals obtained in step (a) above, i.e., $e_{(t-1)}$, $e_{(t-2)}$, ..., $e_{(t-p)}$ [as in equation (12.8)]. Here we take the residual e_t which are estimate of the error u_t, as the error term is not known.

(c) Obtain R^2 from the auxiliary regression (12.8) in the step (b) above.

(d) Now, for large samples, the Breusch and Godfrey test statistic is computed as:

$$(n-p)R^2 \sim \chi_p^2 \quad \ldots (12.10)$$

It is called LM test, as it has a similar form to the LM test described in Unit 8. The BG test statistic follows chi-squares distribution with p degrees of freedom where *p* is the number of regressors in the auxiliary regression (equation (12.8)).

We draw inferences from the BG test as follows:

(i) If $(n-p)R^2 > \chi^2_{critical}$, we reject H_0 and conclude that at least one ρ is statistically different from zero, i.e., there exists autocorrelation.

(ii) If $(n-p)R^2 < \chi^2_{critical}$, we do not reject H_0 and conclude that there exists no autocorrelation.

Check Your Progress 2 [Answer the questions in 50-100 words within the space given]

1) State the methods of detecting autocorrelation.

……

……

……

……

……

2) Specify the test statistic applied in the DW test.

……

……

……

……

……

3) State the assumptions under which the DW test is valid.

……

……

……

……

……

4) Point out the limitations of the DW test.

……

……

……

……

……

5) In what ways the BG test for autocorrelation is an improvement over the DW test?

……

……

……

……

…….

12.6 REMEDIAL MEASURES FOR AUTOCORRELATION

To suggest remedial measures for autocorrelation, we assume the nature of inter-dependence in the error term u_t in a regression model like:

$$Y_t = \beta_1 + \beta_2 X_t + u_t \quad \ldots (12.11)$$

and that the error term is following an AR (1) scheme like:

$$u_t = \rho u_{t-1} + v_t \qquad -1 \leq \rho \leq 1 \quad \ldots (12.12)$$

where v_t is assumed to follow the OLS assumptions. We first consider the case where ρ is known. Here, transforming the model in a certain manner (called as the Cochrane Orcutt procedure) will reduce the equation to an OLS compatible model. When ρ is not known, we need some simple approaches which help us in overcoming the situation of autocorrelation. Let us study these approaches now.

12.6.1 Autoregressive Scheme is Known: Cochrane-Orcutt Transformation

Suppose we know the value of ρ. This helps us to transform the regression model given at (12.11) in a manner that the error term becomes free from autocorrelation. Subsequently, we apply the OLS method to the transformed model. For this, we consider a one-period lag in (12.11) as:

$$Y_{t-1} = \beta_1 + \beta_2 X_{t-1} + u_{t-1} \quad \ldots (12.13)$$

Let us multiply equation (12.13) on both the sides by ρ. We obtain:

$$\rho Y_{t-1} = \rho\beta_1 + \rho\beta_2 X_{t-1} + \rho u_{t-1} \quad \ldots (12.14)$$

Let us now subtract equation (12.14) from equation (12.11) to obtain:

$$(Y_t - \rho Y_{t-1}) = \beta_1(1-\rho) + \beta_2(X_t - \rho X_{t-1}) + v_t \quad \ldots (12.15)$$

Note that we have used v_t for the new disturbance term above. Let us now denote:

$$Y_t^* = (Y_t - \rho Y_{t-1})$$

$$X_t^* = (X_t - \rho X_{t-1})$$

$$\beta_1^* = \beta_1(1-\rho)$$

The transformed model will be

$$Y_t^* = \beta_1^* + \beta_2 X_t^* + v_t \quad \ldots (12.16)$$

Now, the transformed variables $Yt^*$and Xt^*will have the desirable BLUE property. The estimators obtained by applying the OLS method to (12.16) are called the Generalized Least Squares (GLS) estimators. The transformation as suggested above is known as the Cochrane-Orcutt transformation procedure.

