PREFACE

In the curricular structure introduced by this University for students of Post-Graduate degree programme, the opportunity to pursue Post-Graduate course in a subject is introduced by this University is equally available to all learners. Instead of being guided by any presumption about ability level, it would perhaps stand to reason if receptivity of a learner is judged in the course of the learning process. That would be entirely in keeping with the objectives of open education which does not believe in artificial differentiation. I am happy to note that university has been recently accredited by National Assessment and Accreditation Council of India (NAAC) with grade 'A'.

Keeping this in view, the study materials of the Post-Graduate level in different subjects are being prepared on the basis of a well laid-out syllabus. The course structure combines the best elements in the approved syllabi of Central and State Universities in respective subjects. It has been so designed as to be upgradable with the addition of new information as well as results of fresh thinking and analysis.

The accepted methodology of distance education has been followed in the preparation of these study materials. Co-operation in every form of experienced scholars is indispensable for a work of this kind. We, therefore, owe an enormous debt of gratitude to everyone whose tireless efforts went into the writing, editing, and devising of a proper layout of the materials. Practically speaking, their role amounts to an involvement in 'invisible teaching'. For, whoever makes use of these study materials would virtually derive the benefit of learning under their collective care without each being seen by the other.

The more a learner would seriously pursue these study materials, the easier it will be for him or her to reach out to larger horizons of a subject. Care has also been taken to make the language lucid and presentation attractive so that they may be rated as quality self-learning materials. If anything remains still obscure or difficult to follow, arrangements are there to come to terms with them through the counselling sessions regularly available at the network of study centres set up by the University.

Needless to add, a great deal of these efforts is still experimental—in fact, pioneering in certain areas. Naturally, there is every possibility of some lapse or deficiency here and there. However, these do admit of rectification and further improvement in due course. On the whole, therefore, these study materials are expected to evoke wider appreciation the more they receive serious attention of all concerned.

Professor (Dr.) Subha Sankar Sarkar
Vice-Chancellor

Netaji Subhas Open University
Post Graduate Degree Programme
MA in Economics
Course : Basic Econometrics
Code : PGEC-IX

First Print : December, 2021

Printed in accordance with the regulations of the Distance
Education Bureau of the University Grants Commission

Netaji Subhas Open University

Post Graduate Degree Programme

MA in Economics

Course : Basic Econometrics

Code : PGEC-IX

: Course Writer :

Sanchita Daripa
Research Scholar
Burdwan University

: Course Editor :

Dr. Seikh Salim
Associate Professor of Economics
Netaji Subhas Open University

: Format Editor :

Mrs. Priyanthi Bagchi
Assistant Professor of Economics
Netaji Subhas Open University

Notification

Kishore Sengupta
Registrar

Netaji Subhas
Open University

PG : Economics
(PGEC)

PGEC-IX : Basic Econometrics

Unit - 1 ❑ Definition, Scope and Goals of Econometrics

Structure

1.1 Objectives

Reading this chapter, students will get an idea about

- What Econometrics is
- Relationship between Econometrics and Economic Theory
- Relationship between Econometrics and Statistics

- Branches of Econometrics
- Goals of Econometrics
- Methodology of Econometric Research

1.2 Introduction

The term econometrics literally means economic measurement. It can be said as an integration of economic theory, mathematical economics, economic statistics and mathematical statistics. However, the subject has an importance to be studied as a separate discipline. The purpose of econometrics is to provide numerical values for the parameters of economic relationship and to verify economic theories. In econometrics, the general economic theory is formulated in mathematical terms and is combined with empirical measurement of economic phenomenon. It is a special type of economic analysis where the relationships of economic variable as suggested in economic theory are expressed in mathematical terms. This is called econometric model building. Next the statistical methods are used to obtain numerical estimates of coefficients of economic relationships. These methods are called econometric methods.

Economic theory provides various qualitative statements or hypotheses, but does not provide any empirical support regarding the theories. For example, the theory of demand suggests that all other things remaining unchanged there exists an inverse relationship between price and quantity demanded of a commodity i.e., economic theory states the existence of an inverse relationship between price and quantity demanded of a commodity, but it does not express any numerical estimate about how quantity demanded will be affected due to how much change in price. In other words, it does not provide any empirical content to ecomomic theory which is the job of an econometrician.

Mathematical economics expresses economic theories in the form of mathematical equations, but does not take into account empirical verification of the theory. Econometricians take into consideration the equations proposed by the mathematical economists and convert the mathematical equations into econometric equations, thereby contributing to empirical verification of economic theories.

Economic statistics deals with collection, processing and presenting data, but

they do not use these data for testing economic theories. It is the job of an econometrician to use these raw data for empirically verifying economic theories.

Mathematical statistics however provides various tools and techniques that are used by econometrician to analyse economic data. Economic theory tries to postulate an exact relationship among economic variables. But economic relationships always contain random elements. Economic theory ignores this, but econometrics deals with those random components. For example Keynesian consumption function states an exact relationship between consumption and income i.e., $C = C(Y)$. In linear form it is given as $C = a + bY$ where a is autonomous consumption ($a > 0$) and b is marginal propensity to consume (MPC) i.e., $0 < b < 1$. In this model the effect of other variables like wealth etc are ignored. But in econometrics, the effect of these variables are considered by introducing a random component u. It is called error term. So the consumption function considered in econometrics is $C = a + bY + u$. Then econometric methods are applied to estimate the parameters a and b. The choice of econometric method depends on the behaviour of the distribution of the random variable u. This error term may arise due to unpredictable element of randomness in human response, effect of a large number of variables that have been omitted from the functional relation and measurement error.

1.3 Relationship between Econometrics and Economic Theory

Economic theory suggests various qualitative statements or hypotheses but does not verify those empirically. This empirical verification of economic theory is done by econometric methods. For example if we consider the economic theory of law of demand we get an inverse relationship between price of a commodity and quantity demanded of that commodity. Economic theory suggests that this can be represented by the form $q = f(p)$ such that $\frac{dq}{dp} < 0$. This proposition of economic theory can be tested by applying econometric methods and if the results of empirical verification of the theory are found to be consistent with the theory, it is accepted; otherwise it is rejected or modified. If the theory is to be modified, then it should not be rejected, rather it can be modified by including other variables like price of substitute and complementary goods, income of the consumer, taste and preference of the consumer, etc. which can be expressed in the form

$$q = f(p, p_c, p_s, y, t)$$

where, p = own price of the commodity

p_c = price of complementary commodities

p_s = price of substitute commodities

y = income of the consumer

t = taste and preference of the consumer

The signs of the parameters and their relative importance in this new model can also be tested empirically.

Thus, econometric theory provides some hypotheses about economic behaviour. Econometrics tries to test that behaviour by applying some econometric methods.

1.4 Econometrics and Mathematical Economics

Mathematical economics states economic theory in terms of mathematical symbols. There is no essential difference between mathematical economics and economic theory. Both state the same relationship. While economic theory uses verbal approach, mathematical economics employs mathematical symbols. For example, the economic theory of law of demand states that if price of a commodity falls, its quantity demanded will rise and if price rises, its quantity demanded will fall, other things remaining the same. But mathematical economics will state that quantity demanded (q) is a function of price i.e., $q = f(p)$ such that $\frac{dq}{dp} < 0$, ceteris paribus. Both the approaches express economic relationship in an exact or deterministic form. They do not allow for random elements which might affect the relationship and make it stochastic. Further, economic theory and mathematical economics do not provide numerical values for the coefficients of the economic relationships. Relationships in economic theory or in mathematical economics are of non-stochastic form. But econometrics considers the stochastic relationship in mathematical forms unlike mathematical economics. In econometrics, it is assumed that relationships are not exact, rather presence of disturbance term makes deviations from the exact behavioural pattern suggested in economic theory or mathematical economics. Econometric methods incorporate these random disturbance terms and also provide numerical values of coefficients of economic theories. Econometrics combines mathematical formulations of economic theory with empirical data. It thus enables to pass the abstrract theoretical schemes to numerical results in concrete cases.

1.5 Econometrics and Statistics

The concept of econometrics is somehow different from both mathematical statistics as well as economic statistics. Economic statistics is manily concerned with collection, tabulation and representation of economic data. They also explain the pattern of development of the data over time and relationship among various economic variables. But they do not provide the explanation of the development of the data over period of time which is mainly the job of an econometrician. Thus economic statistics is manily a descriptive aspect of economics. It does not provide explanations of the development of various variables. It does not also provide measurement of parameters of economic relationships.

On the other hand, mathematical statistics provides tools and techniques which are developed on the basis of controlled experiment. These techniques cannot be used in the case of economic theories as such experiments are not designed under controlled environment except for few cases. In physics and in some other sciences, researchers can keep all other conditions constant and change only one element in performing an experiment. But this cannot be done in economics where the real world is the laboratory. In the real world, all variables change continuously and simultaneously. Hence controlled experiments are impossible in economics.

Econometrics thus uses statistical methods and helps them adopt the problems of economic life. These particular statistical methods are called econometric methods. They measure economic relationship and take into account the stochastic or random elements as well. The random or stochastic elements that exist in the real world are specified and included in the determination of the observed economic data and empirical verification of economic theories.

1.6 Branches of Econometrics

Econometrics may be broadly divided into two branches : theoretical econometrics and empirical econometrics.

1.6.1 Theoretical Econometrics :

Theoretical econometrics deals with the development of appropriate methods or techniques for testing economic theory empirically. Econometric methods are actually statistical methods which are adapted to the characteristics of economic

relationships. There are two types of methods in theoretical econometrics i.e., single equation and simultaneous equation methods. Single equation techniques are applied to one economic relationship at a particular point of time whereas simultaneous equation techniques are used for all relationships of an economic model.

1.6.2 Applied Econometrics :

Applied econometrics deals with the application of techniques provided by theoretical econometrics to different branches of economic theory. It incorporates the problems of economic life and the findings of applied research in the fields of demand, supply, production, cost, consumption, investment, etc. and other economic theories.

1.7 Goals of Econometrics

Econometrics helps us achieve three main goals i.e., analysis, policy making and forecasting. We will discuss them one by one.

1.7.1 Analysis :

Economic theories provide qualitative statements or hypotheses without empirically verifying them. This analysis of economic theories for providing explanation of the economic system is done by econometrics.

1.7.2 Policy making :

Econometricians carry out analysis of economic theories and obtain estimates of numerical coefficients of the economic relationships. This estimates helps in prescribing appropriate policies to the government. For example, if the price elasticity of demand of certain goods are estimated, then it can be said how much additional revenue the government can raise by imposing tax on those commodities. Alternatively, if the estimates of price elasticities of exports and imports are calculated, it can be said how much the policy for devaluation will be effective in solving the balance of payments deficit problem.

1.7.3 Forecasting :

The numerical estimates of coefficients of economic relationships are used to forecast future values of variables, without which appropriate policies cannot be designed.

It may be mentioned that these goals are not mutually exclusive. Some

combination of all these aims or goals is necessary for a successful econometric application.

1.8 Methodologies of Econometric Research

Applied econometrics deals with the measurement of parameters of economic relationships and prediction of values of economic variables. However, the definitional relationships do not require any such measurements. For example, consider the relationship among national income, consumption expenditure and investment expenditure in a closed economy, i.e., $Y = C + 1$, This mathematical expression of national income in a closed economy does not explain its determination or causes of its variation. It is a definitional equation and does not require any measurement.

There are generally four stages in any econometric research.

Stage A : Specification of the model

Stage B : Estimation of the model

Stage C : Evaluation of estimates

Stage D : Evaluation of the forecasting power of the estimated model.

We will discuss these four stages one by one.

Stage A : Specification of the model :

It is the first stage of econometric research. It involves expressing the hypothesis or the economic theories in their mathematical form. In this stage the dependent and explanatory variables are identified and included in the model. The theoretical expectations about the sign, size of the parameters of the function are also determined and the mathematical form of the model is specified. For example consider the production function of the following form $Y = f(K, L)$ where K and L are the two factors of production, capital and labour, respectively and Y is the level of output. This function can be expressed in its mathematical form as a Cobb-Douglas production function as $Y = K^{\alpha}L^{\beta}$ or its log linear form as $\log Y = \alpha \log K + \beta \log L$. Also, some restrictions are to be imposed on it like $0 < (\alpha, \beta) < 1$, $\alpha + \beta > 1$ if there is increasing returns to scale, $\alpha + \beta < 1$ if there is decreasing returns to scale, $\alpha + \beta = 1$ if there is constant returns to scale. α is the elasticity of output with respect to capital and β is the elasticity of output with respect to labour.

Stage B : Estimation of the model :

It is the second stage of econometric research and deals with the estimation of the model specified in the first stage. This stage includes collection of data on different variables included in the specified model, examining the indetification conditions of the function in which we are interested, examining the aggregation problems of the involved function and examining the degree of correlation among the explanatory variables. For example if we consider the relationship between consumption expenditure and level of income, wealth and prices which is represented by the equation as follows : $C = \alpha Y + \beta W + \gamma P$ where C represents consumption expenditure, Y represents level of income, W represents wealth level and P represents price level. Then we need to check if Y and W, W and P, or Y and P are correlated i.e., the problem of multicollinearity exists or not. The next step involves the selection of appropriate econometric techniques for estimation of the function and examining the assumptions of the technique used for estimation and its economic implication for estimation of the coefficients.

Stage C : Evaluation of estimates :

After the estimation of the model, the next stage in econometric research is to consider the reliability of the estimated results i.e., the evaluation of the estimated results. Evaluation of estimates implies whether the estimated results are theoretically meaningful and statistically significant or not. For evaluation, three major criteria are used, namely, economic criteria, statistical criteria and econometric criteria.

Economic criteria are determined by the principles of economic theory and they refer to the size and sign of the estimated parameters of economic relationships. For example, the Keynesian consumption function is expressed in the mathematical form as follows : $C = a + bY$ where C is the consumption and Y is the level of income and a and b are the parameters whose values and signs are to be estimated on the basis of observed data. The existing theory suggests that $a > 0$ and $0 < b < 1$.

Statistical criteria or first order tests are determined on the basis of statistical theory and they focus on the statistical reliability of the estimated parameters. The most common criteria are correlation coefficients and standard error of estimates.

Econometric criteria or second order tests are determined on the basis of the theory of econometrics. The focus on whether the assumptions of the econometric method employed are satisfied or not. These tests are called secondary tests because

these are actually statistical tests and determine statistical reliability. They establish whether the estimates have the desirable properties of unbiasedness, consistency, etc.

Stage D : Evaluation of the forecasting power of the estimated model :

The objective of any econometric research is to obtain estimates of the coefficients of economic relationships and to use them for predicting future values of economic variables such that appropriate policies can be designed by the policy makers. So, econometricians must test the forecasting power of the estimated model as it plays an important role in designing appropriate policies.

1.9 Summary

To summarize, econometrics literally means economic measurement. It is an integration of economics, mathematics and statistics. But the subject has an importance of its own to be studied as a separate discipline. It does not rely on qualitative hypothesis only like economics, nor does it focus on purely statistical methods only. It adopts statistical methods after adapting them to the problems of economic life. There are three major goals of econometrics, namely, analysis, policy making and forecasting. Econometrics is divided mainly into two branches, namely, theoretical and applied econometrics. Theoretical econometrics deals with development of appropriate methods for measurement of economic relationships while applied econometrics deals with application of those methods in several branches of economic theory. Lastly, there are four stages of econometric research. The first stage involves specifying the model that is to be estimated, the second stage involves collection of data and obtaining the estimates of the parameters of the model, the third stage involves evaluation of the estimated parameters on the basis of economic, statistical and econometric criteria and the last stage involves evaluation of the forecasting power of the model.

1.10 Exercise

Short Answer Type Questions :

1. State true or false :

(a) Econometrics is an integration of economics, mathematics and statistics.

(b) Forecasting is not a goal of econometrics.

2. Choose the correct alternative :

(a) There are _____ stages of econometric research

(i) 3

(ii) 4

(iii) 5

(iv) 1

(b) _____ econometrics deals with the development of appropriate methods for measurement of economic relationship.

(i) Theoretical

(ii) Applied

3. Fill in the blanks :

(a) The term econometrics means _____.

(b) The evaluation of estimated parameters in econometrics is based on _____ criteria, _____ criteria and _____ criteria.

4. Define Econometrics.

5. What are the main branches of Econometrics ?

6. Mention the main goals of Econometrics.

Medium Answer Type Questions :

1. Discuss the relation between econometrics and economic theory.

2. How is statistics related to econometrics ?

3. What is meant by theoretical econometrics ?

4. Write a short note on applied econometrics.

Long Answer Type Questions :

1. What is econometrics? How is it different from mathematical economics and statistics ?

2. What are the two branches of economic theory ? Distinguish between them.

3. Mention the major stages of econometric research. Discuss briefly about them.
4. Briefly discuss the major goals of econometrics.

1.11 References

1. Gujarati, D (2003) : *Basic Econometrics*, McGraw Hill Higher Education.
2. Sarkhel, Jaydeb and Santosh Kumar Dutta (2020) : *An Introduction to Econometrics*, Book Syndicate Private Limited.
3. Koutsoyiannis, A (1996) : *Theory of Econometrics*, ELBS with Macmillan

Unit - 2 ❑ The Classical Linear Regression Model (CLRM)

Structure

2.1 Objectives

Reading this chapter, students will get an idea about

- The Simple Linear Regression Model
- Classical Linear Regression Model and its assumptions
- Ordinary least square (OLS) method of estimating regression parameters
- Properties of OLS estimators

2.2 Introduction

Economic theories deal with relationships among variables and these relationships when expressed in mathematical forms are mostly deterministic relationships or non-stochastic relationships. Deterministic relationships are those relationships where for a given value of the independent variable there exists a definite value of the dependent variable. Consider the economic theory of the law of demand, where we know that there exists an inverse relationship between the own price of a commodity and the quantity demanded of the commodity assuming that other things remain unchanged. If this relationship is expressed in mathematical form, then it can be expressed as $q = f(p)$ assuming ceteris paribus (other things remain unchanged). This functional relationship may be linear, quadratic, logarithmic, exponential or hyperbolic. If we specify the functional form of this ralationship as $q = \alpha + \beta p = 100 - 5p$, then for each value of price level, we get a unique value of quantity demanded. This type of relationship is called deterministic relationship or non-stochastic relationship. But such deterministic relationship is not found in the real world. This deterministic relationship breaks down if the ceteris paribus assumption is relaxed and then we get a stochastic or random relationship. The relationship between two variables, say, X and Y is said to be stochastic if for each value of the independent variable X there exists a probability distribution of the values of the dependent variable Y. In this case we rewrite the earlier demand equation as $q = \alpha + \beta p + u = 100 - 5p + u$ where u is the disturbance term as it disturbs the otherwise deterministic relationship.

2.3 The Simple Linear Regression Model

Economic relationships are actually deterministic relationships among variables, but these relationships are to be tested or verified empirically by econometric techniques, which imply a strong belief of the existence of stochastic variables or random disturbance terms in economic theories. The knowledge of econometrics tries to test these economic theories in terms of stochastic variables. The simplest form of stochastic relation between two economic variables is given by $Y_i = \alpha + \beta X_i + u_i$ where Y is the dependent variable, X is the independent variable, α

and β are regression parameters, u is the disturbance term, i represent the no. of observations and n is the sample size. The stochastic nature of this regression model states that for every value of the independent variable X, there exists a whole probability distribution of the dependent variable Y i.e., the value of Y can never be predicted directly due to the presence of stochastic term u. There are a number of reasons for which this stochastic variable u should be included in the simple linear regression model.

- **Omission of variables from the function :** The disturbance term considers the effect of several variables which are not included in the model. For exmaple, if we consider the simple regression model as mentioned above where Y, the dependent variable be consumption expenditure and X, the independent variable be disposable income, then it might be possible that there are other variables apart from disposable income which affect the consumption expenditure. But such variables are not included in the model and the effects of such variables are captured by the disturbance term.
- **Unpredictable element of randomness in human responses :** Human being does not behave like machines and so there is an unpredictable element in households' consumption expenditure behaviour which is captured by the disturbance term.
- **Imperfect specification of the mathematical model :** It might be possible that we have linearized a non-linear relationship or we might have left out some equations from the model. Such imperfect specifications of the mathematical form of the model are captured by the disturbance term.
- **Aggregation problem :** In economics, we are often faced with the problem of aggregation. We add up magnitudes whose behaviour are dissimilar and these result in disappearance of individual peculiarities from the model. There are other types of aggregation which lead to errors in the relationship of the variables in the model.
- **Due to errors in measurement :** The disturbance term measures the errors in recording or processing of the data on X and Y. It thus reflects the errors in the observations.

2.4 Classical Linear Regression Model and its Assumptions

We consider a stochastic relationship between two variables, X and Y which is given by the model $Y_i = \alpha + \beta X_i + u_i$ for $i = 1, 2, ..., n$ where Y is the dependent variable, X is the independent variable, u is the disturbance term, i denotes item, n denotes number of observations and α and β are parameters whose values are to be estimated on the basis of values of X and Y. This model is called Classical Linear Regression Model if it satisfies the following assumptions :

- **Assumption 1 :** The regression relationship is linear i.e., the variables are linearly related.
- **Assumption 2 :** X is non-stochastic for a given sample, but it may take different values for the given sample.
- **Assumption 3 :** The disturbance term u_i is a random variable and its probability distribution is assumed to be normal.
- **Assumption 4 :** The probability distribution of the disturbance term is such that its mean is zero i.e., $E(u_i) = 0$. Since we have, $Y_i = \alpha + \beta X_i + u_i$. So, $E(Y_i) = E(\alpha + \beta X_i + u_i) = E(\alpha) + \beta E(X_i) + E(u_i) = \alpha + \beta X_i$ as $E(u_i) = 0$ and $E(X_i) = X_i$. This implies that the expectation of the observed value of the dependent variable is its true value i.e., the probability distribution of Y_i is centred around the true relationship.
- **Assumption 5 :** The variance of the disturbance term is a constant and is independent of i where $i = 1, 2, ..., n$ and is denoted by σ_u^2 or σ^2. This implies that Var $(u_i) = \sigma_u^2$ i.e., $E[u_i - E(u_i)^2 = E[u_i]^2 = \sigma_u^2$ since we know from assumption 4 that $E(u_i) = 0$. So both the assumptions together imply that $u_i \sim ID\,(0, \sigma_u^2)$ for $i = 1, 2, ..., n$.
- **Assumption 6 :** Different error terms are independently distributed i.e., $E(u_i, u_j) = E(u_i).E(u_j)$ and Cov $(u_i, u_j) = 0$ for $i \neq j$ and Cov $(u_i, u_j) = \sigma_u^2$ for $i = j$ where, $i, j = 1, 2, ..., n$.
- **Assumption 7 :** The independent variable X is non-stochastic or non-random i.e., X is not a random variable and is measured without error and u_i is independent with explanatory variables i.e., Cov $(X_i, u_i) = 0$.

- **Assumption 8 :** The explanatory variables are independent to each other i.e., Cov $(X_i, X_j) = 0$. If this assumption is violated the problem of multicollinearity arises which will be discussed in Unit 4.
- **Assumption 9 :** The number of observations must be greater than the number of parameters to be estimated i.e., $n > k$.
- **Assumption 10 :** The model is correctly specified i.e., there does not exist any specification error.

2.5 Ordinary Least Square (OLS) Method of Estimating Regression Parameters

There are various methods for estimating regression parameters i.e., the method of moments, the method of ordinary least squares (OLS) and the method of maximum likelihood (MLE). We shall discuss here the method of ordinary least squares (OLS) for estimation of the regression parameters.

