UNIT 6 MULTICOLLINEARITY

Structure

6.0 OBJECTIVES

After going through this unit, you will be in a position to:

- explain the concept of multicollinearity in a multiple regression model;
- identify the implications of multicollinearity;
- detect multicollinearity in a sample data; and
- take remedial measures to correct the problem of multicollinearity.

6.1 INTRODUCTION

The multiple regression model covered in Block 1 is the workhorse of the applied literature in economics. It has been perhaps the single most common tool used for analysis. Therefore it makes sense to consider some of the problems that often arises when using it. This unit as well as the two following will take up some of the more common econometric issues that arise in this context as well as give you the tools to test for and fix some of these. In this unit we take up the issue of multicollinearity. In a multiple regression model, we can include many explanatory variables. These explanatory variables are expected to be unrelated among themselves. In emperical estimation, however, some of these variables may be related.

In such circumstances, are the estimates still unbiased and efficient? We will discuss this issue in the present unit.

6.2 MULTICOLLINEARITY DEFINED

First we take up the technical definition of multicollinearity in the context of the Multiple Linear Regression Model (MLRM) and, then we will discuss a more intuitive explanation of what it actually means for the researcher. Multicollinearity refers to the presence of a *perfect*, or exact, linear relationship among all or some of the explanatory variables in a regression. For instance if $X_1,...,X_k$ (where $X_1 = 1$ is the intercept term for all observations) are the variables being used in a regression, then an exact linear relationship between them is said to exist if the following condition is satisfied:

$$\lambda_1 X_1 + \lambda_2 X_2 + ... + \lambda_k X_k = 0 \qquad ...(6.1)$$

where $\lambda_1,...,\lambda_k$ are constants such that not all of them are zero simultaneously. Equivalently we can say that perfect multicollinearity implies that one of the variables can be exrpessed as a linear combination of the others, as shown below.

$$X_1 = -\frac{\lambda_2}{\lambda_1} X_2 - ... - \frac{\lambda_k}{\lambda_1} X_k \qquad ...(6.2)$$

Exact linear relationship, however, is a pathological case and it almost never arises in the real world. This case also referred to as 'perfect multicollinearity' is therefore only useful for clarifying the basic concept. In reality multicollinearity is commonly used to refer to a broader class of problems arising in MLRM, such that:

$$\lambda_1 X_1 + \lambda_2 X_2 + ... + \lambda_k X_k + v_i = 0 \qquad ...(6.3)$$

where v_i is a stochastic (random) error term, that varies across observations. Note that for this case it is not possible to express any of the variables as a linear combination of the other explanatory variable due to the presence of the random (and therefore indeterminate term) v_i for each observation. We will refer to this as the 'less than perfect multicollinearity' case. Although the implications of perfect multicollinearity are quite serious (see below), technically it is also the least bothersome as will be shown below. The more pervasive type of less than perfect multicollinearity that regularly arises in empirical studies is also technically the more troublesome since there is no easy way to avoid it.

An example of perfect multicolinearity is shown below in Table 6.1:

Table 6.1 : An Example of Perfect Multicollinearity

X_1	X_2	X_3
1	5	15
1	7.5	22.5
1	20	60
1	15	45
1	9	27

Let there be two explanatory variables (apart from the intercept X_1) that we are

from the table that $X_3 = 3X_2$, or using earlier notation if we substitute $\lambda_1 = 0$, $\lambda_2 = 3$ and $\lambda_3 = -1$ in equation (6.1) above, we find that the condition for perfect multicollinearity is fulfilled, i.e., the equation holds.

In the MLRM the parameter β_1 shows the impact of the explanatory variable (X_1) on the dependent variable (Y). Therefore, estimated coefficients are the impact of each variable on Y, keeping all the other variables fixed at a particular level (say their sample mean levels). However, in the presence of perfect multicollinearity this is not possible since in the data (sample) used for estimating the model, there is a perfect correlation between the variables. In the example above there is a linear relationship, ($X_3 = 3X_2$). Therefore there is no way to estimate the impact of X_3 on Y keeping X_2 fixed, since in the data, X_2 and X_3 always change together. In other words we are seeking information that the data does not contain; this can arise due to a number of reasons as discussed below. Multicollinearity therefore often boils down to the fact that we are trying to answer a question that is not technically possible by using the data at hand.

For example, if we are trying to estimate the model:

$$Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon \qquad ...(6.4)$$

where ε_i is a random error term, Y_i is the dependent variable, $X_1=1$ is the intercept term, and X_2, X_3 are other explanatory variables. Let us say that we are trying to estimate the role that income and wealth of an individual play on the consumption expenditure incurred by the person in a given year. Suppose Y_i is the consumption expenditure of individual i in year 2005, X_2 is his income in that year, X_3 is his wealth (total asset value) in the same year. Thus the model before us is

$$C_i = \beta_1 + \beta_2 Y_i + \beta_3 W_i + \varepsilon_i \qquad ...(6.5)$$

where C = consumption, Y = income and W = wealth. Here C is the dependent vaiable while Y and W are explanatory variables.

For the MLRM the estimated coefficients give us the intercept β_1 which is the level of consumption for someone with no income or wealth. β_2 gives us the change in consumption if income level for the person increases by one unit keeping his wealth level fixed (at the mean wealth level for the population). Similarly β_3 gives the impact on consumption of a change in wealth level by one unit keeping his income fixed at the mean level. However, let us assume that after looking at the data we come to realize that those with high incomes are also the ones with a higher wealth level. For simplicity let us assume that this relationship is perfect (linear), i.e., say the wealth level for each person is ten times the level of his current yearly income and this holds for all individuals surveyed (W=10Y). In order to find out how consumption expenditure of a person changes when income changes by a single unit keeping his wealth level fixed, intuitively the data needs to contain observations (people) with the same wealth level and different incomes. However this is not the case since in the data whenever income changes across individuals, the corresponding wealth level also changes proportionately (by a factor of ten). For example, if the second person's income is higher by one unit compared to the first person, then his wealth is ten units more as well, given our assumption of perfect multicollinearity. Therefore this question cannot be answered using the data at hand. In other words, the coefficients of the MLRM cannot be estimated (indeterminate); the technical proof is shown in the next section.

6.3 IMPLICATIONS OF MULTICOLLINEARITY

There are two cases as we mentioned earlier, viz., perfect multicollinearity and less than perfect multicollinearity.

1. *If multicollinearity is perfect in the sense of equation (6.1), the regression coefficients are indeterminate and their standard errors are infinite.*
2. *If multicollinearity is less than perfect as in equation (6.3), the regression coefficients are determinate but possess large standard errors (in relation to the coefficients themselves), which means that the coefficients cannot be estimated with great precision or accuracy.*

6.3.1 Perfect Multicollinearity Case

Let us first consider the case of perfect multicollinearity, defining the deviations from sample mean:

$$y_i = Y_i - \overline{Y}, \qquad x_{2i} = X_{2i} - \overline{X}_2 \text{ and } x_{3i} = X_{3i} - \overline{X}_3$$

where $\overline{Y}, \overline{X}_2, \overline{X}_3$ are the sample means of the variables respectively. By definition (see earlier units), the coefficient β_2 in MLRM is calculated as,

$$\hat{\beta}_2 = \frac{\left(\sum y_i x_{2i}\right)\left(\sum x_{3i}^2\right) - \left(\sum y_i x_{3i}\right)\left(\sum x_{2i} x_{3i}\right)}{\left(\sum x_{2i}^2\right)\left(\sum x_{3i}^2\right) - \left(\sum x_{2i} x_{3i}\right)^2} \qquad ...(6.6)$$

For perfect multicollinearity in our earlier example, we have assumed $X_3 = 10X_2$. Let us make it more general and assume $X_3 = \lambda X_2$, where λ is a constant as before. This implies that in deviations form we can write $x_3 = \lambda x_2$. Substituting this in equation (6.6) we get,

$$\hat{\beta}_2 = \frac{\left(\sum y_i x_{2i}\right)\left(\lambda^2 \sum x_{2i}^2\right) - \left(\lambda \sum y_i x_{2i}\right)\left(\lambda \sum x_{2i}^2\right)}{\left(\sum x_{2i}^2\right)\left(\lambda^2 \sum x_{2i}^2\right) - \left(\lambda \sum x_{2i}^2\right)^2} \qquad ...(6.7)$$

$$= \frac{\left(\lambda^2 \sum x_{2i}^2\right)\left\{\sum y_i x_{2i} - \sum y_i x_{2i}\right\}}{\left(\lambda^2 \sum x_{2i}^2\right)\left\{\sum x_{2i}^2 - \sum x_{2i}^2\right\}}$$

$$= \frac{0}{0}$$

which is an indeterminate expression. Thus the coefficient β_2 is not determinate, i.e., it cannot be estimated. Similarly we can show that the coefficient for β_3 is also indeterminate in the presence of perfect multicollinearity.

6.3.2 Less than Perfect Multicollinearity Case

We now consider the more realistic case of what happens to the parameter estimates of the MLRM when there is high but not perfect multicollinearity. Let us continue with our earlier example where income and wealth determine consumption expenditure. The case of 'less than perfect' multicollinearity arises when wealth for

each individual is not a multiple of the current income of the person concerned. Say you observe that for the first person wealth is 9.5 times his income, for the second person this is 10.75 times his income and so on, such that on the average wealth is ten times the income, but with small deviations for each person (observation) in the data. In terms of equation (6.3) we can say that v is an independent random variable (not correlated with any of the explanatory variables) with mean zero and small variance (compared to λ), and takes the value of –0.5 for the first person and 0.75 for the second person and so on (λ=10 in this example).

The bigger concern for the researcher arises from the impact of multicollinearity on the standard errors of the β estimates. Note that the estimates of the coefficients are as follows (see Unit 4 on MLRM),

$$var(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_{2i}^2 (1 - r_{23}^2)} \qquad ...(6.8)$$

$$var(\hat{\beta}_3) = \frac{\sigma^2}{\sum x_{3i}^2 (1 - r_{23}^2)} \qquad ...(6.9)$$

where r_{23} is the correlation coefficient between X_2 and X_3. Therefore, it is obvious from the equations above that as the level of correlation between the two explanatory variables increases, i.e., the level of multicollinearity increases, or as r_{23} tends towards one, the variance estimates of the estimated coefficients, $var(\hat{\beta})$, increases. In the limiting case of perfect multicollinearity of course $r_{23} = 1$ and the variances are infinite.

6.3.3 Major Implications

The important implications of multicollinearity are given below. In the presence of high, though not perfect, multicollinearity:

1) The OLS coefficient estimates have large variances.
2) Because of large variance the confidence intervals will be very large, which in turn means that there is a high probability of accepting the null hypothesis of zero coefficient, even when the actual parameter is positive.
3) Overall the regression may do very well, i.e., the R^2 may be quite high despite not being able to reject the hypothesis of one or more parameters being equal to zero.
4) The OLS estimates and their standard errors can be very sensitive to small changes in the data.

Note that from a practical standpoint this is an extremely serious problem since most empirical studies try to estimate the impact of one or more economic variables on a particular variable, such as the income consumption relationship discussed above. So if income and wealth are highly correlated then although the regression gives the true coefficient estimates, we cannot reject the null hypothesis that the impact of income on consumption expenditure is zero. This means that the whole purpose of the study is lost, as nothing can be conclusively (statistically speaking) proved.

6.3.4 An Example

Let us consider the regression model given by equation (6.5) earlier, that is,

$$C_i = \beta_1 + \beta_2 Y_i + \beta_3 W_i + \varepsilon_i$$

We take simulated data on consumption expenditure C, income Y and wealth W. The data was generated as follows: first the data for income was generated as a normal variable, $Y \sim N(100,25)$, that is, it is a normal variable with mean 100 and variance of 25. Next, wealth is taken as a multiple of income along with a disturbance term, to ensure multicollinearity, that is $W = 10Y+\gamma$, where γ is a random error term distributed independently and identically as $\gamma \sim N(20,10)$. Note that the correlation coefficient between income and wealth in the data generated is 0.999; therefore there is very strong multicollinearity problem included on purpose. Finally, the consumption expenditure is taken to be a linear function of the two, that is, $C=25+0.75Y+0.1W+\varepsilon$, where ε is a random error term obeying all the standard assumptions (distributed IID), with distribution $\varepsilon \sim N(20,10)$. The data generated is given in Table 6.2.

Table 6.1: Data for Illustration

Observation	Consumption Expenditure (C)	Income Y	Wealth W
1	279.0	131.5	1317.9
2	146.2	65.0	666.8
3	197.2	82.9	865.8
4	177.0	67.1	698.3
5	238.3	109.0	1105.7
6	206.8	85.9	865.5
7	248.1	108.5	1099.4
8	216.8	102.7	1046.8
9	262.3	120.8	1230.2
10	198.1	93.4	954.2

Fitting the standard MLRM to this data yields the following result:

$$\hat{C}_i = 40.22 + 0.971Y_i + 0.084W_i \qquad \ldots(6.10)$$

$$(22.977)\ (4.091)\ (0.417)$$

$$t = [1.750]\ [0.237]\ [0.202]$$

For equation (6.10) we have presented the standard errors of the estimates within parentheses. For example, the intercept estimate of 40.22 has a standard error of 22.977. The corresponding t-ratio is given in the next line as 1.75. You can see the standard error and t-ratio of other estimates as well.

Together income and wealth explain almost 98% of the variation in consumption expenditures ($R^2 = 0.987$ and $= \bar{R}^2\ 0.983$), which is not surprising given the way the simulated data is generated as a linear function of these two variables. Note that as

mentioned before we find that the estimates are not biased, they give estimates of the coefficients that are pretty close to the real ones used for generating the data. For income the estimate is 0.971 and the parameter is 0.75. Similarly, for wealth the estimate is 0.084, compared to the actual one 0.1. The disparity between the actual value and the estimated coefficients arises because of the small sample size; additional data will give a more accurate measure of the true values. Ten observations are really too small a sample to estimate a regression model accurately. Therefore the OLS estimates are not biased in the presence of multicollinearity. However, the problem arises for hypothesis testing since as you can see from the reported results, the standard errors for the estimates are very high, making the t-statistics low, which in turn implies that neither variable is statisticially significant. The F-statistic (not reported here), also shows that the hypothesis that all coefficients are zero is rejected. This is the classic problem that arises with multicollinearity.

Now let us drop the variable 'wealth' from the model and run the same regression on income only. We find that:

$$\hat{C}_i = 43.408 + 1.796Y_i \qquad ...(6.11)$$

$$(15.647) \quad (0.158)$$

$$t = [2.774] \quad [11.348]$$

Earlier the income variable was highly insignificant, but we find that if only income is used in the regression then it is highly significant. The estimated coefficient, however, is quite far off from the actual value. As before, this is entirely due to the small size of the sample. The OLS estimates are still BLUE. The regression gives an R^2 of 0.942, which is also very high as expected, given the nature of the data.

Similarly if we drop income and use only wealth in the regression then we find:

$$\hat{C}_i = 36.632 + 0.183W_i \qquad ...(6.12)$$

$$(16.25) \; (0.016)$$

$$t = [2.254] \; [11.334]$$

We find that earlier wealth was insignificant but now it is highly significant. Moreover, wealth alone explains about 94% of the variation in consumption ($R^2 = 0.941$). Equations (6.11) and (6.12) show that in situations of extreme multicollinearity, dropping one of the highly collinear variables will make the other X variables statistically significant. This suggests that a standard solution for this problem would be to drop the collinear variables in all such situations. We take this up in more details later in this unit.

Check Your Progress 1

1) In data involving economic time series such as GNP, money supply, prices, income, employment, etc., multicollinearity is usually suspected. Why?

.....

.....

.....

.....

2) State with reason whether the following statements are true, false, or uncertain:

a) Despite perfect multicollinearity, OLS estimators are BLUE.

b) In cases of high multicollinearity, it is not possible to assess the individual significance of one or more regression coefficients.

c) If an auxiliary regression shows that a particular R_i^2 is high, there is definite evidence of high collinearity.

d) High pair-wise correlations do not suggest that there is high multicollinearity.

e) Multicollinearity is harmless if the objective of the analysis is prediction only.

f) You will not obtain a high R^2 value in a multiple regression if all the coefficients are individually statistically insignificant on the basis of the usual t test.

.....

.....

.....

.....

.....

6.4 DETECTION OF MULTICOLLINEARITY

Note that multicollinearity is almost always present in most applications, so it is a matter of degree and not whether it is present or not. What most often we are trying to find out in empirical applications is whether it is likely to create problems in terms of hypothesis testing or not (through a higher estimate of standard errors), since the objective of most empirical studies is usually to find and test the exact relationship between variables. There does not exist any formal tests for multicollinearity, however there does exist some informal rules of thumb that most researchers use for this purpose. These are as follows:

1) **High R^2 but few significant t-statistics**. This is the "classic" symptom of multicollinearity. If the R^2 is high, say in excess of 0.8, the F test in most cases will reject the hypothesis that estimated coefficients are all simultaneously zero. The individual t tests, however, will show that none or very few of the slopes (coefficients) estimates are statistically different from zero. This was illustrated earlier in the example shown above.