12.6.2 Autoregressive Scheme is not Known

Suppose we do not know ρ. Thus, we need methods for estimating ρ. We first consider the case where ρ = 1. This amounts to assuming that the error terms are perfectly positively autocorrected. This case is called as the First Difference Method. If this assumption holds, a generalized difference equation can be considered by taking the difference between (12.11) and its first order autoregressive schemes as:

$$Y_t - Y_{t-1} = \beta_2(X_t - X_{t-1}) + v_t \quad \ldots (12.17)$$

i.e., $\Delta Y_t = \beta_2 \Delta X_t + v_t \quad \ldots (12.18)$

where the symbol Δ (read as delta) is the first difference operator. Note that the difference model (12.17) has no intercept. If ρ is not known, then we can estimate ρ by the following two methods.

(i) Durbin Watson Method

From equation (12.6) we see that *d*-statistic and ρ are related. We can this relationship to estimate ρ. The *d*-statistic and ρ are related as:

$$\rho \approx 1 - \frac{d}{2} \quad \ldots (12.19)$$

If the value of *d* is known, then $\hat{\rho}$ can be estimated from the *d*-statistic.

((ii) The OLS Residuals (e_t) Method

Here, we consider the first order autoregression scheme as in (12.12), i.e.,

$u_t = \rho u_{t-1} + v_t$. Since u_t is not directly observable, we use its sample counterpart e_t and run the following regression:

$$e_t = \hat{\rho} e_{t-1} + v_t \quad \ldots (12.20)$$

Note that $\hat{\rho}$ is an estimator of ρ. In small samples, $\hat{\rho}$ is a biased estimator of ρ. As sample size increases, the bias disappears.

12.6.3 Iterative Procedure

This is also called as the Cochrane-Orcutt iterative procedure. We consider the two variable model with the AR(1) scheme for autocorrelation as discussed earlier. That is, we consider: $Y_t = \beta_1 + \beta_2 X_t + u_t$ where $u_t = \rho u_{t-1} + v_t$ with $-1 \leq \rho \leq 1$. We have taken only one explanatory variable for simplicity but we can have more than one explanatory variable too. The iterative procedure suggested by Cochrane-Orcutt has the following steps:

(i) Estimate the equation $u_t = \rho u_{t-1} + v_t$ by the usual OLS method.

(ii) From the above, obtain the residuals e_t.

(iii) Using the residuals e_t, run the regression $e_t = \hat{\rho}\, e_{t-1} + v_t$ and obtain $\hat{\rho}$.

(iv) Use $\hat{\rho}$ obtained in (iii) above to multiply the equation $u_t = \rho u_{t-1} + v_t$.

(v) Now, obtain the generalized difference equation as:

$Y_t^* = \beta_1^* + \beta_2 X_t^* + e_t$ where, $Y_t^* = Y_t - Y_{t-1}$, $X_t^* = X_t - \rho X_{t-1}$ and

$\beta_1^* = \beta_1(1 - \hat{\rho})$

(vi) We are not sure that $\hat{\rho}$ estimated in (iii) above is the best estimate of ρ. Therefore, we repeat the steps (ii) and (iii) to obtain the new residuals e_t^*.

(vii) Now estimate the regression $e_t^* = \hat{\rho} e_{t-1}^* + w_t$to obtain the new estimate of $\hat{\rho}$.

We thus obtain the second-round estimate of ρ. Since we are not sure if the second round estimate of ρ is the best, we go for the third round estimate and so on. We repeat the same steps again and again. Due to this repetitive steps followed, this procedure, suggested by Cochrane-Orcutt, is called the 'iterative procedure'. We stop the iteration when the successive estimates of ρ differ by a small amount (less than 0.01 or 0.005).

12.7 LAGGED DEPENDENT VARIABLE

The Durbin-Watson method is not applicable when the regression model includes lagged value of the dependent variable as one of the explanatory variables. In such models, the *h*-statistic suggested by Durbin is used to identify the presence of autocorrelation in the regression model. Let us consider the regression model as:

$$Y_t = \beta_1 + \beta_2 X_t + \beta_3 Y_{t-1} + v_t \quad \ldots (12.21)$$

In equation (12.21), we have two explanatory variables: X_t and Y_{t-1} with Y_{t-1} as a lagged dependent variable (with one-period lag). For equation (12.21) the *d*-statistic is not applicable to detect autocorrelation. For such models, Durbin suggests replacing the *d*-statistic by the *h*-statistic taken as:

$$h \approx \hat{\rho} = \sqrt{\frac{n}{1 - n\, Var(b_3)}} \quad \ldots (12.22)$$

where, *n* = sample size, $\hat{\rho}$ = the estimator of the autocorrelation coefficient, and

$var(b_3)$ = variance of estimator of β_3, the lagged dependent variable in (12.21).