We consider a two-variable linear regression model as $Y_i + \alpha + \beta X_i + u_i$ where X is the independent variable, Y is the dependent variable and u is the disturbance term. If the assumptions mentioned in the previous section are satisfied, then this model is the classical linear regression model with parameters α and β which are to be estimated using OLS method. Let $\hat{\alpha}$ and $\hat{\beta}$ be the estimated values of α and β. The estimated relation becomes $\hat{Y}_1 = \hat{\alpha} + \hat{\beta} X_i$ and $e_i = Y_i - \hat{Y}_1$ is the residual term which shows the difference between the observed and estimated relation. In OLS method we need to estimate that values of $\hat{\alpha}$ and $\hat{\beta}$ for which $\sum_{i=1}^{n} e_i^2$ is minimum i.e., we need to minimise $\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_1\right)^2 = \sum_{i=1}^{n} \left(Y_i - \hat{\alpha} - \hat{\beta} X_i\right)^2$ through the choice of $\hat{\alpha}$ and $\hat{\beta}$. For minimisation the necessary conditions are

$$\frac{\delta \sum_{i=1}^{n} e_i^2}{\delta \hat{\alpha}} = 0$$

$$\text{or,} \quad \frac{\delta \sum_{i=1}^{n} \left(Y_i - \hat{\alpha} - \hat{\beta} X_i\right)^2}{\delta \hat{\alpha}} = 0$$

$$\text{or,} \quad -2\sum_{i=1}^{n}\left(Y_i-\hat{\alpha}-\hat{\beta}X_i\right)=0 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (1)$$

$$\text{and} \quad \frac{\delta\Sigma_{i=1}^{n} e_i^2}{\delta\hat{\beta}}=0$$

$$\text{or,} \quad \frac{\delta\Sigma_{i=1}^{n}\left(Y_i-\hat{\alpha}-\hat{\beta}X_i\right)^2}{\delta\hat{\beta}}=0$$

$$\text{or,} \quad -2\sum_{i=1}^{n}\left(Y_i-\hat{\alpha}-\hat{\beta}X_i\right)X_i=0 \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (2)$$

These two equations can be written as

$$\sum_{i=1}^{n}\left(Y_i-\hat{\alpha}-\hat{\beta}X_i\right)=0 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (1')$$

and

$$\sum_{i=1}^{n}\left(Y_i-\hat{\alpha}-\hat{\beta}X_i\right)X_i=0 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (2')$$

As $-2\neq 0$

From (1′) and (2′) we can write

$$\sum_{i=1}^{n}Y_i=n\hat{\alpha}+\hat{\beta}X_i \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (1'')$$

$$\sum_{i=1}^{n}X_iY_i=\hat{\alpha}\sum_{i=1}^{n}X_i+\hat{\beta}\sum_{i=1}^{n}X_i^2 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (2'')$$

These two equations are known as normal equations. Solving them we can estimate the values of $\hat{\alpha}$ and $\hat{\beta}$.

From (1″), dividing both sides by n we get,

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

or, $\hat{\alpha} = \overline{Y} - \hat{\beta}\overline{X}$ ..... (A)

which is the OLS estimate of $\hat{\alpha}$

Now, solving (1″) and (2″) using Cramer's rule we have,

$$\hat{\beta} = \frac{\begin{vmatrix} n & \sum Y_i \\ \sum X_i & \sum X_i Y_i \end{vmatrix}}{\begin{vmatrix} n & \sum X_i \\ \sum X_i & \sum X_i^2 \end{vmatrix}}$$

$$\text{or,} \quad \hat{\beta} = \frac{n\sum X_i Y_i - \sum X_i \sum Y_i}{n\sum X_i^2 - \left(\sum X_i\right)^2}$$

$$\text{or,} \quad \hat{\beta} = \frac{Cov\left(X_i, Y_i\right)}{Var\left(X_i\right)}$$

$$\text{or,} \quad \hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2}$$

Where $x_i = X_i - \overline{X}$ and $y_i = Y_i - \overline{Y}$

So, the OLS estimate of $\hat{\beta}$ is $\hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2}$ .....(B)

2.6 Properties of OLS Estimators

The OLS estimates are called BLUE (best, linear, unbiased estimates) provided that the random term *u* satisfies some general assumptions namely that *u* has zero mean and constant variance. This proposition along with the set of assumptions

under which it is true is known as the **Gauss - Markov Least Squares theorem**. The OLS estimates have basically three properties :

- They are linear
- They are unbaised
- They possess the smallest variance.

We now prove them one by one.

2.6.1 The Property of Linearity

The least square estimates of $\hat{\alpha}$ and $\hat{\beta}$ are linear functions of the observed sample values Y_i. We will prove this property as follows.

From the previous section we know, $\hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2}$ where $x_i = X_i - \overline{X}$ and $y_i = Y_i - \overline{Y}$ and $\hat{\alpha} = \overline{Y} - \hat{\beta}\overline{X}$. Now, from (B) we have,

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

or, $$\hat{\beta} = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}$$

or, $$\hat{\beta} = \frac{\sum_{i=1}^{n} Y_i\left(X_i - \overline{X}\right) - \overline{Y}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}$$

or, $$\hat{\beta} = \frac{\sum_{i=1}^{n} Y_i\left(X_i - \overline{X}\right)}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2} \quad \left[\text{since } \sum_{i=1}^{n}\left(X_i - \overline{X}\right) = 0\right]$$

or, $$\hat{\beta} = \frac{\sum_{i=1}^{n} Y_i x_i}{\sum_{i=1}^{n} x_i^2}$$

We assume, $\frac{x_i}{\sum_{i=1}^{n} x_i^2} = K_i$ for $i = 1, 2, \ldots\ldots n$

So, $\hat{\beta} = \sum_{i=1}^{n} K_i Y_i$

$\therefore \hat{\beta} = K_1 Y_1 + K_2 Y_2 + \ldots\ldots\ldots\ldots + K_n Y_n$

i.e., $\hat{\beta}$ is a linear function of Y.

Now, from (A) we know,

$$\hat{\alpha} = \overline{Y} - \hat{\beta}\overline{X}$$

$$\text{or,} \quad \hat{\alpha} = \overline{Y} - \overline{X} \sum_{i=1}^{n} K_i Y_i$$

$$\text{or,} \quad \hat{\alpha} = \frac{1}{n}\sum_{i=1}^{n} Y_i - \overline{X}\sum_{i=1}^{n} K_i Y_i$$

$$\text{or,} \quad \hat{\alpha} = \sum_{i=1}^{n}\left[\frac{1}{n} - K_i \overline{X}\right] Y_i$$

$$\text{or,} \quad \hat{\alpha} = \left[\frac{1}{n} - K_1 \overline{X}\right] Y_1 + \left[\frac{1}{n} - K_2 \overline{X}\right] Y_2 + \ldots\ldots + \left[\frac{1}{n} - K_n \overline{X}\right] Y_n$$

which shows that $\hat{\alpha}$ is also a linear function of Y_i.

2.6.2 The Property of Unbiasedness

The least square estimates of $\hat{\alpha}$ and $\hat{\beta}$ are said to be unbaised if $E(\hat{\alpha}) = a$ and $E(\hat{\beta}) = \beta$.

We know, $\hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2}$ where $x_i = X_i - \overline{X}$ and $y_i = Y_i - \overline{Y}$ and $\hat{\alpha} = \overline{Y} - \hat{\beta}\overline{X}$.

Now, we have,

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

or, $$\hat{\beta} = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}$$

or, $$\hat{\beta} = \frac{\sum_{i=1}^{n} Y_i\left(X_i - \overline{X}\right) - \overline{Y}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}$$

or, $$\hat{\beta} = \frac{\sum_{i=1}^{n} Y_i\left(X_i - \overline{X}\right)}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}$$ [since $\sum_{i=1}^{n}\left(X_i - \overline{X}\right) = 0$]

or, $$\hat{\beta} = \frac{\sum_{i=1}^{n} Y_i x_i}{\sum_{i=1}^{n} x_i^2}$$

or, $$\hat{\beta} = \sum_{i=1}^{n} K_i Y_i$$ where $K_i = \frac{x_i}{\sum_{i=1}^{n} x_i^2}$

Now, putting $Y_i = \alpha + \beta X_i + u_i$ we get,

$$\hat{\beta} = \sum_{i=1}^{n} K_i(\alpha + \beta X_i + u_i)$$

or, $$\hat{\beta} = \alpha\sum_{i=1}^{n} K_i + \beta\sum_{i=1}^{n} K_i X_i + \sum_{i=1}^{n} K_i u_i$$

We know, $K_i = \frac{x_i}{\sum_{i=1}^{n} x_i^2}$

So, $\sum_{i=1}^{n} K_i = \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2} = \frac{0}{\sum_{i=1}^{n} x_i^2} = 0$ since $\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} (X_i - \overline{X}) = 0$

Also, $\sum_{i=1}^{n} K_i X_i = \sum_{i=1}^{n} K_i \left(x_i + \overline{X}\right)$ as $x_i = X_i - \overline{X}$

or, $\sum_{i=1}^{n} K_i X_i = \sum_{i=1}^{n} K_i x_i + \overline{X} \sum_{i=1}^{n} K_i$

or, $\sum_{i=1}^{n} K_i X_i = \sum_{i=1}^{n} K_i x_i + 0$ as $\sum_{i=1}^{n} K_i = 0$

or, $\sum_{i=1}^{n} K_i X_i = \frac{\sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} x_i^2}$

or, $\sum_{i=1}^{n} K_i X_i = 1$

Hence we can write from, $\hat{\beta} = \alpha * 0 + \beta * 1 + \sum_{i=1}^{n} K_i u_i$

So, $E(\hat{\beta}) = E\left(\beta + \sum_{i=1}^{n} K_i u_i\right) = E(\beta) + \sum_{i=1}^{n} K_i E(u_i) = \beta + 0 = \beta$ as $E(\mu_i) = 0$

i.e., $\hat{\beta}$ is the unbiased estimator of β

From equation (A) the previous section we know,

$$\hat{\alpha} = \overline{Y} - \hat{\beta}\overline{X}$$

or, $\hat{\alpha} = \overline{Y} - \overline{X} \sum_{i=1}^{n} K_i Y_i$

or, $\hat{\alpha} = \frac{1}{n} \sum_{i=1}^{n} Y_i - \overline{X} \sum_{i=1}^{n} K_i Y_i$

$$\text{or,} \quad \hat{\alpha} = \sum_{i=1}^{n} \left[\frac{1}{n} - K_i \overline{X}\right] Y_i$$

Now, putting $Y_i = \alpha + \beta X_i + u_i$ we get,

$$\hat{\alpha} = \sum_{i=1}^{n} \left[\frac{1}{n} - K_i \overline{X}\right] [\alpha + \beta X_i + u_i]$$

$$\text{or,} \quad \hat{\alpha} = \sum_{i=1}^{n} \frac{1}{n}\alpha + \sum_{i=1}^{n} \frac{1}{n}\beta X_i + \sum_{i=1}^{n} \frac{1}{n} u_i - \sum_{i=1}^{n} \alpha K_i \overline{X} - \sum \beta K_i X_i \overline{X}$$

$$- \sum_{i=1}^{n} K_i \overline{X} u_i$$

$$\text{or,} \quad \hat{\alpha} = \frac{1}{n}\alpha \sum_{i=1}^{n} 1 + \frac{1}{n}\beta \sum_{i=1}^{n} X_i + \frac{1}{n} \sum_{i=1}^{n} u_i - \alpha \overline{X} \sum_{i=1}^{n} K_i - \beta \overline{X} \sum_{i=1}^{n} K_i X_i$$

$$- \overline{X} \sum_{i=1}^{n} K_i u_i$$

Since we know, $\sum_{i=1}^{n} K_i = 0, \sum_{i=1}^{n} K_i X_i = 1$ and $\sum_{i=1}^{n} 1 = n$, we can rewrite the above expression as

$$\hat{\alpha} = \frac{1}{n}\alpha n + \beta \overline{X} + \frac{1}{n} \sum_{i=1}^{n} u_i - 0 - \beta \overline{X}.1 - \overline{X} \sum_{i=1}^{n} K_i u_i$$

$$\text{or,} \quad \hat{\alpha} = \alpha + \beta \overline{X} + \frac{1}{n} \sum_{i=1}^{n} u_i - \beta \overline{X} - \overline{X} \sum_{i=1}^{n} K_i u_i$$

$$\text{or,} \quad \hat{\alpha} = \alpha + \frac{1}{n} \sum_{i=1}^{n} u_i - \overline{X} \sum_{i=1}^{n} K_i u_i$$

So, $E(\hat{\alpha}) = E(\alpha) + \frac{1}{n}\sum_{i=1}^{n} E(u_i) - \overline{X}\sum_{i=1}^{n} K_i E(u_i)$

or, $E(\hat{\alpha}) = \alpha$ since $E(u_i) = 0$

Hence $\hat{\alpha}$ is also an unbiased estimator of α.

2.6.3 The Property of Smallest Variance

This property indicates that the least square estimates are best i.e., they have the least variance compared with any other linear unbiased estimator obtained from other econometric methods. To prove this property we first estimate the variance of $\hat{\alpha}$ and $\hat{\beta}$ and then prove that their variance is least than any other estimate.

Variance of $\hat{\beta} = \text{Var}(\hat{\beta}) = E\left[\hat{\beta} - E(\hat{\beta})\right]^2 = E\left[\hat{\beta} - \beta\right]^2$ since $E(\hat{\beta}) = \beta$

Again, from the previous property we know,

$$\hat{\beta} = \beta + \sum_{i=1}^{n} K_i u_i$$

or, $$\hat{\beta} - \beta = \sum_{i=1}^{n} K_i u_i$$

or, $$(\hat{\beta} - \beta)^2 = \left[\sum_{i=1}^{n} K_i u_i\right]^2$$

or, $$E(\hat{\beta} - \beta)^2 = E\left[\sum_{i=1}^{n} K_i u_i\right]^2$$

or, $$\text{Var}(\hat{\beta}) = E\left[\sum_{i=1}^{n} K_i^2 u_i^2 + 2\sum_{i \neq j}\sum K_i K_j u_i u_j\right]$$

$$\text{or,} \quad \text{Var}(\hat{\beta}) = \left[\sum_{i=1}^{n} K_i^2 E(u_i^2) + 2\sum_{i \neq j}\sum K_i K_j E(u_i u_j)\right]$$

$$\text{or,} \quad \text{Var}(\hat{\beta}) = \sum_{i=1}^{n} K_i^2 E(u_i^2), \text{ since } E(u_i u_j) = 0$$

$$\text{or,} \quad \text{Var}(\hat{\beta}) = \sum_{i=1}^{n} K_i^2 \sigma_u^2$$

$$\text{or,} \quad \text{Var}(\hat{\beta}) = \frac{\sum_{i=1}^{n} x_i^2}{\left(\sum_{i-1}^{n} x_i^2\right)^2}\sigma_u^2 \text{ since } K_i = \frac{x_i}{\sum_{i=1}^{n} x_i^2}$$

$$\text{or,} \quad \text{Var}(\hat{\beta}) = \frac{\sigma_u^2}{\sum_{i=1}^{n} x_i^2}$$

We now obtain the variance of the OLS estimate of α i.e., $\hat{\alpha}$. So, Variance of $\hat{\alpha}$ = Var($\hat{\alpha}$). From (A) we know, $\hat{\alpha} = \overline{Y} - \hat{\beta}\overline{X}$. Substituting the value of $\hat{\beta} = \sum_{i=1}^{n} K_i Y_i$ in the above expression we get,

$$\hat{\alpha} = \overline{Y} - \overline{X}\sum_{i=1}^{n} K_i Y_i$$

$$\text{or,} \quad \hat{\alpha} = \frac{\sum_{i=1}^{n} Y_i}{n} - \overline{X}\sum_{i=1}^{n} K_i Y_i \qquad \text{or,} \quad \hat{\alpha} = \sum_{i=1}^{n}\left(\frac{1}{n} - \overline{X}K_i\right)Y_i$$

$$\text{So, Variance of } (\hat{\alpha}) = \text{Var}\left[\sum_{i=1}^{n}\left(\frac{1}{n} - \overline{X}K_i\right)Y_i\right]$$

$$\text{or,} \quad \text{Var}(\hat{\alpha}) = \sum_{i=1}^{n}\left[\left(\frac{1}{n} - \overline{X}K_i\right)\right]^2 \text{Var}(Y_i)$$

Now, $$\text{Var}\ (Y_i) = E\,[Y_i - E\,(Y_i)]^2$$

$$= E\,(\alpha + \beta X_i + u_i - \alpha - \beta X_i - E(u_i)]^2$$

$$= E\,(u_i - E\,(u_i)]^2$$

$$= \text{Var}\,(u_i) = \sigma_u^2$$

$$\text{So,} \quad \text{Var}(\hat{\alpha}) = \sum_{o=1}^{n}\left[\left(\frac{1}{n} - \overline{X}K_i\right)\right]^2 \sigma_u^2$$

$$\text{or,} \quad \text{Var}(\hat{\alpha}) = \sigma_u^2 \sum_{i=1}^{n}\left[\left(\frac{1}{n} - \overline{X}K_i\right)\right]^2$$

$$\text{or,} \quad \text{Var}(\hat{\alpha}) = \sigma_u^2 \sum_{i=1}^{n}\left(\frac{1}{n^2} - 2\frac{1}{n}\overline{X}K_i + \overline{X}^2K_i^2\right)$$

$$\text{or,} \quad \text{Var}(\hat{\alpha}) = \sigma_u^2 \sum_{i=1}^{n}\left(\frac{1}{n^2}\right) + 2\sigma_u^2 \frac{1}{n}\overline{X} \sum_{i=1}^{n} K_i + \sigma_u^2 \overline{X}^2 \sum_{i=1}^{n} K_i^2$$

$$\text{or,} \quad \text{Var}(\hat{\alpha}) = \sigma_u^2 \frac{1}{n} + \sigma_u^2 \overline{X}^2 \frac{1}{\sum_{i=1}^{n} x_i^2}$$

$$\left[\text{since} \quad \sum_{i=1}^{n} K_i = 0 \text{ and } \sum_{i=1}^{n} K_i^2 = \frac{1}{\sum_{i=1}^{n} x_i^2}\right]$$

$$\text{or,} \quad \text{Var}(\hat{\alpha}) = \sigma_u^2\left(\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^{n} x_i^2}\right)$$

$$\text{or,} \quad \text{Var}(\hat{\alpha}) = \sigma_u^2 \left(\frac{\sum_{i=1}^{n} x_i^2 + n\overline{X}^2}{n\sum_{i=1}^{n} x_i^2} \right)$$

$$\text{or,} \quad \text{Var}(\hat{\alpha}) = \sigma_u^2 \left[\frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2 + n\overline{X}^2}{n\sum_{i=1}^{n} x_i^2} \right]$$

$$\text{or,} \quad \text{Var}(\hat{\alpha}) = \sigma_u^2 \left[\frac{\sum_{i=1}^{n} X_i^2 - 2\overline{X}\sum_{i=1}^{n} X_i + n\overline{X}^2 + n\overline{X}^2}{n\sum_{i=1}^{n} x_i^2} \right]$$

$$\text{or,} \quad \text{Var}(\hat{\alpha}) = \sigma_u^2 \left[\frac{\sum_{i=1}^{n} X_i^2 - 2n\overline{X}^2 + 2n\overline{X}^2}{n\sum_{i=1}^{n} x_i^2} \right]$$

$$\text{or,} \quad \text{Var}(\hat{\alpha}) = \sigma_u^2 \left[\frac{\sum_{i=1}^{n} X_i^2}{n\sum_{i=1}^{n} x_i^2} \right]$$

Let us show that var $(\hat{\alpha})$ and var $(\hat{\beta})$ are least.

Now, let us assume that $\tilde{\beta}$ is any other linear unbiased estimator such that $\tilde{\beta} = \Sigma_i \omega_i Y_i$ where $\omega_i = K_i + d_i$ for any value of d_i. Here $\tilde{\beta}$ is linear. So, we have

$$\tilde{\beta} = \sum_i \omega_i Y_i$$

$$\text{or,} \quad \tilde{\beta} = \sum_i \omega_i \left(\alpha + \beta X_i + u_i\right)$$

$$\text{or,} \quad \tilde{\beta} = \alpha\sum_i \omega_i + \beta\sum_i \omega_i X_i + \sum_i \omega_i u_i$$

It is also assumed that, $\tilde{\beta}$ is also an unbiased estimator of β like $\tilde{\beta}$. So $E(\tilde{\beta}) = \beta$. This assumption will be fulfilled if $\Sigma_i \omega_i = 0$, $\Sigma_i \omega_i X_i = 1$ and $\Sigma_i \omega_i u_i = 0$.

Now, $\Sigma_i \omega_i = 0$ implies $\Sigma_i d_i = 0$ as $\Sigma_i \omega_i = \Sigma_i (K_i + d_i) = \Sigma_i K_i + \Sigma_i d_i$. Since we know from previous properties that $\Sigma_i K_i = 0$, so $\Sigma_i \omega_i = \Sigma_i d_i$ and if $\Sigma_i \omega_i = 0$ then, $\Sigma_i d_i = 0$.

Similarly, $\Sigma_i \omega_i X_i = 1$ requires that $\Sigma_i d_i X_i = 0$ because $\Sigma_i \omega_i X_i = \Sigma_i (K_i + d_i) X_i = \Sigma_i K_i X_i + \Sigma_i d_i X_i$. Since from previous properties, we know $\Sigma_i K_i X_i = 1$, so we have, $\Sigma_i d_i X_i = \Sigma_i \omega_i X_i - \Sigma_i K_i X_i$ i.e., $\Sigma_i d_i X_i = 1 - 1 = 0$.

So, if $\Sigma_i \omega_i = 0$, $\Sigma_i \omega_i X_i = 1$, $\Sigma_i d_i = 0$ and $\Sigma_i d_i X_i = 0$ then,

$$\tilde{\beta} = \beta + \sum_i \omega_i \mu_i$$

So, $E(\tilde{\beta}) = \beta + \Sigma_i \omega_i E(u_i) = \beta$ since $E(u_i) = 0$

Now, from the first property of OLS estimator, we know

$$\hat{\beta} = \sum_i K_i Y_i$$

So, $\text{Var}(\hat{\beta}) = \text{Var}(\Sigma_i K_i Y_i) = \Sigma_i K_i^2 \text{Var}(Y_i) = \Sigma_i K_i^2 \sigma_u^2$

Similarly, $\tilde{\beta} = \Sigma_i \omega_i Y_i$

So, $\text{Var}(\tilde{\beta}) = \text{Var}(\Sigma_i \omega_i Y_i) = \Sigma_i \omega_i^2 \text{Var}(Y_i) = \Sigma_i \omega_i^2 \sigma_u^2$

Now, $\Sigma_i \omega_i^2 = \Sigma_i (K_i + d_i)^2 = \Sigma_i K_i^2 + \Sigma_i d_i^2 + 2\Sigma_i K_i d_i$

Here, $\sum_i K_i d_i = \dfrac{\sum_i x_i d_i}{\sum_i x_i^2} = \dfrac{\sum_i (X_i - \overline{X}) d_i}{\sum_i x_i^2} = \dfrac{\sum_i x_i d_i - \sum_i \overline{X} d_i}{\sum_i x_i^2}$

$$= \frac{\sum_i x_i d_i - \overline{X} \sum_i d_i}{\sum_i x_i^2} = 0$$

[As $\sum_i X_i d_i = 0$ and $\sum_i d_i = 0$]

Putting the values we get,

$$\text{Var}(\tilde{\beta}) = \sigma_u^2 \sum_{i=1}^{n} \omega_i^2$$

or, $$\text{Var}(\tilde{\beta}) = \sigma_u^2 \sum_{i=1}^{n} (K_i^2 + d_i^2)$$

or, $$\text{Var}(\tilde{\beta}) = \sigma_u^2 \sum_{i=1}^{n} K_i^2 + \sigma_u^2 \sum_{i=1}^{n} d_i^2$$

or, $$\text{Var}(\tilde{\beta}) = \text{Var}(\hat{\beta}) + \sigma_u^2 \sum_{i=1}^{n} d_i^2$$

Since, $\sigma_u^2 \sum_{i=1}^{n} d_i^2 > 0$, we can write from the above expression, Var ($\tilde{\beta}$) > Var ($\hat{\beta}$) i.e., the OLS estimate of β has the lowest variance than any other linear unbaised estimate.

In a similar manner it can be proved that the OLS estimate of α i.e., $\hat{\alpha}$ possesses the least variance than any other linear unbaised estimate of α.

