BHARATHIDASAN UNIVERSITY

Tiruchirappalli- 620024

Tamil Nadu, India.

Programme: M.Sc. Statistics

Course Title : Econometrics

Course Code: 23ST03DEC

Unit-II

Econometrics

Dr. T. Jai Sankar
Associate Professor and Head
Department of Statistics

Ms. E. Devi
Guest Faculty
Department of Statistics

Single equation linear model static case – Ordinary least square model and generalized least squares model: Introduction – estimation and prediction – Problem of multicollinearity and heteroscedasticity – Causes, consequences and solutions of and estimation

SINGLE EQUATION LINEAR MODELS (STATIC CASE)

In econometrics and statistical analysis, the **single equation linear model** forms the foundation for understanding relationships between a dependent variable and one or more independent variables. In the static case, the relationship is assumed to hold at a given point in time without incorporating time lags or dynamic effects. The most widely used estimation method for such models is the **Ordinary Least Squares (OLS)** method, which provides efficient and unbiased parameter estimates under the classical assumptions.

While OLS offers a straightforward and powerful estimation approach, real-world data often deviate from ideal conditions. In such situations, the **Generalized Least Squares (GLS)** method becomes valuable, as it accounts for certain violations of OLS assumptions, such as heteroscedasticity or autocorrelation, thereby improving the efficiency of estimates.

This unit also addresses key issues that can compromise model accuracy, notably **multicollinearity** and **heteroscedasticity**. Multicollinearity arises when independent variables are highly correlated, making it difficult to isolate their individual effects. Heteroscedasticity occurs when the variance of the error term is not constant across observations, leading to inefficient estimates and misleading inference. Understanding their **causes, consequences, and possible solutions** is crucial for accurate estimation and reliable prediction.

By the end of this unit, you will be able to:

- Understand the theoretical basis and assumptions of OLS and GLS models.
- Perform estimation and prediction in single equation static models.
- Identify and address problems like multicollinearity and heteroscedasticity to ensure robust model results.

The Method of Ordinary Least Squares:

The method of ordinary least squares is attributed to Carl Friedrich Gauss, a German mathematician. Under certain assumptions (discussed in Section 3.2), the method of least squares has some very attractive statistical properties that have made it one of the most powerful and popular methods of regression analysis. To understand this method, we first explain the least-squares principle. Recall the two-variable PRF:

$$y_i = \beta_1 + \beta_2 X_i + u_i$$

However, as we noted in Chapter 2, the PRF is not directly observable. We estimate it from the SRF:

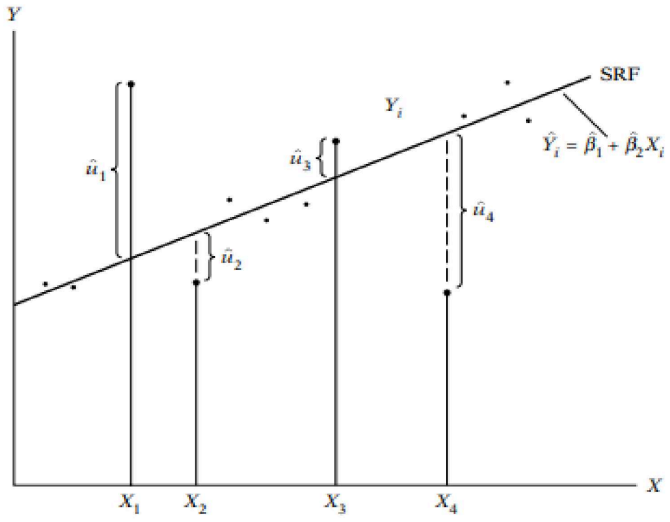
$$y_i = \beta_1 + \beta_2 X_i + u_i$$

$$= \widehat{Y}_i + \widehat{u}_i$$

Where, $\widehat{Y}_i$ is the estimated (conditional mean) value of y_i . But how is the SRF itself determined? To see this, let us proceed as follows.

$$\hat{u}_i = Y_i - \widehat{Y}_i$$

$$= Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i$$



Which shows that the $\hat{u}_i$ (the residuals) are simply the differences between the actual and estimated Y values. Now given n pairs of observations on Y and X, we would like to

determine the SRF in such a manner that it is as close as possible to the actual Y. To this end, we may adopt the following criterion: Choose the SRF in such a way that the sum of the residuals $\sum \hat{u}_i = \sum(Y_i - \hat{Y}_i)$is as small as possible. Although intuitively appealing, this is not a very good criterion, as can be seen in the hypothetical scattergram. If we adopt the criterion of minimizing$\sum \hat{u}_i$, shows that the residuals $\hat{u}_2$ and $\hat{u}_3$ as well as the residuals $\hat{u}_1$ and $\hat{u}_4$ receive the same weight in the sum($\hat{u}_1, +\hat{u}_2 + \hat{u}_3 + \hat{u}_4$), although the first two residuals are much closer to the SRF than the latter two. In other words, all the residuals receive equal importance no matter how close or how widely scattered the individual observations are from the SRF. A consequence of this is that it is quite possible that the algebraic sum of the $\widehat{u}_i$ is small (even zero) although the $\hat{u}_i$are widely scattered about the SRF. To see this, let ($\hat{u}_1, \hat{u}_2$,$\hat{u}_3, \hat{u}_4$), assume the values of 10, −2, +2, and −10, respectively. The algebraic sum of these residuals is zero although $\hat{u}_1$and $\hat{u}_4$are scattered more widely around the SRF than $\hat{u}_2$ and $\hat{u}_3$ We can avoid this problem if we adopt the least-squares criterion, which states that the SRF can be fixed in such a way that :

$$\sum \hat{u}_i = \sum(Y_i - \hat{Y}_i)^2$$

$$= \sum(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)^2$$

is as small as possible, where $\hat{u}_2$ are the squared residuals. By squaring $\hat{u}_i$, this method gives more weight to residuals such as $\hat{u}_1$ and $\hat{u}_4$ than the residuals $\hat{u}_2$ and $\hat{u}_3$