2) **High pair-wise correlations among regressors**. Another suggested rule of thumb is that if the pair-wise correlation coefficient between two regressors is high, say, in excess of 0.8, then multicollinearity is a serious problem. The problem with this criterion is that, this is a sufficient condition for multicollinearity but not a necessary condition. In other words, there can be problems of multicollinearity present even when this condition is not fulfilled.

3) **Auxiliary regressions**. Multicollinearity arises because one or more of the regressors are exact or approximate linear combinations of the other regressors. One way of finding out which X variable is related to other X variables is to regress each X_i on the remaining X variables and compute the corresponding R^2, which we designate as R_i^2. Each one of these regressions is called an auxiliary regression; auxiliary to the main regression of Y on X's. There are subsequently two ways to test for multicollinearity:

   (a) First, we can follow **Klien's rule of thumb**, which suggests that multicollinearity is a troublesome problem only if the R_i^2 obtained from an auxiliary regression is greater than the overall R^2.

   (b) Second, it can be shown that the variable

$$F_i = \frac{R_i^2/(k-2)}{(1-R_i^2)/(n-k+1)} \qquad ...(6.13)$$

follows the F distribution with degrees of freedom k–2 and n–k+1. Here, n stands for sample size, k stands for the number of explanatory variables including the intercept term, and R_i^2 is the explained variation from the regression of X_i on the remaining X variables. If the computed value of F exceeds the critical values at the chosen level of significance, then it is taken to mean that the particular X_i is collinear with the other X's. Conversely, if the computed F is lower than the critical value then we say that it is not collinear and we can retain it in the model.

6.5 REMEDIAL MEASURES

There are two common approaches taken towards the problem of multicollinearity. First, 'do nothing' which is simply to ignore the problem. Alternatively, the researcher can follow some rules of thumb as outlined below. The order in which the measures are given here is roughly by descending order of their popularity. For instance, most researchers choose the 'do nothing' approach if there are no serious problems in interpreting the coefficients, and so on.

1) **Do nothing** : Since multicollinearity is a problem with the data at hand, it is not possible to do anything about it, and therefore we try to make the most of existing situation. In other words, it is a sample problem; the sample available does not contain enough variation to identify certain impacts (for instance, if wealth and income always vary proportionately then the impact of both separately on consumption cannot be identified). Usually not all coefficients are statistically insignificant in a regression model. Hence, we work with what we can get out of the model and hope that availablity of more data will allow a more accurate estimation in future. Moreover, even if one or more regression coefficients cannot be estimated with great precision, a linear combination of them can be estimated relatively efficiently. For instance, from equation (6.4) above if we know that $X_3=\lambda X_2$, then substituting this into the equation we get

$$Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 \lambda X_2 + \varepsilon_i$$

or,
$$Y = \beta_1 X_1 + \alpha X_2 + \varepsilon \qquad ...(6.14)$$

where $\alpha = \beta_2 + \lambda\beta_3$. Note that even if we knew λ, in the case of perfect collinearity, there is no way to get the two coefficients β_2 and β_3 separately, since we have one equation with two unknowns $\hat{a} = \hat{\beta}_2 + \lambda\hat{\beta}_3$. Sometimes we have to accept that this is the best that we can do given the data.

2) **Dropping a variable**. One of the 'simplest' methods is to drop one of the collinear variables. Thus as shown earlier in our consumption expenditure illustration, when we drop the variable wealth, we obtain the result that income is highly significant whereas in the original model, income was insignificant. The problem here is that by dropping variables arbitrarily from the model we are exposed to the risk of *specification bias*, which arises from the incorrect specification of the model for analysis. Thus if economic theory says that both income and wealth should have an impact on consumption then dropping one of them is a misspecification of the model and gives rise to error. Specification error or bias usually leads to biased estimates of the coefficients. Therefore although convenient, this approach has a serious drawback since the remedy may be worse than the disease in some situations. Whereas multicollinearity may prevent precise estimation of the parameters of the model, omitting a variable may seriously mislead us as to the true values of the parameters. We note that with specification error the estimates can be potentially biased. So what you gain from lower standard errors of remaining variables may not be worth paying the price of biased estimates.

3) **A priori information**: Let us consider the earlier example of consumption expenditure. We know that $\beta_2 = 0.13\beta_1$, since $\beta_1 = 0.75$ and $\beta_2 = 0.1$. In other words, the rate of change of consumption with respect to wealth is 0.13 times the rate of change with respect to income. Then the earlier regression becomes

$$\begin{aligned} C_i &= \beta_0 + \beta_1 Y_i + 0.13\,\beta_1 W_i + u_i \qquad (6.15) \\ &= \beta_0 + \beta_1 X_i + u_i \end{aligned}$$

where $X_i = Y_i + 0.13W_i$. Once we obtain β_1, we can estimate β_2 using this estimate. Where does one obtain such a priori information? We obtain it either from an earlier study where the problem of multicollinearity was less severe, or alternatively from theory. For example, if we are estimating a Cobb-Douglass production function and we expect constant returns to scale to prevail then we can impose the restriction that $\beta_1 + \beta_2 = 1$. Note that many times this cannot be done since the objective of many studies is to test the theory. For instance, if the objective is to find out if there exists constant returns to scale or not, then obviously this is not a solution we can use.

4) **Pooling of data** : Another variation of using the *a priori* information method outlined above is to combine cross-sectional and time-series data, also known as pooling the data. Consider the following example – say you are interested in studying the demand for automobiles in India. Assume you have time series data on the number of cars sold, average price of the car and consumer income. The model specified is as follows:

$$lnY_t = \beta_0 + \beta_1 lnP_t + \beta_2 lnI_t + u_t \qquad ...(6.16)$$

where Y =number of cars sold, P = average price, and I = income and the subscript t is the time period. Our objective is to estimate the price elasticity β_1 and the income elasticity β_2. In time series data the price and income variables

generally tend to be highly collinear. Therefore, if we run the regression given at (6.16), we shall be faced with the usual multicollinearity problem. A way out of this has been suggested by Tobin.[1] He says that if we have cross-sectional data (for example, data generated by consumer panels, or budget studies conducted by various private and governmental agencies), we can obtain a fairly reliable estimate of the income elasticity β_2; because in such data, which are at a point of time, the prices do not vary much. Let the cross-sectionally estimated income elasticity be $\hat{\beta}_2$. Using this estimate we can rewrite the preceding time series regression as,

$$lnY_t^* = \beta_0 + \beta_1 lnP_t + u_t \qquad ...(6.17)$$

where $lnY^* = lnY - \hat{\beta}_2 lnI$, that is, Y^* represents the value of Y after removing from it the effect of income. We can now obtain an estimate of the price elasticity from the preceding regression.

This creates a problem of interpretation of the data, because we are assuming that the cross-sectionally estimated income elasticity is the same as that which would be obtained from the time series data. Nevertheless this technique has been used in numerous studies.

5) **Transformation of variables**: Consider the earlier example of consumption expenditure, income and wealth. Here the data is of the time series format, that is, we have data on the same variables for a number of years or months. One reason for high multicollinearity is that over time both income and wealth tend to move together in the data. One way of minimizing this dependence between the variables is to take the *first difference* of the variables. If the relation is

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + u_t \qquad ...(6.18)$$

which holds at time t, then it must also hold at time $t-1$. Therefore, we have

$$Y_{t-1} = \beta_0 + \beta_1 X_{1t-1} + \beta_2 X_{2t-1} + u_{t-1} \qquad ...(6.19)$$

Subtracting equation (6.19) from (6.18) we get,

$$Y_t - Y_{t-1} = \beta_1 (X_{1t} - X_{1t-1}) + \beta_2 (X_{2t} - X_{2t-1}) + v_t \qquad ...(6.20)$$

where $v_t = u_t - u_{t-1}$. In other words, instead of running the regression on the values of the variables we run them on the difference between the successive values of the variables. This reduces the problems associated with multicollinearity since although X_1 and X_2 might be highly correlated there is no reason to believe that the changes in their levels over time are also highly correlated.

Another transformation that is commonly used is the *ratio transformation*. Again consider the earlier model from equation (6.18) above. Let Y be consumption expenditure in real dollars, X_1 be GDP and X_2 be population. Since both GDP and population grow over time they are likely to be correlated. One 'solution' is to express the model in per capita terms, that is, dividing (6.18) throughout by X_2 we get

[1] J.Tobin, "A Statistical Demand Function for Food in the U.S.A.", Journal of the Royal Statistical Society, Series. A, 1950, pp. 113-141.

$$\frac{Y_t}{X_{2t}} = \beta_0\left(\frac{1}{X_{2t}}\right) + \beta_1\left(\frac{X_{1t}}{X_{2t}}\right) + \beta_2 + \left(\frac{u_t}{X_{2t}}\right) \qquad ...(6.21)$$

Such a transformation may reduce collinearity in the original variables. However, note that the remedy might be worse than the disease. Since it is known that the error term in equation (6.21) will suffer from problems of heteroscedasticity even when the original errors in (6.18) are homoscedastic. Similarly, it is also known that the error term obtained in the first difference method v_t in equation (6.20) is going to be serially correlated even when the original error terms u_t were serially uncorrelated. (Issues relating to autocorrelation and heteroscedasticity are covered in Units 7 and 8.) Therefore, such remedies are to be used judiciously by the researcher keeping in mind the risks.

6) **Additional data**: It is possible that with additional data the problem of multicollinearity becomes less severe although still present. This can be seen from equation (6.8) from above; as sample size increases, $\sum x_{2i}^2$ increases. Therefore, for any r_{23} given, the variance of $\hat{\beta}_2$ will decrease, thus decreasing the standard error, which will enable us to estimate β_2 with more precision. Unfortunately, this method is usually not feasible, since economists seldom can obtain additional data without incurring huge costs. Sometimes it may not be possible since the data collected was at a different time and place. Therefore, this is a solution which should be kept in mind but is hardly used in practice.

6.6 EXAMPLE OF TESTING AND REMEDIES

In this section we use the standard data used in this context which was collected by Longley[2]. Originally collected to assess the computational accuracy of least-squares estimates, the Longley data has been widly used to illustrate several econometric problems, including multicollinearity. The data reproduced in Table 6.3, are time series for the years 1947–1962 and pertain to Y = number of people employed, in thousands; X_1 = GNP implicit price deflator; X_2 = GNP, in millions of dollars; X_3 = number of unemployed in thousands; X_4 = number of people in the armed forces; X_5 = noninstitutionalized population over 14 years of age; and X_6 = year, equal to 1 in 1947, 2 in 1948, and so on.

Let us assume that our objective is to predict Y on the basis of the six X variables. Using the econonetric software Stata 9, we obtain the following regression results:

Table 6.3: Longley Data

Obs.	y	X_1	X_2	X_3	X_4	X_5	Time
1947	60,323	830	234,289	2356	1590	107,608	1
1948	61,122	885	259,426	2325	1456	108,632	2
1949	60,171	882	258,054	3682	1616	109,773	3
1950	61,187	895	284,599	3351	1650	110,929	4
1951	63,221	962	328,975	2099	3099	112,075	5
1952	63,639	981	346,999	1932	3594	113,270	6
1953	64,989	990	365,385	1870	3547	115,094	7

[2]Longley, J. "An Appraisal of Least-Squares Programs from the Point of the User," *Journal of the*

1954	63,761	1000	363,112	3578	3350	116,219	8
1955	66,019	1012	397,469	2904	3048	117,388	9
1956	67,857	1046	419,180	2822	2857	118,734	10
1957	68,169	1084	442,769	2936	2798	120,445	11
1958	66,513	1108	444,546	4681	2637	121,950	12
1959	68,655	1126	482,704	3813	2552	123,366	13
1960	69,564	1142	502,601	3931	2514	125,368	14
1961	69,331	1157	518,173	4806	2572	127,852	15
1962	70,551	1169	554,894	4007	2827	130,081	16

Table 6.4 : Result from the Longley Data

Source	SS	df	MS		
				Number of obs	= 16
				F(6, 9)	= 330.29
Model	184172402	6	30695400.3	Prob > F	= 0.0000
Residual	836424.056	9	92936.0062	R-squared	= 0.9955
				Adj R-squared	= 0.9925
Total	185008826	15	12333921.7	Root MSE	= 304.85

y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
X_1	1.506187	8.491493	0.18	0.863	−17.7029	20.71528
X_2	−.0358192	.033491	−1.07	0.313	−.1115811	.0399427
X_3	−2.02023	.4883997	−4.14	0.003	−3.125067	−915393
X_4	−1.033227	.2142742	−4.82	0.001	−1.517949	−.548505
X_5	−.0511041	.2260732	−0.23	0.826	−.5625172	.460309
X_6	1829.151	455.4785	4.02	0.003	798.7875	2859.515
_cons	77270.12	22506.71	3.43	0.007	26356.41	128183.8

A glance at the results shown in Table 6.4 would suggest that we have the multicollinearity problem, for the R^2 value is very high, but quite a few variables are statistically insignificant (X_1, X_2, and X_5), a classic symptom of multicollinearity. To shed more light on this, we show in Table 6.5 the correlation coefficients among the six regressors (obtained from the econometric software Stata 9). Note that the diagonal terms in the table are correlations with itself and therefore are necessarily equal to one.

From Table 6.5 it is clear that several of the pair-wise correlations are quite high, suggesting that there may be a severe collinearity problem. We have mentioned before that high correlations are sufficient but not necessary for multicollinearity to exist. Next we run the auxiliary regressions, that is, the regression of each X variable on the remaining X variables. Only the R^2 values obtained are presented here in Table 6.6. Since all the values presented here are very high, except X_4, it seems we do have a serious collinearity problem.

Table 6.5 : Pair-wise Correlations

	X_1	X_2	X_3	X_4	X_5	X_6
X_1	1.0000					
X_2	0.9916	1.0000				
X_3	0.6206	0.6043	1.0000			
X_4	0.4647	0.4464	–0.1774	1.0000		
X_5	0.9792	0.9911	0.6866	0.3644	1.0000	
X_6	0.9911	0.9953	0.6683	0.4172	0.9940	1.0000

Table 6.6 : Auxiliary R^2 Values

Dependent Variable	R^2 Values
X_1	0.993
X_2	0.999
X_3	0.970
X_4	0.721
X_5	0.997
X_6	0.999

Applying Klien's rule of thumb, we see that the R^2 values obtained from the auxiliary regressions exceed the overall R^2 value (that is, the one obtained from the regression of Y on all the X variables) of 0.99 in 4 out of 6 auxiliary regressions, again suggesting that indeed the Longley data are plagued by the multicollinearity problem. Incidentally, applying the F test given earlier, we can verify that the R^2 values given in Table 6.6 are all statistically significantly different from zero. Another problem that arises with multicollinearity is that the estimates are very sensitive to small changes in the data. For instance, dropping the observation for 1962 significantly changes all results.

Now that we have established the existence of multicollinearity problem, what "remedial" actions can we take? Let us reconsider our original model. First of all, we could express GNP not in nominal terms, but in real terms by deflating for price increase. Second, since noninstitutional population over 14 years of age grows over time because of natural population growth, it will be highly correlated with time, the variable X_6 in our model. Therefore, instead of keeping both these variables, we will keep the variable X_5 and drop X_6. Third, there is no compelling reason to include X_3, the number of people unemployed; perhaps the unemployment rate would have been a better measure of labour market conditions. But we have no data on the latter. So, we will drop the variable X_3. Making these changes, we obtain the following regression results (RGNP = real GNP):

Table 6.7 : Results of the Modified Model

Source	SS	df	MS	Number of obs =	16
				F(3, 12) =	211.10
Model	181568361	3	60522787.1	Prob > F =	0.0000
Residual	3440464.6	12	286705.383	R-squared =	0.9814
				Adj R-squared =	0.9768
Total	185008826	15	12333921.7	Root MSE =	535.45

y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
RGNP	97.36499	17.91551	5.43	0.000	58.33046	136.3995
X_4	-.6879658	.3222373	-2.13	0.054	-1.390061	.0141289
X_5	-.2995369	.1417612	-2.11	0.056	-.608408	.0093342
_cons	65720.39	10624.8	6.19	0.000	42570.94	88869.85

Although the R^2 value has declined slightly compared to before, it is still very high. Now all the estimated coefficients are significant and the signs of the coefficients make economic sense.

Check Your Progress 2

1) Suppose the multiple regression model is given by

   $Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$. It is given that $X_2 = 5X_3$. Show that the parameter β_3 cannot be estimated.

   .....

   .....

   .....

   .....

2) Explain the methods of detecting multicollinearity.

   .....

   .....

   .....

   .....

3) Explain the methods of transformation of variables to remove the problem of multicollinearity.

   .....

6.7 LET US SUM UP

One of the basic assumptions in multiple regression models is the independence between explantory variables. When several explanatory variables are used in a model, none of them should be linear combinations of others. In reality, however, such an assumption is often not fulfilled, though the degree of relationship between explanatory variables differs. In the case of violation of the assumption of independence among explanatory variables, we encounter the problem of multicollinearity.