We assume, $\tilde{\alpha}$ to be a linear unbaised estimate of α other than the OLS estimate with weights ω_i where $\omega_i = K_i + d_i$

We have $$\hat{\alpha} = \sum_{i=1}^{n} \left(\frac{1}{n} - \bar{X} K_i \right) Y_i$$

Similarly, $$\tilde{\alpha} = \sum_{i=1}^{n} \left(\frac{1}{n} - \bar{X} \omega_i \right) Y_i$$

So, $$\tilde{\alpha} = \left(\frac{1}{n} - \bar{X} \omega_i \right) Y_1 + \left(\frac{1}{n} - \bar{X} \omega_2 \right) Y_2 + \cdots + \left(\frac{1}{n} - \bar{X} \omega_n \right) Y_n$$

i.e., $\tilde{\alpha}$ is a linear function of Y_i

Now, $\tilde{\alpha}$ is an unbiased estimator of α if E ($\tilde{\alpha}$) = α

Now, putting $Y_i = \alpha + \beta X_i + u_i$ we get,

$$\tilde{\alpha} = \sum_{i-1}^{n}\left(\frac{1}{n} - \overline{X}\omega_i\right)(\alpha + \beta X_i + u_i)$$

$$\text{or,} \quad \tilde{\alpha} = \alpha\left[1 - \overline{X}\sum_{i=1}^{n}\omega_i\right] + \beta\left[\overline{X} - \overline{X}\sum_{i=1}^{n}\omega_i X_i\right] + \sum_{i=1}^{n}\left[\frac{1}{n} - \overline{X}\omega_i\right]u_i$$

$$\text{So, } E(\tilde{\alpha}) = \alpha\left[1 - \overline{X}\,E\left(\Sigma_{i=1}^{n}\omega_i\right)\right] + \beta\left[X - \overline{X}\,E\left(\Sigma_{i=1}^{n}\omega_i X_i\right)\right] + E\left[\Sigma_{i=1}^{n}\left(\frac{1}{n} - \overline{X}\omega_i\right)u_i\right]$$

Now, $E(\tilde{\alpha}) = \alpha$ if and only if $\Sigma_{i=1}^{n}\omega_i = 0, \Sigma_{i=1}^{n}\omega_i X_i = 1$ and $\Sigma_{i=1}^{n}\omega_i u_i = 0$ and these conditions require $\Sigma_{i=1}^{n} d_i = 0$ and $\Sigma_{i=1}^{n} d_i X_i = 0$.

Now, variance of $\tilde{\alpha}$ is given by

$$\text{Var}(\tilde{\alpha}) = \text{Var}\left[\sum_{i=1}^{n}\left(\frac{1}{n} - \overline{X}\omega_i\right)Y_i\right]$$

$$\text{or,} \quad \text{Var}(\tilde{\alpha}) = \sum_{i=1}^{n}\left(\frac{1}{n} - \overline{X}\omega_i\right)^2 \text{Var}\left[Y_i\right]$$

$$\text{or,} \quad \text{Var}(\tilde{\alpha}) = \sigma_u^2\sum_{i=1}^{n}\left[\frac{1}{n^2} - 2\frac{1}{n}\overline{X}\omega_i + \overline{X}^2\omega_i^2\right]$$

$$\text{or,} \quad \text{Var}(\tilde{\alpha}) = \sigma_u^2\left[\frac{n}{n^2} - 2\frac{1}{n}\overline{X}\sum_{i=1}^{n}\omega_i + \overline{X}^2\sum_{i=1}^{2}\omega_i^2\right]$$

$$\text{or,} \quad \text{Var}(\tilde{\alpha}) = \sigma_u^2\left[\frac{1}{n} + \overline{X}^2\sum_{i=1}^{n}\omega_i^2\right] \text{since} \sum_{i=1}^{n}\omega_i = 0$$

$$\text{or,} \quad \text{Var}(\tilde{\alpha}) = \sigma_u^2\left[\frac{1}{n} + \overline{X}^2\left(\sum_{i=1}^{n}K_i^2 + \sum_{i=1}^{n}d_i^2\right)\right] \quad \text{since } \Sigma K_i d_i = 0$$

$$\text{or,}\quad \text{Var}(\tilde{\alpha}) = \sigma_u^2\left[\frac{1}{n}+\frac{\overline{X}^2}{\sum_{i=1}^{n} x_i^2}\right]+\left[\sigma_u^2\overline{X}^2\sum_{i=1}^{n} d_i^2\right] \text{ since } \sum K_i^2 = \frac{1}{\sum_{i=1}^{n} x_i^2}$$

$$\text{or,}\quad \text{Var}(\tilde{\alpha}) = \text{Var}(\hat{\alpha})+\left[\sigma_u^2\overline{X}^2\sum_{n-1}^{n} d_i^2\right]$$

Since $\sum_{i=1}^{n} d_i^2 > 0$ because all $d_i's$ are not zero, so $\text{Var}(\tilde{\alpha}) > \text{Var}(\hat{\alpha})$ i.e., the OLS estimate of α has the least variance.

2.7 Goodness of Fit of the Multiple Correlation Coefficient (R^2)

So far we were concerned mainly about the estimation and precision of the regression parameters α and β. But we need to consider the regression line as a whole and examine its goodness of fit. We know that the error of estimate, $e_i = Y_i - \hat{Y}_i$ So, $Y_i = \hat{Y}_i + e_i$

i.e., observed value = estimated value and the error of estimate.

From this we can write $Y_i - \overline{Y} = (\hat{Y}_i - \overline{Y}) + e_i$

Now, squaring both sides and summing for all i we have,

$$\sum\left(Y_i - \overline{Y}\right)^2 = \sum\left[\left(\hat{Y}_i - \overline{Y}\right)+e_i\right]^2$$

$$\text{or,}\quad \sum\left(Y_i - \overline{Y}\right)^2 = \sum\left(\hat{Y}_i - \overline{Y}\right)+2\sum\left(\hat{Y}_i - \overline{Y}\right)+\sum e_i^2$$

$$\text{Now,}\quad \sum\left(\hat{Y}_i - \overline{Y}\right)e_i = \sum(\hat{\alpha}+\hat{\beta}X_i - \overline{Y})e_i = \hat{\alpha}\sum e_i + \hat{\beta}\sum X_i e_i - \overline{Y}\sum e_i$$

From the first and second normal equations we know,

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

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

Also, $e_i = Y_i - \hat{Y}_i$

$= Y_i - \hat{\alpha} - \hat{\beta}X_i$

$\therefore \sum e_i = \sum\left(Y_i - \hat{\alpha} - \hat{\beta}X_i\right)$

$= \sum Y_i - n\hat{\alpha} - \hat{\beta}\sum X_i$

$= \sum Y_i - \sum Y_i$

$= 0$

and $\sum X_i e_i$

$= \sum X_i\left(Y_i - \hat{Y}_i\right)$

$= \sum X_i\left(Y_i - \hat{\alpha} - \hat{\beta}X_i\right)$

$= \sum X_i Y_i - \hat{\alpha}\sum X_i - \hat{\beta}\sum X_i^2$

$= 0$

So, $\sum\left(\hat{Y}_i - \overline{Y}\right)e_i = 0$

Hence we can write, $\sum(Y_i - \overline{Y})^2 = \sum(\hat{Y}_i - \overline{Y})^2 + \sum e_i^2$ or, $\sum(Y_i - \overline{Y})^2 = \sum(\hat{Y}_i - \overline{Y})^2 + \sum(e_i - \overline{e})^2$ $[\because \overline{e} = 0]$ i.e., $\text{Var}(Y) = \text{Var}(\hat{Y}) + \text{Var}(e)$

or, Total sum of squares (TSS) = Explained sum of squares (ESS) + Residual sum of squares (RSS). Let us express our result in deviational form.

Our first term on the RHS $= \sum\left(\hat{Y}_i - \overline{Y}\right)^2$

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

$$= \sum\left[\left(\overline{Y} - \hat{\beta}\overline{X}\right) + \hat{\beta}X_i - \overline{Y}\right]^2 \quad \left[\because \hat{\alpha} = \overline{Y} - \hat{\beta}X\right]$$

$$= \sum\left[\hat{\beta}\left(X_i - \overline{X}\right)\right]^2$$

$$= \hat{\beta}^2 \sum\left(X_i - \overline{X}\right)^2$$

$$= \hat{\beta}^2 \sum x_i^2$$

Hence we can write $\sum\left(Y_i - \overline{Y}\right)^2 = \hat{\beta}^2 \sum\left(X_i - \overline{X}\right)^2 + \sum e_i^2$

or, $\sum y_i^2 = \hat{\beta}^2 \sum x_i^2 + \sum e_i^2$

i.e. $\text{Var}(Y) = \text{Var}(\hat{Y}) + \text{Var}(e)$

or, TSS = ESS + RSS

Thus the total variations in Y can be decomposed into two parts :

(i) The estimated effect of X on the variations in Y which we call the explained sum of squares.

(ii) $\sum e_i^2$ or the unexplained variation in Y in the estimated relationship between Y and X.

We call it residual or unexplained sum of squares (RSS)

Thus, ESS + RSS = TSS

The ratio of ESS and TSS is taken as a measure of *goodness of fit* of the regresion line. It is also called coefficient of determination and is denoted by R^2.

$$\text{Thus, } R^2 = \frac{\text{Explained Variation (ESS)}}{\text{Total Variation (TSS)}}$$

$$= \frac{\text{Var}\,(\hat{Y})}{\text{Var}\,(Y)}$$

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

$$= \frac{\sum y_i^2 - \sum e_i^2}{\sum y_i^2}$$

$$= 1 - \frac{\sum e_i^2}{\sum y_i^2}$$

$$= 1 - \frac{\text{RSS}}{\text{TSS}}$$

From this we can write, $\frac{\text{RSS}}{\text{TSS}} = 1 - R^2 = 1 - \frac{\text{ESS}}{\text{TSS}}$

In symbols, ESS + RSS = TSS

or, $\frac{\text{ESS}}{\text{TSS}} + \frac{\text{RSS}}{\text{TSS}} = 1$

$\therefore \quad \frac{\text{ESS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}$

and $\frac{\text{RSS}}{\text{TSS}} = 1 - \frac{\text{ESS}}{\text{TSS}}$

Let us determine the range of value of R^2.

We have, Var (Y) = Var $(\hat{Y}) + \text{Var}(e)$

Since Var $(e) \geq 0$

ee may write, $0 \leq$ Var $(\hat{Y}) \leq$ Var (Y)

Thus, $0 \leq \frac{\text{Var}(\hat{Y})}{\text{Var}(Y)} \leq 1$

or, $0 \leq R^2 \leq 1$

When Var $(\hat{Y}) = 0$, $R^2 = 0$

i.e. When $\sum y_i^2 = \sum e_i^2$

or, Var (Y) = Var (e)

Here observed values are as much variable as the error of estimates. Then the estimated line has no "goodness of fit". On the other hand when, $R^2 = 1$ Var $(\hat{Y})$ = Var (Y) i.e., $\sum e_i^2 = 0$ or Var $(e) = 0$ i.e., the regrssion line has the best fit or, the highest goodness of fit.

2.8 Some Numerical Examples

1. Consider the data on advertising expenditures (X) and sales revenue (Y) for an athletic sportswear store for 5 months.

The observations are as follows :

Month	Sales Revenue (Y) (in '000 ₹)	Advertising Expenditure (X) (in '00 ₹)
1	3	1
2	4	2
3	2	3
4	6	4
5	8	5

Estimate the regression equation.

Solution :

Let $Y_i = \alpha + \beta X_i + u_i$ be the regression equation.

The two normal equations for estimating the regression coefficients are :

$$\sum_{i=1}^{n} Y_i = n\hat{\alpha} + \hat{\beta} \sum_{i=1}^{n} X_i \quad \text{.....(1)}$$

$$\sum_{i=1}^{n} X_i Y_i = \hat{\alpha} \sum_{i=1}^{n} X_i + \hat{\beta} \sum_{i=1}^{n} X_i^2 \quad \text{.....(2)}$$

Calculation for estimating $\hat{\alpha}$ and $\hat{\beta}$ i.e., estimates of α and β.

Month	X_i	Y_i	X_i^2	X_iY_i	$\hat{u}_i = Y_i - \hat{\alpha} - \hat{\beta}X_i$
1	1	3	1	3	0.8
2	2	4	4	8	0.6
3	3	2	9	6	– 2.6
4	4	6	16	24	0.2
5	5	8	25	40	1.0
Total	$\sum_{i=1}^{n} X_i = 15$	$\sum_{i=1}^{n} Y_i = 23$	$\sum_{i=1}^{n} X_i^2 = 55$	$\sum_{i=1}^{n} X_iY_i = 81$	$\sum \hat{u}_i = 0$

Here $n = 5$

So, from (1) and (2) we have

$5\hat{\alpha} + 15\hat{\beta} = 23$ .....(1a)

$15\hat{\alpha} + 55\hat{\beta} = 81$ ..... (2a)

Solving (1a) and (2a) by Cramer's rule, we get,

$$\hat{\alpha} = \frac{\begin{vmatrix} 23 & 15 \\ 31 & 55 \end{vmatrix}}{\begin{vmatrix} 5 & 15 \\ 15 & 55 \end{vmatrix}} = \frac{1265-1215}{275-225} = \frac{50}{50} = 1$$

$$\hat{\beta} = \frac{\begin{vmatrix} 5 & 23 \\ 15 & 81 \end{vmatrix}}{\begin{vmatrix} 5 & 15 \\ 15 & 55 \end{vmatrix}} = \frac{405-345}{275-225} = \frac{60}{50} = 1.2$$

The estimated regression equation is, $\hat{Y} = \hat{\alpha} + \hat{\beta}X \Rightarrow \hat{Y} = 1.0 + 1.2X$

This result states that if advertisement expenditure (X) is ₹ 0 then sales revenue is ₹ 1000.

If advertisement expenditure increases by 1 unit i.e., ₹ 100 then sales revenue on an average rises by ₹ 1,200. The errors are also estimated by $\hat{u}_i = Y_i - 1.0 - 1.2X_i$ as shown in the last column of the table.

2. The following table includes the price and quantity demanded of the product of a monopolist over a six year period :

Year	Quantity ('000 Kg)	Price ('00 ₹)
1990	8	2
1991	3	4
1992	4	3
1993	7	1
1994	8	3
1995	0	5

(i) Estimate lthe demand function, assuming a linear demand function. Comment on the values of the estimated coefficients ($\hat{\alpha}$ and $\hat{\beta}$) on the basis of economic theory.

(ii) Forecast the level of demand if price rises from ₹ 4 to ₹ 5. Comment on your forecast.

Solution :

Let $Y_i = \alpha + \beta X_i$ for $i = 1, 2, ... 6$, be the linear demand function. By the OLS method we can get the estimators of α and β.

Here Y = demand, X = price, α, β are two parameters.

Theoretically we may assume $\alpha > 0$, $\beta < 0$. By OLS method

$$\hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2} \text{ when } x_i = X_i - \overline{X},$$

$$y_i = Y_i - \overline{Y} \text{ and } \hat{\alpha} = \overline{Y} - \hat{\beta}\,\overline{X}$$

$$\overline{X} = \frac{\sum X_i}{n}, \overline{Y} = \frac{\sum Y_i}{n}$$

Calculation for $\hat{\alpha}$ and $\hat{\beta}$

Year (n)	Quantity ('000 Kg) Y_i	Price ('00 ₹) X_i	$y_i = Y_i - \overline{Y}$	$x_i = X_i - \overline{X}$	$x_i y_i$	x_i^2
1990	8	2	3	–1	–3	1
1991	3	4	–2	1	–2	1
1992	4	3	–1	0	0	0
1993	7	1	+2	–2	–4	4
1994	8	3	3	0	0	0
1995	0	5	–5	2	–10	4
Total	$\Sigma Y_i = 30$	$\Sigma X_i = 18$	$\Sigma y_i = 0$	$\Sigma x_i = 0$	$\Sigma x_i y_i = -19$	$\Sigma x_i^2 = 10$

Here $n = 6$

$$\overline{Y} = \frac{\sum Y_i}{n} = \frac{30}{6} = 5$$

$$\overline{X} = \frac{\sum X_i}{n} = \frac{18}{6} = 3$$

Now, $$\hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{-19}{10} = -1.9$$

$$\hat{\alpha} = \overline{Y} - \hat{\beta}\overline{X} = 5 - (-1.9) \times 3$$

$$= 5 + 5.7 = 10.7$$

So, the estimated regression equation is $Y = 10.7 - 1.9X$

This is consistent with the economic theory where we assume $\alpha > 0$ and $\beta < 0$ and it clearly shows an inverse relationship between price and demand (i.e., the law of demand holds true).

From the estimated demand function, we have to forecast the level of demand when price rises to ₹ 5 i.e., $X = 5$.

The demand function is $\hat{Y} = 10.7 - 1.9X$

When $X = 4$, $\hat{Y} = 10.7 - (1.9 \times 4)$

$= 7.09$

If $X = 5$, $\hat{Y} = 10.7 - (1.9 \times 5)$

$= 10.7 - 9.5$

$= 1.2$

This shows that if price rises from ₹ 4 to ₹ 5, quantity demanded falls from 7.09 to 1.2. i.e., from 7,090 Kg to 1,200 Kg.

3. The following table shows ten pairs of observation on X (price) and Y (quantity supplied).

No. of Observations (n)	Quantity (in tons) (Y)	Price (in '00 ₹) (X)
1	69	9
2	76	12
3	52	6
4	56	10
5	57	9
6	77	10
7	58	7
8	55	8
9	67	12
10	53	6
11	72	11
12	64	8

(i) Assuming a linear supply function, estimate the supply function. Comment on the values of the estimated coefficients ($\hat{\alpha}$ and $\hat{\beta}$) on the basis of economic theory.

Solution :

Let $Y_i = \alpha + \beta X_i$ for $i = 1, 2 \ldots\ldots\ldots 12$ be the linear supply function. By OLS method we can get the estimates of α and β.

Here Y = supply, X = price, α and β are two parameters. Theoretically we may assume $\alpha \gtrless 0$ and $\beta > 0$.

By OLS method we may get,

$$\hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2} \quad \text{where} \quad x_i = X_i - \overline{X},$$

$$y_i = Y_i - \overline{Y}$$

and $\hat{\alpha} = \overline{Y} - \hat{\beta}\overline{X}$

where $\overline{X} = \frac{\sum X_i}{n}$

and $\overline{Y} = \frac{\sum Y_i}{n}$

Calculation for (α, β)

Observation (n)	Y_i Quantity (in tons)	X_i Price (in '00 ₹)	$x_i = X_i - \overline{X}$	$y_i = Y_i - \overline{Y}$	$x_i y_i$	x_i^2
1	69	9	0	6	0	0
2	76	12	3	13	39	9
3	52	6	–3	–11	33	9
4	56	10	1	–7	–7	1
5	57	9	0	–6	0	0
6	77	10	1	14	14	1
7	58	7	–2	–5	10	4
8	55	8	–1	–8	8	1

Observation (n)	Y_i Quantity (in tons)	X_i Price (in '00 ₹)	$x_i = X_i - \overline{X}$	$y_i = Y_i - \overline{Y}$	$x_i y_i$	x_i^2
9	67	12	3	4	12	9
10	53	6	–3	–10	30	9
11	72	11	2	9	18	4
12	64	8	–1	1	–1	1
Total n = 12	ΣY_i = 756	ΣX_i = 108	Σx_i = 0	Σy_i = 0	$\Sigma x_i y_i$ = 156	Σx_i^2 = 48

$$\therefore \quad \overline{X} = \frac{\sum X_i}{n} = \frac{108}{12} = 9$$

$$\overline{Y} = \frac{\sum Y_i}{n} = \frac{756}{12} = 63$$

Now the OLS estimators of the regression coefficients α and β can be obtained by,

$$\hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{156}{48} = 3.25$$

and $\hat{\alpha} = \overline{Y} - \hat{\beta}\overline{X}$

$= 63 - 3.25 \times 9$

$= 63 - 29.25 = 33.75$

Thus the estimated supply function is

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

or, $\hat{Y} = 33.75 + 3.25\,X$

Here we see $\hat{\alpha} > 0$ and $\hat{\beta} > 0$. This means that there is a direct positive relation between supply and price. The intercept of the supply function is positive here. Hence our results are consistent with the theory.

4. Find the value of R^2 from the following information and comment. $\sum_{i=1}^{n} x_i y_i = 3347.60$, $\sum_{i=1}^{n} x_i^2 = 604.80$, $\sum_{i=1}^{n} y_i^2 = 19837$, $n = 20$ where $x_i = X_i - \overline{X}$ and $y_i = Y_i - \overline{Y}$

Solution :

$$\text{Since } R^2 = \frac{\hat{\beta}^2 \sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} y_i^2} \text{ where } \hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

$$\therefore \hat{\beta}^2 = \left[\frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}\right]^2$$

$$= \left(\frac{3347.60}{604.80}\right)^2$$

$$= (5.54)^2 = 30.69$$

$$\text{Now } R^2 = \frac{\hat{\beta}^2 \sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} y_i^2}$$

$$= \frac{30.69 \times 604.80}{19837} = \frac{18561.312}{19837} = 0.935$$

$$\therefore R^2 = 0.935$$

This suggests that 93.5 per cent of the sample observations of Y can be attributed to the variations of the fitted value of Y i.e., $\hat{Y}_i$ or we say that our regression line fits the given data well.

Thus R^2 measures the proportion of variations in the dependent variable that is explained by the independent variable.

5. A sample of 20 observations corresponding to the regression model $Y_i = \alpha + \beta X_i + u_i$ where u_i is normally distributed with mean zero and unknown variance σ_u^2, gives the following data :

$$\sum_{i=1}^{n} Y_i = 21.9, \sum_{i=1}^{n} (Y_i - \overline{Y})^2 = 86.9$$

$$\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y}) = 106.4$$

$$\sum_{i=1}^{n} X_i = 186.2$$

$$\sum_{i=1}^{n} (X_i - \overline{X})^2 = 215.4, \; n = 20.$$ Obtain the usual regression results.

Solution : On the basis of the given information we have to fit a linear relation between Y (dependent variable) and X (explanatory variable).

(i) Estimation of $\hat{\alpha}$ and $\hat{\beta}$:

We know that $\hat{\beta} = \dfrac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} = \dfrac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n} (X_i - \overline{X})^2}$

$\therefore \; \hat{\beta} = \dfrac{106.4}{215.4} = 0.494$ and $\hat{\alpha} = \overline{Y} - \beta \overline{X}$

$= 1.095 - 0.494 \times 9.31$

$= 1.095 - 4.60 = -3.505$

where $\overline{Y} = \dfrac{\sum_{i=1}^{n} Y_i}{n} = \dfrac{21.9}{20} = 1.095$ and $\overline{X} = \sum_{i=1}^{n} \dfrac{X_i}{n} = \dfrac{186.2}{20} = 9.31$

Thus we have $\hat{\alpha}$ = –3.505 and $\hat{\beta}$ = 0.494 and our estimated regression line is $\hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i \Rightarrow \hat{Y}_i = -3.503 + 0.494X_i$

Estimation of variances :

We know that, $\text{Var}(\hat{\alpha}) = \sigma_u^2 \left[\frac{\sum_{i=1}^{n} X_i^2}{n\sum_{i=1}^{n} x_i^2} \right]$ and $\text{Var}(\hat{\beta}) = \frac{\hat{\sigma}_u^2}{\sum_{i=1}^{n} x_i^2}$

Here we see that σ_u^2 is not known and hence we replace it by its unbiased estimator $\hat{\sigma}_u^2 = \sum_{i=1}^{n} e_i^2 / n-2$

Thus we have, $\text{Var}(\hat{\alpha}) = \hat{\sigma}_u^2 \frac{\sum_{i=1}^{n} X_i^2}{n\sum_{i=1}^{n} x_i^2}$ and $\text{Var}(\hat{\beta}) = \frac{\hat{\sigma}_u^2}{\sum_{i=1}^{n} x_i^2}$

Again, we know that $\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} y_i^2 - \hat{\beta}\sum_{i=1}^{n} x_i^2$

$\therefore \sum_{i=1}^{n} e_i^2 = 86.9 - (0.494)^2 \times 215.4$

$= 86.9 - 52.56 = 34.34$

Now, $\hat{\sigma}_u^2 = \sum_{i=1}^{n} e_i^2 / n-2 = \frac{34.34}{20-2} = \frac{34.34}{18} = 1.908$

where $\sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} (Y_i - \overline{Y})^2$ and $\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} (X_i - \overline{X})^2$.