In other words, for a given sample, the method of least squares provides us with unique estimates of β1 and β2 that give the smallest possible value of $\sum \hat{u}_i^{\ 2}$ .How is this accomplished? This is a straightforward exercise in differential calculus. As shown in Appendix 3A, Section 3A.1, the process of differentiation yields the following equations for estimating β1 and β2:

$$\sum Y_i = n\hat{\beta}_1 + \hat{\beta}_2 \sum X_i$$

$$\sum Y_i X_i = \hat{\beta}_1 \sum X_i + \hat{\beta}_2 \sum X_i^2$$

where n is the sample size. These simultaneous equations are known as the normal equations. Solving the normal equations simultaneously, we obtain

$$\hat{\beta}_2 = \frac{n\sum X_i Y_i - \sum X_i \sum Y_i}{n\sum X_i^2 - \left(\sum X_i\right)^2}$$
$$= \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$$
$$= \frac{\sum x_i y_i}{\sum x_i^2}$$

where $\bar{X}$ and $\bar{Y}$ are the sample means of X and Y and where we define xi = (Xi $-\bar{X}$) and yi = (Yi $-\bar{Y}$). Henceforth, we adopt the convention of letting the lowercase letters denote deviations from mean values.

$$\hat{\beta}_1 = \frac{\sum X_i^2 \sum Y_i - \sum X_i \sum X_i Y_i}{n\sum X_i^2 - \left(\sum X_i\right)^2}$$
$$= \bar{Y} - \hat{\beta}_2 \bar{X}$$

Incidentally, note that, by making use of simple algebraic identities, formula for estimating β2 can be alternatively expressed as,

Numerical properties of estimators obtained by the method of OLS:

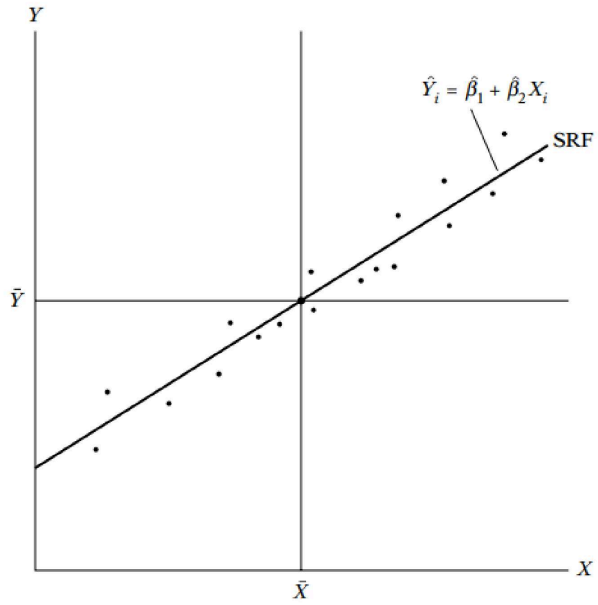
The estimators obtained previously are known as the least-squares estimators, for they are derived from the least-squares principle. Note the following numerical properties of estimators obtained by the method of OLS: "Numerical properties are those that hold as a consequence of the use of ordinary least squares, regardless of how the data were generated." Shortly, we will also consider the statistical properties of OLS estimators, that is, properties "that hold only under certain assumptions about the way the data were generated.

1. The OLS estimators are expressed solely in terms of the observable (i.e., sample) quantities (i.e., X and Y). Therefore, they can be easily computed.

2. They are point estimators; that is, given the sample, each estimator will provide only a single (point) value of the relevant population parameter.

3. Once the OLS estimates are obtained from the sample data, the sample regression line (Figure 3.1) can be easily obtained. The regression line thus obtained has the following properties:

1. It passes through the sample means of Y and X. This fact is obvious, for the latter can be written as $\bar{Y} = \hat{\beta}_1 + \hat{\beta}_2 \bar{X}$



2. The mean value of the estimated $Y = \hat{Y}_i$ is equal to the mean value of the actual Y for

$$\begin{aligned} \hat{Y}_i &= \hat{\beta}_1 + \hat{\beta}_2 X_i \\ &= (\bar{Y} - \hat{\beta}_2 \bar{X}) + \hat{\beta}_2 X_i \\ &= \bar{Y} + \hat{\beta}_2 (X_i - \bar{X}) \end{aligned} \tag{3.1.9}$$

Summing both sides of this last equality over the sample values and dividing through by the sample size n gives

$$\bar{\hat{Y}} = \bar{Y} \tag{3.1.10}$$[5]

where use is made of the fact that $\sum(X_i - \bar{X}) = 0$. (Why?)

3. The mean value of the residuals $\hat{u}_i$ is zero. From Appendix 3A, Section 3A.1, the first equation is

$$-2\sum(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i) = 0$$

But since $\hat{u}_i = Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i$, the preceding equation reduces to $-2\sum \hat{u}_i = 0$, whence $\bar{\hat{u}} = 0$.[6]

As a result of the preceding property, the sample regression

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{u}_i \tag{2.6.2}$$

can be expressed in an alternative form where both Y and X are expressed as deviations from their mean values. To see this, sum (2.6.2) on both sides to give

$$\begin{aligned} \sum Y_i &= n\hat{\beta}_1 + \hat{\beta}_2 \sum X_i + \sum \hat{u}_i \\ &= n\hat{\beta}_1 + \hat{\beta}_2 \sum X_i \qquad \text{since } \sum \hat{u}_i = 0 \end{aligned} \tag{3.1.11}$$

Dividing Equation 3.1.11 through by n, we obtain

$$\bar{Y} = \hat{\beta}_1 + \hat{\beta}_2 \bar{X} \tag{3.1.12}$$

which is the same as Eq. (3.1.7). Subtracting Equation 3.1.12 from Eq. (2.6.2), we obtain

$$Y_i - \bar{Y} = \hat{\beta}_2 (X_i - \bar{X}) + \hat{u}_i$$

or

$$y_i = \hat{\beta}_2 x_i + \hat{u}_i \tag{3.1.13}$$

where y_i and x_i, following our convention, are deviations from their respective (sample) mean values.