In the presence of multicollinearity the regression model exhibits very high coefficient of determination, R^2, which implies that the explanatory variables under consideration explain the variation in the dependent variable very well. On the basis of the high R^2, when we undertake a F-test, we find that all coefficients in the model are significant. But when we look into individual estimates and test for their significance on the basis of t-test, we find that very few (or, even none) of the coefficients are statistically significant. This happens because of large variance (also the standard error) of the parameter estimates.

Multicollinearity is considered to be a sample problem in the sense that as sample size increases the standard error of the estimates declines and the estimates may show their significance. In order to remove the problem of multicollinearity often we drop a variable but it may result in misspecification problem. At times we use *a priory* information on the relationship between variables or transform the variables to remove this problem.

6.8 KEY WORDS

Coefficient of determination : It is denoted by the symbol R^2. It is given by the ratio of RSS to TSS. A high value of R^2 shows that the explantory variables explain the change in dependent variable very well.

Multiple regression model : It denotes a single equation regression model which contains one explained variable and more than one explanatory variables.

Perfect multicollinearity : In the case of perfect multicollinearity there is exact linear relationship between the explanatory variables.

Specification bias : It denotes the bias that results due to mis-specification of a model. It can result due to dropping of an important variable or inclusion of an unimportant variable in the model.

Standard error : It is the standard deviation of the sampling distribution of the estimate under consideration. Thus a high standard error of β_1 implies that the value of β_1 fluctuates widely from sample to sample.

6.9 SOME USEFUL BOOKS/REFERENCES

Farrar, D.E., and R. R. Glauber, 1967 "Multicollinearity in Regression Analysis: The Problem Revisited," *Review of Economics and Statistics*, vol. 49, pp. 92–107

Johnston, J., 1984, *Econometric Methods*, 3rd ed., McGraw-Hill, New York

Judge, George G., Carter R. Hill, William E. Griffiths, Helmut Lutkepohl, and Tsoung-Chao Lee,: 1980, *Theory and Practice of Econometrics*, John Wiley & Sons, New York

Kennedy, Peter, 1998, *A Guide to Econometrics*, 4th ed., MIT Press, Cambridge, Mass.

Klien, Lawrence R., 1962, *An Introduction to Econometrics*, Prentice-Hall, Englewood Cliffs, N.J.

6.10 ANSWERS/HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) These variables move together in an economy over time. Usually we observe that as economic growth occurs these is an increase in GNP, money supply, prices, income, employment, etc. These variables also influence one another. Thus there is a fair chance that these are related and problem of multicollinearity may take place.

2) a) False b) True c) Uncertain

   d) False e) True f) False

Check Your Progress 2

1) See Section 6.2 and answer.

2) The methods of detection are given at Section 6.4 Your answer should include three points given in that section.

3) Refer to Section 6.5 and answer.

UNIT 7 AUTOCORRELATION

Structure

7.0 OBJECTIVES

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

- explain the concept of autocorrelation;
- explain theoretical and practical implications of autocorrelation;
- test for autocorrelation using actual data; and
- find remedial measures for the problem of autocorrelation.

7.1 INTRODUCTION

It was shown in Unit 2 that the Ordinary Least Squares (OLS) estimator is the Best Linear Unbiased Estimator (BLUE) under a set of assumptions. One of the assumptions was that the error terms are independent of each other. If we denote the error term as u_i for the i-th observation, then we assumed that:

$$cov(u_i,u_j) = 0 \ \forall \ i \neq j \qquad ...(7.1)$$

where $cov(x,y)$ is the covariance between the two variables x and y.[1] However if this assumption is violated then there exists autocorrelation, that is,

[1]Note that we use correlation and covariance interchangeably since no correlation means zero covariance and vice versa. These two terms, however, are different by definition.

$$cov(u_i,u_j) \neq 0 \qquad ...(7.2)$$

There are typically two types of data used for empirical studies: cross section and time series. In the first case data on various variables are collected across a number of households or firms at a given point in time. Examples in India include the Economic Census, and the Census of India. This type of data can be used to test for any relationship between variables across firms or households (depending on the sampling unit). For instance, are bigger firms (in terms of revenue) more productive (Per unit of labour) is a question that can be answered using the Economic Census data. Data may also be collected over time for a set of variables for the same sampling unit (usually at country or state level, may also be for households or firms). This type of data is known as time series data. This allows us to check for any relationship between the variables for the same unit, over time. For instance, if aggregate consumption and income data are collected at the country level, then we can estimate the consumption function, that is, the relationship between income and consumption. As you know from macroeconomic analysis, here we can answer questions such as: Does consumption increases with income and if so by what amount? Another form of data that has gained prominence in recent times is when time series data is available for multiple units, then we say this is pooled (time series and cross-section data), also called panel data.

7.2 PRESENCE OF AUTOCORRELATION

Note that the problem of autocorrelation primarily arises in the case of time series data and not for cross-sectional data. Panel data involves a different set of tools and are not covered in this course. It is easiest to make this point using the consumption function example. Say consumption and income data are collected across households (cross-sectional data) for August, 2006. It is extremely unlikely that the error terms will be correlated. That is because correlation usually occurs when there is a common event that creates a deviation from the 'true' model. For instance, one household holding a party may be a reason why their error term may be high (consumption expenditure much higher than predicted by income and other factors, which are the regressors or independent variables in the model estimated). However it is extremely unlikely that just because household *i* had a party, household *j* will also have a party during the same time period, when data was collected. Note that if all households are having a party during the survey period then this is a common factor for all data points and gets captured in the intercept term in the model (given income and other factors, consumption is predicted to be higher for all), and this does not affect the error terms.

For time series data, however, there is usually considerable inertia; most macroeconomic time series data follow a cyclical behaviour due to business cycles. For several periods income and consumption will be low; then income starts recovering. However, consumption shows an inertia and recovers after a lag as people become more confident that the rise in income is not temporary but permanent and change their consumption patterns slowly over time. In terms of the error term this will show up as several consecutive data points with low errors (deviations from the estimated model), then several periods when the errors are large (due to the inertia in economic behaviour), and so on. Therefore, there is a good chance that consecutive error terms will be highly correlated. Another case where autocorrelation arises is when lagged value of a variable is used as a regressor. For instance, if past consumption (for last period and/or the period before that) is used as an explanatory variable or regressor in the MRM. This almost always results in autocorrelated error

terms, and adequate steps must be taken to correct this. Therefore one needs to be very cautious while using any kind of time series data for the problem of autocorrelation. However this does not mean that you can be lax if you are dealing with cross-section data; you cannot simply assume the problem away, appropriate tests still need to be conducted.

In Fig.7.1, Fig.7.2 and Fig.7.3 we show different error terms, the true values. Note that the true values of error terms are not observable and are given here for illustrative purposes. In reality you will observe only the estimated residuals from the MRM. First, in Fig.7.1 there is no autocorrelation. In Fig.7.2, there is positive autocorrelation in the error structure. Finally, in Fig.7.3 there is negative autocorrelation. In the first figure you will find that the error terms are scattered around zero, which is expected, since it is a normal variable drawn independently with mean zero. Hence, there should be no correlation between consecutive error terms. So a plot should not reveal any pattern at all. In Fig.7.2 on the other hand there is substantial correlation between the error terms in consecutive periods; there are several periods when the error is high followed by several periods when it is low. So there is a cyclical pattern to be observed in the plot of residuals. This is true for most time series obtained from the real world, i.e., this is the most common pattern. Fig. 7.3 shows negative autocorrelation. This shows period to period fluctuations, that is, if this period the error is highly positive the next period it will be highly negative and the period after that it will be highly positive again. In other words, consecutive terms are correlated but the sign changes from positive to negative and back and so on.

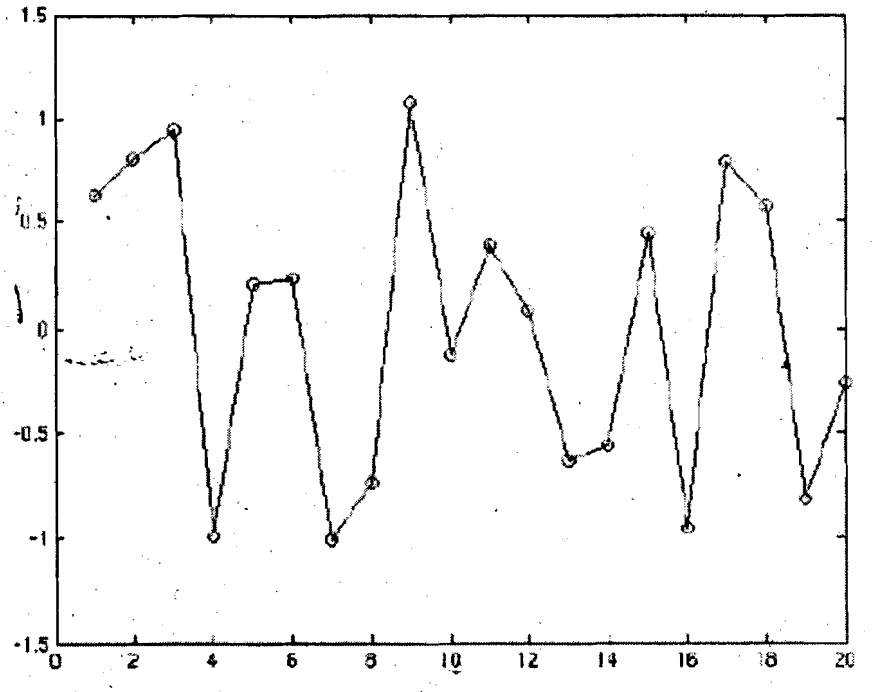

Fig.7.1: Error Term with No Autocorrelation

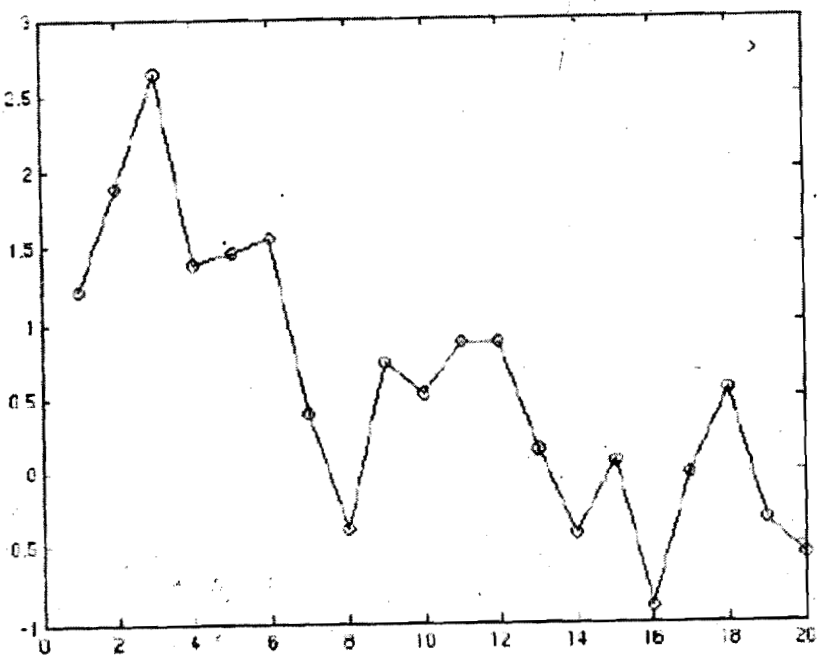

Fig.7. 2: Error Term with Positive Autocorrelation

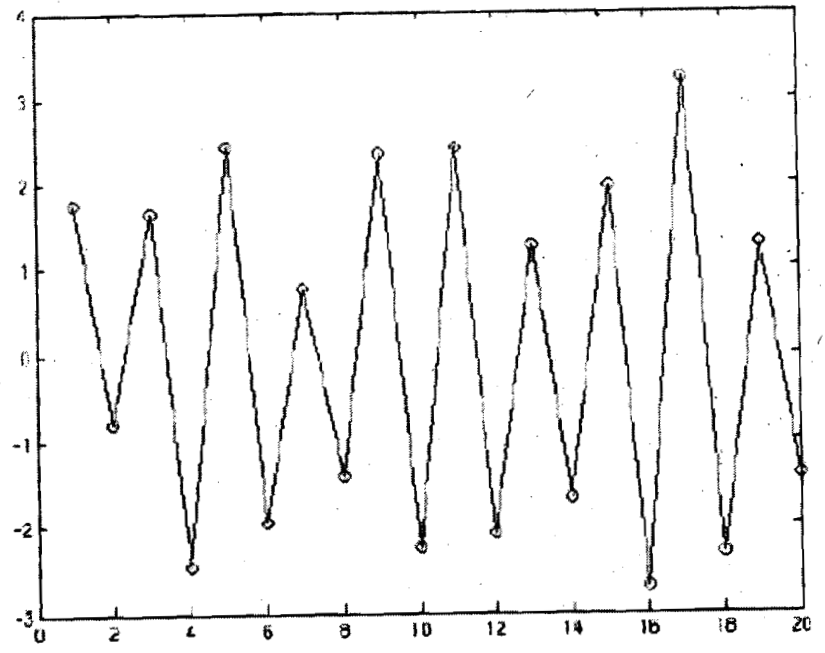

Fig.7. 3: Error Term with Negative Autocorrelation

7.3 IMPLICATIONS OF AUTOCORRELATION

The good news is that even in the presence of autocorrelation, the OLS estimators are still linear unbiased as well as consistent and asymptotically normally distributed. However they are no longer efficient, that is, they are not minimum variance. It means that since the variance of the OLS estimator is larger than alternatives, the confidence interval for the estimated coefficients in the MRM are likely to be larger. Recall from Unit 3 that confidence intervals are constructed using estimated coefficients and plus/minus a constant depending on the level of confidence chosen (e.g., 95%) times the estimated variance of the estimate. Also recall that hypothesis testing is done, in

particular whether an estimated coefficient is significant or not (different from zero or not) by finding out whether the *t*-statistic for the estimate falls outside the confidence interval or not. So a larger confidence interval means a greater likelihood of falsely accepting the null hypothesis of insignificant coefficient. This is a major flaw, since usually the objective of empirical studies is to find relationship between variables; so anything that makes it more likely to draw wrong inferences is a problem.

7.3.1 Example of Autocorrelated Error

To make this discussion more concrete, we now consider a specific example. Let us consider the two-variable regression model discussed in Unit 2. In order to be more specific that we are dealing with time series data we use the subscript 't' instead of '*i*'.

$$Y_t = \beta_1 + \beta_2 X_t + u_t \quad t = 1,2,...,T \qquad ...(7.3)$$

Let us assume that the error term is generated by the following mechanism

$$u_t = \rho u_{t-1} + \varepsilon_t \qquad ...(7.4)$$

where ρ is also known as the coefficient of autocovariance and where ε_t is the stochastic error term such that it satisfies the standard OLS assumptions, namely,

$$\begin{aligned} E(\varepsilon_t) &= 0 \\ var(\varepsilon_t) &= \sigma_\varepsilon^2 \\ cov(\varepsilon_t, \varepsilon_{t+s}) &= 0 \ \forall \ s \neq 0 \end{aligned} \qquad ...(7.5)$$

which says that this term has constant mean, fixed variance and is independent, not correlated with any other value of the time series ε_s. The technical term for the error generation scheme that we have assumed is a *first order autoregressive scheme*, usually denoted as *AR*(1) (more details on this are given in Unit 16 on time series). Given this scheme, it can be shown (derivations have been skipped for simplicity), that

$$\begin{aligned} var\ (u_t) &= \frac{\sigma_\varepsilon^2}{1-\rho^2} \\ cov(u_t, u_{t+s}) &= \sigma_\varepsilon^2 \\ corr(u_t, u_{t+s}) &= 0 \ \forall \ s \neq 0 \end{aligned} \qquad ...(7.6)$$

where *corr*(*x*,*y*) denotes correlation between variables *x* and *y*. Note first that the variance of the error terms are fixed (Unit 8 takes up the case when the error terms have variable variance). Second, the error term is correlated not only with its immediate past value but also several period's past value as well. It is important to note that it always needs to be that $|\rho|<1$, that is, the absolute value of ρ is less than one. If for example $\rho=1$ then the variances and covariances above are not defined.[2] Under this assumption also note that the value of the covariance declines as we go further into the past.

[2] Note that the assumption that $|\rho|<1$ is the same as assuming a stationary *AR*(1) process. See unit 16 for definition. In short, it means that the mean, variance and covariance of u_t do not change over time.

There are two reasons for using this process (first order autoregressive). First, it is the simplest form of a correlated error term structure. Second, it is the most widely used in terms of applications. A considerable amount of theoretical and empirical work has been done using such processes.