Now, $\text{Var}(\hat{\alpha}) = \hat{\sigma}_u^2 \left[\frac{\sum_{i=1}^{n} X_i^2}{n\sum_{i=1}^{n} x_i^2} \right] = \frac{1.908 \times 1948.922}{20 \times 215.4} = 0.8631$

We have, $\sum_{i=1}^{n}(X_i - \overline{X})^2 = 215.4$ or, $\sum_{i=1}^{n} X_i^2 - n\overline{X}^2 = 215.4$

or, $\sum_{i=1}^{n} X_i^2 = 215.4 + n\overline{X}^2 = 215.4 + 20 \times (9.3)^2$

$= 215.4 + 1733.522$

$= 1948.922$. We have put this value to determine Var $(\hat{\alpha})$.

$\therefore$ Var $(\hat{\alpha}) = 0.8631$

Similarly, Var $(\hat{\beta}) = \dfrac{\hat{\sigma}_u^2}{\sum_{i=1}^{n} x_i^2} = \dfrac{1.908}{215.4} = 0.0089$

Now SE$(\hat{\alpha}) = \sqrt{\text{Var}(\hat{\alpha})} = \sqrt{0.8631} = 0.929$

SE$(\hat{\beta}) = \sqrt{\text{Var}(\hat{\beta})} = \sqrt{0.0089} = 0.094$

Construction of confidence intervals :

Now we like to set up a confidence interval for α and β at (a) $P = 0.95$ (i.e., 5% level of significance) and (b) $P = 0.99$ (i.e., 1% level of significance)

In other words, we like to find the value of 't' that cuts off (a) 0.025 and (b) 0.005 of the area at the tail end of the distribution on both sides. From table value : $t_{0.025,\ n-2} = t_{0.025,\ 18} = 2.101$ and $t_{0.025,\ n-2} = t_{0.025,\ 18} = 2.878$

Therefore 95% confidence internal for α are : $\hat{\alpha} \pm t_{0.025,\ n-2}$ SE$(\hat{\alpha})$ i.e., $P[\hat{\alpha} - t_{0.025}\ n-2$ SE$(\hat{\alpha}) \le \alpha \le \hat{\alpha} + t_{0.025,\ n-2}$ SE$(\hat{\alpha}) = 0.95$ and 99% confidence interval for α are : $\hat{\alpha} \pm t_{0.0005,\ n-2}$ SE$(\hat{\alpha})$. i.e., $P[\hat{\alpha} - t_{0.005}\ n-2$ SE$(\hat{\alpha}) \le \alpha \le \hat{\alpha} + t_{0.005,\ n-2}$ SE$(\hat{\alpha})] = 0.99$

Therefore 95% confidence interval for α would be : $\hat{\alpha} \pm t_{0.025,\ n-2}$ SE$(\hat{\alpha})$.

$\Rightarrow -3.505 \pm 2.101 \times 0.929$

or, -3.505 ± 1.9518

Similarly, 99% confidence interval for α would be :

$-3.505 \pm 2.878 \times 0.929$

or, -3.505 ± 2.6736

Similarly 95% confidence interval of β are :

$\hat{\beta} \pm t_{0.025,\ n-2}\ \text{SE}(\hat{\beta})$

i.e., $P[\hat{\beta} - t_{0.025,\ n-2}\ \text{SE}(\hat{\beta}) \leq \beta \leq \hat{\beta} + t_{0.025,\ n-2}\ \text{SE}(\hat{\beta})]$
$= 0.95$

and 99% confidence interval for β are : $\hat{\beta} \pm t_{0.005,\ n-2}\ \text{SE}(\hat{\beta})$

i.e., $P[\hat{\beta} - t_{0.005,\ n-2}\ \text{SE}(\hat{\beta}) \leq \beta \leq \hat{\beta} + t_{0.005,\ n-2}\ \text{SE}(\hat{\beta})]$
$= 0.99$

Thus 95% confidence interval for β would be :

$\hat{\beta} \pm t_{0.025,\ n-2}\ \text{SE}(\hat{\beta})$

or, $0.494 \pm 2.101 \times 0.494$

or, 0.494 ± 0.1974

where $\hat{\beta} = 0.494$ and

$t_{0.025,\ n-2} = t_{0.025,\ 18}$
$= 2.101$

$\text{SE}(\hat{\beta}) = 0.094$

Hypothesis testing : Suppose we like to test $H_0 = \beta = 0$ against the alternative $H_1 : \beta \neq 0$. Now on the basis of the given sample $H_0 : \beta = 0$ will be rejected at 5% level of significance if $|t_{n-2}| = \left|\frac{\hat{\beta}}{\text{SE}(\hat{\beta})}(\text{observed}\right| > t_{0.025,\ n-2}$ (table value) and will be accepted otherwise.

Here $t_{n-2} = \frac{\hat{\beta}}{\text{SE}(\hat{\beta})} = \frac{0.494}{0.094} = 5.255$ (where $n = 20$)

Thus we see that, $|t_{n-2}| = 5.255 > t_{0.025}$, 18 (=2.101) and hence $H_0 : \beta = 0$ is rejected (alternative $H_1 : \beta \neq 0$ is accepted) at 5% level of significance. So, the hypothesis of no relationship between X and Y is to be rejected at 5% level of significance. Similarly, it can be tested for 1% level of significance.

2.9 Summary

Economic relationships are mainly of two types as discussed in this chapter-deterministic relationships and stochastic relationships. Deterministic relationships are those relationships where for a given value of the independent variable there exists a definite value of the dependent variable whereas stochastic relationships are those where for each value of the independent variable there exists a probability distribution of the value of the dependent variable. The simplest form of stochastic relation between two economic variables is given by $Y_i = \alpha + \beta X_i + u_i$ where Y is the dependent variable, X is the independent variable, α and β are regression parameters, u is the disturbance term, i represents the number of observations and n is the sample size. This relationship is called Classical Linear Regression Model if it satisfies some assumptions like parameters are linearly related, probability distribution of the distrubance term is normal i.e., mean is zero and variance is constant, different error terms are independently distributed, disturbance term is independent with explanatory variables, explanatory variables are independent to each other, number of observations must be greater than the number of parameters to be estimated and the model is correctly specified. The most popular method of estimating the regression parameters of the CLRM is the method of least squares. Lastly, the properties of the OLS estimators state that the OLS estimators are the BLUE i.e., best, linear and unbiased estimators.

2.10 Exercise

Short Answer Type Questions :

1. State true or false :

(a) Deterministic relationship breaks down if the ceteris paribus assumption is relaxed.

(b) The probability distribution of the disturbance term is such that its mean is zero in the Classical Linear Regression Model.

2. Choose the correct alternative :

(a) The relationship between two variables say X and Y is said to be _____ if for each value of the independent variable X there exists a probability distribution of the values of the dependent variable Y.

(i) Deterministic

(ii) Stochastic

(iii) Non-stochastic

(iv) None of the above

(b) The OLS estimators of α and β in a two-variable CLRM satisfies the property of

(i) Linearity

(ii) Unbiasedness

(iii) Minimum variance

(iv) All of the above

3. Fill in the blanks :

(a) The OLS estimate of α in a 2-variable CLRM is given by _____.

(b) The OLS estimate of β in a 2-variable CLRM is given by _____.

4. What is Gauss-Markov theorem on least squares?

5. What are the properties of OLS estimates ?

Medium Answer Type Questions :

1. Show that the regression coefficient ($\hat{\beta}$) is linear.

2. Prove that the intercept estimate ($\hat{\alpha}$) is linear.

3. State the assumptions of classical linear regression model.

4. Write a short note on ordinary least square method.

Long Answer Type Questions :

1. State and explain the assumptions of a classical linear regression model.

2. Describe briefly the method of least squares for estimating the regression parameters in a two-variable CLRM.

3. State and prove the properties of the least square estimators in a two-variable CLRM.
4. State and prove the Gauss Markov Least Square Theorem with reference to CLRM.

2.11 References

1. Gujarati, D. (2003) : *Basic Econometrics*, McGraw Hill Higher Education.
2. Johnston, J. (1996) : *Econometric Methods*, McGraw Hill.
3. Madalla, G. S. (2005) : *Introduction to Econometrics*, John Wiley and Sons Ltd.
4. Koutsoyiannis, A. (1996) : *Theory of Econometrics*, ELBS with Macmillan.

Unit - 3 ❑ General Linear Model : K Variable CLRM

Structure

3.1 Objectives

Reading this chapter, students will get an idea about

- Estimation technique for more than two variables or GLM
- BLUE property of regression coefficient

- Log likelihood estimation
- Testing of hypothesis
- Prediction of regressand

3.2 Introduction

The two-variable model discussed in the previous chapter is somewhat inadequate for estimation. In the two-variable model we have taken only one explanatory variable. In case of more than one explanatory variable or independent variable this process will not work at all. The dependent variable which is in consideration (suppose quantity demanded) may depend on so many factors (like income of the consumer, own price of the commodity, price of related goods, etc.) but not on a single factor. So to involve all those independent variables in the model of estimation we have to extend the two-variable classical linear regression model. So in this chapter we are going to add more explanatory variables or regressors on which the explained variable or regressand depends. This type of estimation is called 'Multiple Regression Analysis' and the econometric model is called 'General Linear Model'. The simplest form of multiple-regression model is the three-variable regression model where there is one dependent variable and two independent variables. We should note one important thing that in this chapter we will analyze only linear regression model i.e., linear in parameters but, may or may not be linear in variables.

3.3 General Model Specification

Let us consider 'k' number of variables, where there are $(k - 1)$ explanatory variables like X_{2i}, X_{3i}.....X_{ki} and one explained variable Y_i. So the regression equation can be written as : $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + ...\beta_k X_{ki} + u_i$

Or, in matrix from, $Y = X\beta + u$ where, 'u' is the disturbance term, 'Y' is the explained variable, 'X' is the explanatory variable and 'β' is the parameter.

3.4 Assumptions

- Zero mean value for u, i.e., $E(u) = 0$

- $E(uu') = \sigma^2 I_n$, i.e., there is no problem of autocorrelation and heteroscedasticity in the model.
- Set X remains fixed for a given sample, i.e., X is non-stochastic.
- Rank of $X = \rho(X) = k < n$
- Here k denotes the number of parameters and n denotes the number of observation. So, number of observation must be greater than the number of parameters to be estimated. Otherwise parameters cannot be estimated using OLS. Further, $\rho(X) = k$, implies that the columns of X are independent to each other, i.e., there is no multicollinearity problem.
- There is no autocorrelation between two disturbance terms. So,

  $\text{cov}(u_i, u_j) = 0$ for all $i \neq j$

  or, $E(u_i, u_j) = 0$ for all $i \neq j$
- Zero covariance between X_i and u_i, i.e., X_i and u_i are independent to each other, or, $\text{cov}(X_i, u_i) = 0$
- There is no multicollinearity problem, or,

  $\text{cov}(X_i, X_j) = 0$
- Model is correctly specified. So, there is no specification error.

3.5 Estimation of Parameters

We can start with general regression equation in matrix form. Then we can specify the regression model depending on the number of variables.

3.5.1 Estimation of Parameters for General Regression Equation

The model used for estimation is $Y = X\beta + u$, where the order of Y matrix is $n*1$, order of X is $n*k$, order of β is $k*1$ and order of u is $n*1$.

$$E(Y) = E(X\beta) + E(u)$$

or, $\hat{Y} = X\hat{\beta}$

This is the estimated regression equation.

So the error term (e) = $Y - \hat{Y} = Y - X\hat{\beta}$

Applying OLS method we shall have to minimize the sum of square values of errors, So, $\sum_{i=1}^{n} e_i^2 = [e_1 e_2 \cdots e_n] \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} = e'e = (Y - X\hat{\beta})'(Y - X\hat{\beta})$

or, $e'e = Y'Y - 2\hat{\beta}'X'Y + \hat{\beta}'X'X\hat{\beta}$.

For OLS, $e'e$ is to be minimized with respect to $\hat{\beta}$.

Minimization requires $\frac{\partial(e'e)}{\partial\hat{\beta}} = -2X'Y + 2X'X\hat{\beta} = 0$

or, $X'X\hat{\beta} = X'Y$

or, $\hat{\beta} = (X'X)^{-1}X'Y$ .....(1)

This is the OLS estimator of β.

3.5.2 Estimation of Parameters for Three-variable Regression Equation

The simplest form of multiple regression equation (for more than two variables) is three variable regression equation which can be specified as follows :

$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$

or, $E(Y_i) = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i}$ (given the assumptuion of multiple regression model)

or, $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i}$ and $\overline{Y} = \hat{\beta}_1 + \hat{\beta}_2 \overline{X}_2 + \hat{\beta}_3 \overline{X}_3$

Subtracting we have, $\hat{y}_i = \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i}$ (variables are in deviational form)

So, error $e_i = Y_i - \hat{Y}_i = (Y_i - \overline{Y}) - (\hat{Y}_i - \overline{\hat{Y}}) = y_i - \hat{y}_i$

$= y_i - \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i}$

So, $\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i})^2$

For minimization or error term (applying OLS) $\sum_{i=1}^{n} e_i^2$ with respect $\hat{\beta}_2$ and $\hat{\beta}_3$ requires

$$\frac{\partial \sum_i e_i^2}{\partial \hat{\beta}_2} = 0 \text{ and } \frac{\partial \sum_i e_i^2}{\partial \hat{\beta}_3} = 0$$

Again, these two equations can be rewritten as

$\hat{\beta}_2 s_{22} + \hat{\beta}_3 s_{23} = s_{2y}$ and $\hat{\beta}_2 s_{23} + \hat{\beta}_3 s_{33} = s_{3y}$

Where, $s_{ii} = \sum_i x_i^2$, $i = 2, 3$

$s_{kl} = \sum_i x_{ki} . x_{li}$, $k \neq l = 2, 3$, $s_{jy} = \sum_i x_{ji} . y_i$

So, by applying Cramer's rule we have

$$\hat{\beta}_2 = \frac{\begin{vmatrix} s_{2y} & s_{23} \\ s_{3y} & s_{33} \end{vmatrix}}{\begin{vmatrix} s_{22} & s_{23} \\ s_{23} & s_{33} \end{vmatrix}} = \frac{s_{2y} s_{33} - s_{23} s_{3y}}{s_{22} s_{33} - s_{23}^2}$$

$$\hat{\beta}_3 = \frac{\begin{vmatrix} s_{22} & s_{2y} \\ s_{23} & s_{3y} \end{vmatrix}}{\begin{vmatrix} s_{22} & s_{23} \\ s_{23} & s_{33} \end{vmatrix}} = \frac{s_{22} s_{3y} - s_{2y} s_{23}}{s_{22} s_{33} - s_{23}^2}$$

and $\hat{\beta}_1 = \overline{Y} - \hat{\beta}_2 \overline{X}_2 - \hat{\beta}_3 \overline{X}_3$

3.6 BLUE Property of $\hat{\beta}$

Now we have to see whether $\hat{\beta}$ is Best, Linear, Unbaised Estimator (BLUE) or not. Here $\hat{\beta}$ is the OLS estimator of β.

From equation (1) we have, $\hat{\beta} = (X'X)^{-1} X'Y$ which is the linear function of Y.

From the above equation we can write

$\beta = (X'X)^{-1} X' (X\beta + u)$

$= (X'X)^{-1} (X'X)\beta + (X'X)^{-1} X'u$

or, $\hat{\beta} = \beta + (X'X)^{-1} X'u$

or, $E(\hat{\beta}) = \beta + (X'X)^{-1} (X'E(u)$

or, $E(\hat{\beta}) = \beta$

So, we can say that $\hat{\beta}$ is an unbiased estimator of β.

Now we have to calculate the variance of $\hat{\beta}$. Then we have to prove that variance of $\hat{\beta}$ is minimum than the variance of any other unbiased estimator of β.

Now, the variance matrix of $\hat{\beta}$ is denoted by

var $(\hat{\beta}) = E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)']$

We know that $\hat{\beta} = \beta + (X'X)^{-1}X'u$

or, $(\hat{\beta} - \beta) = (X'X)^{-1} X'u$

or, $(\hat{\beta} - \beta)' = u'X (X'X)^{-1}$

So, var$(\hat{\beta}) = E [(X'X)^{-1}X'u * u' X(X'X)^{-1}]$

$= (X'X)^{-1} X'E (u * u') X (X'X)^{-1} = \sigma^2 (X'X)^{-1} X'X (X'X)^{-1}$

or, var$(\hat{\beta}) = \sigma^2(X'X)^{-1}$

This is the variance of $\hat{\beta}$. Now we have to prove that this variance is minimum than the variance of any other unbiased estimator of β.

Let us assume that another linear unbiased estimator of β is b, where, $b = a'Y$ for any vector a. So b is a linear function of Y. Further $b = a'Y = a'(X\beta + u)$

So, $b = a'X\beta + a'u$

or, $E(b) = a'X\beta + a'E(u)$

or, $E(b) = a'X\beta$

Now if $a'X = 1$, we can write $E(b) = \beta$

Therefore b will be unbaised if and only if $a'X = 1$.

Now variance of b can be rewritten as

$\text{var}(b) = E[(b - \beta)(b - \beta)']$

$= E[(a'u)(u'a)] = a'E[uu']a = \sigma^2 a'a$

Now, $\text{var}(b) - \text{var}(\hat{\beta})$

$= \sigma^2 a'a - \sigma^2(X'X)^{-1}$

$= \sigma^2 a'a - \sigma^2 a'X(X'X)^{-1}X'a$

$= \sigma^2 a'[1 - X(X'X)^{-1}X']a$

$= \sigma^2 a'Ma$ where $M = 1 - X(X'X)^{-1}X'$

'M' matrix is a symmetrical, idempotent and a positive semi-definite matrix.

As 'M' is a positive semi-definite matrix, $a'Ma \geq 0$

So, $\text{var}(b) - \text{var}(\hat{\beta}) \geq 0$

or, $\text{var}(b) \geq \text{var}(\hat{\beta})$

So, $\hat{\beta}$ is BLUE.

3.7 Estimation of σ^2

We want to estimate σ^2 in GLM, so that this estimator becomes unbiased.

Let $\hat{\sigma}^2$ be the unbiased estimator of σ^2 in GLM, i.e. $E(\hat{\sigma}^2) = \sigma^2$

Now by definition $e = Y - \hat{Y} = Y - X\hat{\beta}$

or, $e = X\beta + u - X(X'X)^{-1}X'(X\beta + u)$ [putting the value of Y and β]

or, $e = X\beta + u - X(X'X)^{-1}X'X\beta - X(X'X)^{-1}X'u$

or, $e = X\beta + u - X\beta - X(X'X)^{-1}X'u$

or, $e = u - X(X'X)^{-1}X'u$

or, $e = [1 - X(X'X)^{-1}X']u$

or, $e = Mu$ where we have $M = I - X(X'X)^{-1}X'$

And, $\sum_i e_i^2 = e'e = (Mu)'(Mu) = u'M'Mu = u'Mu$ [as M is idempotent matrix $M'M = M$]

So, $\sum_i e_i^2 = u'[I - X(X'X)^{-1}X']u$

So, $E(e'e) = E(u'u)\ tr\ [I - X(X'X)^{-1}X']$

$= \sigma^2\ [tr(I_n) - tr\ \{X(X'X)^{-1}X'\}]$

$= \sigma^2\ [n - tr\ \{(X'X)^{-1}X'X\}]$

$= \sigma^2\ [n - tr\ (I_n)]$

$= \sigma^2\ [n - k]$

or, $\sigma^2 = \dfrac{E(e'e)}{n-k}$

or, $\sigma^2 = E\left(\dfrac{e'e}{n-k}\right) = E\left[\dfrac{\sum e_i^2}{n-k}\right]$

or, $\sigma^2 = E(\hat{\sigma}^2) = E\left[\dfrac{\sum e_i^2}{n-k}\right]$

This is the unbiased estimator of σ^2.

3.8 Maximum Likelihood Estimator

For maximum likelihood estimator (MLE) we additionally assume that, u independently follows the normal distribution with zero mean and σ^2 variance. The

pdf of u_i is given as $f(u_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{u_i^2}{2\sigma^2}}$

The joint pdf of $u_1, u_2, u_3, \ldots\ldots u_n$ is given as $f(u_1, u_2, u_3, \ldots\ldots u_n) =$

$f(u_1) . f(u_2) \ldots\ldots\ldots\ldots f(u_n) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\frac{\sum u_i^2}{2\sigma^2}}$

Therefore, the likelihood function is $L = (2\pi\sigma^2)^{-n}/2.e^{-\frac{u'u}{2\sigma^2}}$

Where, $u = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$

We have defined our model as $Y = X\beta + u$

or, $u = Y - X\beta$

So, $u'u = (Y' - \beta'X')(Y - X\beta)$

or, $u'u = Y'Y - 2\beta'X'Y + \beta'X'X\beta$

So the likelihood function reduces to

$$L = (2\pi\sigma^2)^{-n}/2.e^{-\frac{(YY-2\beta'XY+\beta'XX\beta)}{2\sigma^2}}$$

This has to be minimized with respect to β and σ^2 where MLE estimate of β is similar to the case of OLS method. So the MLE of β can be written as

$\tilde{\beta} = \hat{\beta} = (X'X)^{-1}X'Y$

So, $\tilde{\beta}$ is BLUE.

However MLE estimate of σ^2 will be different which is derived below :

$$ln(L) = -\frac{n}{2}ln(2\pi) - \frac{n}{2}ln(\sigma^2) - \frac{1}{2\sigma^2}\left(\sum e_i^2\right)$$

Differentiating with respect to σ^2 we have $\frac{\partial ln(L)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum e_i^2 = 0$

So, $\tilde{\sigma}^2 = \frac{\sum e_i^2}{n}$

But this estimator is not unbaised whereas OLS estimator of σ^2 is unbiased.

3.9 Testing of Hypothesis

3.9.1. Point Estimation

In GLM $E(\hat{\beta}) = \beta$ and $\text{var}(\hat{\beta}) = \sigma^2(X'X)^{-1}$

So, for any particular β, say β_j, $E(\hat{\beta}_j) = \beta_j$ and var $(\hat{\beta}_j) = \hat{\sigma}^2 a_{jj}$ where a_{jj} is the j-th diagonal element of $(X'X)^{-1}$.

So, the standard error of $\hat{\beta}_j = \text{SE}(\hat{\beta}_j) = \hat{\sigma}\sqrt{a_{jj}}$, where, $\hat{\sigma}^2 = \frac{\sum e_i^2}{n-k} = \frac{(e'e)}{n-k}$.

Therefore the test statistic is $\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}\sqrt{a_{jj}}} \sim t_{n-k}$

Here the null hypothesis is $H_0 : \beta_j = 0$ aganist the alternative $H_1 : \beta_j \neq 0$.

Under H_0 the test statistic is : $t_0 = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{a_{jj}}}$

If $|t_0| > t_{\frac{\alpha}{2},n-k}$ then H_0 is rejected i.e., β_j is statistically significant or significantly different from zero, otherwise H_0 is accepted.

3.9.2 Interval Estimation

Interval estimation of β_j is derived as follows :

$$P\left[-t_{\frac{\alpha}{2},n-k} \le t \le t_{\frac{\alpha}{2},n-k}\right] = 1-\alpha$$

where, α is the level of significance.

or, $P\left[-t_{\frac{\alpha}{2},n-k} \le \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}\sqrt{a_{jj}}} \le t_{\frac{\alpha}{2},n-k}\right] = 1-\propto$

or, $P\left[\left(\hat{\beta}_j - t_{\frac{\alpha}{2},n-k}\hat{\sigma}\sqrt{a_{jj}}\right) \le \beta_j \le \left(t_{\frac{\alpha}{2},n-k}\hat{\sigma}\sqrt{a_{jj}} + \hat{\beta}_j\right)\right] = 1-\propto$

So, $\hat{\beta}_j \pm t_{\frac{\alpha}{2},n-k}\hat{\sigma}\sqrt{a_{jj}}$ is the interval estimation of β_j

3.10 Prediction in GLM

Now we want to predict the mean value or expected value of Y for given value of X, say $C' = \{1, X_{2(n+1)}, X_{3,(n+1)}, \ldots\ldots X_{k,(n+1)}\}$, for $i = n + 1$

So, $E(Y) = X\beta$

Here for given values of X_j $E(Y) = c'\beta$.