[5]Note that this result is true only when the regression model has the intercept term β_1 in it. As **Appendix 6A, Sec. 6A.1** shows, this result need not hold when β_1 is absent from the model.

[6]This result also requires that the intercept term β_1 be present in the model (see **Appendix 6A, Sec. 6A.1).**

Equation 3.1.13 is known as the **deviation form.** Notice that the intercept term $\hat{\beta}_1$ is no longer present in it. But the intercept term can always be estimated by Eq. (3.1.7), that is, from the fact that the sample regression line passes through the sample means of Y and X. An advantage of the deviation form is that it often simplifies computing formulas.

In passing, note that in the deviation form, the SRF can be written as

$$\hat{y}_i = \hat{\beta}_2 x_i \tag{3.1.14}$$

whereas in the original units of measurement it was $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$, as shown in Eq. (2.6.1).

4. The residuals $\hat{u}_i$ are uncorrelated with the predicted Y_i. This statement can be verified as follows: using the deviation form, we can write

$$\begin{aligned} \sum \hat{y}_i \hat{u}_i &= \hat{\beta}_2 \sum x_i \hat{u}_i \\ &= \hat{\beta}_2 \sum x_i (y_i - \hat{\beta}_2 x_i) \\ &= \hat{\beta}_2 \sum x_i y_i - \hat{\beta}_2^2 \sum x_i^2 \\ &= \hat{\beta}_2^2 \sum x_i^2 - \hat{\beta}_2^2 \sum x_i^2 \\ &= 0 \end{aligned} \tag{3.1.15}$$

where use is made of the fact that $\hat{\beta}_2 = \sum x_i y_i / \sum x_i^2$.

5. The residuals $\hat{u}_i$ are uncorrelated with X_i; that is, $\sum \hat{u}_i X_i = 0$. This fact follows from Eq. (2) in Appendix 3A, Section 3A.1.

3.2 The Classical Linear Regression Model: The Assumptions Underlying the Method of Least Squares

If our objective is to estimate β_1 and β_2 only, the method of OLS discussed in the preceding section will suffice. But recall from Chapter 2 that in regression analysis our objective is not only to obtain $\hat{\beta}_1$ and $\hat{\beta}_2$ but also to draw inferences about the true β_1 and β_2. For example, we would like to know how close $\hat{\beta}_1$ and $\hat{\beta}_2$ are to their counterparts in the population or how close $\hat{Y}_i$ is to the true $E(Y \mid X_i)$. To that end, we must not only specify the functional form of the model, as in Eq. (2.4.2), but also make certain assumptions about the manner in which Y_i are generated. To see why this requirement is needed, look at the PRF: $Y_i = \beta_1 + \beta_2 X_i + u_i$. It shows that Y_i depends on both X_i and u_i. Therefore, unless we are specific about how X_i and u_i are created or generated, there is no way we can make any statistical inference about the Y_i and also, as we shall see, about β_1 and β_2. Thus, the assumptions made about the X_i variable(s) and the error term are extremely critical to the valid interpretation of the regression estimates.

The **Gaussian, standard,** or **classical linear regression model (CLRM),** which is the cornerstone of most econometric theory, makes 7 assumptions.[7] We first discuss these assumptions in the context of the two-variable regression model; and in Chapter 7 we extend them to multiple regression models, that is, models in which there is more than one regressor.

[7]It is classical in the sense that it was developed first by Gauss in 1821 and since then has served as a norm or a standard against which may be compared the regression models that do not satisfy the Gaussian assumptions.

bias. Knowing the consequences of autocorrelation, we may therefore want to take some corrective action. We will do so shortly.

Incidentally, for all the wages–productivity regressions that we have presented above, we applied the **Jarque–Bera test of normality** and found that the residuals were normally distributed, which is comforting because the d test assumes normality of the error term.

12.9 Correcting for (Pure) Autocorrelation: The Method of Generalized Least Squares (GLS)

Knowing the consequences of autocorrelation, especially the lack of efficiency of OLS estimators, we may need to remedy the problem. The remedy depends on the knowledge one has about the nature of interdependence among the disturbances, that is, knowledge about the structure of autocorrelation.

As a starter, consider the two-variable regression model:

$$Y_t = \beta_1 + \beta_2 X_t + u_t \tag{12.9.1}$$

and assume that the error term follows the AR(1) scheme, namely,

$$u_t = \rho u_{t-1} + \varepsilon_t \quad -1 < \rho < 1 \tag{12.9.2}$$

Now we consider two cases: (1) ρ is known and (2) ρ is not known but has to be estimated.

When ρ Is Known

If the coefficient of first-order autocorrelation is known, the problem of autocorrelation can be easily solved. If Eq. (12.9.1) holds true at time t, it also holds true at time $(t-1)$. Hence,

$$Y_{t-1} = \beta_1 + \beta_2 X_{t-1} + u_{t-1} \tag{12.9.3}$$

Multiplying Eq. (12.9.3) by ρ on both sides, we obtain

$$\rho Y_{t-1} = \rho\beta_1 + \rho\beta_2 X_{t-1} + \rho u_{t-1} \tag{12.9.4}$$

Subtracting Eq. (12.9.4) from Eq. (12.9.1) gives

$$(Y_t - \rho Y_{t-1}) = \beta_1(1-\rho) + \beta_2(X_t - \rho X_{t-1}) + \varepsilon_t \tag{12.9.5}$$

where $\varepsilon_t = (u_t - \rho u_{t-1})$

We can express Eq. (12.9.5) as

$$Y_t^* = \beta_1^* + \beta_2^* X_t^* + \varepsilon_t \tag{12.9.6}$$

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

Since the error term in Eq. (12.9.6) satisfies the usual OLS assumptions, we can apply OLS to the transformed variables Y^* and X^* and obtain estimators with all the optimum properties, namely, BLUE. In effect, running Eq. (12.9.6) is tantamount to using generalized least squares (GLS) discussed in the previous chapter—recall that GLS is nothing but OLS applied to the transformed model that satisfies the classical assumptions.