The reasons for using the first order autoregressive process are both because it is the simplest form of a correlated error term structure, as well as it is the most widely used in terms of applications. A considerable amount of theoretical and empirical work has been done using such processes. Note from Unit 4, that the OLS estimate for the slope coefficient is

$$\hat{\beta}_2 = \frac{\sum x_t y_t}{\sum x_t^2} \qquad ...(7.9)$$

and its variance is given by

$$var(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_t^2} \qquad ...(7.9)$$

where y_t is the deviation from the sample mean, that is $y_t = (Y_t - \bar{Y})$ given that $\bar{Y}$ is the mean value of y_t, and similarly for x_t. Now it can be shown under this error scheme that the variance of the OLS estimator is

$$var(\hat{\beta}_2)_{AR(1)} = \frac{\sigma^2}{\sum x_t^2} + \frac{2\sigma^2}{\sum x_t^2}\left[\rho\frac{\sum x_t x_{t+1}}{\sum x_t^2} + \rho^2\frac{\sum x_t x_{t+2}}{\sum x_t^2} + ... + \rho^{n-1}\frac{\sum x_t x_n}{\sum x_t^2}\right] ...(7.11)$$

Note the difference between the variance of the OLS estimator under standard assumptions of no autocorrelation in (7.10) and the new variance of the estimator with autocorrelation (in the form of a first order autoregressive error term), as shown in (7.11). The former is added by a factor which is the multiple of ρ as well as the sample autocorrelation between the values taken by the regressor X at various lags. Note that if ρ is zero then there is no autocorrelation which means that the two values will coincide as expected. We cannot say whether the former is smaller or bigger than the latter, but that they are different. Note in this context that for most economic time series data it is not unreasonable to assume that the regressors are positively correlated (consumption high in one period usually means it will be high in the next period as well), rarely do economic data follow a period to period fluctuations. Usually they follow a cycle (high values for a while during boom, followed by average and then low values during recession) which leads to positive correlation between successive values of the time series.

7.3.2 Problems Arising from Autocorrelation

Let us see what happens if we ignore autocorrelation, when it exists. The standard software package after running a regression usually not only gives the estimate $\hat{\beta}_2$ but also an estimate of its variance (under the standard assumption of no autocorrelation) $var(\hat{\beta}_2)$. Although the estimate $\hat{\beta}_2$ is consistent, that is it converges to the true value of β_2 as sample size increases, some serious issues remain as follows:

1) The residual variance $\hat{\sigma}^2 = \sum \hat{u}_t^2 / (n-2)$ is likely to underestimate the true σ^2.

2) As a result, we are likely to overestimate R^2.

3) Even if σ^2 is not underestimated, from equation (7.11) we saw that $var(\hat{\beta}_2)$ will underestimate $var(\hat{\beta}_2)_{AR(1)}$ (given the factor in parentheses there which is non zero as long as $\rho \neq 0$. It is strictly larger in most cases arising in the real world since the regressors are usually positively autocorrelated themselves (see above).

4) Therefore, the usual t and F tests of significance are no longer valid, and if applied, are likely to give seriously misleading conclusions about the statistical significance of the estimated regression coefficients. It is more likely that we find an insignificant relationship, falsely.

7.4 TESTING FOR AUTOCORRELATION

There are a multitude of methods for testing for autocorrelation, the simple reason being that no particular test has yet been judged to be unequivocally best (i.e., more powerful in a statistical sense). Thus the analyst is in the unenviable position of considering a varied collection of test procedures for detecting the presence of autocorrelation. The three most commonly used tests are discussed here, starting with the simplest to the most computationally complex.

7.4.1 Graphical Method

Although the true values of the error terms u_t are not available to the researcher, the estimated values of the same terms $\hat{u}_t$, are their proxies and can be used. Usually these residuals are generated automatically after an OLS estimation by most statistical software packages. Very often a visual examination of the residuals gives us an indication of the presence of autocorrelation. One can simply plot the estimated residuals $\hat{u}_t$ against time, also called the *time sequence plot*. If autocorrelation exists this will usually show up as several small (large) followed by several large (small) errors. That is, the consecutive values of the residuals will be highly correlated. Alternatively we can also plot the values of $\hat{u}_t$ against $\hat{u}_{t-1}$. If this shows up as a linear pattern then there is autocorrelation. A positive relationship implies positive autocorrelation and vice versa. This is arguably a crude method of examination. The following methods are statistically more powerful in detecting autocorrelation.

7.4.2 Durbin-Watson Test

A popular test for detecting autocorrelation is the Durbin and Watson test statistic defined as

$$d = \sum_{t=2}^{T} (\hat{u}_t - \hat{u}_{t-1})^2 / \sum_{t=1}^{T} \hat{u}_t^2 \qquad \ldots(7.13)$$

which is simply the ratio of the sum of squared differences in successive residuals to the overall error sum of squares (ESS).

If this was based on the true u_t and T was very large then d can be shown to tend in the limit as T gets large to $2(1-\rho)$. This means that if $\rho \to 0$, then $d \to 2$; if $\rho \to 1$ then $d \to 0$ and if $\rho \to -1$ then $d \to 4$. Therefore, a test for $H_0: \rho = 0$, can be based on whether d is close to 2 or not. Unfortunately, the critical values of d depend upon the X, and these vary from one data set to another. To get around this, Durbin and Watson established upper (d_U) and lower (d_L) bounds for this critical value.

- It is obvious that if the observed d is less than the lower bound d_L, or larger than $(4 - d_L)$, then we can safely reject H_0.

- On the other hand if the observed d is between d_U and $(4 - d_U)$ then we do not reject the hypothesis.
- If d lies in any of the two indeterminate areas, then one should compute the critical values depending on X. Most econometric software packages report the Durbin-Watson statistic automatically along with the p-value, that is the level of significance of this statistic (in general for the 5% level of testing if the p-value is less than 0.05 then you can reject the null hypothesis of no autocorrelation).
- If we are interested in a one-tailed test, say $H_0: \rho = 0$ versus $H_1: \rho > 0$, that is positive autocorrelation only, then we would reject H_0 if $d < d_L$, and not reject H_0 if $d < d_U$. If on the other hand $d_L < d < d_U$ then the test is inconclusive. Note that with a one tailed test only one side of the earlier test is operational. The conditions are still the same; the only difference is we ignore one set of conditions. With a two-tailed test we reject H_0 based on both where d lies and where (4-d) lies. Here we are interested only in the value of d since we are testing only for positive autocorrelation versus the null of no autocorrelation. Earlier the alternate hypothesis was any type of autocorrelation.
- Similarly, for testing $H_1: \rho = 0$ versus $H_1: \rho < 0$, we compute (4–d) and follow the same steps as before while testing for negative autocorrelation. Here our alternate hypothesis in only negative autocorrelation, versus the null of no autocorrelation.

These conditions are illustrated in Fig. 7.4.

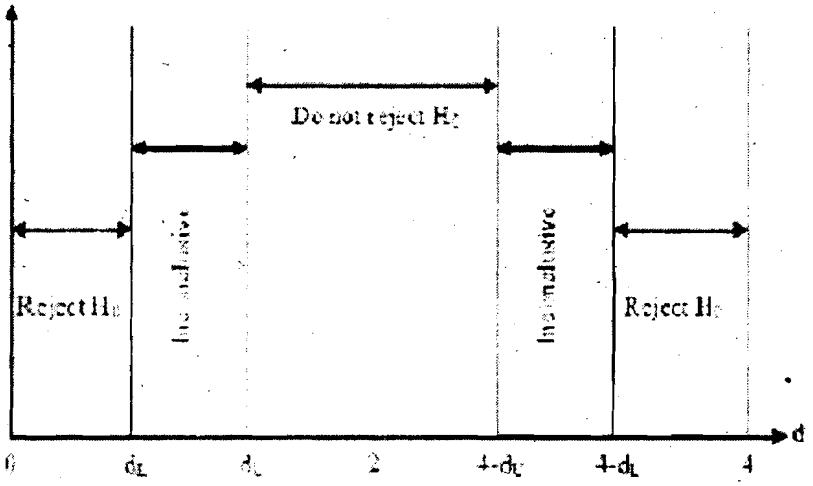

Fig.7.4: Durbin-Watson Test Statistic

The D-W statistic depends upon the number of explanatory variables (excluding the intercept) in the model and the number of observations (n). The critical values of D-W test are given at the end of this Unit in **Annexure-1**. In the table you have to select the row corresponding the number of observations and the desired level of significance. If the observed value is less than d_L or greater than d_u, then there is autocorrelation. If it remains between d_L and d_u, the test is inconclusive.

Limitations : The D-W test has several shortcomings. First, there are certain regions where the d statistic does not provide conclusive evidence. Second, for drawing inferences the critical values need to be calculated. Third, the Durbin-Watson statistic is valid only when there is an intercept term in the regression. Fourth, it is also

invalid in the presence of lagged values of the dependent variable among the regressors. Fifth, this test can only detect a first order autoregression structure of the error term. For instance, if there are higher order autoregression terms, that is, if the error term not only depended on its last period value but on earlier periods values as well, then this test is not valid. Similarly, if the error generation mechanism is a Moving Average (MA) process, a different class of time series data(see unit 15), then also this test is not applicable. A more diverse test is the Breusch-Godfrey test discussed next.

7.4.3 Breusch-Godfrey Test

A more powerful test that is also commonly used in empirical applications is the Breusch-Godfrey (BG) test, also known as the LM test. Using our earlier example of a single variable with intercept model (equation 7.3 above) the test proceeds in the following way. Note in this context that there can be many more variables as regressors, one variable case is taken for simplicity only. Similarly, lagged values of the dependent variable Y_{t-1} etc. can also be there as a regressor. Assume that the error term follows the p-th order autoregressive, $AR(p)$ scheme (extension of the earlier $AR(1)$ scheme), as follows

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \ldots + \rho_p u_{t-p} + \varepsilon_t \qquad \ldots(7.14)$$

where ε_t is a stochastic error as before. The null hypothesis to be tested is:

$$H_0: \rho_1 = \rho_2 \ \ldots = \rho_p = 0 \qquad \ldots(7.15)$$

In (7.15) it is hypothesized that there is no autocorrelation of any order.

The BG test involves the following steps:

1) Estimate (7.3) by OLS and obtain the residuals $\hat{u}_t$.

2) Regress $\hat{u}_t$ on the original x_t (as well as any other variables used as regressors earlier) and $\hat{u}_{t-1}, \hat{u}_{t-2}, \ldots, \hat{u}_{t-p}$, where the latter are lagged values of the estimated residuals in step 1. Thus if p=4, we will introduce four lagged values of the residuals as additional regressors in the model. In short, run the following regression:

$$\hat{u}_t = \beta_1 + \beta_2 X_t + \rho_1 \hat{u}_{t-1} + \rho_2 \hat{u}_{t-2} + \cdots + \rho_p \hat{u}_{t-p} + \varepsilon_t \qquad \ldots(7.16)$$

and obtain from this the *auxiliary* regression.

3) If the sample size is large then Breusch and Godfrey have shown that

$$(T - p)R^2 \sim \chi^2_p \qquad \ldots(7.17)$$

where T is the total number of observations. That is the BG test statistic is asymptotically (as the sample size increases to infinity) distributed as a chi-squared distribution with p degrees of freedom. Therefore, once the test statistic has been calculated this way it can be compared to a standard chi-squared distribution to accept or reject the null hypothesis of no autocorrelation of any order.

Note that unlike the Durbin-Waton test the BG test can test for both higher order schemes of the error process $AR(p)$ not just $AR(1)$. It also is applicable when alternate time series for the error term is specified such as a Moving Average scheme. You will learn in Unit 16 that moving average of order p, that is $MA(p)$, is defined as

follows:

$$u_t = \varepsilon_t + \lambda_1 \varepsilon_{t-1} + \lambda_2 \varepsilon_{t-2} + \cdots + \lambda_p \varepsilon_{t-p} \qquad ...(7.18)$$

Check Your Progress 1

1) Define autocorrelation, when is it more likely to arise in applied work?

.....

.....

.....

.....

2) Consider the following scenarios, which do you think is more likely to exhibit autocorrelation and why?

   a) You are trying to estimate the relationship between industrial production and investment, you have data on several firms collected recently.

   b) You are estimating the same model with aggregate data for India, over the last fifty years.

   c) You are trying to estimate a consumer demand model for a particular commodity, using data on income and expenditure using a household survey data.

   d) You have data on several households over a period of time, and you fit the same model.

.....

.....

.....

.....

.....

3) Explain the applications of Breusch-Godfrey test.

.....

.....

.....

.....

7.5 CORRECTING FOR AUTOCORRELATION

After testing for autocorrelation you find that you cannot accept the null hypothesis of no autocorrelation. What is the next course of action? It would be very simple if we knew the actual value of ρ. If the model in (7.3) holds for time t then it also holds for time (t-1). Therefore,

$$Y_{t-1} = \beta_1 + \beta_2 X_{t-1} + u_{t-1} \qquad ...(7.19)$$

Multiplying (7.19) throughout by ρ and subtracting from (7.3) we get

$$(Y_t - \rho Y_{t-1}) = \beta_1(1-\rho) + \beta_2(X_t - \rho X_{t-1}) + \varepsilon_t \qquad ...(7.20)$$

where $\varepsilon_t = u_t - \rho u_{t-1}$ by definition. Note that the first observation is lost here since a lag has been introduced in the model and therefore you are left with t=2,3,...,T. This transformation is known as the *Cochrane-Orcutt (1949)* transformation and it reduces the error terms to classical errors, that is, since ε_t is distributed identically and independently, all the OLS assumptions are valid for this new regression with transformed variables. Therefore, the OLS estimator has all the optimal properties, that is, it is the best linear unbiased estimator or BLUE. However, in most practical applications ρ is not known and has to be estimated. Some commonly used methods for estimating it are given below.

7.5.1 Non-iterative Methods of Obtaining ρ

A couple of simple methods of obtaining a crude estimate of ρ are as follows:

- *Using the Durbin-Watson statistic*: Since d has been estimated earlier while conducting the test, and we know that as the sample gets large that is as $T\to\infty$, this statistic tends to 2(1–ρ). We can use this fact to construct the estimator:

$$\hat{\rho} = 1 - \frac{d}{2} \qquad (7.21)$$

  Note that this estimator is only good for very large samples and in small samples this gives highly inaccurate results and should be avoided.

- *From residuals* : If the $AR(1)$ scheme $u_t = \rho u_{t-1} + \varepsilon_t$ is true, then ρ can be obtained easily by running a regression of the residuals $\hat{u}_t$ on $\hat{u}_{t-1}$. That is,

$$\hat{u}_t = \rho \hat{u}_{t-1} + \varepsilon_t \qquad t = 2,3,...,T \qquad ...(7.22)$$

  Note that the equation (7.22) does not have a constant. The estimate of ρ obtained is consistent since the $\hat{u}_t$ are consistent estimators of the true u_t. Again, this estimate can be used in the transformation to obtain the BLUE estimates of β_1 and β_2, respectively, after applying the Cochrane-Orcutt transformation given at (7.20).

  Durbin's Method: One can rearrange (7.20) by moving the Y_{t-1} term to the right hand side, so that

$$Y_t = \rho Y_{t-1} + \beta_1(1-\rho) + \beta_2 X_t - \rho\beta_2 X_{t-1} + \varepsilon_t \qquad ...(7.23)$$

  and running OLS on (7.23) instead. Note that no prior estimate of ρ is required here. The error in (7.23) is classical, and we can show that the estimate of ρ obtained as the coefficient of Y_{t-1} in this case is biased but consistent. This is the Durbin estimate of ρ; let us call it $\hat{\rho}_D$. Next, the second step of the Cochrane-Orcutt procedure is performed by using this estimate of ρ. That is, we estimate the regression (7.20) by using this $\hat{\rho}_D$, to get the estimates of β_1 and β_2.

7.5.2 Iterative Methods

Apart from the non-interative methods mentioned above, there are a few iterative methods of obtaining ρ. These are discussed below.

The Cochrane-Orcutt Method

This method starts with an initial estimate of ρ, the most convenient being 0, and the OLS estimates of β_1 and β_2 from (7.3), since we have assumed that $\rho = 0$. Note that if we start with an assumption other than $\rho = 0$ then we need to use equation (7.20) and not (7.3). Once we have the OLS estimates we can calculate the residuals $\hat{u}_t$ also. Then we run the same regression as in (7.22) above. The resulting estimate of ρ is $\hat{\rho}_{CO} = \sum \hat{u}_t \hat{u}_{t-1} / \sum \hat{u}_t^2$ where both summations run over t=2,3,...,T since one observation is lost due to lags. The second step of the Cochrane-Orcutt method is to perform the regression in (7.20) with $\hat{\rho}_{CO}$ instead of ρ.

We can also iterate this procedure by computing new residuals based on the new estimates of β_1 and β_2 from (7.20), which are then used to calculate a new value of $\hat{\rho}_{CO}$ and so on until *convergence* takes place. It means that successive values of the ρ that is estimated do not change substantially, say the difference is less than 0.01.

The Hildreth-Lu Search Procedure

Since we know that ρ lies between 1 and -1, this procedure searches over this range, i.e., using values of ρ say between -0.9 and 0.9 in intervals of 0.1. For each ρ we compute the regression in (7.3) and report the residual sum of squares corresponding to that ρ. The minimum residual sum of squares gives us our choice of ρ and the corresponding regression gives us the estimates of β_1,β_2 and σ^2. We can refine this procedure around the best ρ found in the first stage of the search. For example, suppose that ρ=0.6 gave the minimum residual sum of squares. We can search next between 0.51 and 0.69 in intervals of 0.01. This search procedure guards against a local minimum.