We know that if $\hat{\beta}$ is BLUE of β then $c'\hat{\beta}$ will also be the BLUE of $c'\beta$. Therefore $c'\hat{\beta}$ is the best linear unbiased predictor of $c'\beta$ i.e., the predictor is $\hat{Y} = c'\hat{\beta}$ which is the best linear unbaised point predictor with variance $\sigma^2 c'(X'X)^{-1}c$.

We can also derive the interval predictor as follows :

We have $\hat{\beta} \sim N[\beta, \sigma^2(X'X)^{-1}]$

or, $c'\hat{\beta} \sim N[c'\beta, \sigma^2 c'(X'X)^{-1}c]$

So, we can write $\frac{c'\hat{\beta} - c'\beta}{\sigma\sqrt{c'(X'X)^{-1}c}} \sim N(0,1)$

So, $\frac{c'\hat{\beta}-c'\beta}{s\sqrt{c'(X'X)^{-1}c}} \sim t_{n-k}$

Where $s^2 = \hat{\sigma}^2 = \frac{\sum e_i^2}{n-k} = \frac{e'e}{n-k}$

So, $P\left[-t_{\frac{\alpha}{2},n-k} \leq \frac{c'\hat{\beta}-c'\beta}{s\sqrt{c'(X'X)^{-1}c}} \leq t_{\frac{\alpha}{2},n-k}\right] = 1-\propto$

Therefore the interval predictor is :

$$c'\hat{\beta} \pm t_{\frac{\alpha}{2},n-k} s\sqrt{c'(X'X)^{-1}c}$$

3.11 Summary

In a two-variable regression model we can take one explanatory variable. So for more than one explanatory variable this process will not work. For the estimation of regression model involving more than one explanatory variable GLM is introduced. This type of estimation is called multiple regression analysis and the econometric model is called general linear model (GLM). Here we have estimated the general model and also the simplest form of multiple-regression model i.e., the three-variable regression model with one dependent variable and two independent variables. These models are linear in parameters. Given the assumptions of the model (which are similar to the assumptions of CLRM) we have estimated the parameters following OLS method. We have also estimated the regression parameters following the technique of MLE (maximum likelihood estimator). We have estimated the variance in GLM and derived the unbiased estimator of the variance. Comparing the OLS technique and the MLE technique we have found that regression parameter is the Best Linear Unbaised Estimator in both the processes. But the MLE estimate of variance is different from the OLS estimate. We have found that MLE estimator of variance is not unbaised whereas the OLS estimator of variance is unbiased. Lastly, we have predicted the mean value of the dependent variable (Y), given the value of explanatory variable (X) in GLM.

3.12 Exercise

Short Answer Type Questions :

(a) Choose the correct answer :

(i) In GLM variance of the disturbance term is considered as

(1) Zero

(2) One

(3) Not equal to zero

(4) Greater than one

(ii) $\hat{\beta}$ will be the unbiased estimator of β if

(1) $E(\hat{\beta}) = \beta$

(2) $E(\hat{\beta}) < \beta$

(3) $E(\hat{\beta}) > \beta$

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

(b) Fill in the blanks :

(i) Number of observations must be _____ than the number of parameters to be estimated.

(ii) If $|t_0| > t_{\frac{\alpha}{2},n-k}$ then β_j, the regression coefficient is _____.

(iii) Statistically _____ and the null hypothesis is rejected.

(c) Identify whether the statements are true or false :

(a) If $\text{cov}(X_i, X_j) = 0$ then the multicollinearity problem is present.

(b) Estimation in GLM is done by applying OLS method.

(d) What is multiple regression analysis?

(e) What is general linear model?

(f) Specify a general linear regression model.

(g) What is meant by BLUE of an estimate?

Medium Answer Type Questions :

1. State the assumptions of GLM.
2. What is point estimation?
3. What is interval estimation?

Long Answer Type Questions :

(a) What are the assumptions under GLM?

(b) Consider a model = $X\beta + u$, where the order of Y matrix is $n*1$, order of X is $n*k$, order of b is $k*1$ and order of u is $n*1$. The estimated regression equation is $\hat{Y} = X\hat{\beta}$. Prove that $\hat{\beta}$ is BLUE.

(c) Consider a three-variable regression equation $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$. Estimate the parameters.

(d) How can you predict the expected value of dependent variable, given the value of explanatory variable in GLM?

3.13 References

1. Maddala, G. S. (2005) : *Introduction to Econometrics*, John Wiley and Sons Ltd.
2. Johnston, J. (1996) : *Econometric Methods*, McGraw Hill.
3. Kmenta, J. (1991) : *Elements of Econometrics*, Macmillan Publishing Company.

Unit - 4 ❑ Violating the Assumptions of the CLRM : I-Multicollinearity

Structure

4.1 Objectives

Reading this chapter, students will get an idea about

- Meaning of Multicollinearity
- Sources of Multicollinearity
- Consequences of Multicollinearity
- Detection of Multicollinearity
- Solution of Multicollinearity

4.2 Introduction

Multicollinearity problem was first identified by Ragnar Frisch. Collinearity means linear relationship among explanatory variables and multicollinearity means a number of linear relations among explanatory variables. But broadly in econometrics, multicollinearity implies either single or multiple linear relations among explanatory

variables. In classical linear regression model, it is assumed that explanatory variables are independent to each other. But in reality if they are found to be linearly dependent to each other, we get this multicollinearity problem. Let us assume that the explanatory variables are x_1, x_2, ...x_k. If $\delta_1 x_1 + \delta_2 x_2 + \ldots\ldots\ldots + \delta_k x_k = 0$ for any $\delta_j \neq 0$ we can say that there is a multicollinearity problem. Let us assume that $\delta_2 \neq 0$ then we can write, $x_2 = -\frac{\delta_1}{\delta_2} x_1 - \ldots\ldots\ldots - \frac{\delta_k}{\delta_2} x_k$ i.e., x_2 is linearly dependent on other explanatory variables. This dependence is prefect. So there exists a multicollinearity problem. We further assume $\in$ to be any random variable for which we get the following relationship : $x_2 = -\frac{\delta_1}{\delta_2} x_1 - \ldots\ldots\ldots\ldots\ldots - \frac{\delta_k}{\delta_2} x_k - \frac{\in}{\delta_2}$

where also x_2 is linearly dependent on other explanatory variables. This dependence is however not perfect. Yet there still exists multicollinearity. Therefore it can be inferred that multicollinearity implies either perfect or near perfect relationship among explanatory variables.

4.3 Sources of Multicollinearity

Multicollinearity may arise due to a number of reasons which are noted as follows :

(i) Model Specification : If any explanatory variable is included in different polynomial forms in the regression model, then multicollinearity problem may arise. For example if we consider the model $Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{2t}^2 + \in_t$ then X_{2t} and X_{2t}^2 may be correlated.

(ii) Overdependence : If the number of explanatory variables is relatively more than the number of observations, then the model is known as overdependent and there we get the problem of multicollinearity. For instance, medical research data may be collected on a number of parameters from a limited number of patients and there we may get the problem of multicollinearity.

(iii) Data collection : If data are collected from a limited range of samples, then multicollinearity problem is very likely to be observed.

(iv) Parametric constraint : In some studies, variables are related by definition

i.e., there exists constraints amongst variables and that leads to multicollinearity problem. For example, if we consider a regression analysis where we are estimating electricity bill on income level and expenditure in residential house, then income and expenditure in residential house are positively related on the presumption that rich people reside in bigger houses.

4.4 Consequences of Multicollinearity

In case of perfect dependence of explanatory variables i.e., if the multicollinearity problem is perfect, we cannot compute $(X'X)^{-1}$ and hence we cannot estimate $\hat{\beta}$. So, the regression parameters will remain indeterminate. However, if multicollinearity problem is not perfect, but the explanatory variables are strongly correlated, then the following consequences may arise :

- The precision level will fall. The effect of one independent variable is entangled with the effect of other independent variables. The variance of the estimator and the co variance of the estimators will increase with the increase in degree of dependence of explanatory variables.
- If there is multicollinearity problem, then the estimates will be very sensitive to the size of the sample. By increasing observation or by deleting one observation one can change the magnitude and sign of the estimates.
- The estimate of sample variance would be affected in the presence of multicollinearity i.e., the inferences would be erroneous.

The main consequence of multicollinearity problem is that the precision level will fall i.e., errors in estimation will increase due to increase in collinearity of independent variables. The co-variance of the errors of estimates will also increase here. We can establish this with the help of an example :

Suppose we consider a model $Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} = u_t$

In deviational form the model is written as

$$y_t = \beta_2 x_{2t} + \beta_3 x_{3t} + u_t - \bar{u}$$

Here we assume that x_{2t} and x_{3t} are correlated as follows :

$$x_{3t} = \alpha x_{2t} + \upsilon t$$

To establish α as their correlation coefficient, we assume that $\sum x_{3t}^2 = \sum x_{2t}^2 = 1$ and $\sum \upsilon_t = 0, \sum \upsilon_t x_{2t} = 0$

The correlation coefficient between x_{2t} and x_{3t} is given by

$$\frac{\text{Covariance}(x_{2t}, x_{3t})}{\sqrt{\text{var}(x_{2t})\,\text{var}(x_{3t})}}$$

$$= \frac{\sum x_{2t}x_{3t}}{\sqrt{\sum x_{2t}^2 \sum x_{3t}^2}}$$

$$= \frac{\sum x_{2t}x_{3t}}{1}$$

$$= \sum x_{2t}x_{3t}$$

$$= \sum x_{2t}(\alpha x_{2t} + \upsilon_t)$$

$$= \alpha \sum x_{2t}^2 + \sum x_{2t}\upsilon_t$$

$$= \alpha.\ 1 + 0$$

$$= \alpha$$

So, $(X'X) = \begin{bmatrix} \sum x_{2t}^2 & \sum x_{2t}x_{3t} \\ \sum x_{2t}x_{3t} & \sum x_{3t}^2 \end{bmatrix} = \begin{bmatrix} 1 & \alpha \\ \alpha & 1 \end{bmatrix}$

And

$$(X'X)^{-1} = \frac{\text{Adj}(X'X)}{|X'X|} = \frac{1}{1-\alpha^2}\begin{bmatrix} 1 & -\alpha \\ -\alpha & 1 \end{bmatrix}$$

The variance-covariance matrix is

$$\text{Var } (\hat{\beta}) = \sigma^2 (XX)^{-1} = \sigma^2 \frac{1}{1-\alpha^2} \begin{bmatrix} 1 & -\alpha \\ -\alpha & 1 \end{bmatrix}$$

$$\text{So, } \text{Var}(\hat{\beta}_2) = \sigma^2 \frac{1}{1-\alpha^2} = \text{Var}(\hat{\beta}_3)$$

As α increases. i.e., as degree of multicollinearity increases, variance of $\hat{\beta}_2$ and variance of $\hat{\beta}_3$ also increase i.e., the error of estimate will increase. Hence the precision level will fall due to increase in degree of multicollinearity. Similarly, covariance will also increase due to increase in degree of multicollinearity.

4.5 Detection of Multicollinearity

It is to be noted that multicollinearity is a problem of sample (not of population) and its degree is the main concern, not its existence. It is also to be noted that there is no single method for detection of Multicollinearity. Rather, there are few popular methods used by econometricians for the detection of multicollinearity. We will discuss these methods one by one in this section as follows :

- **Value of R^2 :** If it is found that the overall R^2 is very high and significant, but individual regression parameters are mostly insignificant then that gives us a signal of the existence of high multicollinearity problem where individual effects of explanatory variables cannot be disentangled.
- **Value of zero order correlation coefficient :** Sometimes simple product-moment or zero-order correlation coefficient is used for detecting multicollinearity. If the zero-order correlation coefficient is found to be high and significant then that signifies the existence of multicollinearity problem. Here product-moment correlation coefficients are calculated for every pair of independent variables. But this method is applicable only when there are two explanatory variables. Further, this method is sufficient but not necessary for detecting multicollinearity problem.
- **Partial correlation coefficient :** Farrar and Glauber argued that if there are more than two independent variables, then partial correlation coefficient should be used for detecting multicollinearity problem. Let us assume that

there are four independent variables, x_1, x_2, x_3 and x_4. The partial correlation coefficients are $r_{12.34}$, $r_{13.24}$, $r_{14.23}$, and so on. If any of these partial correlation coefficients is found to be significant, it can be said that there is multicollinearity problem. But this technique has also been criticised by econometricians because often it fails to detect multicollinearity in reality.

- **Auxillary regression method :** It is more powerful to detect multicollinearity compared to the correlation methods. Here auxillary regressions are fitted taking all explanatory variables into account. More specifically, in auxillary regression, one explanatory variable is regress and on remaining explanatory variables at a time and its coefficient of determination is calculated. Let the Auxillary Regression be X, $f(X_1, X_2, X_3, \ldots\ldots X_{i-1}, X_{i+1}, \ldots\ldots X_k)$ and its coefficient of determination is R_i^2. If R_i^2 is found to be significant and high we can say that there is Multicollinearity. Klien prescribed a rule of thumb here : If any $R_i^2 > R^2$ i.e., overall coefficient of determination obtained from regression of y on x_i's, it can be said that there is multicollinearity problem.

- **Variance Inflation Factor test :** Another popularly used test to detect multicollinearity is Variance Inflation Factor test or VIF test. In this method,

  $(VIF)_i = \dfrac{1}{(1-R_i^2)}$ where R_i^2 is the coefficient of determination obtained from regressing X_i on remaining explanatory variables. Here also rule of thumb says that if any VIF is more than 10 we can say that there is multicollinearity problem.

- **Condition number test (CN test) :** In the case of VIF we get multiple tests for multiple explanatory variables. But to get a single test, Condition number test can be applied. So, CN test is more preferable to VIF test and that's why it is used widely. For the matrix, $(X'X)$, eigen values are computed. Let δ_1 be the highest eigen value and δ_2 be the lowest engien value, then condition number = $\dfrac{\delta_1}{\delta_2}$ and CN index = $\sqrt{\dfrac{\delta_1}{\delta_2}}$

Here, the rule of thumb is that if *CN* index lies between 10 and 30 we can say that there is multicollinearity and if *CN* index > 30 we can say that there is severe multicollinearity.

4.6 Solution of Multicollinearity Problem

The solution techniques applied for solving the problem of multicollinearity can be broadly classified into two categories :

- Preliminary techniques
- Specialised solutions

Preliminary techniques of solving the multicollinearity problem are started as follows :

- Dropping of variables
- Transformation of variables
- Change of observations

Under specialised solutions, we get mainly two techniques for solving multicollinearity problem. They are :

- Ridge regression technique
- Principal Component Analysis

These techniques of solving Multicollinearity problem are discussed as follows :

- **Dropping of variables :** If the explanatory variables are found to be correlated among themselves, then the variable which is causing multicollinearity and which is not very much relevant, that variable can be dropped from the regression analysis to solve the multicollinearity problem. For instance, in estimating consumption function, two independent variables are income and wealth which are correlated. To solve the multicollinearity problem we can drop wealth from the regression analysis considering it less essential than income.
- **Transformation of variables :** In time series data, trend component may cause multicollinearity among all the variables. If this trend component is eliminated, then the multicollinearity problem can be avoided. For that,

the variables may be transformed in their difference forms (not in the level form) in the regression analysis Suppose the regression is given by $Y_t = a + bt$. So, $Y_{t-1} = a + b(t-1)$. Subtracting the second equation from the first we get, $\Delta Y_t = b$ i.e., here trend element is eliminated and original variables are taken.

- **Change of observations :** Multicollinearity is a sample problem and not a problem of population. So, it can be avoided by changing the size of the sample i.e., by increasing the number of observations. For that the size of the sample can be increased (if possible), cross section data can be combined with time series data i.e., by using pooled data the problem of multicollinearity can be solved. For example, from budget study one can establish relationship between wealth and consumption and to determine relationship between income and consumption, time series data can be used.
- **Ridge regression analysis :** To avoid multicollinearity i.e., for getting lower values of correlation coefficient some constant say λ can be added to variances and thus the multicollinearity problem can be solved from regression analysis. This is a mechanical procedure which has been rationalised by econometricians considering different variants of ridge regression. One such variant is ridge regression with prior information. Let the model be

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

where X_1 and X_2 are found to be highly correlated.

We know that $\beta_1^2 + \beta_2^2 = c$ (any constant).

Here for OLS we have to minimise

$$L = \sum e_i^2 + \lambda[c - \beta_1^2 - \beta_2^2]$$

$$= \sum \left(Y_i - \hat{\alpha} - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i}\right)^2 + \lambda\left[c - \hat{\beta}_1^2 - \hat{\beta}_2^2\right]$$

$$= \sum_i \left(y_i - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i}\right)^2 + \lambda\left[c - \hat{\beta}_1^2 - \hat{\beta}_2^2\right] \text{ [in deviational form]}$$

It is to be minimised with respect to $\hat{\beta}_1$, $\hat{\beta}_2$ and λ.

For minimisation,

$$\frac{\delta L}{\delta \hat{\beta}_1} = -2\sum\left(y_i - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i}\right)x_{1i} - 2\lambda\hat{\beta}_1 = 0$$

$$\text{or, } S_{y1} = (S_{11} + \lambda)\hat{\beta}_1 + S_{12}\hat{\beta}_2 \quad \ldots\ldots\ldots\ldots\ldots\ldots(1)$$

Similarly,

$$\frac{\delta L}{\delta \hat{\beta}_2} = -2\sum\left(y_i - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i}\right)x_{2i} - 2\lambda\hat{\beta}_2 = 0$$

$$\text{or, } S_{y2} = \hat{\beta}_1(S_{12}) + \hat{\beta}_2(S_{12} + \lambda) \quad \ldots\ldots\ldots\ldots\ldots\ldots(2)$$

Solving (1) and (2), we get estimated values of parameters from ridge regression. Here, λ can be solved from the constraint. In ridge regression, estimators may be biased but they have lowered mean square errors.

- **Principal component analysis :** It is the most widely used method for solving the problem of multicollinearity. We can explain the principal component analysis using a suitable example. Let the model be $Y_i = \alpha + \beta_1 X_{1i} + \beta_{2i} X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + u_i$

  Among the explanatory variables, suppose X_1, X_2, X_3 are correlated among themselves. Those correlated variables are to be clubbed to get principal component of those variables. For clubbing, zero order correlation coefficients are to be computed of X_1, X_2, X_3 and to be arranged in the following matrix form :

X	X_1	X_2	X_3	Row totals
X_1	$r_{11} = 1$	r_{12}	r_{13}	$\sum_j r_{1j}$
X_2	r_{21}	$r_{22} = 1$	r_{23}	$\sum_j r_{2j}$
X_3	r_{31}	r_{32}	$r_{33} = 1$	$\sum_j r_{3j}$
Column totals	$\sum_i r_{i1}$	$\sum_i r_{i2}$	$\sum_i r_{i3}$	$\sum_i \sum_j r_{ij}$

Next, row totals and grand totals are to be computed and the following formulae are to be applied.

$$a_1 = \frac{\sum r_{1j}}{\sum_i \sum_j r_{ij}}$$

$$a_2 = \frac{\sum r_{2j}}{\sum_i \sum_j r_{ij}}$$

$$a_3 = \frac{\sum r_{3j}}{\sum_i \sum_j r_{ij}}$$

And $p_{1i} = a_1x_{1i} + a_2x_{2i} + a_3x_{3i}$ is the first principal component derived from assimilation of 3 variables x_1, x_2 and x_3 in their standardisded forms. Likewise, other principal components are calculated. For instance, for second principal components, new correlation matrix is formed where new correlation is equal to (old correlation — product of row and column totals) i.e., $r_{11}^* = r_{11} - (r_{ij} * r_{i1})$. In a similar manner the second principal component is calculated i.e., p_{2i}. It is to be noted that the number of principal components is equal to the number of variables clubbed and those principal components are selected for regression whose latent root or eigen value is greater than one where eigen value = $\frac{a_i^2 + a_2^2 + a_3^2}{3}$. Principal component analysis is applied only when variables are homogeneous and they carry economic meaning.

4.7 Summary

If the assumption of independence of explanatory variables in the Classical Linear Regression Model is violated, then the problem of multicollinearity arises. This problem was first identified by Ragnar Frisch. It may arise in econometric models from improper model specification, limited sample size or if there is any

parametric constraint. The presence of multicollinearity leads to a fall in precision level, standard errors of estimate will be large and inferences will be erroneous and the estimates will be very sensitive to the size of the sample. It can be detected by a very high value of R^2 but none of the regression coefficients being significant or by high value of zero order correlation coefficient. It is also tested by auxillary regression method, variance inflation factor test or by condition number test. The problem of multicollinearity can be solved by either dropping or transformation of some variables, or there are some sophisticated techniques like ridge regression analysis or principal component analysis which are used to solve the problem of multicollinearity.

4.8 Exercise

Short Answer Type Questions :

(1) Choose the correct answer :

(a) Which of the following is a test to detect multicollinearity?

(i) CN test

(ii) Ridge regression analysis

(iii) Principal Component Analysis

(iv) Durbin-Watson test

(b) A sure way of removing multicollinearity from the model is

(i) Work with panel data

(ii) Drop variables that cause multicollinearity in the first place

(iii) Transform the variables by first differencing them

(iv) Obtaining additional sample data

(2) Identify whether the statements are True or False

(a) Multicollinearity is essentially a sample phenomenon

(b) The precision level will fall due to the presence of multicollinearity

(3) Fill in the blanks

(a) In a regression model $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$, F test is seen to

satistically significant at less than 5 percent level of singinifcance, but the coefficients are seen to be statistically insignificant. This means that the variables are _____

(b) _____ is the most widely used technique to solve the problem of multicollinearity.

Medium Answer Type Questions :

1. Mention the main sources of multicollinearity problem.
2. What are the major consequences of multicollinearity in a linear regression model?
3. Who did first identify the problem of multicollinearity?
4. What is multicollinearity?

Long Answer Type Questions :

1. What do you mean by the problem of multicollinearity ? Discuss the sources from which multicollinearity may arise and explain the various consequences of multicollinearity.
2. How can the problem of multicollinearity be detected ? Discuss different ways to solve the problem of multicollinearity.

4.9 References

1. Gujarati, D (2003) : *Basic Econometrics*, McGraw Hill Higher Education
2. Johnston, J. (1996) : *Econometric Methods*, McGraw Hill.
3. Maddala, G. S. (2005) : *Introduction to Econometrics*, John Wiley and Sons Ltd.
4. Kmenta, J. (1991) : *Elements of Econometrics*, Macmillan Publishing Company.

Unit - 5 ❑ Violating the Assumptions of the CLRM : II-Heteroscedasticity

Structure

5.1 Objectives

Reading this chapter, students will get an idea about

- Meaning of heteroscedasticity
- Source of heteroscedasticity
- Consequences of heteroscedasticity
- Detection of heteroscedasticity
- Remedial measures of heteroscedasticity

5.2 Introduction

If we drop the assumption of constant variance of the disturbance term, i.e., homoscedasticity in the population regression function from the CLRM then the variance of the disturbance term becomes heteroscedastic. In this chapter we will examine the validity of this assumption and also we will find out what happens if the assumption of homoscedasticity is not fulfilled.

5.3 Meaning of Heteroscedasticity

Homoscedasticity means equal spread. When variances of the disturbance term 'u' are equal then it is known as homoscedasticity. But when variances of the disturbance term 'u' are different then it is known as heteroscedasticity. Therefore homoscedasticity means var $(u_i) = \sigma^2$ but heteroscedasticity means var $(u_i) = \sigma_i^2$ which is different of different i. It is postulated that σ_i^2 is a function of X_i, i.e., $\sigma_i^2 = f(X_i)$. As X_i changes, σ_i^2 also changes.

5.4 Sources of Heteroscedasticity

Heteroscedasticity arises due to different reasons which are mentioned below :

- **Error learning model :** With the process of acquiring experience or the error learning process the workers will come nearer to efficiency and their incorrectness will be less and less. So the variance will be lower. For example, typing error will be less as the typist acquires more experience.
- **Improvement in data processing technique :** With the introduction of computer and other sophisticated electronic devices, data processing technique, has been improved enormously. With the improvement in data processing technique we observed less error in data compilation and data correction process. That leads to lowering the variance in disturbance term. For example, heteroscedasticity is observed in studies related to banking data where data processing technique is improving day-by-day.