Regression (12.9.5) is known as the **generalized,** or **quasi, difference equation.** It involves regressing Y on X, not in the original form, but in the **difference form,** which is obtained by subtracting a proportion ($=\rho$) of the value of a variable in the previous time period from its

value in the current time period. In this differencing procedure we lose one observation because the first observation has no antecedent. To avoid this loss of one observation, the first observation on Y and X is transformed as follows:[35] $Y_1\sqrt{1-\rho^2}$ and $X_1\sqrt{1-\rho^2}$. This transformation is known as the **Prais–Winsten transformation.**

When ρ Is Not Known

Although conceptually straightforward to apply, the method of generalized difference given in Eq. (12.9.5) is difficult to implement because ρ is rarely known in practice. Therefore, we need to find ways of estimating ρ. We have several possibilities.

The First-Difference Method

Since ρ lies between 0 and ± 1, one could start from two extreme positions. At one extreme, one could assume that $\rho = 0$, that is, no (first-order) serial correlation, and at the other extreme we could let $\rho = \pm 1$, that is, perfect positive or negative correlation. As a matter of fact, when a regression is run, one generally assumes that there is no autocorrelation and then lets the Durbin–Watson or other test show whether this assumption is justified. If, however, $\rho = +1$, the generalized difference equation (12.9.5) reduces to the **first-difference equation:**

$$Y_t - Y_{t-1} = \beta_2(X_t - X_{t-1}) + (u_t - u_{t-1})$$

or

$$\Delta Y_t = \beta_2 \Delta X_t + \varepsilon_t \qquad \textbf{(12.9.7)}$$

where Δ is the first-difference operator introduced in Eq. (12.1.10).

Since the error term in Eq. (12.9.7) is free from (first-order) serial correlation (why?), to run the regression (12.9.7) all one has to do is form the first differences of both the regressand and regressor(s) and run the regression on these first differences.

The first-difference transformation may be appropriate if the coefficient of autocorrelation is very high, say in excess of 0.8, or the Durbin–Watson d is quite low. Maddala has proposed this rough rule of thumb: *Use the first-difference form whenever* $d < R^2$.[36] This is the case in our wages–productivity regression (12.5.2), where we found that $d = 0.2176$ and $r^2 = 0.9845$. The first-difference regression for our illustrative example will be presented shortly.

An interesting feature of the first-difference model (12.9.7) is that **there is no intercept in it.** Hence, to estimate Eq. (12.9.7), you have to use the **regression through the origin** routine (that is, suppress the intercept term), which is now available in most software packages. If, however, you forget to drop the intercept term in the model and estimate the following model that includes the intercept term

$$\Delta Y_t = \beta_1 + \beta_2 \Delta X_t + \varepsilon_t \qquad \textbf{(12.9.8)}$$

[35]The loss of one observation may not be very serious in large samples but can make a substantial difference in the results in small samples. Without transforming the first observation as indicated, the error variance will not be homoscedastic. On this, see Jeffrey Wooldridge, op. cit., p. 388. For some Monte Carlo results on the importance of the first observation, see Russell Davidson and James G. MacKinnon, *Estimation and Inference in Econometrics,* Oxford University Press, New York, 1993, Table 10.1, p. 349.

[36]Maddala, op. cit., p. 232.

then the original model must have a *trend* in it and β_1 represents the coefficient of the trend variable.[37] Therefore, one "accidental" benefit of introducing the intercept term in the first-difference model is to test for the presence of a trend variable in the original model.

Returning to our wages–productivity regression (12.5.2), and given the AR(1) scheme and a low d value in relation to r^2, we rerun Eq. (12.5.2) in the first-difference form without the intercept term; remember that Eq. (12.5.2) is in the *level form*. The results are as follows:[38]

$$\widehat{\Delta Y}_t = 0.6539\Delta X_t$$
$$t = (11.4042) \qquad r^2 = 0.4264 \qquad d = 1.7442 \tag{12.9.9}$$

Compared with the level form regression (12.5.2), we see that the slope coefficient has not changed much, but the r^2 value has dropped considerably. This is generally the case because by taking the first differences we are essentially studying the behavior of variables around their (linear) trend values. Of course, we cannot compare the r^2 of Eq. (12.9.9) directly with that of the r^2 of Eq. (12.5.2) because the dependent variables in the two models are different.[39] Also, notice that compared with the original regression, the d value has increased dramatically, perhaps indicating that there is little autocorrelation in the first-difference regression.[40]

Another interesting aspect of the first-difference transformation relates to the *stationarity* properties of the underlying time series. Return to Eq. (12.2.1), which describes the AR(1) scheme. Now if in fact $\rho = 1$, then it is clear from Eqs. (12.2.3) and (12.2.4) that the series u_t is *nonstationary,* for the variances and covariances become infinite. That is why, when we discussed this topic, we put the restriction that $|\rho| < 1$. But it is clear from Eq. (12.2.1) that if the autocorrelation coefficient is in fact 1, then Eq. (12.2.1) becomes

$$u_t = u_{t-1} + \varepsilon_t$$

or

$$(u_t - u_{t-1}) = \Delta u_t = \varepsilon_t \tag{12.9.10}$$

That is, it is the first-differenced u_t that becomes stationary, for it is equal to ε_t, which is a white noise error term.

The point of the preceding discussion is that if the original time series are nonstationary, very often their first differences become stationary. And, therefore, first-difference transformation serves a dual purpose in that it might get rid of (first-order) autocorrelation and also render the time series stationary. We will revisit this topic in **Part 5,** where we discuss the econometrics of time series analysis in some depth.