Prais-Winsten Procedure

A problem with the Cochrame-Orcutt transformation (7.20) is the loss of one observation due to lagging. The Prais-Winsten method tries to fill this gap.

It has been shown by Rao and Griliches (1969), using simulated data (randomly generated data – also called Monte Carlo experiments) that the best result can be obtained when one uses the Durbin estimate of ρ discussed above with this procedure. The basic objective is to retain the one lost observation due to lagging. The Prais-Winsten method uses the following steps:

1) Multiply the first observation of the regression equation (7.3) by $\sqrt{1-\rho^2}$, which gives us (7.24). Note that we are multiplying throughout by $\sqrt{1-\rho^2}$

$$\sqrt{1-\rho^2}Y_1 = \beta_1\sqrt{1-\rho^2} + \beta_2\sqrt{1-\rho^2}X_1 + \sqrt{1-\rho^2}u_1 \qquad ...(7.24)$$

2) Add this transformed first observation to the Cochrane-Orcutt transformed observations for t=2,3,...,T and run the regression on all T observations as in (7.20) above. It is easier if we define variables $\tilde{Y}_t = (Y_t - \rho Y_{t-1})$ and $\tilde{X}_t = (X_t - \rho X_{t-1})$ for all t=2,3,...,T. However, for the first observation we define $\tilde{Y}_t = (Y_t - \sqrt{1-\rho^2}Y_{t-1})$ and similarly for $\tilde{X}_t = (X_t - \sqrt{1-\rho^2}X_{t-1})$. Unlike the standard Cochrane-Orcutt method in the second step we can think of this as a regression model without a constant but with two variables: the

first one (earlier constant) now takes the value of $\sqrt{1-\rho^2}$ for the observation and takes the value of $(1-\rho)$ for the remaining as before. The second variable is the standard $X_t - \sqrt{1-\rho^2}X_{t-1}$) for the first observation and $X_t - \rho X_{t-1}$ for other observations.

7.6 NUMERICAL EXAMPLE

In this section we illustrate the methods for testing and correcting for autocorrelation. The standard example is the estimation of the Keynesian consumption function $C_t = \beta_0 + \beta_1 Y_t + \varepsilon_t$. where C_t is consumption expenditure in period t, and Y_t is the total personal disposable income in the same period. Note that we consider the real value of both variables, that is, they are measured in constant prices (1987 dollars). The data used for this exercise was obtained for the United States, for the period 1950—1993 from the Economic Report of the president. It is summarized in Table 7.1.

Table 7.1 : U.S. Consumption Data, 1950–1993
C = Real Personal Consumption Expenditure (in 1987 dollars)
Y = Real Disposable Personal Income (in 1987 dollars)

Year	Y	C	Year	Y	C
1950	6284	5820	1972	10414	9425
1951	6390	5843	1973	11013	9752
1952	6476	5917	1974	10832	9602
1953	6640	6054	1975	10906	9711
1954	6628	6099	1976	11192	10121
1955	6879	6325	1977	11406	10425
1956	7080	6440	1978	11851	10744
1957	7114	6465	1979	12039	10876
1958	7113	6449	1980	12005	10746
1959	7256	6658	1981	12156	10770
1960	7264	6698	1982	12146	10782
1961	7382	6740	1983	12349	11179
1962	7583	6931	1984	13029	11617
1963	7718	7089	1985	13258	12015
1964	8140	7384	1986	13552	12336
1965	8508	7703	1987	13545	12568
1966	8822	8005	1988	13890	12903
1967	9114	8163	1989	14005	13029
1968	9399	8506	1990	14101	13093
1969	9606	8737	1991	14003	12899
1970	9875	8842	1992	14279	13110
1971	10111	9022	1993	14341	13391

Source: Economic Report of the President

The GLS regression using this data is,

$$C_t = -70.957 + 0.916Y_t \qquad \ldots(7.25)$$

$$(90.57) \quad (0.009)$$

where the standard errors are given in parentheses. Note that the negative intercept for the consumption function is not significant (t-statistic is –0.78), whereas the coefficient for income is highly significant (t-statistic is 106.41). The R^2 is also very high at 0.996; together these imply that the statistical relation between the variables is very strong.

7.6.1 Testing for Autocorrelation

In order to check for autocorrelation the first method is the graphical method. The estimated residuals from equation (7.25) are plotted in Fig.7.5. From the figure it can be clearly seen that the residuals are serially correlated; that is, each value of the error term is strongly influenced by the term immediately before it. Note that if there was no serial correlation then you would not see any pattern at all. It for each period the error term is independent of all other errors, it would be a random scatter plot; the residuals would be all over the place. As there is an identifiable pattern in this plot it gives you very strong indication that there is autocorrelation. This pattern is a standard $AR(1)$ pattern, which is the easiest to identify visually. Note that there are other types of autocorrelation that you cannot detect so easily, however they have all the same negative implications for your results. Therefore it is best to say that visual evidence of autocorrelation is sufficient but not necessary for there to exist autocorrelation. That is why you should always test for autocorrelation using more formal methods as well.

Next we consider the Durbin-Watson test for autocorrelation. Note that the lower bound d_L in this case with 44 observations and one regressor is 1.468, The critical values of D-W statistic for various number of variables and observations are given in *Annexure-1* at the end of this unit. The actual statistic obtained in this case (each software package gives this automatically, but you can also calculate it manually using the formula from (7.13) above) is 0.457, which is lower than the lower bound. Therefore, we reject the hypothesis of no autocorrelation.

Similarly, we can use the Breusch-Godfrey test by regressing the residuals on the past values of the same, along with the regressors. For simplicity let us assume a single period lag for the error term, and an autoregressive process. In other words, assume an $AR(1)$ scheme of autocorrelation and then run the regression equation (7.16) from above. The result that we obtain is

$$\hat{u}_t = -53.993 + 0.006Y_t + 0.804\hat{u}_{t-1} \qquad \ldots(7.26)$$

$$(61.4844) \quad (0.006) \quad (0.108)$$

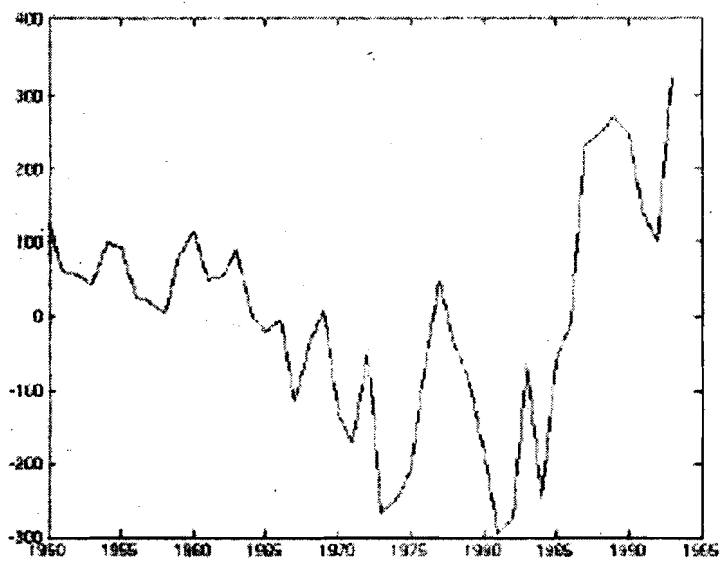

Fig.7. 5: Residuals from OLS Regression

The R^2 of this equation is 0.583. Note that here T, the number of observations, is 44 and the lag value p is 1. Therefore, the BG statistic is $(T-p)R^2$, which is (44–1)0.583=25.069. We know that this is distributed as $\chi_p^2 = \chi_1^2$, with degree of freedom 1, since p is 1. At the 5% level of significance the critical value is 0.49. Thus the observed value of χ^2 is more than the critical value. Therefore, we reject the null hypothesis of no autocorrelation by this test as well. Note that sometimes in applied work it may so happen that the lag structure for the error term is more complicated. In that case you need to try different values of p. The best model selection then has to be using some information criterion such as Bayesian information criterion (BIC) which is beyond the scope of this unit.

7.6.2 Correcting for Autocorrelation

In order to get the correct estimates of β_1 and β_2 in the presence of autocorrelation we need to have some estimate of ρ, the first order autoregressive term in the error mechanism. The easiest but not the best way to obtain it is from the Durbin-Watson statistic, which in this case is 0.457. We know from (7.21) that approximately $\hat{\rho} = (1-d/2)$, which gives us $\hat{\rho} = 0.772$. We can also run the regression of $\hat{u}_t$ on $\hat{u}_{t-1}$ (without a constant) as per(7.22) to give us an estimate of ρ as follows:

$\hat{u}_t = 0.794\,\hat{u}_{t-1}$ with standard error 0.106 for $\hat{\rho}$.

Note that this value of ρ is not substantially different from, the one obtained above.

This only happens when the error structure is a simple one, as chosen here.

Next the Durbin's method involves regressing C_t on last periods value of consumption C_{t-1} as well as Y_t and Y_{t-1}. This gives an estimate of $\hat{\rho}$ as the coefficient obtained of the term C_{t-1} as follows:

$$\begin{matrix} C_t \\ \\ \end{matrix} = \begin{matrix} -41.829 + 0.802C_{t-1} + 0.718Y_t - 0.5296Y_{t-1} \\ (59.11) \quad (0.102) \quad (0.086) \quad (0.127) \end{matrix} \qquad ...(7.27)$$

Therefore, the estimate of ρ obtained by this method is 0.802.

For iterative methods, first consider the Cochrane-Orcutt method. The first step of

second step is to run the regression of $(C_t - 0.794\, C_{t-1})$ on $(Y_t - 0.794\, Y_{t-1})$, which gives

$$(C_t - 0.794C_{t-1}) = \underset{(62.169)}{-35.667} + \underset{(0.027)}{0.927}\,(Y_t - Y_{t-1}) \qquad ...(7.28)$$

So the autocorrelation corrected estimate of the marginal propensity to consume (the coefficient of the Y_t term) is 0.927 as opposed to the uncorrected one obtained earlier of 0.916. Note that this is called the two step estimator or 2SCO. We can continue iterating on this, as follows :

Use the residuals from this new equation, to calculate a new value for $\hat{\rho}$ and then use that instead of the 0.794 we have used in the second step above. This will yield a new set of estimates for β_1 and β_2, if these are very close to the one obtained in the first iteration as reported here, then stop; otherwise do one more round of iteration. The estimator obtained by this method is also known as the iterated Cochrane-Orcutt (ITCO), which can perform better when accuracy is very important.

To illustrate the Prais-Winsten procedure (recall eq. (7.24))one needs to calculate the value of $\sqrt{1-\rho^2}$. Using our earlier value of $\hat{\rho} = 0.794$, we find that $\sqrt{1-\rho^2} =$ 0.608. So for the first observation we have $C_t - 0.608C_{t-1}$, and similarly for the regressor $Y_t - 0.608Y_{t-1}$, for the first observation the constant term is also 0.608. Keep in mind that we have to run a regression with two variables and no constant. For all other observations t=2,3,...,T the earlier procedure of Cochrane-Orcutt is applied. Thus we take $C_t - 0.794C_{t-1}$ as the dependent variable, and for the regressor we have $Y_t - 0.794Y_{t-1}$, the constant term which we are treating as a variable now is (1–0.794)=0.206. Running this regression gives, β_2 or the MPC as 0.927 with standard error of 0.024.

Check Your Progress 2

1) Explain the following methods for correction of autocorrelation in data.

   a) Cochrane - Orcutt method

   b) Prais - Winsten method

   .....

   .....

   .....

   .....

   .....

   .....

2) Explain the non-iterative methods of correcting for the problem of autocorrelation.

   .....

   .....

   .....

7.7 LET US SUM UP

If the assumption of independent errors of the classical linear regression model is violated, in particular if the error terms are correlated across consecutive observations, then the problem of autocorrelation arises. The OLS estimators continue to be unbiased but they are no more BLUE, that is, there are more efficient estimators available. The main problem that arises is that the standard errors of the estimated coefficients are higher, which in turn means that the confidence interval for the estimates are wider. This implies that the researcher is more likely to find an insignificant relationship even when there is a statistically significant relationship between the variables. There are several tests that one can use to detect the presence of autocorrelation, however, no test performs uniformly better than the others. Therefore there is a need to test for autocorrelation using a number of alternative testing procedures, since the underlying error generation process in not known a priori to the researcher. Once autocorrelation is detected, there are several methods for correcting for it, the standard one being the Cochrane-Orcutt method. This method allows a more efficient estimator compared to the OLS and solves many of the problems mentioned before of falsely rejecting a significant statistical relationship.

7.8 KEY WORDS

Cross-section data : It refers to data collected from various units (households, firms, individuals, etc.) at a point of time.

Cochrane-Orcutt transformation : In this method the original regression equation $y_t = \beta_1 + \beta_2 x_t + u_t$ is transformed into the following:

$$(y_t - \rho y_{t-1}) = \beta_1 (1-\rho) + \beta_2 (x_t - \rho x_{t-1}) + \varepsilon_t$$

The Cochrane-Orcutt transformation is used for correction of autocorrelation in data.

First order auto regressive scheme : Here the error term follows the following pattern : $u_t = \rho u_{t-1} + \varepsilon_t$.

Multiple regression model : It refers to a single equation regression model comprising one dependent and many explanatory variables.

Time series data : It refers to data collected from the same unit over a period of time.

7.9 SOME USEFUL BOOKS/REFERENCES

Breusch, T.S. (1978), " Testing for Autocorrelation in Dynamic Linear Models," *Australian Economic Papers*, 17: 334-355

Durbin, J. (1960), "Estimation of Parameters in Time-Series Regression Model," *Journal of the Royal Statistical Society*, Series B, 22: 139-153

Durbin, J. and G. Watson (1950), "Testing for Serial Correlation in Least Squares Regression-I," *Biometrika*, 37: 409-428

Durbin, J. and G. Watson (1951), "Testing for Serial Correlation in Least Squares Regression-II," *Biometrika*, 38: 159-178

Godfrey, L.G. (1978), "Testing Against General Autoregressive and Moving Average Error Models When the Regressors Include Lagged Dependent Variables," *Econometrica* 46:1293-1302

Hildreth, C. and J. Lu (1960), "Demand Relations with Autocorrelated Disturbances," Technical Bulletin 276 (Michigan State University, Agriculture Experiment Station)

Newey, W.K. and K.D.West (1987), "A Simple, Positive Semi-definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, 55:703-708

Prais, S. and C. Winsten (1954), "Trend Estimation and Serial Correlation," Discussion Paper 383 (Cowles Commission: Chicago)

Theil, H. (1978), *Introduction to Econometrics*, (Prentice-Hall: Englewood Cliffs, NJ)

7.10 ANSWERS/HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) See Section 7.2 and answer

2) Problem of autocorrelation is likely to arise in the case of (b) and (d) as these are time series data.

3) See Sub-section 7.4.3 and answer.

Check Your Progress 2

1) See Sub-section 7.5.2 and answer

2) See Sub-section 7.5.1 and answer.

ANNEXURE 1 : DURBIN WATSON CRITICAL VALUES

X variables, excluding the intercept

Observations		1		2		3		4		5	
N	Prob.	D-L	D-U	D-L	D-U	D-L	D-U	D-L	D-U	D-L	D-U
15	0.05	1.08	1.36	0.95	1.54	0.82	1.75	0.69	1.97	0.56	2.21
	0.01	0.81	1.07	0.7	1.25	0.59	1.46	0.49	1.70	0.39	1.96
20	0.05	1.20	1.71	1.10	1.54	1.00	1.68	0.90	1.83	0.79	1.99
	0.01	0.95	1.15	0.86	1.27	0.77	1.41	0.68	1.57	0.60	1.74
25	0.05	1.29	1.45	1.21	1.55	1.12	1.66	1.04	1.77	0.95	1.89
	0.01	1.05	1.21	0.98	1.30	0.90	1.41	0.83	1.52	0.75	1.65
30	0.05	1.35	1.49	1.28	1.57	1.21	1.65	1.14	1.74	1.07	1.83
	0.01	1.13	1.26	1.07	1.34	1.01	1.42	0.94	1.51	0.88	1.61
40	0.05	1.44	1.54	1.39	1.60	1.34	1.66	1.29	1.72	1.23	1.79
	0.01	1.25	1.34	1.20	1.40	1.15	1.46	1.10	1.52	1.05	1.58
50	0.05	1.50	1.59	1.46	1.63	1.42	1.67	1.38	1.72	1.34	1.77
	0.01	1.32	1.40	1.28	1.45	1.24	1.49	1.20	1.54	1.16	1.59
60	0.05	1.55	1.62	1.51	1.65	1.48	1.69	1.44	1.73	1.41	1.77
	0.01	1.38	1.45	1.35	1.48	1.32	1.52	1.28	1.56	1.25	1.60
80	0.05	1.61	1.66	1.59	1.69	1.56	1.72	1.53	1.74	1.51	1.77
	0.01	1.47	1.52	1.44	1.54	1.42	1.57	1.39	1.60	1.36	1.62
100	0.05	1.65	1.69	1.63	1.72	1.61	1.74	1.59	1.76	1.57	1.78
	0.01	1.52	1.56	1.50	1.58	1.48	1.60	1.46	1.63	1.44	1.65

UNIT 8 HETEROSCEDASTICITY

Structure

8.0 OBJECTIVES

After going through this unit, you will be in a position to:

- explain the concept of heteroscedasticity;
- identify the consequences of heteroscedasticity;
- test for the presence of heteroscedasticity;
- take remedial measures for the problem of heteroscedasticity; and
- compare between the problems of autocorrelation and heteroscedasticity.