- **Presence of outlier in data :** An outlier is an extreme observation whose probability distribution is different from the probability distribution of other observations. The outlier may occur due to drought, flood, war, famine, etc. Due to the presence of outlier the variance of the disturbance term will increase and that leads to heteroscedasticity problem.
- **Existence of discretionary income :** Discretionary income is that part of income which can be used by the individuals as per their desire. This income is alternatively known as transitory income or the income from 'windfall gain'. Due to the existence of this discretionary income, variation in the human behavior is observed and that leads to the heteroscedasticity problem in the model.
- **Mis-specification error :** In the specified regression model, all the relevant variables may not be included. Some important variables may remain omitted whose influence would be reflected in the disturbance term causing the heteroscedasticity problem. For example, in a demand function, prices of other goods or other variables may not be inserted and these omitted variables will create more variance in the disturbance term.
- **Incorrect data transformation or functional form :** Let us suppose that variation of the data will be lessened if the data are considered either in ratio scale or in first difference form. Similarly, instead of original variables if we estimate the log function, the variation in the data can be minimized. Therefore, incorrect data transformation or the erroneous regression function selection would lead to heteroscedasticity in the disturbance term.
- **Skewed distribution in data :** In Economics there are number of variables like income, consumption, saving, etc. whose distribution is skewed or asymmetric in the society. Due to skewness in the distribution, every section of the society cannot enjoy same type of freedom in their choices. Naturally, for some people we get less information and for others we get more variation in the resultant data and that leads to heteroscedasticity in the problem.

5.5 Consequences of Heteroscedasticity

Following are the consequences of heteroscedasticity :

- In the presence of heteroscedasticity, the OLS estimator of regression parameter is still unbaised (consistent) but inefficient.
- The OLS estimate of the variance of the estimator of regression parameter is biased. Consequently, the usual t, χ^2 and F tests cannot be efficiently applied. We also face practical difficulty in applying hypothesis test under changing variance of disturbance term.

Now we will prove the above mentioned consequences of heteroscedasticity.

Let us consider a model $Y_i = \beta X_i + u_i$ .....(1)

where $E(u_i) = 0$, $E(u_i, u_j) = 0$ and $E(u_i^2) = \sigma_i^2$ i.e., there is heteroscedasticity.

Applying OLS we have the estimator of β

$$= \hat{\beta} = \frac{\sum X_i Y_i}{\sum X_i^2} = \frac{\sum X_i (\beta X_i + u_i)}{\sum X_i^2} = \beta + \frac{\sum X_i u_i}{\sum X_i^2}$$

or, $E(\hat{\beta}) = \beta + \frac{\sum X_i E(u_i)}{\sum X_i^2}$ $= \beta$ as $E(u_i) = 0$

So OLS estimator is unbiased here.

Now the variance of $\hat{\beta} = \text{var}(\hat{\beta}) = E(\hat{\beta} - \beta)^2 = E\left[\frac{\sum X_i u_i}{\sum X_i^2}\right]^2 = \frac{\sum X_i \sigma_i^2}{\left(\sum X_i^2\right)^2}$

Now we have to prove that this variance is not least, i.e., $\hat{\beta}$ is not efficient.

For this we now apply WLS (weighted least squares) estimation technique.

We assume that $\sigma_i^2 = \sigma^2 Z_i^2$ where Z is any unknown variable. Now dividing each component of the model by Z_i we have $\frac{Y_i}{Z_i} = \beta \frac{X_i}{Z_i} + \frac{\mu_i}{Z_i} = \beta \frac{X_i}{Z_i} + v_i$

To get WLS estimator, we have to apply OLS on the above equation.

The WLS estimator of β is $\tilde{\beta} = \dfrac{\sum \frac{X_i}{Z_i} \cdot \frac{Y_i}{Z_i}}{\sum \left(\frac{X_i}{Z_i}\right)^2}$

or, $(\tilde{\beta})$ = β, i.e., WLS estimator is unbiased.

But variance of $\tilde{\beta} = \text{var}(\tilde{\beta}) = E(\tilde{\beta} - \beta)^2 = E\left[\dfrac{\sum \frac{X_i}{Z_i} \cdot \frac{u_i}{Z_i}}{\sum \left(\frac{X_i}{Z_i}\right)^2}\right]^2 = \dfrac{\sigma^2}{\sum \left(\frac{X_i}{Z_i}\right)^2}$

Now, $\dfrac{\text{var}(\tilde{\beta})}{\text{var}(\hat{\beta})} = \dfrac{\sum (a_i b_i)^2}{\sum a_i^2 \sum b_i^2} \leq 1$

or, $\text{var}(\tilde{\beta}) \leq \text{var}(\hat{\beta})$ where $\dfrac{X_i}{Z_i} = a_i$, $X_i Z_i = b_i$

So, $\hat{\beta}$ is inefficient.

Now we want to prove the second consequence i.e., estimated variance of $\hat{\beta}$ is biased.

Under heteroscedasticity, $\text{var}(\hat{\beta}) = \dfrac{\sigma^2}{\sum X_i^2}$ and its estimated value is $\dfrac{\hat{\sigma}^2}{\sum X_i^2}$

where $\hat{\sigma}^2 = \dfrac{\sum e_i^2}{n-1} = \dfrac{\text{RSS}}{n-1}$

So, the estimated variance of $\hat{\beta}$ under heteroscedasticity is $\dfrac{\text{RSS}}{n-1} \cdot \dfrac{1}{\sum X_i^2}$

Now, consider $Y_i = \beta X_i + u_i$ or, $\hat{Y}_i = \hat{\beta} X_i$

So, $\sum e_i^2 = \sum \left(Y_i - \hat{Y}_i\right)^2 = \sum u_i^2 - \left(\hat{\beta} - \beta\right)^2 \sum X_i^2 = \text{RSS}$

So, $E(\text{RSS}) = \sum E\left(u_i^2\right) - E\left(\hat{\beta} - \beta\right)^2 \sum X_i^2 = \dfrac{\sum X_i^2 \sum \sigma_i^2 - \sum X_i^2 \sigma_i^2}{\sum X_i^2}$

So the estimated variance of $\hat{\beta}$ under heteroscedasticity is

$$\frac{\text{RSS}}{n-1} \cdot \frac{1}{\sum X_i^2} = \frac{\sum X_i^2 \sum \sigma_i^2 - \sum X_i^2 \sigma_i^2}{\sum X_i^2} \cdot \frac{1}{n-1} \frac{1}{\sum X_i^2}$$

This is different from the true variance of $\hat{\beta}$.

So, the estimated variance is biased. The usual t, χ^2 and F tests will give us erroneous result.

5.6 Detection of Heteroscedasticity

There are different methods of detecting heteroscedasticity. The popular methods are given below :

- Anscombe and Ramsey method
- White method
- Glejser method

These three methods are applied for small sample. But if the sample size is large we can apply the following three methods.

- Likelihood Ratio test
- Goldfeld-Quandt test
- Breusch-Pagan test

These methods are discussed below.

(i) Anscombe and Ramsey Test :

First, using OLS method, residuals (e_i) are computed and the residuals are regressed on different forms of the predicted part of the explained variable such as $\hat{Y}_i^2, \hat{Y}_i^3$ and so on. So we have $e_i = \beta_1 + \beta_2 \hat{Y}_i^2 + \beta_3 \hat{Y}_i^3 + \beta_4 \hat{Y}_i^4 + \in$, where $\in \sim iid(0, \sigma^2)$.

If any coefficient of the above equation is found to be significant then there is heteroscedasticity, otherwise there is homoscedasticity.

(ii) White Test :

Using OLS method, first the residuals are computed from the original regression equation. Let the residual be denoted as e_i. Let us take the square of residual which is to be regressed on different froms of explanatory variables. If there are three explanatory variables X_1, X_2, X_3, X_1X_2, X_2X_3, X_1X_3 and so on. There are second type of regression equation to be estimated is

$$e_i^2 = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{1i}X_{2i} + \beta_5 X_{1i}X_{3i} + \beta_6 X_{2i}X_{3i} + \in$$

where $\in \sim iid(0, \sigma^2)$.

If any coefficient of the above equation is found to be significant then there is heteroscedasticity; otherwise there is homoscedasticity.

(iii) Glejser Test :

This test also belongs to the category of other two tests described above. In all the three tests it is assumed that $\sigma_i^2 = f(Z_i)$ for any value of Z. In Glejser test also, original regression equation is estimated using OLS method and absolute value of residual is computed. The absolute value of residual is then regressed on different forms of explanatory variables as follows :

$$|e_i| = a + bX_i \text{ or, } |e_i| = a + \frac{1}{X_i} \text{ or, } |e_i| = a + b\sqrt{X_i} \text{ etc.}$$

If 'b' is found to be significant in any case, then there is heteroscedasticity otherwise there is homoscedasticity.

These three tests are generally applied for small samples. If the sample size is large we have to apply other tests of heteroscedasticity which is discussed below :

(i) Likelihood Ratio Test :

Using OLS method, original regression equation is estimated and residuals are computed. Then the residuals are arranged according to the ascending order of predicted value of Y. Next, rediduals are grouped in k groups. For each group variance of the residual is estimated and for all groups taken together overall variance is also estimated. Let $\hat{\sigma}^2$ be the overall variance taking n observations into account and $\hat{\sigma}_i^2$ be the variance of residuals in the i-th group where $i = 1, 2, ...k$ and in each group there are n_i observations such that $\Sigma n_i = n$. Next compute λ as follows : $\lambda = \prod_{i=1}^{k} \frac{(\hat{\sigma}_i)^{n_i}}{(\hat{\sigma})^n}$

Here the test statistics is $-2ln(\lambda) \sim \chi^2_k$

If this test statistic is found to be significant then there is heteroscedasticity otherwise the regression is free from heteroscedasticity problem.

(ii) Goldfield-Quandt Test :

Unlike likelihood ratio test, the whole sample is divided into equal three groups in this type of test. If there are two residuals in deriving three groups, the first and last observations are to be deleted. Likewise, if there is one residual then the first observation is deleted. Before getting groups, observations are arranged in ascending order of X values (explanatory variable). Middle group is ignored and two separate regressions are to be estimated for the first and the last group. For each regression, variance of the residual is estimated. The estimates are $s_1^2 = \frac{\sum e_{1i}^2}{n_1 - k}$ and $s_3^2 = \frac{\sum e_{3i}^2}{n_3 - k}$

Then the relevant test statistic for the variance ratio test is $\frac{s_3^2}{s_2^2} \sim F_{(n_1-k),(n_3-k)}$.

If the test statistic is found to be significant then there is heteroscedasticity; otherwise there is homoscedasticity.

(iii) Breusch-Pagan Test :

In the Breusch-Pagan test the relevant test statistic is $\lambda = \frac{s_0}{2\hat{\sigma}^4}$ which follows χ^2_r, r being the number of explanatory variables in the regression equation used for heteroscedasticity test. Here $\hat{\sigma}^2 = \frac{\sum e_i^2}{n}$ = overall estimated value of the variance of disturbance term and s_0 is the explained sum of squares derived from regression $e_i^2 = f(Z_1, Z_2, ...)$ where Z_j may be x, e^x, $\sqrt{x}$ and so on. If λ is found to be statistically significant then there is heteroscedasticity, otherwise there is homoscedasticity in the disurbance term.

5.7 Solutions of Heteroscedasticity Problem

We have two types of solution of heteroscedasticity problem. In one type, a specific assumption is made regarding the form of heteroscedasticity. In another type, no apriori assumption is made regarding the form of heteroscedasticity. The first type is divided into two parts — one is WLS (weighted least square method) and other is MLE method. Further, we have different variations of WLS method— one-setp WLS, two-step WLS, iterated WLS and so on.

In the second category we have two methods—one is long transformation method and other is ratio method or deflating method. We now discuss the WLS method as a remedial measure of heteroscedasticity problem.

Let us assume the form of heteroscedasticity as $\sigma_i^2 = \sigma^2 X_i^2$. In this case regression equation to be estimated is in the following form : $\frac{Y_i}{X_i} = \alpha \frac{1}{X_i} + \beta + \frac{u_i}{X_i}$

or, $Y_i^* = \alpha X_i^* + \beta + \upsilon_i$ ..... (1)

On the above expression we can apply OLS to get WLS estimate of the parameters which will be unbiased and efficient.

If we assume that $\sigma_i^2 = \sigma^2 X_i$ is the form of heteroscedasticity then to apply WLS we have to divide the original model by $\sqrt{x_i}$

Therefore the deflated model is $\frac{c}{\sqrt{X_i}} = \beta \frac{X_i}{\sqrt{X_i}} + \frac{u_i}{\sqrt{X_i}}$ or, $Y_i^* = \beta X_i^* + \upsilon_i$ ....(2)

Applying OLS on (2) we get WLS estimator of β as follows

$$\tilde{\beta} = \frac{\sum Y_i^* X_i^*}{\sum X_i^{*2}} = \frac{\overline{Y}}{\overline{X}}$$

So the ratio of mean values will be the WLS estimator of β.

We generally assume that $\sigma_i^2 = \sigma^2 Z_i^2$. But in reality it is difficult to identify Z_i, i.e., the form of heteroscedasticity. To get rid of this problem, Paris and Houthakker prescribed the general form of assumed heteroscedasticity as $\sigma_i^2 = \sigma^2 [E(Y_i)]$

For practical application of Paris and Houthakker's method first OLS is applied to estimate the parameters of the regression equation. So the new form of heteroscedasticity becomes $\sigma_i^2 = \sigma^2 \left[\hat{\alpha} + \hat{\beta} X_i\right]^2$ and for WLS we estimate

$$\frac{Y_i}{\hat{\alpha} + \hat{\beta} X_i} = \frac{\alpha}{\hat{\alpha} + \hat{\beta} X_i} + \frac{\beta X_i}{\hat{\alpha} + \hat{\beta} X_i} + \upsilon_i$$

Then we have to apply OLS on the above equation to get two-step WLS method. To get accurate values of α and β we shall have to repeat the earlier steps of this method until we get convergence. This method is known as Iterated WLS method.

5.8 Some Numerical Problems

A researcher using time series data for the period 1954-65, estimated the following consuption function : $\hat{c} = 3 + 0.9272X$

The following table includes the data used and the residual errors.

Year	Consumption (c) (Billions of $)	Income (X) (Billions of $)	e_i
1954	236	257	0.52
1955	254	275	1.82
1956	267	293	1.87
1957	281	309	2.71
1958	290	319	2.99
1959	311	337	1.30
1960	325	350	3.25

Year	Consumption (C) (Billions of $)	Income (X) (Billions of $)	e_i
1961	335	364	0.26
1962	355	385	0.78
1963	375	405	2.23
1964	401	437	1.45
1965	431	469	1.14

(i) Test for heteroscedasticity, using Spearman's rank correlation coefficient.

(ii) Outline the corrective solution which you would adopt if heteroscedasticity is found significant.

Solution :

To apply Spearman's rank correlation test we rank X's and $|e$'s$|$ in ascending order. The rankings are shown in the following table.

Rank of X	Rank of e	D_i	D_i^2
1	9	–8	64
2	7	–5	25
3	4	–1	1
4	10	–6	36
5	12	–7	49
6	1	·5	25
7	8	–1	1
8	2	·6	36
9	3	6	36
10	5	5	25
11	6	5	25
12	11	1	1
Total			$\sum D_i^2 = 324$

The rank correlation coefficient estimated from the above data is

$$r'_{ex} = 1 - \frac{6\sum D_i^2}{n(n^2-1)} = 1 - \frac{6\times 324}{12(12^2-1)}$$

$$= 1 - \frac{1944}{1716} = \frac{1716-1944}{1716} = -\frac{228}{1716} \simeq -0.139$$

Now we have to test the null hypothesin that the value of correlation coefficient is zero against the alternative hypothesis that it is not equal to zero. i.e., we have to test $H_0 : \rho = 0$ against $H_1 : \rho \neq 0$. The appropriate test statistic is then given by $t = \frac{r'\sqrt{n-2}}{\sqrt{1-r'^2}} \sim t_{n-2}$ i.e., it follows a 't' distribution with $(n - 2)$ degrees of freedom. Here we have,

$$t = \frac{-0.139\times\sqrt{12-2}}{\sqrt{1-(0.139)^2}} = \frac{-0.139\times 3.162^3}{\sqrt{0.9806}} = \frac{-04396}{0.9902}$$

$\therefore\ t = -0.444$

Now the null hypothesis will be accepted for the given sample at 5% level of significance if $-t_{0.025,\ n-2} \le t \le t_{0.025,\ n-2}$ and will be rejected otherwise.

From the table value we see that $t_{0.025,\ n-2} = t_{0.025,\ 10} = 2.228$ (as $n = 12$) Here $t = -0.444$ lies in the range – 2.228 and 2.228 and hence the null hypothesis should be accepted here. We may thus conclude that there is no problem of heteroscedasticity.

(ii) We assume that the pattern of heteroscedasticity is $E\left(u_i^2\right) = \sigma_u^2 . x_i$ so that the appropriate transformation of the original model $c_t = \beta\sqrt{x_t} + u_t$ will become

$$\frac{c_t}{\sqrt{x_t}} = \beta\sqrt{x_t} + \frac{u_t}{\sqrt{x_t}}$$

Here $c_t = C_t - \overline{C}_t$

$x_t = X_t - \overline{X}_t$ and $\overline{u}_t = 0$

So applying OLS to the new variables we can obtain

$\hat{\beta} = \dfrac{\sum c_t}{\sum x_t} = \dfrac{\overline{c}}{\overline{x}}$ & $\hat{\alpha} = \overline{c}_t - \hat{\beta}\overline{X}_t$ and correspondingly we can estimate SE($\hat{\alpha}$), SE($\hat{\beta}$) and value of R^2.

2. The estimated saving function for a 31 years period is given by

$$\hat{S}_t = \underset{(117.6)}{-644.1} + \underset{(0.005)}{0.085} X_t \qquad R^2 = 0.903$$

After arranging the X's in ascending order and omitting nine central observations we are left with two subsets of data, one with the lower values of X and the other with the higher values of X.

Applying OLS to each subset, we obtain,

(i) For subset 1 $\hat{S}_1 = \underset{(189.4)}{-738.84} + \underset{(0.015)}{0.088} X$

R^2 = 0.787 and $\sum e_1^2 = 144,771.5$

(ii) For subset 2 $\hat{S}_2 = \underset{(709.8)}{1141.07} + \underset{(0.022)}{0.029X}$

R^2 = 0.152 and $\sum e_2^2 = 769,899.2$

By using Goldfeld and Quandt test, examine whether the problem of heteroscedasticity exists or not in this problem.

Solution :

For Goldfeld and Quand test, we use the test statistic

$$F^* = \frac{\sum e_2^2}{\sum e_1^2} \text{ with } df = \{n - 2 - 2k\}/2[v_1 = v_2]$$

Where n = total number of observations.

c = number of central observations omitted

k = number of parameters to be estimated

Here n = 31, c = 9, k = 2

$$\therefore \frac{n-c-2k}{2} = \frac{31-9-4}{2} = \frac{18}{2} = 9$$

Now, $F^* = \frac{\sum e_2^2}{\sum e_1^2} = \frac{769,899.2}{144,771.6} \approx 5$ with d.f (9, 9)

From the table value $F^* > F_{0.05}$; 9, 9 (= 3.18) and hence the null hypothesis is rejected at 5% level of significance. Thus the problem is involved with heteroscedasticity.

5.9 Summary

Heteroscedasticity problem arises when variance of the disturbance term is not constant, rather it is varying. In this chapter we have discussed the meaning of the heteroscedasticity problem. We have identified different sources of heteroscedasticity problem which includes error learning model, improvement in data processing technique, presence of outlier in data, existence of discretionary income, misspecification error, incorrect data tranasformation or functional forms, skewed distribution in data, etc. There are some procedures to detect the heteroscedasticity problem which are different depending on the sample size. Under heteroscedasticity, the OLS estimator of regression parameter is unbaised but indifferent. Funtuer, the OLS estimatae of the variance of of the estimator of regression parameter is biased.

5.10 Exercise

Short Answer Type Questions :

(a) Choose the correct answer :

(i) Problem of heteroscedasticity is observed when variance of the disturbance term is

(1) zero

(2) one

(3) constant

(4) varying

(ii) Which is not a method of detecting heteroscedasticity ?

(1) White test

(2) Glejser test

(3) Likelihood ratio test

(4) PE test

(b) Fill in the blanks :

(i) Due to the presence of outlier the variance of the disturbance term will increase and that leads to _____ problem.

(ii) When the variances of the disturbance term 'u' are _____ then that is known as heteroscedasticity.

(c) Identify whether the statements are true or false :

(i) In the presence of heteroscedasticity, the OLS estimator of regression parameter is still unbiased (consistent) but inefficient.

(ii) Likelihood Ratio test is applied to detect the heteroscedasticity problem when the sample size is small.

Medium Answer Type Questions :

1. Mention two major sources of heteroscedasticity.
2. Explain one major consequence of heteroscedasticity.
3. Discuss the likelihood ratio test of detecting heteroscedasticity.
4. State the Goldfeld-Quandt test for detecting the problem of heteroscedasticity.

Long Answer Type Questions :

1. What are the sources of heteroscedasticity ?
2. What are the consequences of heteroscedasticity problem ?
3. How can you detect the presence of heteroscedasticity problem ?
4. How can you solve the heteroscedasticity problem ?

5.11 References

1. Koutsoyiannis, A (1996) : *Theory of Econometrics*, ELBS with Macmillan
2. Gujarati, D (2003) : *Basic Econometrics*, McGraw Hill Higher Education
3. Sarkhel, J and Santosh Kumar Dutta (2020) : *An Introduction to Econometrics*, Book Syndicate Private Limited.

Unit - 6 ❑ Violating the Assumptions of the CLRM : III-The Problem of Autocorrelation

Structure

6.1 Objectives

Reading this chapter, students will get an idea about

- Meaning of autocorrelation
- Sources of autocorrelation
- Consequences of autocorrelation
- Tests for detecting autocorrelation
- Remedial measures for autocorrelation

6.2 Introduction

Autocorrelation is a speical case of correlation. It refers to the relationship between successive values of the same variable, while correlation refers to the relationship between two or more different variables. In the Classical Linear Regression Model, it is assumed that the disturbance terms are independent of each other i.e., Cov $(u_i, u_j) = E(u_i, u_j) = E(u_i).E(u_j) = 0$ for all $i \neq j$. This assumption implies that successive values of disturbance term u are temporarily independent, i.e., disturbance term occurring at one period is not related to any other disturbance. In other words, when observations are taken over time, the effect of disturbance occurring at one period does not carry over into another period. However, if the above assumption is not satisfied, i.e., if the successive values of disturbance term are dependent to each other, we say that there is autocorrelation problem. This autocorrelation problem is mainly observed in time series data unlike heteroscedasticity which is observed in cross section data also.