We mentioned that the first-difference transformation may be appropriate if ρ is high or d is low. Strictly speaking, the first-difference transformation is valid only if $\rho = 1$. As a

[37]This is easy to show. Let $Y_t = \alpha_1 + \beta_1 t + \beta_2 X_t + u_t$. Therefore, $Y_{t-1} = \alpha + \beta_1(t-1) + \beta_2 X_{t-1} + u_{t-1}$. Subtracting the latter from the former, you will obtain: $\Delta Y_t = \beta_1 + \beta_2 \Delta X_t + \varepsilon_t$, which shows that the intercept term in this equation is indeed the coefficient of the trend variable in the original model. Remember that we are assuming that $\rho = 1$.

[38]In Exercise 12.38 you are asked to run this model, including the constant term.

[39]The comparison of r^2 in the level and first-difference form is slightly involved. For an extended discussion on this, see Maddala, op. cit., Chapter 6.

[40]It is not clear whether the computed d in the first-difference regression can be interpreted in the same way as it was in the original, level form regression. However, applying the runs test, it can be seen that there is no evidence of autocorrelation in the residuals of the first-difference regression.

matter of fact, there is a test, called the **Berenblutt–Webb test,**[41] to test the hypothesis that $\rho = 1$. The test statistic they use is called the ***g* statistic,** which is defined as follows:

$$g = \frac{\sum_2^n \hat{e}_t^2}{\sum_1^n \hat{u}_t^2} \tag{12.9.11}$$

where $\hat{u}_t$ are the OLS residuals from the original (i.e., level form) regression and e_t are the OLS residuals from the first-difference regression. Keep in mind that in the first-difference form there is no intercept.

To test the significance of the g statistic, assuming that the level form regression contains the intercept term, we can use the Durbin–Watson tables except that now the null hypothesis is that $\rho = 1$ rather than the Durbin–Watson hypothesis that $\rho = 0$.

Revisiting our wages–productivity regression, for the original regression (12.5.2) we obtain $\sum \hat{u}_t^2 = 0.0214$ and $\sum \hat{e}_t^2 = 0.0046$. Putting these values into the g statistic given in Eq. (12.9.11), we obtain

$$g = \frac{0.0046}{0.0214} = 0.2149 \tag{12.9.12}$$

Consulting the Durbin–Watson table for 45 observations (the number closest to 45 observations) and 1 explanatory variable (Appendix D, Table D.5), we find that $d_L = 1.288$ and $d_U = 1.376$ (5 percent level). Since the observed g lies below the lower limit of d, we do not reject the hypothesis that true $\rho = 1$. *Keep in mind that although we use the same Durbin–Watson tables, now the null hypothesis is that* $\rho = 1$ *and not that* $\rho = 0$. In view of this finding, the results given in Eq. (12.9.9) may be acceptable.

ρ Based on Durbin–Watson d Statistic

If we cannot use the first-difference transformation because ρ is not sufficiently close to unity, we have an easy method of estimating it from the relationship between d and ρ established previously in Eq. (12.6.10), from which we can estimate ρ as follows:

$$\hat{\rho} \approx 1 - \frac{d}{2} \tag{12.9.13}$$

Thus, in reasonably large samples one can obtain ρ from Eq. (12.9.13) and use it to transform the data as shown in the generalized difference equation (12.9.5). Keep in mind that the relationship between ρ and d given in Eq. (12.9.13) may not hold true in small samples, for which Theil and Nagar have proposed a modification, which is given in Exercise 12.6.

In our wages–productivity regression (12.5.2), we obtain a d value of 0.2176. Using this value in Eq. (12.9.13), we obtain $\hat{\rho} \approx 0.8912$. Using this estimated ρ value, we can estimate regression (12.9.5). All we have to do is subtract 0.8912 times the previous value of Y from its current value and similarly subtract 0.8912 times the previous value of X from its current value and run the OLS regression on the variables thus transformed as in Eq. (12.9.6), where $Y_t^* = (Y_t - 0.8912Y_{t-1})$ and $X_t^* = (X_t - 0.8912X_{t-1})$.

ρ Estimated from the Residuals

If the AR(1) scheme $u_t = \rho u_{t-1} + \varepsilon_t$ is valid, a simple way to estimate ρ is to regress the residuals $\hat{u}_t$ on $\hat{u}_{t-1}$, for the $\hat{u}_t$ are consistent estimators of the true u_t, as noted previously. That is, we run the following regression:

$$\hat{u}_t = \rho \,.\, \hat{u}_{t-1} + v_t \tag{12.9.14}$$

[41] I. I. Berenblutt and G. I. Webb, "A New Test for Autocorrelated Errors in the Linear Regression Model," *Journal of the Royal Statistical Society,* Series B, vol. 35, no.1, 1973, pp. 33–50.

where $\hat{u}_t$ are the residuals obtained from the original (level form) regression and where v_t are the error term of this regression. Note that there is no need to introduce the intercept term in Eq. (12.9.14), for we know the OLS residuals sum to zero.

The residuals from our wages–productivity regression given in Eq. (12.5.1) are already shown in Table 12.5. Using these residuals, the following regression results were obtained:

$$\begin{aligned} \widehat{\hat{u}_t} &= \quad 0.8678\hat{u}_{t-1} \\ t &= (12.7359) \qquad\quad r^2 = 0.7863 \end{aligned} \tag{12.9.15}$$

As this regression shows, $\hat{\rho} = 0.8678$. Using this estimate, one can transform the original model as per Eq. (12.9.6). Since the ρ estimated by this procedure is about the same as that obtained from the Durbin–Watson d, the regression results using the ρ of Eq. (12.9.15) should not be very different from those obtained from the ρ estimated from the Durbin–Watson d. We leave it to the reader to verify this.