8.2 INTRODUCTION

In the classical least squares approach that we discussed in Block 1 we assumed that the error terms are identically independently distributed (IID) with mean 0 and variance σ^2. Moreover, it was assumed that the variance remains constant for all observations. This assumption of constant variance, however, does not remain valid always. It may so happen that the errors are mutually uncorrelated (that is, no serial correlation) but have different variances. Thus when the variance of the error term increases or decreases with the dependent variable, we have the case of heteroscedasticity. In other words, we say that the problem of heteroscedasticity arises when the assumption of homoscedasticity – that the variances of the stochastic disturbance term are finite and constant over the sample – is not met. Thus, with heteroscedasticity

$$Var(\varepsilon_i) \neq Var(\varepsilon_j) \qquad i \neq j$$

8.2 SOURCES OF HETEROSCEDASTICITY

Heteroscedasticity, or unequal variance, often occurs in cross-section data. For example, let us consider the savings behaviour of households. We know that savings is relatively higher among households with higher income. It is also likely that the variance of savings among high-income families may be larger than the variance among low-income families. Low-income families do not have much to save, so they cannot differ much in their levels of savings. High-income families, by contrast, exhibit wide ranges, from the miser who saves virtually all his income to the spendthrift who saves virtually nothing. Secondly, there is more freedom of choice for high income families while low income families have to spend a major part of their income on consumption goods. Clearly, a cross-section sample of data on savings and income behaviour to estimate an equation may not satisfy the homoscedasticity assumption. More generally, there is typically a problem of heteroscedasticity in cross-section studies in which there is a large variation in the size of the entities for which data are obtained. These entities could be households with widely different income levels, firms with widely different scales of operation, and nations with widely different levels of output.

There could be several reasons for the presence of heteroscedasticity: There may be certain outliers in the data which would increase or decrease error variance. Secondly, the problem of heteroscedasticity often arises because the scale of a variable varies enormously within the sample. For example, in India the large states have relatively large state domestic product (SDP) compared small states. If we take SDP as a regressor in a model then the variability is likely to increase enormously as we move from small to large states. However, if we take per capita SDP then the variation would not be so prominent. For instance, while Maharashtra's SDP is 80 times that of Manipur, Maharashtra's per capita SDP is just about two times that of Manipur. The point we try to make is that, heteroscedasticity in many cases could be due to incorrect data transformation. Thus variables should be appropriately transformed before running a regression.

8.3 CONSEQUENCES

Heteroscedasticity has two important implications for estimation. The first is that the least squares estimators, while still being linear and unbiased (in the case of finite but differing variances), are no longer efficient. They no longer provide minimum variance estimators among the class of linear unbiased estimators (that is, they are not BLUE). The second implication is that the estimated variance of the least squares estimates is biased; so the usual tests of statistical significance, such as t and F tests are no longer valid. It is thus important to test for and remove heteroscedasticity from the data.

Let us discuss how the changing variance affects the desirable properties of regression coefficients. We will consider two aspects: unbiasedness and minimum variance of the OLS estimate of β. Let the model in deviation form be

$$y_i = \beta x_i + \varepsilon_i \qquad ...(8.1)$$

Then the least squares estimator of $\hat{\beta}$, as we learnt from Unit 2, is

$$\hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2} \qquad ...(8.2)$$

By substituting the value of y_i from (8.1) in the above equation (8.2) we obtain

$$\hat{\beta} = \frac{\sum x_i(\beta x_i + \varepsilon_i)}{\sum x_i^2}$$

By rearranging terms in the above we have

$$\hat{\beta} = \beta \frac{\sum x_i^2}{\sum x_i^2} + \frac{\sum x_i \varepsilon_i}{\sum x_i^2}$$

Under the standard assumption of $E(\varepsilon_i) = 0$, we find that

$$E(\hat{\beta}) = \beta + \frac{E(\sum x_i \varepsilon_i)}{\sum x_i^2} = \beta + \frac{\sum x_i E(\varepsilon_i)}{\sum x_i^2} = \beta \qquad ...(8.3)$$

Notice that variances of the error terms play no role in the proof that least-squares estimators are unbiased and consistent. Thus β is unbiased even in the presence of heteroscedasticity.

The problem lies with the variance of the estimated parameters $\hat{\beta}$. The variance of $\hat{\beta}$ under the assumption of homoscedasticity, as we know from Unit 2, is

$$Var(\hat{\beta}) = \frac{\sigma^2}{\sum x_i^2} \qquad ...(8.4)$$

When there is heteroscedasticity the variance of $\hat{\beta}$ is quite different from (8.4). Let us derive the variance of $\hat{\beta}$ again.

$$Var(\hat{\beta}) = E(\hat{\beta} - \beta)^2$$

$$= E\left(\frac{\sum x_i \varepsilon_i}{\sum x_i^2}\right)^2$$

$$= E\left[\frac{1}{\left(\sum x_i^2\right)^2}(x_1^2\varepsilon_1^2 + x_2^2\varepsilon_2^2 + \cdots + x_n^2\varepsilon_n^2 + 2x_1x_2\varepsilon_1\varepsilon_2 + \cdots\right]$$

$$= \frac{1}{\left(\sum x_i^2\right)^2}\left[x_1^2 E(\varepsilon_1^2) + x_2^2 E(\varepsilon_2^2) + \cdots + x_n^2 E(\varepsilon_n^2)\right] \qquad ...(8.5)$$

The variance of ε_i is not constant in the presence of heteroscedasticity. Suppose that $E(\varepsilon_1)^2 = k_1\sigma^2$, $E(\varepsilon_2)^2 = k_2\sigma^2$ and so on. In general terms we say that $E(\varepsilon_i)^2 = k_i\sigma^2$.

By using the above in (8.5) we find that

$$\hat{Var}(\hat{\beta}) = \frac{\sigma^2}{\left(\sum x_i^2\right)^2}\left(k_1 x_1^2 + k_2 x_2^2 + \cdots + k_n x_n^2\right)$$

$$= \frac{\sigma^2}{\left(\sum x_i^2\right)} \frac{\sum k_i x_i^2}{\sum x_i^2} \qquad \ldots(8.6)$$

The difference between (8.4) and (8.6) is the factor $\frac{\sum k_i x_i^2}{\sum x_i^2}$. If k_i and x_i are positively correlated and $\left(\sum k_i x_i^2 / \sum x_i^2\right) > 1$, then the classical least-squares estimation for the variance of $\hat{\beta}$ will be overestimated.

Thus the variance is different from that when the disturbance was homoscedastic. It means that the least squares estimates are not efficient (that is, not BLUE). It can also be shown that the OLS estimates are not asymptotically efficient and therefore, the tests of significance and confidence limits do not apply.

Heteroscedastic disturbance, therefore, gives us false results whose significance is not liable to test. Hence, before testing for a hypothesis, the homoscedasticity of the disturbance term should be detected.

Check Your Progress 1

1) Explain the concept of heteroscedasticity.

.....

.....

.....

.....

2) State whether the following statements are true or false.

   a) Heteroscedasticity leads to estimates of the regression coefficients that are biased.

   b) The presence of heteroscedasticity in an equation leads to biased estimates of the variances of the estimated parameters.

   c) The tests of statistical significance can be applied in the presence of heteroscedasticity.

3) Show that when there is heteroscedasticity, the variance of the estimated parameter is different from that when it is homoscedastic.

.....

.....

.....

.....

8.4 METHODS OF DETECTION

Although heteroscedasticity does not destroy the unbiased and consistency properties of the OLS estimators, they are no longer efficient. This lack of efficiency makes the usual hypothesis testing procedure faulty. Therefore, remedial measures are clearly called for. Before we make any corrections for heteroscedasticity we should first try to detect the presence of heteroscedasticity in the error terms. Two common methods are used to detect the presence of heteroscedasticity.

1) We plot the residuals ($\hat{\varepsilon}_i$) against the predicted values of the dependent variable and examine the patterns of residuals. If the variance of residuals is not constant but increases or decreases with the dependent variable, then there is a possibility of heteroscedasticity in the data.

2) A more careful method is to break the sample into two or more sub-groups, each corresponding to a single value of the independent variable X, and then compute the error variance of each group. Our null hypothesis is that there is no difference among the variances of these groups. To test this hypothesis we can use the chi-square test assuming that the error terms are normally and independently distributed. However, selection of the break point in the data for formation of these groups is rather arbitrary. Therefore, the test is viewed to be merely an indication of the presence of heteroscedasticity.

8.5 CORRECTIONS FOR HETEROSCEDASTICITY

Once we have detected that there is heteroscedasticity in the error terms, we have to correct this problem. A simple method for correction of heteroscedasticity is the method of weighted least squares. We discuss this estimation technique in two conceptually separate cases based on our knowledge about the error variance. In the first case we assume that error variance is known in advance to us. In the second case we assume that the error variance increases or decreases with one of the explanatory variables.

i) When Error Variance is known

We assume that sufficient prior knowledge is available for values of each of the error variances. The case of known variances occurs occasionally in econometric work but is important here in illustrating how to correct for heteroscedasticity.

Recall that in the OLS method appropriate estimator is obtained by minimising the expression

$$\sum_{i=1}^{n}\left(Y_1 - \hat{\alpha} - \hat{\beta}X_i\right)^2$$

For weighted least squares we have to minimise

$$\sum_{i=1}^{n}\frac{1}{\sigma_i^2}\left(Y_i - \hat{\alpha} - \hat{\beta}X_i\right)^2 \qquad \ldots(8.7)$$

Thus each squared residual is weighted by $\frac{1}{\sigma_i^2}$. Solving for the least-squares parameter estimates and using the variables in deviation form, we obtain

$$\hat{\beta} = \frac{\sum x_i y_i / \sigma_i^2}{\sum x_i^2 / \sigma_i^2}$$

$$= \frac{\sum (x_i / \sigma_i)(y_i / \sigma_i)}{\sum (x_i / \sigma_i)^2}$$

$$= \frac{\sum x_i^* y_i^*}{\sum (x_i^*)^2} \qquad ...(8.8)$$

where $x_i^* = x_i / \sigma_i$, and $y_i^* = y_i / \sigma_i$

The desired estimation procedure is accomplished by weighting the original data and then performing ordinary least-squares estimation on the transformed model. Here the transformed model will be

$$\frac{Y_i}{\sigma_i} = \alpha\left(\frac{1}{\sigma_i}\right) + \beta\left(\frac{X_i}{\sigma_i}\right) + \frac{\varepsilon_i}{\sigma_i} \qquad ...(8.9)$$

The transformed error term is homoscedastic as

$$Var(\varepsilon_i^*) = Var\left(\frac{\varepsilon_i}{\sigma_i}\right) = \frac{1}{\sigma_i^2} Var(\varepsilon_i) = \frac{\sigma_i^2}{\sigma_i^2} = 1 \qquad ...(8.10)$$

This procedure yields efficient estimators as the transformed model by construction satisfies all the assumptions of the classical linear regression model. A major difficulty with this estimation method, however, is that it involves prior information on the error variances which is not possible always. This difficulty, however, can be overcome by making certain assumptions about σ_i^2. For example, we can estimate σ_i^2 from the sample, as we will see in the case of White's heteroscedasticity-consistent estimator (HCE) given in Section 8.6.

ii) When Error Variance Varies Directly with an Independent Variable

The most commonly used assumption is that σ_i^2 is associated with a variable. If the variance of the i^{th}observation is proportional to the square of the explanatory variable, then deflation by this variable results in a model exhibiting homoscedasticity. Suppose the variance of the i^{th} observation is proportional to the square of the explanatory variable, X_i. Thus

$$Var(\varepsilon_i) = \sigma^2 X_i^2 \qquad ...(8.11)$$

In the case of two-variable model, the transformed equation will be

$$\frac{Y_i}{X_i} = \frac{\alpha}{X_i} + \beta + \frac{\varepsilon_i}{X_i}$$

$$= \beta + \frac{\alpha}{X_i} + \frac{\varepsilon_i}{X_i} \qquad ...(8.12)$$

Hence, the variance of the disturbance term is now constant, and we can proceed to apply OLS to the transformed equation (8.12), by regressing $\frac{Y_i}{X_i}$ on $\frac{1}{X_i}$.

In our original regression model α is the intercept and β is the slope parameter. In the transformed regression model (8.12), however, β is the intercept and α is the slope. Therefore, to get back the original model we need to multiply the estimated transformed model by X_i. Notice that by applying OLS method to (8.12) we obtain estimators which are BLUE.

8.6 CONSISTENT ESTIMATES OF VARIANCES

With heteroscedasticity, the biased and inconsistent estimates of the variances of the OLS parameter estimates causes statistical inferences to be invalid. White (1980) has suggested a method for obtaining consistent estimates of variance and covariance of OLS estimates which provides valid statistical tests for large samples. The heteroscedasticity-consistent estimator (HCE) is based on the principle of maximum likelihood.

In the two-variable regression model

$$Var\left(\hat{\beta}\right)=\frac{\sigma^2}{\sum x_i^2}$$

generates a biased estimate of the variance of β. An unbiased estimator is

$$Var\left(\hat{\beta}\right)=\frac{\sum x_i^2\sigma_i^2}{\left(\sum x_i^2\right)^2} \qquad ...(8.14)$$

The heteroscedasticity-consistent estimator uses equation (8.14) as its basis, replacing the unknown σ_i^2 with the squares of the residuals $\hat{\varepsilon}_i$. With HCE estimation, the R^2 for the regression will be the same, but all estimates of standard errors and related statistics will change because they are now consistent estimates. While the use of the HCE estimator does generate consistent variance estimates, it does not provide the most efficient parameter estimates. For efficient estimation, one of the weighted least squares estimation procedures must be used.

8.7 TESTS OF HETEROSCEDASTICITY

It is natural for us to consider whether appropriate statistical procedures can be found to test for heteroscedasticity. We wish to find a test statistic for testing the null hypothesis of homoscedasticity, that is, $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = ... = \sigma_n^2$, where x is the number of observations. The specific alternative hypothesis against which the null hypothesis is to be tested depends on the estimation procedure that is considered to yield the most desirable correction for heteroscedasticity.

While there are a number of specific tests for heteroscedasticity, a useful first procedure is the informal one of examining the pattern of the residuals to see whether estimated variances differ from observation to observation. To do this, we suggest calculating the squares of the residuals, $\hat{\varepsilon}_i^2$ and plot these against one or several variables in the model.

8.7.1 White's Test

We now consider the White's test of heteroscedasticity proposed by Hal White (1980). From the OLS estimates we can find out the residuals $\hat{\varepsilon}_i$. Suppose we use the regression residuals to run the following regression:

$$\hat{\varepsilon}_i^2 = \gamma + \delta Z_i + v_i \qquad ...(8.15)$$

from which we calculate the measure of goodness of fit, R^2. The White's test is based on the fact that when there is homoscedasticity,

$$NR^2 \sim \chi^2$$

with 1 degree of freedom. More generally, when there are p independent Z variables, the distribution will have p degree of freedom.

Note that White's test is very general. To carry it out we need not make any specific assumptions about the nature of heteroscedasticity. Obviously it gives us some advantage – we can test for every form of heteroscedasticity through this test. However, we are exposed to the risk that the power of the White's test is often bad so that the null hypothesis gets too often rejected. The White's test is also sensitive to specification bias, that is, if any relevant variable is omitted from the model. In addition, the White's test is non-constructive in the sense that when we reject the null hypothesis this test does not offer any suggestion regarding future course of action.

8.7.2 Goldfeld-Quandt Test

The Goldfeld-Quandt test is applicable if heteroscedasticity is related to only one of the explanatory variables. Suppose in our regression model the error variance increases with the regressor x_1. In order to carry out the Goldfeld-Quandt test we proced as follows:

- arrange the observations in increasing order of x_1
- omit some observations (say, r out of n observations) in the middle of the data series
- run a regression on the first $n_1 = (n-r)/2$ observations, obtaining the error sum of squares, ESS_1
- run a regression on the last $n_2 = (n-r)/2$ observations, obtaining the error sum of squares, ESS_2
- test the null hypothesis $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = ... = \sigma_n^2$, by using the F-test for the test statistic $F = ESS_2/ESS_1$ for degrees of freedom $\frac{1}{2}(n_1 - r - 2k), \frac{1}{2}(n_2 - r - 2k)$

The omission of r observations from the data set is required for the G-Q test. There is no hard and fast rule for the exact value of r and the choice is quite aubitrary. In practice, about one fourth observations are omitted.