6.3 Structure of Autocorrelation Problem

To explain the structure of autocorrelation problem, we are explaining a simple two variable model as follows : $Y_t = \alpha + \beta X_t + u_t$ where subscript t denotes time series data. In this model, all the assumptions of CLRM are satisfied except no autocorrelation assumption which is replaced by the assumption $u_t = \rho u_{t-1} + \epsilon_t$ i.e., the disturbance term follows first order auto regressive scheme. Here, $|\rho| < 1$ and ϵ_t is a white noise, $E(\epsilon_t) = 0$, $E(\epsilon_t^2) = \sigma_\epsilon^2$ and $E(\epsilon_t . \epsilon_{t-s}) = 0$ for $s \neq 0$

So, it can be written

$$u_t = \rho(\rho u_{t-2} + \epsilon_{t-1}) + \epsilon_t$$

$$\text{or, } u_t = \epsilon_t + \rho\epsilon_{t-1} + \rho^2(\rho u_{t-3} + \epsilon_{t-2})$$

$$\text{or, } u_t = \epsilon_t + \rho\epsilon_{t-1} + \rho^2\epsilon_{t-2} + \rho^3\epsilon_{t-3} + \ldots\ldots$$

$$\text{or, } u_t = \sum_{r=0}^{\infty} \rho^r \epsilon_{t-r}$$

$$\text{So, } E(u_t) = E(\epsilon_t) + \rho E(\epsilon_{t-1}) + \rho^2 E(\epsilon_{t-2}) + \rho^3 E(\epsilon_{t-3}) + \ldots..$$

$$\text{So, } E(u_t) = 0$$

$$\text{and } E\left(u_t^2\right) = E\left(\sum_{r=0}^{\infty} \rho^r \epsilon_{t-r}\right)^2$$

$$\text{or, } E\left(u_t^2\right) = E(\epsilon_t + \rho\epsilon_{t-1} + \rho^2\epsilon_{t-2} + \rho^3\epsilon_{t-3} + \ldots..)^2$$

$$\text{or, } E\left(u_t^2\right) = E(\epsilon_t^2) + \rho^2 E(\epsilon_{t-1}^2) + \rho^4 E(\epsilon_{t-2}^2) + \ldots\ldots\ldots$$

[since, $E(\epsilon_t, \epsilon_{t-s}) = 0$ for $s \neq 0$]

$$\text{or, } E\left(u_t^2\right) = \sigma_\epsilon^2 [1 + \rho^2 + \rho^4 + \ldots\ldots..]$$

$$\text{or, } E\left(u_t^2\right) = \sigma_\epsilon^2 \frac{1}{1-\rho^2}$$

$$\text{or, } E\left(u_t^2\right) = \sigma_\epsilon^2$$

i.e. constant variance

Now, $E(u_t\ u_{t-1}) = E[\{\epsilon_t + \rho\epsilon_{t-1} + \rho^2\epsilon_{t-2} + \rho^3\epsilon_{t-3} + ...\}\{\epsilon_{t-1} + \rho\epsilon_{t-2} + \rho^2\epsilon_{t-3} + \rho^3\epsilon_{t-4} + ...\}]$

or, $E(u_t\ u_{t-1}) = E[\rho(\epsilon_{t-1} + \rho\epsilon_{t-2} + \rho^2\epsilon_{t-3} + \rho^3\epsilon_{t-4} + ...)^2]$

or, $E(u_t\ u_{t-1}) = \rho E[(\epsilon_{t-1} + \rho\epsilon_{t-2} + \rho^2\epsilon_{t-3} + \rho^3\epsilon_{t-4} + ...)^2]$

or, $E(u_t\ u_{t-1}) = \rho[E(\epsilon_{t-1}^2) + \rho^2 E(\epsilon_{t-2}^2) + \rho^4 E(\epsilon_{t-3}^2) + ...]$

or, $E(u_t\ u_{t-1}) = \rho\sigma_\epsilon^2[1 + \rho^2 + \rho^4 + ...]$

or, $E(u_t\ u_{t-1}) = \dfrac{\rho\sigma_\epsilon^2}{1-\rho^2}$

or, $E(u_t\ u_{t-1}) = \rho\sigma_u^2 \neq 0$

or, $\dfrac{E(u_t, u_{t-1})}{\sigma_u^2} = \rho$

Which is the first order autocorrelation coefficient.

Thus, we have

$$u_t = \epsilon_t + \rho\epsilon_{t-1} + \rho^2\epsilon_{t-2} + \rho^3\epsilon_{t-3} + ...$$

Similarly we can write in a generalised version as

$$u_{t-s} = \epsilon_{t-s} + \rho\epsilon_{t-s-1} + \rho^2\epsilon_{t-s-2} + \rho^3\epsilon_{t-s-3} + ...$$

So, $E(u_t\ u_{t-s}) = E[\{\epsilon_t + \rho\epsilon_{t-1} + ...+ \rho^{s-1}\epsilon_{t-s+1} + (\rho^s\epsilon_{t-s} + \rho^{s+1}\epsilon_{t-s-1} + ...)\}$
$\{\epsilon_{t-s} + \rho\epsilon_{t-s-1} + \rho^2\epsilon_{t-s-2} + ...\}]$

or, $E(u_t\ u_{t-s}) = E[\rho^s(\epsilon_{t-s} + \rho\epsilon_{t-s-1} + \rho^2\epsilon_{t-s-2} + ...)^2]$

[since, $E(\epsilon_t,\ \epsilon_{t-s}) = 0$ for $s \neq 0$]

or, $E(u_t\ u_{t-s}) = \rho^s E[(\epsilon_{t-s} + \rho\epsilon_{t-s-1} + \rho^2\epsilon_{t-s-2} + ...)^2]$

or, $E(u_t\ u_{t-s}) = \rho^s[E(\epsilon_{t-s}^2) + \rho^2 E\left(\epsilon_{t-s-1}^2\right) + ...]$

or, $E(u_t\ u_{t-s}) = \rho^s\sigma_\epsilon^2\ [1 + \rho^2 + \rho^4 + ...]$

$$\text{or, } E(u_t u_{t-s}) = \frac{\rho^s \sigma_\epsilon^2}{1-\rho^2}$$

$$\text{or, } E(u_t u_{t-s}) = \rho^s \sigma_\epsilon^2$$

$$\text{or, } \frac{E(u_t, u_{t-s})}{\sigma_u^2} = \rho^s$$

Therefore, the s-th order autocorrelation coefficient is ρ^s

6.4 Sources, of Autocorrelation or Serial Correlation

Autocorrelation in time series data may arise due to a number of reasons as pointed out below :

- **Inerita :** Autocorrelation in times serices data may arise due to inertia or sluggishness in the data. In economics, there is the existense of business cycles, due to which variables move simultaneously as per the cycle. In recession, variables will decline and in recovery variables will move upward. Therefore, one movement depends on preceding movements of the data and hence autocorrelation arises.
- **Omission of explanatory variables :** The regression equation may not be specified properly in all cases. A few explanatory variables may be omitted whose influence remains present in the disturbance term. Due to this, the successive values of the disturbance term remain dependent to each other causing autocorrelation problem.
- **Incorrect functional form :** The regression equation may not be always correctly specified. Non-linear regression may be specified as a linear one. For instance, cost function is actually a quadratic function. But if we take its linear form, the quadratic part will be incorporated in the disturbance term which may be the cause of autocorrelation problem.
- **Presence of lag :** In a number of economic specifications, the lagged form of dependent variable is used as an independent variable in regression

equation. For instance, present consumption depends on past consumption level. In the presence of this lagged variable, the disturbance term may be correlated to each other.

- **Manipulation of data :** Time series data are mainly collected from secondary sources. But all data from secondary source may not be available. In that case, interpolation or extrapolation is made to find out the missing values. Further, to avoid short term volatility in data, the smoothing technique is applied. Due to all these manipulations, autocorrelation problem may crop up.
- **Transformation of data :** If we estimate the regression equation $Y_t = \alpha + \beta X_t + u_t$, then the equation is expressed in its level form. But sometimes we are interested to express the regression equation in its difference form i.e., $\Delta Y_t = \beta \Delta X_t + \upsilon_t$ where $\Delta Y_t = Y_t - Y_{t-1}$ and $\Delta X_t = X_t - X_{t-1}$. If u_t follows the assumption of CLRM, then, υ_t will be autocorrelated.
- **Cobweb Model :** In economics, we observe cobweb phenomenon in supply function when supply of any period depends on price of previous period i.e., $S_t = a + bP_{t-1} + u_t$. If in any period price decreases, in the next period, supply will also decrease. So the disturbance term is non-random and that may lead to the existence of autocorrelation problem.
- **Non–stationary data :** A series is said to be non-stationary if its mean or variance is dependent on time and covariance depends on time lag. If X and Y are both non-stationary in regression equation $Y_t = \alpha + \beta X_t + u_t$, then u_t, will also be non-stationary. If u_t is found to be non stationary, then it will be autocorrelated.

6.5 First Order Autoregressive Scheme/Markov Process

Generally the problem of autocorrelation arises in the presence of time series data. We counder a model : $Y_t = \alpha + \beta X_t + u_t$ based on time series data for t = 1, 2, 3, ..., ∞. We now assume that $u_t = \rho u_{t-1} + \epsilon_t$ with $|\rho| < 1$

This is called the first order autoregressive scheme.

Here, ρ = the coefficient of autocorrelation

ϵ_t = a random term with usual assumption of a random variable, i.e., $E(\epsilon) = 0$, $E(\epsilon^2) = \sigma_\epsilon^2$ and $E(\epsilon_i \epsilon_j) = 0$ for $i \neq j$.

The complete form of the first order Markov process (the pattern of autocorrelation for the values of u) is as follows :

$u_t = f(u_{t-1}) = \rho u_{t-1} + \epsilon_t$

$u_{t-1} = f(u_{t-2}) = \rho u_{t-2} + \epsilon_{t-1}$

$u_{t-2} = f(u_{t-3}) = \rho u_{t-3} + \epsilon_{t-2}$

. . . . . .

. . . . . .

. . . . . .

$u_{t-r} = f(u_{t-(r+1)}) = \rho u_{t-(r+1)} + \epsilon_{t-r}$

In order to define the error term in any particular period t, we follow the method given below. We follow the autocorrelation relationship in period t : $u_t = \rho u_{t-1} + \epsilon_t$ and then perform continuous substitutions of the lagged values of u. The process is shown below.

If we substitute u_{t-1} in the expression of u_t, we get,

$u_t = \rho[\rho u_{t-2} + \epsilon_{t-1}] + \epsilon_t = \rho^2 u_{t-2} + (\rho\epsilon_{t-1} + \epsilon_t)$

If we substitute u_{t-2}, then we get,

$u_t = \rho^2[\rho u_{t-3} + \epsilon_{t-2}] + (\rho\epsilon_{t-1} + \epsilon_t) = \rho^3 u_{t-3} + (\rho^2\epsilon_{t-2} + \rho\epsilon_{t-1} + \epsilon_t)$

If we substitute u_{t-3}, then we get,

$u_t = \rho^3[\rho u_{t-4} + \epsilon_{t-3}] + (\rho^2\epsilon_{t-2} + \rho\epsilon_{t-1} + \epsilon_t)$

$= \rho^4 u_{t-4} + (\rho^3\epsilon_{t-3} + \rho^2\epsilon_{t-2} + \rho\epsilon_{t-1} + \epsilon_t)$

Let us continue the substitution process for r periods where r is quite large. Then $u_t = \epsilon_t + \rho\epsilon_{t-1} + \rho^2\epsilon_{t-2} + \rho^3\epsilon_{t-3} +$ .....

Here, as the power of ρ increases to infinity with the lagged u, then $\rho^r u_{t-r} +$ ..... tends to zero as $|\rho| <$ | Thus, $u_t = \sum_{r=0}^{\alpha} \rho^r \epsilon_{t-r}$

This is the value of the error term when it is autocorrelated with a first order autoregressive scheme.

6.6 A note on Autocorrelation Coefficient ρ or $\rho u_t u_{t-1}$

We write, $u_t = \rho u_{t-1} + \epsilon_t$ where ρ is the true auto correlation co-efficient. With usual assns about ϵ_t (i.e., $E(\epsilon_t) = 0$, $E(\epsilon_t \epsilon_{t-1}) = 0$, $E\left(\epsilon_t^2\right) = \sigma_\epsilon^2$ We have $\rho = \rho u_t u_{t-1} = \frac{\sum u_t u_{t-1}}{\sqrt{\sum u_t^2 \sum u_{t-1}^2}}$. But $u_t s$ are not observable. Hence we estimate ρ by replacing u's by e's, where e is the error of estimate. Then $\hat{\rho} = \frac{\sum e_t e_{t-1}}{\sqrt{\sum e_t^2 \sum e_{t-1}^2}}$.

Here, $\hat{\rho}$ is an estimate of true antocorrelation coefficient, ρ.

Now, when u in any period depends on its own value of the previous period, we say that u follows a first order autoregressive scheme or first order Markov process. Then $u_t = f(u_{t-1})$. Let there be a simple linear relation : $u_t = a_1 u_{t-1} + \epsilon_t$. ϵ_t is a random variable with usual assumptions $E(\epsilon_t) = 0$, $E\left(\epsilon_t^2\right) = \sigma_t^2$, $E(\epsilon_i \epsilon_j) = 0$. Here a_1 is the coefficient of autocorrelation relationship. From OLS, $\hat{a}_1 = \frac{\sum u_t u_{t-1}}{\sum u_{t-1}^2}$. On the other hand, the autocorrelation coefficient, $\rho = \frac{\sum u_t u_{t-1}}{\sqrt{\sum u_t^2 \sum u_{t-1}^2}}$

For large examples, $\sum u_t^2 \approx \sum u_{t-1}^2$. So, $\rho \approx \hat{a}$.

Hence the simple first order autoregressive scheme is $u_t = \rho u_{t-1} + \epsilon_t$. If $\rho = 0$, $u_t = \epsilon_t$ and there is no autocorrelation.

6.7 Mean, Variance and Covariance of the Autocorrelated Disturbance Variable

1. Mean of the autocorrelated disturbance term (u)

We have, $u_t = \sum_{r=0}^{\alpha} \rho^r \epsilon_{t-r}$

Now, mean of u_t = $E(u_t)$ = $E\left[\sum_{r=0}^{\alpha} \rho^r \in_{t-r}\right] = \sum_{r=0}^{\alpha} \rho^r E(\in_{t-r})$

But by assumption of the distribution of $\in$, we have, $E(\in_{t-r})= 0$. So, $E(u_t) = 0$. Thus mean of autocoorrelated u's is zero.

2. Variance of the autocorrelated *u*'s

By definition, var (u_t) = $E[u_t - E(u_t)]^2$

As $E(u_t) = 0$, var (u_t) = $E(u_t^2) = E\left[\sum_{r=0}^{\alpha} \rho^r \in_{t-r}\right]^2$

$$= \sum_{r=0}^{\alpha} (\rho^r)^2 E(\in_{t-r})^2 = \sum_{r=0}^{\alpha} \rho^{2r} \text{Var}(\in_{t-r})$$

$$= \sum_{r=0}^{\infty} \rho^{2r}\sigma_{\in}^2 = \sigma_{\in}^2 (1+\rho^2+\rho^4+\rho^6+\cdots)$$

$$= \sigma_{\in}^2 \cdot \frac{1}{1-\rho^2} \text{ as } |\rho| < 1.$$

So, Var (u_t) = $\sigma_u^2 = \frac{\sigma_{\in}^2}{1-\rho^2}$ for $|\rho| < 1$.

3. Covariance of the autocorrelated *u*'s

We know, $u_t = \in_t + \rho \in_{t-1} + \rho^2 \in_{t-2} + \cdots$

and so, $u_{t-1} = \in_{t-1} + \rho \in_{t-2} + \rho^2 \in_{t-3} + \cdots$

Now, Cov $(u_t\ u_{t-1})$ = $E[u_t - E(u_t)][u_{t-1} - E(u_{t-1})]$

= $E[u_t\ u_{t-1}]$ as $E(u_t) = 0$ and $E(u_{t-1}) = 0$

$= E[(\in_t + \rho\in_{t-1} + \rho^2\in_{t-2} + \ldots)(\in_{t-1} + \rho\in_{t-2} + \rho^2\in_{t-3} + \ldots)]$

$= E[\{\in_t + \rho(\in_{t-1} + \rho\in_{t-2} + \ldots)\}(\in_{t-1} + \rho\in_{t-2} + \rho^2\in_{t-3} + \ldots)]$

$= E[\in_t (\in_{t-1} + \rho\in_{t-2} + \rho^2\in_{t-3} + \ldots)] + E[\rho(\in_{t-1} + \rho\in_{t-2} + \rho^2\in_{t-3} + \ldots)^2]$

$$= 0 + \rho E\ (\epsilon_{t-1} - \rho\epsilon_{t-2} + \rho^2\epsilon_{t-3} + \ldots..)^2$$

$$= \rho E\left(\epsilon_{t-1}^2 + \rho^2\,\epsilon_{t-2}^2 + \rho^4\,\epsilon_{t-3}^2 + \cdots + \text{cross products}\right)$$

$$= \rho\left(\sigma_\epsilon^2 + \rho^2\sigma_\epsilon^2 + \rho^4\sigma_\epsilon^2 + \cdots + 0\right) \text{as } E(\text{cross products}) = 0$$

$$= \rho.\sigma_\epsilon^2\ (1 + \rho^2 + \rho^4 + \ldots.) = \rho\sigma_\epsilon^2.\frac{1}{1-\rho^2} \text{ as } |\rho| < 1$$

$$= \rho\sigma_u^2\left[as\ \sigma_\epsilon^2 \Big/ 1-\rho^2 = \sigma_u^2 = \text{Var}(u)\right]$$

So, Cov $(u_t, u_{t-1}) = \rho\sigma_u^2$

Similarly, Cov $(u_t, u_{t-2}) = \text{E}(u_t\ u_{t-2}) = \rho^2\sigma_u^2$, etc.

In general, Cov $(u_t, u_{t-s}) = \rho^s\sigma_u^2$ (for $s \neq$ t)

6.8 Consequences of Autocorrelation

If there is autocorrelation, applying the OLS method, we can get the unbiased estimators of the paramters but cannot get the minimum variance of the estimators. In other words, under autocorrelation or serial correlation of the disturbance term, the value as well as the standard errors of the parameter estimates are affected. In particular we get the following results in the presence of autocorrelation in the disturbance term :

1. OLS estimates are unbiased

In the simple regression model, the value of the slope parameter in deviation form is given by :

$$\hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{\sum x_i\left(Y_i - \bar{Y}\right)}{\sum x_i^2} = \frac{\sum x_i Y_i - \bar{Y}\sum x_i}{\sum x_i^2} = \frac{\sum x_i Y_i - 0}{\sum x_i^2} \text{ as } \sum x_i = 0$$

$= \sum K_i Y_i$ where $K_i = \dfrac{x_i}{\sum x_i^2}$

$= \sum K_i (\alpha + \beta X_i + u_i) = \alpha \sum K_i + \beta \sum K_i X_i + \sum K_i u_i$

or, $\hat{\beta} = \beta + \sum K_i u_i$ [as $\sum K_i = \dfrac{\sum x_i}{\sum x_i^2} = 0$ and $\sum K_i x_i = \dfrac{\sum x_i (x_i + \bar{X})}{\sum x_i^2}$

$= \dfrac{\sum x_i^2 + \bar{X} \sum x_i}{\sum x_i^2} = \dfrac{\sum x_i^2}{\sum x_i^2} = 1$]

$E(\hat{\beta}) = \beta + E(\sum K_i u_i) = \beta + \sum K_i E(u_i) = \beta + 0 = \beta$ [as $E(u_i) = 0$]

So, Bias in $\hat{\beta} = E(\hat{\beta}) - \beta = \beta - \beta = 0$ Thus, whether there is autocorrelation or not, the slope parameter's estimate has no statistical bias. Similarly we can show that the estimate of intercept parameter ($\hat{\alpha}$) has also no bias in the presence of autocorrelation. We have, $\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} = \sum \left[\frac{1}{n} - \bar{X} K_i\right] Y_i$ where $\hat{\beta} = \sum K_i Y_i$

and $K_i = \dfrac{x_i}{\sum x_i^2}$

$= \sum \left[\frac{1}{n} - \bar{X} K_i\right] (\alpha + \beta X_i + u_i)$

$= \alpha + \beta \bar{X} + \frac{1}{n} \sum u_i - \bar{X} . \alpha \sum K_i - \bar{X} \beta \sum K_i X_i - \bar{X} \sum K_i u_i$

$= \alpha + \beta \bar{X} + \frac{1}{n} \sum u_i - \beta \bar{X} - \bar{X} \sum K_i u_i \left(\text{as} \sum K_i = \dfrac{\sum x_i}{\sum x_i^2} = 0 \right)$

$= \alpha + \frac{1}{n} \sum u_i - \bar{X} \sum K_i u_i$

$\therefore E(\hat{\alpha}) = \alpha + \frac{1}{n} \sum E(u_i) - \bar{X} \sum K_i E(u_i) = \alpha \left(\because E(u_i) = 0\right)$

$\therefore$ Bias in $\hat{\alpha} = E(\hat{\alpha}) - \alpha = \alpha - \alpha = 0.$

Thus, even in the presence of autocorrelation, $\hat{\alpha}$ has no bias.

2. The variances of OLS estimates are underestimated.

When u's are autocorrelated, the variance of estimate $\hat{\beta}$ in simple regression model will be biased downwards (i.e., underestimated).

Proof : We have $\hat{\beta} - \beta = \sum K_i u_i$ and the Var $(\hat{\beta}) = E\left[\sum K_i u_i\right]^2$ where

$$K_i = \frac{x_i}{\sum x_i^2}$$

or, Var $(\hat{\beta}) = E\left[\sum K_i^2 u_i^2 + 2\sum_i \sum_j K_i K_j u_i u_j\right]$

$$= \sum K_i^2 E(u_i^2) + 2\sum\sum K_i K_j E\left(u_i u_j\right)$$

$$= \sigma_u^2 \frac{\sum x_i^2}{\left(\sum x_i^2\right)^2} + 2 \times 0 \quad (\because E(u_i u_j) = 0) = \frac{\sigma_u^2}{\sum x_i^2}$$

This is the value of Var $(\hat{\beta})$ in the absence of autocorrelation in u. However, with u's related with a first order autoregressive scheme, $E(u_t^2) = \sigma_u^2$ and

$E\left(u_t u_{t-s}\right) = \rho^s . \sigma_u^2$. So, now,

$$Var(\hat{\beta}) = \sigma_u^2 \Big/ \sum x_i^2 + 2\frac{\sum x_i x_j}{\left(\sum x_i^2\right)^2} . \rho^s . \sigma_u^2 \quad \text{.....(1)}$$

Expanding the second term on the RHS, we get,

$$Var(\hat{\beta}) = \frac{\sigma_u^2}{\sum x_i^2} + 2\sigma_u^2 \left[\rho . \frac{\sum_{i=1}^{n-1} x_i x_{i+1}}{\left(\sum_{i=1}^{n} x_i^2\right)^2} + \rho^2 \frac{\sum_{i=1}^{n-2} x_i x_{i+2}}{\left(\sum_{i-1}^{n} x_i^2\right)^2} + \cdots\cdots\right]$$

$$= \frac{\sigma_u^2}{\sum x_i^2}\left[1 + 2\rho . \frac{\sum_{i}^{n-1} x_i x_{i+1}}{\sum x_i^2} + 2\rho^2 \frac{\sum_{i}^{n-2} x_i x_{i+2}}{\sum x_i^2} + \cdots + 2\rho^{n-1} \frac{x_i x_n}{\sum x_i^2}\right]$$

In the absence of autocorrelation, $\text{Var}(\hat{\beta}) = \sigma_u^2 / \sum x_i^2$.

But in the presence of autocorrelation ($\rho > 0$) and if X is also positively correlated (i.e., $\sum x_i, x_j \neq 0$), the expression in the bracket is almost certainly greater than one (or the second term in equation (1) is positive). This proves that the estimate of variance will have downward bias due to positive autocorrelation. If the explanatory variable X of the model is random, the co-variance of successive values is zero (i.e., $\sum x_i x_j = 0$). Under such circumstances, the bias in Var ($\hat{\beta}$) will not be serious even though μ is auto correlated.

3. In the presence of autocorrelation, in u_i's, the predictions will be inefficient.

If the values of u are autocorrelated, the predictions based on least square estimates will be inefficient. This means that the predictions will have a larger variance as compared with predictions based on estimates obtained from other econometric techniques like GLS (Generalised Least Squares).

6.9 Test, for Autocorrelation

There are various ways of testing autocorrelation. Two traditionally applied tests are : Durbin-Watson test and Von Neumann ratio method. There are however, some other tests as well. We have considered seven such tests.

6.9.1 Durbin-Waston Test

J. Durbin and G. S. Watson have suggested a test for autocorrelation. The test is applicable to small samples. However, the test is appropriate only for the first order autoregressive scheme :

$$u_t = \rho u_{t-1} + \epsilon_t$$

In this method, we test the null hypothesis

H_0 : $\rho = 0$ against the alternative

H_1 : $\rho \neq 0$

In language, our null hypothesis is : H_0 : the u's are not autocorrelated with a first order scheme. We shall test this hypothesis against the alternative hypothesis : H_1 : the u's are autocorrelated with a first order scheme.

To test the null hypothesis we use the Durbin-Watson statistic, say, d. It is given by :

$d = \dfrac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}$ where e_t (for t = 1, 2, ...n) are the OLS residual terms.