Iterative Methods of Estimating ρ

All the methods of estimating ρ discussed previously provide us with only a single estimate of ρ. But there are the so-called **iterative methods** that estimate ρ iteratively, that is, by successive approximation, starting with some initial value of ρ. Among these methods the following may be mentioned: the **Cochrane–Orcutt iterative procedure,** the **Cochrane–Orcutt two-step procedure,** the **Durbin two–step procedure,** and the **Hildreth–Lu scanning or search procedure.** Of these, the most popular is the Cochran–Orcutt iterative method. To save space, the iterative methods are discussed by way of exercises. Remember that the ultimate objective of these methods is to provide an estimate of ρ that may be used to obtain GLS estimates of the parameters. One advantage of the Cochrane–Orcutt iterative method is that it can be used to estimate not only an AR(1) scheme, but also higher-order autoregressive schemes, such as $\hat{u}_t = \hat{\rho}_1\hat{u}_{t-1} + \hat{\rho}_2\hat{u}_{t-2} + v_t$, which is AR(2). Having obtained the two ρs, one can easily extend the generalized difference equation (12.9.6). Of course, the computer can now do all this.

Returning to our wages–productivity regression, and assuming an AR(1) scheme, we use the Cochrane–Orcutt iterative method, which gives the following estimates of ρ: 0.8876, 0.9944, and 0.8827. The last value of 0.8827 can now be used to transform the original model as in Eq. (12.9.6) and estimate it by OLS. Of course, OLS on the transformed model is simply the GLS. The results are as follows:

Stata can estimate the coefficients of the model along with ρ. For example, if we assume the AR(1), Stata produces the following results:

$$\begin{aligned} \hat{Y}_t^* &= \ 43.1042 \ + \ 0.5712X_t \\ \text{se} &= \ (4.3722) \quad (0.0415) \\ t &= \ (9.8586) \quad (13.7638) \qquad r^2 = 0.8146 \end{aligned} \tag{12.9.16}$$

From these results, we can see that the estimated rho ($\hat{\rho}$) is ≈0.8827, which is not very much different from the $\hat{\rho}$ in Eq. (12.9.15).

As noted before, in the generalized difference equation (12.9.6) we lose one observation because the first observation has no antecedent. To avoid losing the first observation, we can use the *Prais–Winsten transformation.* Using this transformation, and using STATA (version #10), we obtain the following results for our wages–productivity regression:

$$\begin{aligned} \text{Rcompb}_t &= \ 32.0434 \ + \ 0.6628 \ \text{Prodb}_t \\ \text{se} &= \ (3.7182) \quad (0.0386) \qquad r^2 = 0.8799 \end{aligned} \tag{12.9.17}$$

In this transformation, the ρ value was 0.9193, which was obtained after 13 iterations. *It should be pointed out that if we do not transform the first observation à la Prais–Winsten and drop that observation, the results sometimes are substantially different, especially in small samples.* Notice that the ρ obtained here is not much different from the one obtained in Eq. (12.9.15).

General Comments

There are several points about correcting for autocorrelation using the various methods discussed above.

First, since the OLS estimators are consistent despite autocorrelation, in large samples, it makes little difference whether we estimate ρ from the Durbin–Watson d, or from the regression of the residuals in the current period on the residuals in the previous period, or from the Cochrane–Orcutt iterative procedure because they all provide consistent estimates of the true ρ. *Second,* the various methods discussed above are basically two-step methods. In step 1 we obtain an estimate of the unknown ρ and in step 2 we use that estimate to transform the variables to estimate the generalized difference equation, which is basically GLS. But since we use $\hat{\rho}$ instead of the true ρ, all these methods of estimation are known in the literature as **feasible GLS (FGLS)** or **estimated GLS (EGLS)** methods.

Third, it is important to note that whenever we use an **FGLS** or **EGLS** method to estimate the parameters of the transformed model, the estimated coefficients will not necessarily have the usual optimum properties of the classical model, such as BLUE, especially in small samples. Without going into complex technicalities, it may be stated *as a general principle that whenever we use an estimator in place of its true value, the estimated OLS coefficients may have the usual optimum properties asymptotically, that is, in large samples. Also, the conventional hypothesis testing procedures are, strictly speaking, valid asymptotically. In small samples, therefore, one has to be careful in interpreting the estimated results.*

Fourth, in using EGLS, if we do not include the first observation (as was originally the case with the Cochrane–Orcutt procedure), not only the numerical values but also the efficiency of the estimators can be adversely affected, especially if the sample size is small and if the regressors are not strictly speaking nonstochastic.[42] Therefore, in small samples it is important to keep the first observation *à la* Prais–Winsten. Of course, if the sample size is reasonably large, EGLS, with or without the first observation, gives similar results. Incidentally, in the literature EGLS with Prais–Winsten transformation is known as the **full EGLS,** or **FEGLS,** for short.

12.10 The Newey–West Method of Correcting the OLS Standard Errors

Instead of using the FGLS methods discussed in the previous section, we can still use OLS but correct the standard errors for autocorrelation by a procedure developed by Newey and West.[43] This is an extension of White's heteroscedasticity-consistent standard errors that we discussed in the previous chapter. The corrected standard errors are known as **HAC (heteroscedasticity- and autocorrelation-consistent) standard errors** or simply **Newey–West standard errors.** We will not present the mathematics behind the

[42]This is especially so if the regressors exhibit a trend, which is quite common in economic data.

[43]W. K. Newey and K. West, "A Simple Positive Semi-Definite Heteroscedasticity and Autocorrelation Consistent Covariance Matrix, *Econometrica,* vol. 55, 1987, pp. 703–708.

MULTICOLLINEARITY

Definition:

Multicollinearity is the occurrence of high intercorrelations among two or more independent variables in a multiple regression model. Multicollinearity can lead to skewed or misleading results when a researcher or analyst attempts to determine how well each independent variable can be used most effectively to predict or understand the dependent variable in a statistical model.