8.7.3 Breusch-Pagan Test

The Gold feld-Quandt test is sensitive to the number of central observations omitted from the data set. Secondly, it is important to identify the variable X with which the error variance is related, as the observations are arranged in increasing order of X.

These limitations of G-Q test is overcome by the B-P test.

The B-P test assumes the error term to be normally distributed. The steps given below are followed while carrying out the test.

- Estimate the regression model by OLS and obtain the estimated residuals
- Obtain $\hat{\sigma}^2 = \frac{1}{n}\sum \hat{\varepsilon}_i^2$, which is the maximum likelihood estimator of variance.
- Construct variable p_i defined as

$$p_i = \frac{\varepsilon_i^2}{\hat{\sigma}^2}$$

- Regress p_i on Z, where Z is a non-stochastic variable (any explanatory variable or combination of some explanatory variables can serve as Z) such that

$$p_i = \alpha_1 + \alpha_2 Z_{2i} + \ldots\ldots\ldots + \alpha_m Z_{mi} + v_i$$

- Obtain ESS from the above regression and take $\theta = \frac{1}{2}$ ESS as the test statistic.

In the above θ follows chi-square distribution with (m–1) degrees of freedom. We compare the θ value obtained above with the critical value of chi-square. If the observed value of χ^2 exceeds the critical value then the hypothesis of homoscedasticity is rejected.

Most of the econometric software give results of the B-P test statistic.

8.8 COMPARISON BETWEEN AUTOCORRELATION AND HETEROSCEDASTICITY

Autocorrelation occurs when the assumption of classical linear regression that errors corresponding to different observations are uncorrelated breaks down. It occurs in both time-series as well as cross-section analysis. Heteroscedasticity, on the other hand, occurs when the assumption of constant error variance, or homoscedasticity, does not satisfy. The existence of heteroscedasticity often occurs in cross-section data.

In both autocorrelation and heteroscedasticity, the least-squares estimators are linear and unbiased but inefficient. While autocorrelation tends to make the variance of the error term relatively large, heteroscedasticity makes estimated variances of least-squares biased. The usual tests of statistical significance such as t and F are, therefore, no longer valid in both the cases.

Check Your Progress 2

1) Discuss the methods of detecting heteroscedasticity.

.....

.....

.....

.....

2) Show how the problem of heteroscedasticity can be corrected.

.....

.....

.....

.....

3) Distinguish between autocorrelation and heteroscedasticity.

.....

.....

.....

.....

8.9 LET US SUM UP

In the classical least-squares approach we assume that the variances of the error terms are finite and constant. When this assumption does not hold, the problem of heteroscedasticity arises. This problem occurs often in cross-section of data. When this problem arises, the least-squares estimators are not efficient, no longer providing minimum variance estimators among the class of linear unbiased estimators, and the estimated variances become biased. In this case the usual tests of statistical significance are no longer valid.

We can test for the presence of heteroscedasticity by using White's test of heteroscedasticity and correct for the problem of heteroscedasticity by using the weighted least-squares estimation method, which has been discussed in this unit. We have also discussed the method for obtaining consistent estimates of variances and covariance in this unit.

8.10 KEY WORDS

Autocorrelation : Autocorrelation occurs when the assumption of classical linear regression that errors corresponding to different observations are uncorrelated breaks down.

Heteroscedasticity : Heteroscedasticity occurs when the assumption of constant error variance, or homoscedasticity, does not satisfy.

Homoscedasticity : When the variances of the stochastic disturbance term are finite and constant over the sample are called homoscedasticity.

8.11 SOME USEFUL BOOKS/REFERENCES

Maddala, G.S. (2002); *Introduction to Econometrics*, John Wiley & Sons Ltd., Delhi.

Pindyck, Robert S. and Daniel L. Rubinfeld (1998); *Econometric Models and Economic Forecasts*, Irwin/McGraw Hill, Singapore.

Hu, The-wei (1975); *Econometrics: An Introductory Analysis*, University Park Press, Baltimore, USA.

8.12 ANSWERS/HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) Go through Sections 8.1 and 8.2.

2) a) False b) True. c) False

3) Go through Section 8.3.

Check Your Progress 2

1) Go through Section 8.4.

2) Go through Section 8.5.

3) Go through Section 8.8.

UNIT 9 ERRORS IN VARIABLES

Structure

9.0 OBJECTIVES

After going through this unit, you should be in a position to:

- explain the concept of errors in variables;
- indetify the consequences of errors in variables;
- explain the concept of instrumental variables;
- test for the presence of measurement errors; and
- take remedial measures to solve for the problem of measurement error.

9.1 INTRODUCTION

In ordinary least squares model we assume that sample observations are measured accurately. All our formulae are based upon the presumption that variables (both explained and explanatory) are measured without error. The only form of error admitted to our model is in the form of disturbance term. Here the error term represents the influence of various explanatory variables that have not accurately been included in the model. However, the assumption may not be realistic, particularly in the case of secondary data.

Variables, both dependent and independent, are measured subject to error. In particular, the available data may not refer to the variable as specified, as in the case of proxy variable, or there may be systematic biases in the collection or publication of data. If the measurement errors are systematic, in general, auxiliary equations can be specified to capture these errors. This unit will focus only on the impact of random measurement errors on the regression model.

9.2 CONSEQUENCES OF ERRORS IN VARIABLES

Most of the published data or summary information contain errors of summarizing or misrepresenting by informer. When these data are used, one of the assumptions of classical least-squares method is violated. In this case, the classical least-squares estimator will be biased even when sample size increases, which is alternatively known as asymptotic biases or inconsistency.

Under the classical assumptions the ordinary least squares (OLS) estimators are best linear unbiased. One of the major underlying assumptions is the interdependence of regressors from the disturbance term. If this condition does not hold, OLS estimators are biased and inconsistent. This statement may be illustrated by simple errors in variables.

We discuss about the consequences in the following, if the error appears in the measurement of dependent variable, independent variable or both.

9.2.1 Measurement Error in *Y*

Let us assume that only the dependent variable contains error of measurement. Assume that the true regression model (written in deviation form) is

$$y_i = \beta x_i + \varepsilon_i \qquad ...(9.1)$$

where ε_i represents errors associated with the specification of the model (the effects of omitted variables, etc.).

Assume that the variable y^*, instead of y, is obtained in the measurement process such that

$$y_i^* = y_i + u_i, \qquad ...(9.2)$$

The measurement error u_i is not associated with the regressor. Thus we have

$$\text{Cov}\,(u_i, x_i) = 0$$

The regression model is estimated with y^* as the dependent variable, with no accoun being taken of the fact that y^* is not an accurate measure of y. Therefore, instead o estimating Eq. (9.1), we estimate

$$y^* = y_i + u_i = (\beta x_i + \varepsilon_i) + u_i$$

$$\text{or,} \quad y^* = \beta x_i + (\varepsilon_i + u_i)$$

$$\text{or,} \quad y^* = \beta x_i + e_i \qquad ...(9.3)$$

where e_i is a composite error term, containing the population disturbance term ε_i (which may be called the error term in the equation) and the measurement error term u_i.

For simplicity let us assume that $E(\varepsilon_i) = E(u_i) = 0$, Cov $(x_i, \varepsilon_i) = 0$ (which is the assumption of classical linear regression), and Cov $(x_i, u_i) = 0$, i.e., the errors of

measurement in y^* are uncorrelated with x_i, and Cov (ε_i, u_i) = 0, the equation error and measurement error are uncorrelated.

With these assumptions, it can be shown that β estimated from either (9.1) or (9.3) will be an unbiased estimator of the true β. Thus, the errors of measurement in dependent variable y^* do not destroy the unbiased property of the OLS estimators. However, the variances and standard errors of β estimated from (9.1) and (9.3) will be different because, employing the usual formula, we obtain

For (9.1), $$\text{Var}(\hat{\beta}) = \frac{\sigma_\varepsilon^2}{\sum x_i^2} \quad ...(9.4)$$

For (9.3), $$\text{Var}(\hat{\beta}) = \frac{\sigma_e^2}{\sum x_i^2} = \frac{\sigma_\varepsilon^2 + \sigma_u^2}{\sum x_i^2} = \frac{\sigma_\varepsilon^2}{\sum x_i^2} + \frac{\sigma_u^2}{\sum x_i^2} \quad ...(9.5)$$

Obviously, the variance given at (9.5) is larger than the variance given at (9.4). Therefore, although the errors of measurement in the dependent variable still give unbiased estimates of the parameters and their variables, the estimated variances are now larger than the case where there are no such errors of measurement.

9.2.2 Measurement Error in *X*

We have taken the true regression model deviation form to be

$$y_i = \beta x_i + \varepsilon_i \quad ...(9.1)$$

Now let us assume that explanatory variable x_i is measured with error and the observed value becomes x_i^* such that

$$x_i^* = x_i + v_i$$

or, $$x_i = x_i^* - v_i$$

Putting the value of x_i in (9.6), we have

$$y_i = \beta(x_i^* - v_i) + \varepsilon_i$$

$$= \beta x_i^* - \beta v_i + \varepsilon_i$$

$$= \beta x_i^* + (\varepsilon_i - \beta v_i)$$

$$= \beta x_i^* + w_i$$

where $w_i = \varepsilon_i - \beta v_i$. We have assumed earlier that the measurement error in x is normally distributed with zero mean, it has no serial correlation, and it is uncorrelated with ε_i. However, we can no longer assume that the composite error term w_i is independent of the explanatory variable x_i^*.

$$\begin{aligned}
\text{Cov}(w_i, x_i^*) &= E[w_i - E(w_i)][x_i^* - E(x_i^*)] \\
&= E[w_i - 0][x_i^* - x_i)] \\
&= E[\varepsilon_i - \beta v_i] v_i \\
&= E[(\varepsilon_i - v_i) - \beta v_i^2)] \\
&= 0 - \beta\sigma_v^2 \\
&= -\beta\sigma_v^2 \qquad \ldots(9.6)
\end{aligned}$$

The composite error term w_i has mean zero as $E(w_i) = E[\varepsilon_i - \beta v_i] = E(\varepsilon_i) - \beta E(v_i) = 0$.

Thus, the explanatory variable and the error term are correlated, which violates the crucial assumption of the classical linear regression model that the explanatory variable is uncorrelated with the stochastic disturbance term. If this assumption is violated, then OLS estimates are not only biased but also inconsistent; they remain biased asymptotically. Let us now show this.

Instead of being able to estimate true relationship (9.1), we are forced to replace x_i by x_i^*. Now the OLS estimator is

$$\begin{aligned}
\hat{\beta} &= \frac{\sum x_i^* y_i}{\sum x_i^{*2}} \\
&= \frac{\sum (x_i + v_i) y_i}{\sum (x_i + v_i)^2} \\
&= \frac{\sum (x_i + v_i)(\beta x_i + \varepsilon_i)}{\sum (x_i + v_i)^2} \\
&= \frac{\beta\sum x_i^2 + \beta\sum x_i v_i + \sum x_i \varepsilon_i + \sum v_i \varepsilon_i}{\sum x_i^2 + \sum v_i^2 + 2\sum x_i v_i}
\end{aligned}$$

From above we find that $\hat{\beta}$ is a biased estimate as $E(\hat{\beta}) \neq \beta$. Let us see the asymptotic properties of $\hat{\beta}$.

Since v_i and ε_i are stochastic and are uncorrelated with each other as well as with x_i, we can say that

$$\text{plim } \hat{\beta} = \text{plim } \frac{\beta\sum x_i^2}{\sum x_i^2 + \sum v_i^2}$$

$$= \frac{\beta Var(x)}{Var(x)+\sigma_v^2} \text{ (Since } x_i = X_i - \bar{x} \text{ by definition)}$$

$$= \frac{\beta\sigma_x^2}{\sigma_x^2+\sigma_v^2}$$

$$= \beta \frac{\beta\sigma_x^2}{\sigma_x^2+\sigma_v^2}$$

$$= \beta\frac{1}{1+\frac{\sigma_v^2}{\sigma_x^2}} < \beta \qquad ...(9.7)$$

Thus, $\hat{\beta}$ is biased even for an infinite sample and $\hat{\beta}$ is an inconsistent estimator of β.

9.2.3 Measurement Errors in both *X* and *Y*

Now let us assume that both X and Y have errors of measurement. The true model is as before

$$y_i = \beta x_i + \varepsilon_i$$

Since X and Y have errors of measurement, we observe x_i^* and y_i^* instead of x_i and y_i, such that

$$y_i^* = y_i + u_i, \qquad x_i^* = x_i + v_i,$$

where u_i and v_i present the errors in the values of y_i and x_i respectively. We make the following assumptions about the error terms:

(i) There is no correlation between the error term and corresponding variable, i.e.,

$$E(y_i, u_i) = 0, E(x_i, v_i) = 0$$

(i) There is no correlation between error of one variable and measurement of the other variable, i.e.,

$$E(u_i, x_i) = 0, E(v_i, y_i) = 0$$

(i) There is no correlation between errors in measurement of both the variables, i.e.,

$$E(u_i, v_i) = 0$$

On the basis of the above assumptions, our estimated regression equation will be

$$y_i = \beta x_i$$

or, $$y_i^* - u_i = \beta (x_i^* - v_i)$$

or, $$y_i^* = \beta x_i^* - \beta v_i + u_i$$

$$\text{or,} \qquad y_i^* = \beta x_i^* + (u_i - \beta v_i) \qquad \ldots(9.8)$$

Equation (9.8) shows that if we model y^* as a function of x^*, the transformed disturbance contains the measurement error in β. We then write the OLS estimator '$\hat{\beta}$' as

$$\hat{\beta} = \frac{\sum x_i^* y_i^*}{\sum x_i^{*2}}$$

By substiuing the values of x^* and y^* in terms of x and y we find that

$$\hat{\beta} = \frac{\sum (x_i + v_i)(y_i + u_i)}{\sum (x_i + v_i)^2}$$

$$= \frac{\sum (x_i + v_i)(\beta x_i + u_i)}{\sum (x_i + v_i)^2}$$

$$= \frac{\beta \sum x_i^2 + \beta \sum x_i v_i + \sum x_i u_i + \sum v_i u_i}{\sum x_i^2 + \sum v_i^2 + 2\sum x_i v_i}$$

From above we observe that $E(\hat{\beta}) \neq \beta$.

$$\text{plim } \hat{\beta} = \text{plim } \frac{\beta \sum x_i^2}{\sum x_i^2 + \sum v_i^2}$$

$$= \frac{\beta Var(x)}{Var(x) + \sigma_v^2}$$

$$= \frac{\beta \sigma_x^2}{\sigma_x^2 + \sigma_v^2}$$

$$= \beta \frac{\sigma_x^2}{\sigma_x^2 + \sigma_v^2}$$

$$= \beta \frac{1}{1 + \frac{\sigma_v^2}{\sigma_x^2}} < \beta \qquad \ldots(9.9)$$

Thus, $\hat{\beta}$ will not be a consistent estimator of β. The presence of measurement error of the type in question will lead to an underestimate of the true regression parameter if ordinary least squares are used.

Example 9.1

A specific illustration of this problem in economics is the *permanent income theory of the consumption function*, which is an errors-in-variables model. According to this model measured income, Y^*, has a permanent (true) component, Y, and a

transitory component, Y_ε such that

$$Y^* = Y + Y_\varepsilon \qquad ...(9.10)$$

Measured consumption expenditure, C^*, similarly, has a permanent (true) component, C, and a transitory component, C_ε such that

$$C^* = C + C_\varepsilon \qquad ...(9.11)$$

As you know form MEC-002, Block 4, the permanent consumption (C) is proportional to permanent income (Y). Due to transitary components, however, the average propensity to consume declines as income of a household increases. Let us analyse this issue through errors in variables.

The transitory components represent accidental or chance factors, including cyclical variations. They can be treated as the errors in measuring income and consumption. It is usually assumed that errors, on the average, vanish:

$$E\ (Y_\varepsilon) = E\ (C_\varepsilon) = 0 \qquad ...(9.12)$$

It is also usually assumed that the errors are correlated neither with each other nor with the permanent components:

$$\text{Cov}\ (Y, Y_\varepsilon) = \text{Cov}\ (C, C_\varepsilon) = \text{Cov}\ (Y_\varepsilon, C_\varepsilon) = 0 \qquad ...(9.13)$$

According to permanent-income hypothesis,

$$C = \beta_1 Y + \beta_2 + u \qquad ...(9.14)$$

that is, permanent consumption is a linear function of permanent income, where

β_1 and β_2 are constant parameters and u is a stochastic disturbance term. Notice that the parameter β_1 in (9.14) represents marginal propensity to consume (MPC). Combining (9.14) with (9.10) and (9.11) yields

$$C^* = \beta_1 Y^* + \beta_2 + (v - \beta_1 Y_\varepsilon), \quad \text{where} \quad v = u + C_\varepsilon \qquad ...(9.15)$$

Equation (9.15) is a simple linear regression model of the same form as (9.8). Then,

$$p\text{lim}\ \hat{\beta}_1 < \beta_1 \quad \text{(as we know from (9.9))}$$

Since the ratio β_1 is less than unity, we observe that the above term is less than the true parameter.