The null hypothesis is H_0: $\rho = 0$. Durbin has shown that for large samples the *h*-statistic is distributed as h~$N(0,1)$. For normal distribution, we know that the critical value at 5 per cent level of significance is 1.96 and at 1 per cent level of significance it is 2.58. Using this information, we can draw inference from equation (12.22) as follows:

(i) If the computed value of h is greater than the critical value of h, we reject H_0. We interpret the result as existence of no autocorrelation.

(ii) If the computed value of h is less than the critical value of h, we do not reject H_0. We interpret the result as existence of autocorrelation.

Check Your Progress 3 [Answer the questions in 50-100 words within the space given]

1) Outline the transformation procedure suggested by Cochrane-Orcutt to resolve the problem of autocorrelation.

……

……

……

……

……

2) State how the iterative procedure of Cochrane-Orcutt is applied in the case of autocorrelation in a dataset. Why is it called iterative procedure?

……

……

……

……

……

3) What is the advantage of using the h-statistic in regression model having autocorrelation problem?

……

……

……

……

……

11.8 LET US SUM UP

The unit has discussed the concept of autocorrelation in regression models. The consequences of the presence of autocorrelation, its detection and techniques that provide remedial measures for such situations have been explained. The unit also discusses the case of autocorrelation in regression models with lagged dependent variables.

11.9 ANSWERS/ HINTS TO CHECK YOUR PORGRESS EXERCISES

Check Your Progress 1

1) Autocorrelation refers to the presence of correlation between the error terms of any two observations. This means if U_i and U_j are the error terms, then *Corr (Ui, Uj)* $\neq 0$ for $i \neq j$. In the CLRM, one of our assumptions is that the *Corr (Ui, Uj)* $= 0$. This means the two error terms are not correlated. Violation of this assumptions is a situation of autocorrelation.

2) The problem of autocorrelation is more common in time series data. This is because a phenomena affecting the error term in one point of time is more likely to be influencing the error term in the next point of time. This is especially identified as the factor of 'inertia or sluggishness'. Across units of cross section this is less likely. But it cannot be ruled out even in cross section data. In such cases, due to the spatial effect in cross section data, which is more like a demonstration effect, it is distinctly termed as spatial correlation.

3) Inertia or sluggishness, specification error in the model, cobweb phenomenon and data smoothening.

4) The consequences are: (i) least squares estimators are not efficient, (ii) the estimated variances of OLS estimates are biased, (iii) the standard error of true variances are underestimated, (iv) we are more likely to commit an error in deciding on the hypothesis of 'no statistical significance' of a particular estimated coefficient i.e. the decisions based on t and F tests would be unreliable, (v) estimated error variance would be biased and (vi) the value of R^2 would be misleading or unreliable.

Check Your Progress 2

1) Time sequence plotting (graphical method), Durbin-Watson test and Breusch-Godfrey (BG) Test.

2) $\boldsymbol{d} = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}$. It is the ratio of the sum of the squared differences in the successive residuals to the residual sum of squares.

3) The regression model includes an intercept term, the X variables are non-stochastic, the error term follows the following mechanism $u_t = \rho u_{t-1} + v_t$, $-1 \leq \rho \leq 1$, and the regression does not contain any lagged values of the dependent variable as one of the explanatory variables.

4) The one drawback of the d-test is that it has two zones of indecision, viz., $d_L < d < d_U$ and $(4 - d_U < d < 4 - d_L)$.

5) (i) It can handle non-stochastic regressors as well as the lagged values of Y_t, (ii) it can deal with higher-order autoregressive schemes such as AR(2)… etc. and (iii) it can also handle simple or higher order moving averages.

Check Your Progress 3

1) In this method we lag the regression equation by one period; multiply it by $\boldsymbol{\rho}$; and subtract it from the original regression equation. This gives us a transformed regression model. When estimated by OLS method, the estimators of the transformed model are BLUE.

2) In Sub-Section 12.6.3 we have outlined steps of the Cochrane-Orcutt iterative procedure. You should go through it and answer.

3) The *h*-statistic can be used in regression models having lagged dependent variables as explanatory variables.