In general, multicollinearity can lead to wider confidence intervals that produce less reliable probabilities in terms of the effect of independent variables in a model.

In technical analysis, multicollinearity can lead to incorrect assumptions about an investment. It generally occurs because multiple indicators of the same type have been used to analyze a stock.

Cause of Multicollinearity

There are various reasons for Multicollinearity, and a few of them are as follows:

- Poor experiment design, data manipulation error, or incorrect data observations.

  For example, we interchange the data for some rows of income and salary columns.

- Creating a new variable that is dependent on other available variables. For instance, if the income variable is with monthly salary and other incomes.
- Having identical variables in the dataset. For example, If the age variable is taken as year and in months too.
- Including both columns in the dataset while creating dummy variables from categorical features will cause Multicollinearity.

Impact of Multicollinearity:

While multicollinearity does not affect the overall fit of the regression model, it has several undesirable consequences:

- It becomes difficult to identify the relative importance of individual predictors in explaining the dependent variable.
- The estimated coefficients can be unstable and difficult to interpret.
- Standard errors of the coefficients are inflated, leading to wider confidence intervals and reduced statistical significance.
- Multicollinearity can make it difficult to make accurate predictions or draw reliable conclusions from the model.

Measuring Multicollinearity:

Several statistical measures can help assess the degree of multicollinearity in a regression model:

Variance Inflation Factor (VIF): VIF quantifies how much the variance of a coefficient is inflated due to multicollinearity.

- A VIF value of 1 indicates no multicollinearity.
- Values above 1 suggest increasing multicollinearity.
- Generally, VIF values greater than 5 or 10 are considered problematic.

Condition Number: Measures the sensitivity of the regression coefficients to small changes in the data. High condition numbers indicate multicollinearity.

Eigenvalues and Eigenvectors: The eigenvalues and eigenvectors of the correlation matrix can provide insights into the presence and extent of multicollinearity.

Estimation in presence of perfect and imperfect multicollinearity problems with measuring multicollinearity

<u>Perfect Multicollinearity:</u>

Perfect multicollinearity occurs when there is an exact linear relationship among two or more independent variables in a regression model. In this case, one variable can be expressed as a

linear combination of the others, making it impossible to estimate the regression coefficients using Ordinary Least Squares (OLS) estimation. Perfect multicollinearity leads to an undefined or singular matrix of the independent variables, and the model cannot be fit.

Handling Perfect Multicollinearity:

To handle perfect multicollinearity, one of the correlated variables must be removed from the model. It is crucial to identify and remove the variable causing the perfect multicollinearity. Alternatively, if the variables are theoretically necessary for the model, one can use techniques such as data aggregation or differencing to create new variables that do not have perfect multicollinearity.

Imperfect Multicollinearity:

Imperfect multicollinearity, also known as high multicollinearity, occurs when there is a high correlation among independent variables but not an exact linear relationship. Imperfect multicollinearity can lead to unstable and unreliable coefficient estimates, as well as inflated standard errors, making it difficult to determine the true significance of predictors.

Handling Imperfect Multicollinearity:

Variable Selection: Consider removing or combining variables that are highly correlated, especially if they represent similar information or measure the same underlying concept.

Data Transformation: Transforming variables (e.g., taking differences or percentage changes) can sometimes reduce multicollinearity.

Regularization Techniques: Ridge regression and Lasso regression are regularization methods that can help reduce multicollinearity by adding a penalty term to the loss function, shrinking the coefficient estimates.

Principal Component Analysis (PCA): PCA is a dimensionality reduction technique that can transform correlated variables into a new set of uncorrelated variables (principal components), which can be used as predictors.

Measuring Multicollinearity:

Several statistical measures can help assess the degree of multicollinearity in a regression model:

Variance Inflation Factor (VIF): VIF quantifies how much the variance of a coefficient is inflated due to multicollinearity. A VIF value of 1 indicates no multicollinearity, while values above 1 suggest increasing multicollinearity. Generally, VIF values greater than 5 or 10 are considered problematic.

Condition Number: The condition number measures the sensitivity of the regression coefficients to small changes in the data. High condition numbers indicate multicollinearity.

Eigenvalues and Eigenvectors: The eigenvalues and eigenvectors of the correlation matrix can provide insights into the presence and extent of multicollinearity.

By examining these measures, researchers can identify the presence and severity of multicollinearity in the regression model. If multicollinearity is a concern, appropriate remedial actions can be taken to improve the reliability of the regression results.

Solution to Multicollinearity problem

Addressing multicollinearity is essential to obtain reliable and interpretable regression results. Here are some strategies to mitigate the multicollinearity problem:

Variable Selection: Carefully select the relevant variables for your regression model. If two or more variables are highly correlated and represent similar information, consider removing one of them from the model.

Data Transformation: Transform the variables to reduce multicollinearity. For example, taking differences or percentage changes between variables can sometimes help reduce correlation.

Combine Variables: If appropriate, consider creating composite variables or indices that combine correlated variables into a single predictor. This can help reduce multicollinearity.

Use Interaction Terms Sparingly: Interaction terms, created by multiplying two or more variables, can introduce multicollinearity if the original variables are highly correlated. Use interaction terms only when necessary and with caution.

Centering Variables: Centering variables by subtracting their mean from each observation can sometimes reduce multicollinearity.

Regularization Techniques: Consider using regularization methods like Ridge regression and Lasso regression. These techniques add penalty terms to the regression objective function, which helps to shrink the coefficient estimates and mitigate multicollinearity.

Principal Component Analysis (PCA): PCA is a dimensionality reduction technique that can transform correlated variables into a new set of uncorrelated variables (principal components), which can be used as predictors.

Stepwise Regression: Stepwise regression is an iterative process that adds or removes variables based on their significance, effectively selecting the most relevant predictors while minimizing multicollinearity.

Collect More Data: Increasing the sample size can help reduce the impact of multicollinearity, as the correlation between variables may become less pronounced in larger datasets.