Thus,.the estimated marginal propensity to consume, $\hat{\beta}_1$, even in the probability limit, systematically underestimates the true β_1. The greater the variance of transitory income, the greater will be the underestimation. Only when the variance of transitory (error) component of income is zero, the least squares estimator of β_1 is a consistent estimator. To solve this problem, it is necessary to construct a measure of permanent income Y before estimating the consumption function.

Check Your Progress 1

1) Explain the concept of errors in variable. What are its consequences?

.....

.....

.....

.....

2) State whether the following statements are true or false.

   a) Errors in variables lead to estimates of the regression coefficients that are biased.

   b) The presence of measurement error in an equation leads to an underestimate of the true regression parameter if ordinary least squares are used.

.....

.....

.....

.....

3) Explain briefly why measurement error in the explanatory variables leads to biased and inconsistent parameter estimates while measurement error in the dependent variable does not.

.....

.....

.....

.....

9.3 INSTRUMENTAL VARIABLES METHOD

The problem of errors in the measurement of variables in regression models is quite important, yet econometricians do not have much to offer in the way of useful solutions. As a general rule, we tend to pass over the problem of measurement error, hoping that the errors are small enough which will not destroy the validity of the estimation procedure. One technique which is available and can solve the measurement error problem is the technique of instrumental variables estimation. We briefly outline the concept of instrumental variables because it is likely to be useful with measurement errors.

The method of instrumental variables involves the search for a new variable Z which is highly correlated with the independent variable X and at the same time uncorrelated with the error term in the equation (as well as the errors of measurement of both variables). In practice, we are concerned with the consistency of parameter estimates and therefore concentrate on the relationship between the variable Z and the remaining variables in the model when the sample size gets large. We define the random variable Z to be an *instrument* if the following conditions are met.

1) The correlations between Z and ε, u, and v, respectively, in the regression approach zero, as the sample size gets large.

2) The correlation between Z and X is nonzero, as the sample size gets large.

We simply select one instrument (or combination of instruments) that has the highest correlation with the X variable.

Assuming for the moment that such a variable can be found, we can alter the least squares regression procedure to obtain estimated parameters that are consistent. Unfortunately, there is no guarantee that the estimation process will yield unbiased parameter estimates. To simplify the matter, let us consider the case of measurement errors in the independent variable such that $y_i = \beta x_i + \varepsilon_i$ and only x is measured with error (as $x^* = x + v$). In order to solve the problem we take the regression equation $y_i = \beta z_i + \varepsilon_i$ Where Z is the instrumental variable. The instrumental variables estimator of the regression slope in the above model is

$$\hat{\beta} = \frac{\sum y_i z_i}{\sum x_i^* z_i} \qquad ...(9.16)$$

The choice of this particular slope formula is made so that the resulting estimator will be consistent. To see this, we can derive the relationship between the instrumental-variables estimator and the true slope parameter as follows:

$$\begin{aligned}
\hat{\beta} &= \frac{\sum y_i z_i}{\sum x_i^* z_i} \\
&= \frac{\sum \left(\beta x_i^* + \varepsilon_i^*\right) z_i}{\sum x_i^* z_i} \\
&= \frac{\beta \sum x_i^* z_i + \sum z_i \varepsilon_i^*}{\sum x_i^* z_i} \\
&= \beta + \frac{\sum z_i \varepsilon_i^*}{\sum x_i^* z_i}
\end{aligned}$$

Clearly, the choice of Z as an instrument guarantees that $\hat{\beta}$ will approach β as the sample size gets large [Cov (z, ε^*) approaches 0] and will therefore be a consistent estimator of β. Remember that the variable x_i^* in (9.16) was not replaced by z_i in the denominator of instrumental variables estimator, as the estimator $\frac{\sum y_i z_i}{\sum z_i^2}$ does not yield a consistent estimator of β. Let us show this.

$$\begin{aligned}
\hat{\beta} &= \frac{\sum y_i z_i}{\sum z_i^2} \\
&= \frac{\sum y_i z_i}{\sum z_i^2} \\
&= \frac{\sum \left(\beta x_i^* + \varepsilon_i^*\right) z_i}{\sum z_i^2}
\end{aligned}$$

$$= \frac{\beta\sum x_i^* z_i + \sum z_i \varepsilon_i^*}{\sum z_i^2}$$

$$= \frac{\beta\sum (x_i + v_i) z_i + \sum z_i \varepsilon_i^*}{\sum z_i^2}$$

$$= \frac{\beta\sum x_i z_i + \beta\sum z_i v_i + \sum z_i \varepsilon_i^*}{\sum z_i^2}$$

$$p\lim \hat{\beta} = p\lim \frac{\beta\sum x_i z_i}{\sum z_i^2}$$

$$= \beta \frac{Cov(x_i, z_i)}{\sigma_z^2} \neq \beta$$

The above shows that $\frac{\sum y_i z_i}{\sum z_i^2}$ will not yield a consistent estimator of β.

The technique of instrumental variables appears to provide a simple solution to a difficult problem. We have defined an estimation technique, which yields consistent estimates if we can find an appropriate instrument.

A few concluding comments may be instructive.

i) First, the OLS estimation technique is actually a special case of instrumental variables. This follows because in the classical regression model X is uncorrelated with the error term and because X is perfectly correlated with itself.

ii) Second, if we generalize the measurement error problem to errors in more than one independent variable, one instrument is needed to replace each of the designated independent variables.

iii) Finally, we repeat that instrumental-variables estimation guarantees consistent estimation but does not guarantee unbiased estimation.

9.4 TEST OF MEASUREMENT ERRORS

Suppose we are estimating the two-variable regression model

$$y_i = \beta x_i + \varepsilon_i \qquad ...(9.1)$$

There is a possibility that x might be measured with error.

If $x_i = x_i^* - v_i$, then the actual least squares regression would be

$$y_i = \beta (x_i^* - v_i) + \varepsilon_i$$

$$= \beta\, x_i^* - \beta\, v_i + \varepsilon_i$$

$$= \beta\, x_i^* + (\varepsilon_i - \beta\, v_i)$$

$$= \beta x_i^* + \varepsilon_i^* \qquad ...(9.17)$$

where $\varepsilon_i^* = \varepsilon_i - \beta v_i$.

If x is measured with error, we have seen that consisent estimator of β can be obtained by using an instrument z which is correlated with x^* but uncorrelated with ε and v. Suppose the relationship between z and x^* is given by

$$x_i^* = \gamma z_i + w_i \qquad ...(9.18)$$

When (9.18) is estimated using least squares, we obtain

$$\hat{x}_i^* = \hat{\gamma} z_i$$

or $$x_i^* = \hat{x}_i^* + \hat{w}_i \qquad ...(9.19)$$

where $\hat{w}_i$ are the regression residuals. Substituting the value of eq. (9.19) into eq. (9.17) we have

$$y_i = \beta \left(\hat{x}_i^* + \hat{w}_i\right)$$

$$y_i = \beta \hat{x}_i^* + \beta \hat{w}_i + \varepsilon_i^* \qquad ...(9.20)$$

Whether or not there is measurement error, the coefficient of $\hat{x}_i^*$ will be consistently estimated by ordinary least squares, since

$$\text{plim} \frac{\sum \hat{x}_i^* \varepsilon_i^*}{N} = \text{plim} \frac{\hat{\gamma} \sum z_i (\varepsilon_i - \beta v_i)}{N} = 0.$$

In fact, the least-squares estimator of the coefficient of $\hat{x}_i^*$ in eq. (9.20) is identically equal to the instrumental-variables estimator, which is given by

$$\hat{\beta} = \frac{\sum y_i z_i}{\sum x_i^* z_i}.$$

To look at the coefficient of the variable $\hat{w}_i$, note that

$$\text{plim} \frac{\sum \hat{w}_i \varepsilon_i^*}{N} = \text{plim} \frac{\sum \left(x_i^* - \hat{\gamma} z_i\right)\left(\varepsilon_i - \beta v_i\right)}{N}$$

$$= \text{plim} \frac{-\beta \sum v_i \left(x_i + v_i\right)}{N}$$

$$= -\beta \sigma_v^2$$

When there is no measurement error, $\sigma_v^2 = 0$, so that OLS applied to eq. (9.20) will generate a consistent estimator of the coefficient of $\hat{w}_i$. However, when there is measurement error, the coefficient will be estimated inconsistently.

This suggests a relatively easy measurement error test. Let δ represent the coefficient of the variable $\hat{w}_i$ in eq. (9.20).

Substituting $\hat{x}_i^* = x_i^* - \hat{w}_i$, we get

$$y_i = \beta(x_i^* - \hat{w}_i) + \delta\hat{w}_i + \varepsilon_i^*$$

or $$y_i = \beta x_i^* + (\delta - \beta)\hat{w}_i + \varepsilon_i^* \qquad ...(9.21)$$

With no measurement error, $\delta = \beta$, so that the coefficient of $\hat{w}_i$ should equal zero. However, with measurement error, $\delta \neq \beta$, and the coefficient will (in general) be different from zero. We can test for measurement error by doing a simple two-stage procedure. First, we regress x^* on z to obtain the residuals $\hat{w}$. Then, we regress y on x^* and $\hat{w}$ and perform a t test on the coefficient of the $\hat{w}$ variable. If we are concerned with measurement error in more than one variable of a multiple regression model, an equivalent F test could be applied.

The test just described is a special case of a more general test for specification error proposed by Hausman. The Hausman specification test relies on the fact that under the null hypothesis the ordinary least-squares estimator of the parameters of the original eq. (9.17) is consistent and (for large samples) efficient, but is inconsistent if the alternative hypothesis is true. However, the instrumental-variables estimator [the least-squares estimator of eq. (9.20)] is consistent whether or not the null hypothesis is true, although it is inefficient if the null hypothesis is not valid.

The Hausman specification test is as follows: If there are two estimators $\hat{\beta}_1$ and $\hat{\beta}_2$ that converge to the true value β under the null but converge to different values under the alternative, the null hypothesis can be verified by testing whether the probability limit of the difference of the two estimators is zero.

Example 9.2

The expenditure of states and local governments (EXP) in US vary substantially by state and region. Among the important variables that explain differences in spending levels are federal grants-in-aid (AID), the income of states (INC), and the population of states (POP). When a model that relates the dependent variable EXP to the independent variables AID, INC, and POP was estimated by ordinary least squares, using census data for 50 states, the following results were obtained (with t statistics in parentheses):

$$\begin{array}{lcccc} \text{EXP} & = & -46.81 + 0.00324\,\text{AID} + 0.00019\,\text{INC} - 0.597\,\text{POP} \\ & & (-0.56) \quad (13.64) \qquad (8.12) \qquad (-5.71) \\ & & R^2 = 0.993 \qquad F = 2190 \end{array}$$

There is an important possible source of measurement error in the AID variable. This component can be subject to substantial error as the state aid programmes involve fixed sums of money, and the money involved are easy to measure even before state and local budgets are set.

We can test the presence of measurement error by using a Hausman specification test. To perform the test, we use the population of primary and secondary school children (PS) as an instrument. The test proceeds in two stages. In the first stage AID is regressed on PS, and the residual variable $\hat{w}_i$ is calculated as follows:

$$\hat{w}_i = \text{AID} - 77.95 + 0.845\ \text{PS} \qquad R^2 = 0.87$$

$$(-1.28) \quad (18.02)$$

In the second stage $\hat{w}_i$ is added to the original regression to correct for measurement error. The resulting equation is

$$\text{EXP} = -138.51 + 0.00174\ \text{AID} + 0.00018\text{INC} - 0.275\ \text{POP} + 1.372\ \hat{w}_i$$

$$(-1.41) \quad (1.94) \qquad (7.55) \qquad (-1.29) \qquad (1.73)$$

A two-tailed t test of the null hypothesis that there is no measurement error would be accepted at the 5 per cent level, since $1.73 < 1.96$. However, measurement error would be important if we were using either a one-tail test, or a two-tailed test at 10 per cent significance level. Note that correcting for the possibility of measurement error has substantially lowered the coefficient on the AID variable, suggesting that measurement error causes the effect of AID on spending to be overstated.

9.5 INVERSE REGRESSION

When we have the variables y and x both measured with errors (the observed values being y^* and x^*), we consider two regreesion equations:

1) Regression of y^* on x^*, which is called the "direct" regression.

2) Regression of x^* on y^*, which is called the "inverse" or "reverse" regression.

Inverse regression has been frequently advocated in the case of analysis of salary discrimination. Since the problem is one of usual errors in variables, both regressions need to be computed and whether inverse regression alone gives the correct estimates depends on the assumptions we make.

The usual model, in its simplest form, is that of two explanatory variables, one of which is measured with error

$$y = \beta_1 x_1 + \beta_2 x_2 + \varepsilon \qquad \ldots(9.22)$$

where y = salary

x_1 = true qulifications

x_2 = gender

What we are interested in is the coefficient of x_2. The problem is that x_1 is measured with error.

Let x_i^* = measured qualifiations

$$x_i^* = x_i + v_i$$

Suppose we adopt the notion that

$$x_2 = \begin{cases} 1, \text{ for men} \\ 0, \text{ for women} \end{cases}$$

Then $\beta_2 > 0$ implies that men are paid more than women with the same qualifications and thus there is gender discrimination. A direct least squares estimation of eq. (9.22) with x_1^* substituted for x_1, and $\hat{\beta}_2 > 0$ has been frequently used as evidence of gender discrimination.

In the inverse regression we take

$$x_1^* = \gamma_1 y + \gamma_2 x_2 + w \qquad ...(9.23)$$

We are asking whether men are more or less qualified than women having the same salaries. The proponents of inverse regression argue that to establish discrimination, one has to have $\hat{\gamma}_2 < 0$ in eq. (9.23); that is, among men and women receiving equal salaries, men possess lower qualifications.

The usual errors in variables results however show that we should not make inferences on the basis of $\hat{\beta}_2$ and $\hat{\gamma}_2$ but obtain bounds for β_2, from the direct regression and the inverse regression estimates.

Check Your Progress 2

1) . Show that in the two-variable model $\hat{\beta} = \frac{\sum y_i z_i}{\sum z_i^2}$.(where z is an instrument) will not yield a consistent estimate of the true slope parameter.

.....

.....

.....

.....

2) What is the Hausman specification test? How would you carry out a Hausman specification test to evaluate the presence or absence of measurement error?

.....

.....

.....

.....

9.6 LET US SUM UP

In ordinary least squares model we assume that sample observations are measured without error, which is always not true. When this assumption does not hold, OLS estimators are biased and inconsistent. Errors may appear in the measurement of dependent variable, independent variable or both. When there is error in dependent variable, this does not destroy the unbiased property of the OLS estimators but the estimated variances are larger than the case where there is no such errors of measurement. On the other hand, when there is error in the independent variable, or both dependent and independent variables, then OLS estimates are not only biased but also inconsistent.

We can test for the presence of measurement error by using Hausman specification test. One technique which is available and can solve the measurement error problem is the technique of instrumental variables estimation.

9.7 KEY WORDS

Asymptotic bias : When the classical least-squares estimator is biased even with the increase in sample size, it is known as asymptotic bias or inconsistency.

Equation error : It represents the influence of various explanatory variables that have not accurately been included in the model and is in the form of disturbance term.

Errors in Variables : When errors are found in the measurement of variables, we say that there are errors in variables. These errors can substantially alter the properties of the estimated regression parameters.

Explanatory variable : It is the independent variable which explains the dependent variable.

Instrumental variable : It is a random variable, which is highly correlated with the independent variable and at the same time uncorrelated with the error term in the equation.

Inverse regression : It is the reverse of direct regression. If we consider regression of Y on X as direct regression, then regression of X on Y is inverse regression.

Measurement error : When either dependent variable, or independent variable or both, are measured subject to error, we say that there is measurement error.

Parameter : It is a measure of some characteristics of the population.

Proxy variable : It is a variable, which is proxy for the true variable.

Specification error : Errors found in the specification of the model.

Stochastic disturbance term : It is the random error term whose value is based on an underlying probability distribution.

9.8 SOME USEFUL BOOKS/REFERENCES

Intriligator, Michael, Ronald Bodkin and Cheng Hsiao (1996); *Econometric Models, Techniques, and Applications*; Prentice-Hall International Inc., New Jersey.

Johnston, Jack and John Dinardo (1997); *Econometric Methods*, The McGraw-Hill Companies Inc., Singapore.

Maddala, G.S. (2002); *Introduction to Econometrics*, John Wiley & Sons Ltd., Delhi.

Pindyck, Robert S. and Daniel L. Rubinfeld (1998): *Econometric Models and Economic Forecasts*, Irwin/McGraw-Hill, Singapore.

9.9 ANSWERS/HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) Go through Section 9.1 and 9.3.

2) a) true b) true.

3) Go through Section 9.3

Check Your Progress 2

1) Go through Section 9.4.

2) Go through Section 9.5 and Example 9.2.