Consider Causality: When possible, prioritize causal relationships when selecting variables for the model. Including variables that are not causally related to the dependent variable can introduce spurious correlations and exacerbate multicollinearity.

Check Model Specification: Reassess the model specification and theoretical assumptions. Ensure that the model makes sense in the context of the data and the research question.

Before applying any of these strategies, it is essential to thoroughly analyze the nature and severity of multicollinearity in the regression model. Techniques such as examining the variance inflation factor (VIF) and the condition number can help identify problematic multicollinearity. By taking appropriate actions to address multicollinearity, researchers can obtain more reliable and interpretable regression results and draw meaningful conclusions from their analysis.

HETEROSCEDASTICITY

Definition:

In statistics, heteroskedasticity (or heteroscedasticity) happens when the standard deviations of a predicted variable, monitored over different values of an independent variable or as related to prior time periods, are non-constant. With heteroskedasticity, the tell-tale sign upon visual inspection of the residual errors is that they will tend to fan out over time, as depicted in the image below.

Heteroskedasticity often arises in two forms: conditional and unconditional. Conditional heteroskedasticity identifies nonconstant volatility related to prior period's (e.g., daily) volatility. Unconditional heteroskedasticity refers to general structural changes in volatility that are not related to prior period volatility. Unconditional heteroskedasticity is used when future periods of high and low volatility can be identified.

Causes Of Heteroscedasticity:

Some of the most common causes of heteroscedasticity are:

Outliers: outliers are specific values within a sample that are extremely different (very large or small) from other values. Outliers can also alter the results of regression models and cause heteroscedasticity. That is, outlying observations can often lead to a non-constant variance of residuals.

Mis-specification of the model: incorrect specification can further lead to heteroscedastic residuals. For example, if an important variable is excluded from the model, its effects get captured in the error terms. In such a case, the residuals might exhibit non-constant variance because they end up accounting for the omitted variable.

Wrong Functional form: Misspecification of the model's functional form can cause heteroscedasticity. Suppose, the actual relationship between the variables is non-linear in nature. If we estimate a linear model for such variables, we might observe its effects in the residuals in the form of heteroscedasticity.

Error-learning: let us consider an example for this case. Errors in human behaviour become smaller over time with more practice or learning of an activity. In such a case, therefore, errors will tend to decrease. For example, with the skill development of labour, their error

will decrease leading to lower defective products in the manufacturing process or a rise in their productivity. Hence, error variance will decrease in such a setup.

Nature of variables: for instance, an increase in income is accompanied by an increase in choices to spend that extra income. This further leads to discretionary expenditure. In such a case, error variance will increase with an increase in income. A model with consumption as a dependent variable and income as an independent variable can have an increasing error variance. Hence, the nature of variables and their relationships can play a huge role in this phenomenon.

Types of Heteroscedasticity

There are two types of Heteroscedasticity that we may generally encounter:

1. Pure Heteroscedasticity – When we use proper independent features, the residual plot shows the non-non-constant variance. In other words, we observe unequal residual variance despite specifying the correct model.

2. Impure Heteroscedasticity – This scenario is typical where we specify an incorrect model, and the number of input features also gets the unequal residual variance.

Consequences of heteroscedasticity:

The major consequences of heteroscedasticity can be summarized as follows:

Validity of statistical inference: standard errors, confidence intervals, p-values and other tests of significance are no longer reliable in the presence of heteroscedasticity. This is because OLS standard errors assume constant variance of residuals. The tests of significance are based on these standard errors. In heteroscedasticity, error variance is non-constant, therefore, OLS standard errors are not applicable. As a result, it is not advisable to rely on confidence intervals and p-values.

The variance of estimates: OLS estimates no longer have minimum variance property because the variance of residuals is not constant. The coefficients end up having larger standard errors and lower precision in the presence of heteroscedasticity. Hence, OLS estimators become inefficient in the presence of heteroscedasticity.

Predictions: the forecasted or predicted values of the dependent variable based on a heteroscedastic model will have high variance. This is because the OLS estimates are no longer efficient. The variance of residuals is not minimum in presence of heteroscedasticity due to which the variance of predictions is also high.

Biasedness: the unbiasedness property of OLS estimates does not require a constant variance of residuals. However, the predictions from a heteroscedastic model can still end up being biased, especially in the case of observations with large residuals.

Solutions of the problem of Heteroscedasticity:

Several approaches can be adopted to counter heteroscedasticity. Some of these methods are as follows:

Robust Standard Errors: The usual OLS standard errors assume constant variance of residuals and cannot be used in heteroscedasticity. Instead, robust standard errors can be employed in such cases. These allow a non-constant variance of residuals to estimate the standard errors of coefficients. Moreover, the HC3 Robust standard errors have been observed to perform well in heteroscedasticity.

Weighted Least Squares: The Weighted Least Squares technique is a special application of Generalized Least Squares. The variables in the model are transformed by assigning weights in such a manner that the variance of residuals becomes constant. The weights are determined by understanding the underlying relationship or form of heteroscedasticity. Glejser's heteroscedasticity test can also be used to determine the heteroscedastic relationship between residuals and independent variables.

Transforming the variables: Transforming the variables can help reduce or even completely eliminate heteroscedasticity. For example, using log transformation reduced the scale of all the variables. As a result, the scale of residual variance also decreases which helps mitigate the problem of heteroscedasticity. In some cases, the problem of heteroscedasticity may be totally eliminated. In addition, log transformation provides its own benefits in economics by letting us study the elasticities of variables.

Non-parametric methods: Non-parametric regression techniques make no assumptions about the relationship between dependent and independent variables. In OLS, the assumption of constant residual variance is violated by heteroscedasticity. However, there are no

assumptions to violate in the case of non-parametric methods. Estimation techniques such as Kernel regression, Splines and Random Forest fall under the category of non-parametric methods. These provide a lot more flexibility in estimating complex relationships.