Now, $d = \dfrac{\sum_{t=2}^{n} e_t^2 + \sum_{t=2}^{n} e_{t-1}^2 - 2\sum_{t=2}^{n} e_t e_{t-1}}{\sum_{t=1}^{n} e_t^2}$

Now, for very large values of n i.e., for large samples, $\sum_{t=2}^{n} e_t^2, \sum_{t=2}^{n} e_{t-1}^2$ and $\sum_{t=1}^{n} e_t^2$ are approximately equal.

So $d \approx \dfrac{2\sum e_{t-1}^2 - 2\sum e_t e_{t-1}}{\sum e_{t-1}^2} \approx 2 - 2\dfrac{\sum e_t e_{t-1}}{\sum e_{t-1}^2} \approx 2\left(1 - \dfrac{\sum e_t e_{t-1}}{\sum e_{t-1}^2}\right)$

But in the model $e_t = \rho e_{t-1} + \in_t$ for t = 2, 3, ..., n,

$\hat{\rho} = \dfrac{\sum e_t e_{t-1}}{\sum e_{t-1}^2}$ where $\hat{\rho}$ is OLS estimator of ρ.

$\therefore d \approx 2(1 - \hat{\rho})$. Since $-1 < \hat{\rho} < 1$, $0 < d < 4$. Thus, d lies between 0 and 4.

Let us consider different values of $\hat{\rho}$ and the corresponding values of d where $d = 2\ (1 - \hat{\rho})$.

First, if there is no autocorrelation, $\hat{\rho} = 0$ and $d = 2$. Thus if from the sample data are find $d^* = 20$, we accept that there is no autocorrelation in the scheme.

Second, if $\hat{\rho} = +1$, $d = 0$ and we have perfect positive autocorrelation.

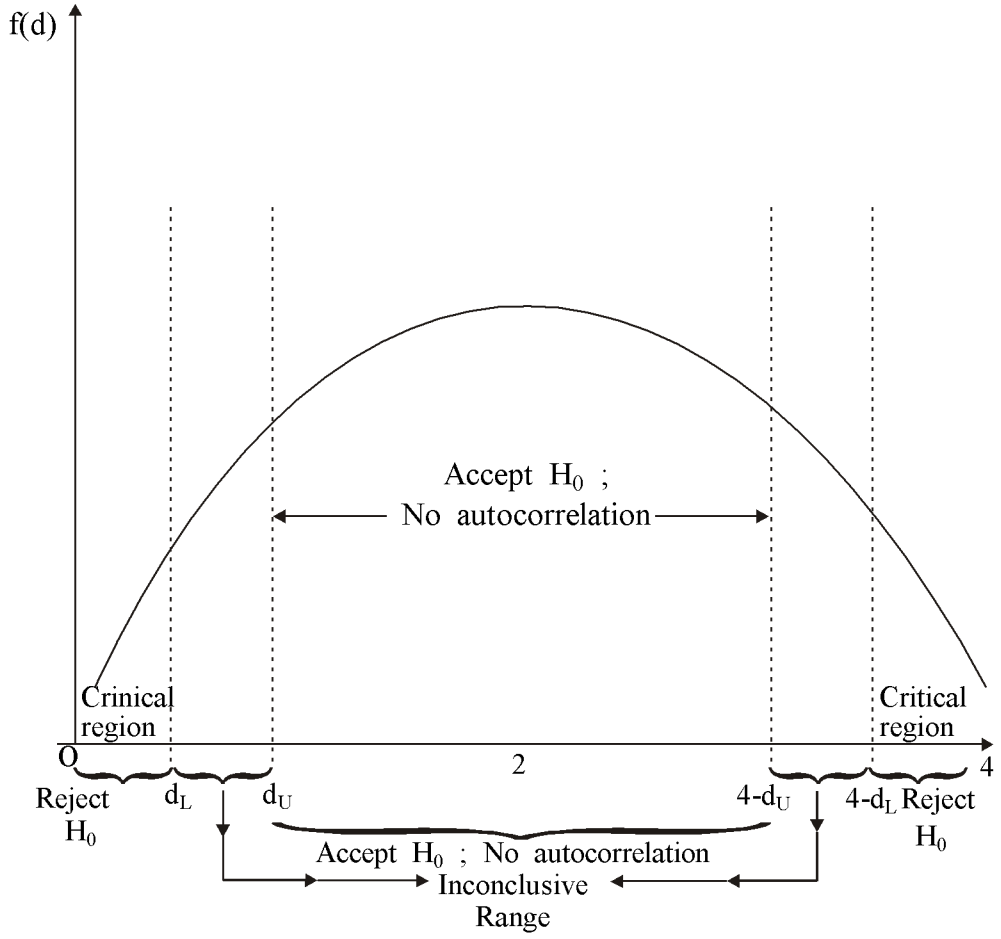
So, if $0 < d^* < 2$, there is some degree of positive autocorrelation. If d^* is closer to zero, the autocorrelation is stronger. If d^* is closer to 2, the autocorrelation is weaker.

Third, if $\hat{\rho} = -1$, $d = 4$ and we have perfect negative autocorrelation. So, if $2 < d^* < 4$, there is some degree of negative autocorrelation. If d^* is closer to 4, (negative) autocorrelation is stronger. If d^* is closer to 2, (negative) autocorrelation is weaker.

If there is no problem of autocorrelation, $\hat{\rho}$ should be zero and $d \approx 2$. Thus, to test the null hypothesis $H_0 : \rho = 0$ against the alternative $H_1 : \rho \neq 0$ means to test the null hypothesis $H_0 : d \approx 2$ against the alternative $H_1 : d \not\approx 2$.

Here the problem is that the exact sampling distribution of the statistic 'd' is not known. What Durbin and Watson have doen is to specify one upper limit and one lower limit of d. Let d_U stand for the upper limit of d and d_L stand for the lower limit of d. With the help of d_L and d_U we have to determine whether autocorrelation exists or not. The values of d_L and d_U are available at the 5% and 1% levels of significance.

The whole argument is shown in the following diagram.



From the sample, we calculate $d^* = 2(1-\hat{\rho})$ where $\hat{\rho} = \frac{\sum_{t=2}^{n} e_t e_{t-1}}{\sum_{t=2}^{n} e_{t-1}^2}$

We may now consider the following cases.

1. If it is found that $d^* < d_L$, we reject the null hypothesis of no autocorrelation and accept that there is positive autocorrelation of first order.

2. If $d^* > (4 - d_L)$, we reject the null hypothesis of no autocorrelation and accept that there is negative autocorrelation of the first order.

3. If $d_U < d^* < (4 - d_U)$, we accept the null hypothesis of no autocorrelation.

4. If $d_L < d^* < d_U$ or if $(4 - d_U) < d^* < (4 - d_L)$, the test is inconclusive.

Limitations of Durbin Waston Test

There are some limitations of Durbin-Watson test.

1. There are some inconclusive ranges in the test. If d^* lies between d_L and d_U or between $(4 - d_U)$ and $(4 - d_L)$, we cannot conculde whether there is autocorrelation or not in the given set of data.

2 This test method is appropriate only when the nature of the autocorrelation is of first order autoregressive type. When autocorrelation is of higher order and non-linear type, this test is inappropriate.

3. If there is any lagged independent variable in the model, the Durbin-Watson d-statistic is inappropriate for testing autocorrelation.

6.9.2 Von Neumann Ratio

In general Von Neumann ratio is the ratio of the variance of the first difference of any variable x over the variance of x, i.e., $\frac{\delta^2}{s_x^2} = \frac{\sum_{t=2}^{n}(X_t - X_{t-1})^2/n-1}{\sum_{t=1}^{n}(X_t - \overline{X})^2/n}$. It is applicable when X is directly observable and its successive values are not autocorrelated.

Thus, Von Neumann ratio is another traditionally applied test for detecting autocorrelation in regression analysis. The Von Neumann ratio is :

$$VN = \frac{\delta^2}{s_e^2} = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2 \Big/ n-1}{\sum_{t=1}^{n} e_t^2 \Big/ n}$$ where e_t is the value of the residual in period t and n is the sample size. As the values of the random variable (u's) are not directly observable, they are estimated from OLS residuals (e's). The VN ratio is applicable for large samples ($n > 30$).

The VN ratio is related to the Durbin-Watson d-statistic by the formula :

$\frac{\delta^2}{s^2} = \frac{n}{n-1}.d$, and follows approximately a normal distribution for large values of n.

The VN test statistic is used for testing the presence of autolcorrelation in the same manner as d-statistic does the same.

The VN ratio is however not applicable for testing the autocorrelation of the u's, especially if the sample is small ($n < 30$).

6.9.3 Berenblut and Webb Test

This test is applicable when the absolute value of ρ is very high. Other conditions required for this test are same as DW test. The test statistic is here—

$$g = \frac{\sum_{t=1}^{n} e_{1t}^2}{\sum_{t=1}^{n} e_t^2}$$

Where e_t is the OLS residual obtained from regressing Y_t on the original explanatory variables with constant term.

But e_{1t} is the OLS residual obtained from regressing ΔY_t on the first differences of explanatory variables with no intercept term.

g is tested using the same bound of DW statistic.

If $g > d_u$ or $g > 4 - d_u$, then there is no autocorrelation. Otherwise either there is autocorrelation or we cannot derive any conclusion regarding the presence of autocorrelation.

6.9.4 Wallis Test

Wallis test is applied for testing fourth order autocorrelation. Fourth order autocorrelation may be found in seasonal data. i.e., quarterly data. It is applied when we have quarterly time series data. Here the test statistic is

$$d_4 = \frac{\sum_{t=5}^{n}(e_t - e_{t-4})^2}{\sum_{t=1}^{n} e_t^2}$$

Here d_4 is also tested on the basis of DW bounds.

If $d_4 > d_u$ or $d_4 > 4 - d_u$ then there is no autocorrelation. Otherwise we get either inconclusive range or there is autocorrelation problem in the disturbance term.

6.9.5 Durbin's h Test

When lagged value of dependent variable is used as an explanatory variable we use Durbin's h test to detect autocorrelation.

Let the model be $y_t = \alpha y_{t-1} + \beta X_t + u_t$

where $u_t = \rho u_{t-1} + \epsilon_t$

Here ϵ_t is white noise and u_t depends on u_{t-1} and y_{t-1} depends on u_{t-1}

So, y_{t-1} and u_{t-1} are not independent.

Therefore the OLS estimates of regression parameters will not be consistent in the presence of autocorrelation when y_{t-1} is used as an independent variable. Here the test statistic is $h = \hat{\rho}\sqrt{\frac{n}{1 - n\hat{v}(\hat{\alpha})}}$ where h approximately follows standard normal distribution. With the null hypothesis $H_0 : \rho = 0$ against. $H_1 : \rho \neq 0$ i.e., if $h \leq 1.96$, there is no autocorrelation problem; otherwise there is autocorrelation problem. This test can be further illustrated with an example.

Numerical Example :

An equation of demand for food estimated from 50 observations gets the following results (figures in the paranthesis are standard errors) :

$$\log q_t = \text{constant} - \underset{(0.05)}{0.31}\log p_t + \underset{(0.20)}{0.45}\log y_t + \underset{(0.14)}{0.65}\log q_{t-1}$$

$R^2 = 0.90$ & DW = 1.8

Where q_t = food consumption per capita

p_t = food price

y_t = per capita disposable income

Apply Durbin's h test for examining the presence of 1st order autocorrelation at 1% level.

Solution :

Here the following information are given,

$$n = 50$$

$$\hat{\alpha} = 0.65$$

$$\text{S.E.}(\hat{\alpha}) = 0.14$$

$$\text{Var}(\hat{\alpha}) = (0.14)^2 = 0.0196$$

$$\text{D. W.} = 1.8$$

We know that

$$\text{DW} = 2(1 - \hat{\rho})$$

$$\text{or, } \hat{\rho} = 1 - \frac{\text{DW}}{2}$$

$$= 1 - \frac{1.8}{2} = 1 - 0.9 = 0.1$$

$$\therefore h = 0.1 \sqrt{\frac{50}{1-50(0.0196)}} = 5.0$$

Here, $H_0 : \rho = 0$ against $H_1 : \rho \neq 0$

h approximately follows standard normal distribution.

At 1% level, table value is 2.58

As observed value of h is greater than table value, H_0 is rejected, So there is autocorrelation in the disturbance term of the problem.

6.9.6 Durbin's t Test

The Durbin's h test cannot be applied if n $\hat{v}(\hat{\alpha}) > 1$. In such a situation, Durbin prescribes an alternative test known as Durbin's t test. This test is explained as follows :

Let the model be $y_t = \alpha y_{t-1} + \beta_1 x_{1t} + \beta_2 x_{2t} + u_t$ .....(1)

where $u_t = \rho u_{t-1} + \epsilon_t$

Here ϵ_t is a white noise.

Here equation (1) is to be estimated using OLS and the OLS residuals are to be computed.

Let the OLS residuals be e_t

Using e_t we estimate the following relation

$$e_t = \rho e_{t-1} + \alpha' y_{t-1} + \beta_1' x_{1t} + \beta_2' x_{2t} + \epsilon_t \quad .....(2)$$

and test the regression parameter of e_{t-1} using t test where $H_0 : \rho = 0$.

If H_0 is accepted using 't' test, there is no autocorrelation problem, otherwise there is autocorrelation problem in the disturbance term. This type of test is known as Durbin's t test.

6.9.7 BG Test or LM Test

Higher order autocorrelation test has been devised by Breusch and Godfrey and following their names, this test is known as BG test. In this test, the general principle of Lagrange multiplier test is available and so this test is also known as LM test.

Let the model be

$$y_t = \alpha + \sum_{j=1}^{s} r_j y_{t-j} + \sum_{i=1}^{k} \beta_i x_{it} + u_t \quad .....(1)$$

Where $u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \cdots\cdots + \rho_p u_{t-p} + \epsilon_t$

Where ϵ_t is white noise. Equation (1) is estimated using OLS and OLS residuals are estimated which be e_t.

Next, the regression equation of e_t on e_{t-1}, e_{t-2}, ..., y_{t-1}, y_{t-2}, ..., x_1, x_2, ... are estimated as follows :

$$\therefore \; e_t = \sum_{i=1}^{p} \rho_i e_{t-1} + \sum_{j=1}^{s} r_j y_{t-j} + \sum_{e=1}^{k} \beta_e x_{et} + \epsilon_t$$

and test the null hypothesis $H_0 = \rho_1 = \rho_2 = \ldots\ldots\ldots = \rho_p = 0$.

This test can be applied using F test based on R^2.

But for large sample, Breusch and Godfrey prescribed that one can apply LM test where test statistic is :

$p.\ F = (n - p)\ R^2 \sim \chi_p^2$

If this test statistic is found to be significant, we can say that there is autocorrelation problem.

6.10 Remedial Measures of Autocorrelation Problem

There are various remedial meaures for solving autocorrelation problem. They are discussed as follows :

6.10.1 Estimating First Difference Equation :

In the presence of autocorrelation, instead of estimating regression equation in level form, the equation should be estimated in its difference form, specially when the value of DW statistic is very low. As a rule of thumb, 1st difference equation can be estimated if DW value is less than R^2 i.e., DW value < R^2 value.

Let the regression model be

$$Y_t = \alpha + \beta X_t + u_t$$

This equation's one period lagged form is

$$Y_{t-1} = \alpha + \beta X_{t-1} = u_{t-1}$$

Deducting this from the first, we get,

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

On the assumption that u_t & u_{t-1} are independent to each other, one can apply OLS to equation (1) to get the estimate of parameters in the presence of autocorrelation problem. It is to be noted that in the difference form, no intercept term is used. The intercept term can be added if it is assumed that there is a trend component.

$$\text{i.e.,}\ Y_t = \alpha + \delta t + \beta X_t + u_t$$

In lagged form,

$$Y_{t-1} = \alpha + \delta\ (t-1) + \beta X_{t-1} + u_{t-1}$$

Subtracting we get,

$$Y_t - Y_{t-1} = \delta + \beta\ (X_t - X_{t-1}) + (u_t - u_{t-1}) \quad \ldots\ldots\ldots\ldots (2)$$

Hence in the presence of autocorrelation either equation (1) or (2) can be estimated as its solution instead of estimating the equation in its level form.

The problem before the researcher is to make a choice between the level form and difference form estimation. We should choose that form whose Residual sum of squares is least. But the RSS of these equations cannot be compared as R^2 cannot be compared with different forms of dependent variable.

The variance of u_t is assumed to be known : Var $(u_t) = \sigma^2$

It is assumed that there is no heteroscedasticity

$$\begin{aligned} &\text{Var}\ (u_t - u_{t-1}) \\ &= \text{Var}\ (u_t) + \text{Var}\ (u_{t-1}) - 2\text{Cov}\ (u_t, u_{t-1}) \\ &= \sigma^2 + \sigma^2 - 2\rho\sigma^2 \\ &= 2\sigma^2 - 2\rho\sigma^2 \\ &= 2\sigma^2\ (1 - \rho) \end{aligned}$$

where $\rho = \dfrac{\text{Cov}(u_t, u_{t-1})}{\text{Var}(u_t)}$

So, Var $(u_t - u_{t-1}) = 2\sigma^2\ (1 - \rho)$

$$= \sigma^2\ DW$$

as $DW = 2\ (1 - \rho)$

σ^2 is estimated from respective RSS

$$E\left(\frac{RSS \text{ Difference form}}{n-k-1}\right) = \sigma^2\ DW \quad \ldots\ldots\ldots (3)$$

Where RSS_D is obtained from difference equations.

$$E\left(\frac{RSS_L}{n-k}\right) = \sigma^2 \quad \text{.....(4)}$$

From (3) we get,

$E(RSS_D) = (n - k - 1)\ \sigma^2\ DW$

From (4) we get,

$$E\left[\frac{RSS_L}{n-k}(n-k-1)DW\right] = (n - k - 1)\ \sigma^2\ DW$$

Hence RSS_D can be compared with adjusted RSS i.e., $\frac{(n-k-1)}{(n-k)}$ $(RSS_L)(DW)$

$$= \frac{n-k-1}{n-k}\ RSS_L\ 2\ (1-\rho)$$

Hence RSS obtained from level form is to be adjusted to make it comparable with RSS obtained from difference form and that equation is to be selected whose RSS after adjustment is minimum.

6.10.2 Estimating Quasi-difference Equation :

When the autocorrelation coefficient ρ is known, quasi-difference equation can be estimated to solve the autocorrelation prbolem.

Let the model be,

$$Y_t = \alpha + \beta X_t + u_t \text{ where } \text{.....(1)}$$

$$u_t = \rho u_{t-1} + \epsilon_t$$

Here ρ is assumed to be known and ϵ_t is a white noise.

Taking one period lag and multiplying by ρ, expression (1) is written as

$$\rho Y_{t-1} = \rho\alpha + \beta\rho X_{t-1} + \rho u_{t-1} \text{ .....(2)}$$

Deducting (2) from (1) we get,

$$(Y_t - \rho Y_{t-1}) = \alpha(1-\rho) + \beta[X_t - \rho X_{t-1}] + [u_t - \rho u_{t-1}] \text{ .....(3)}$$

which is now difference equation or quasi-difference equation.

Expression (3) can be written as $Y_t^* = \alpha' + \beta X_t^* + \epsilon_t$ .....(3′)

As ϵ_t is a white noise, we can apply OLS on this transformed variables to avoid autocorrelation problem.

If (3) is estimated, Y_1 and X_1 are ommited, but those should be included with the following transformation

$$Y_1^* = Y_1\sqrt{1-\rho^2}$$

$$X_1^* = X_1\sqrt{1-\rho^2}$$

Application of OLS on transformed variable is GLS.

Where, $Y_t^* = Y_t - \rho Y_{t-1}, \ \forall \ t = 1, 2, \ldots\ldots n$

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

6.10.3 Durbin's Two-Step Procedures

This method is applied to slove autocorrelation problem when ρ is unknown.

Let the model be

$$Y_t = \alpha + \beta X_t + u_t$$

where $u_t = \rho u_{t-1} + \epsilon_t$ .....(1)

and μ_t follows first order autoregressive scheme and ϵ_t is a white noise. Therefore there is autocorrelation problem in the disturbance term.

So, we can write,

$$\rho Y_{t-1} = \rho\alpha + \rho\beta X_{t-1} + \rho u_{t-1} \quad \ldots\ldots\ldots\ldots\ldots(2)$$

Deducting (2) from (1) we get,

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

i.e., $Y_t = \alpha(1-\rho) + \rho Y_{t-1} + \beta X_t + (-\rho\beta)X_{t-1} + \epsilon_t$ .....(3)

Now, estimating equation (3) by OLS we can estimate the coefficient of Y_{t-1} which is $\hat{\rho}$.

Estimating $\hat{\rho}$ we can transform the variables as follows :

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

$$X_t^* = X_t - \hat{\rho} X_{t-1} \qquad t = 2, 3, \ldots\ldots, n$$

$$Y_1^* = \sqrt{1-\hat{\rho}^2}\, Y_1$$

$$\& \; X_1^* = \sqrt{1-\hat{\rho}^2}\, X_1$$

Now, we can estimate finally the relation

$$Y_t^* = \alpha^* + \beta X_t^* + \epsilon_t$$

Using OLS method we can estimate the parameters.

Here intercept term is to be adjusted as follows :

$$\alpha = \frac{\alpha^*}{1-\hat{\rho}}$$

This method is known as Durbin's two-step procedure.

But this procedure has one limitation. If there are a number of explanatory variables, number of parameters to be estimated will be unduly large, lowering the degree of freedom.

6.10.4 Cochrane Orcutt Iterative Procedure

This method is applied to solve autocorrelation problem when ρ is unknown.

Let the model be

$$Y_t = \alpha + \beta X_{t-1} + u_t \ldots\ldots\ldots\ldots\ldots\ldots(1)$$

where $u_t = \rho u_{t-1} + \epsilon_t$

and ϵ_t is a white noise.

Therefore there is first order autocorrelation problem in the disturbance term.

Using OLS method the residuals are computed and ρ is estimated as follows :

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

Next, the variables are transformed as follows :

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

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

for $t = 2, 3, \ldots\ldots\ldots, n$

$$Y_1^* = \sqrt{1-\hat{\rho}^2}\, Y_1$$

$$X_1^* = \sqrt{1-\hat{\rho}^2}\, X_1$$

The relation (1) with transformed varibles is to be estimated by OLS method. The residuals are then to be calculated and $\hat{\rho}$ is estimated. For a new value of $\hat{\rho}$, the variables are given transformed values and the whole process is repeated until $\hat{\rho}$ converges and after convergence of $\hat{\rho}$ we shall finally estimate the parameters using OLS.

This procedure also has two limitations.

(i) $\hat{\rho}$ may not coverage for successive estimations.

(ii) With this procedure, only first order autocorrelation can be cured. Higher order autocorrelation cannot be cured.

6.10.5 Grid Search Technique

The earlier method i.e., Cochrane-Orcutt iterative procedure suffers from the limitation that $\hat{\rho}$ may not converge. To avoid this limitation Hilderth and *Lu* prescribed Grid Search technique as a solution of autocorrelation problem.

We know that $-1 < \rho < +1$

Let the model be $Y_t = \alpha + \beta X_t + u_t$ .....(1)

where $u_t = \rho u_{t-1} + \epsilon_t$

Let us take all values of ρ with the interval of 0.1 in the range –1 to +1 and estimate the following relation for each $\hat{\rho}$:

$$Y_t^* = \alpha' + \beta X_t^* + \epsilon_t$$

where $Y_t^* = Y_t - \hat{\rho} Y_{t-1}$

and $X_t^* = X_t - \hat{\rho} X_{t-1}$

and finally we select that $\hat{\rho}$ for which *RSS* of equation (1) is minimum.

Let $\hat{\rho}$ be – 0.4

Next, consider all the $\hat{\rho}$ values in the interval 0.01 for the range – 0.5 to – 0.3 and estimate equation (1).

Next, select that $\hat{\rho}$ for which *RSS* is minimum and the same process is repeated for smaller intervals like 0.001, 0.0001, etc. and we shall stop until *RSS* remains constant.

6.10.6 Durbin's Higher Order Technique

Let the model be

$$Y_t = \alpha + \beta X_t + u_t \text{ .....(1)}$$

where $u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \epsilon_t$

and ϵ_t is a white noise i.e., there is autoregressive process of order two in the disturbance term.

According to Durbin's prescription, the following relation is to be estimated.

$$Y_t = A + \rho_1 Y_{t-1} + \rho_2 Y_{t-2} + \beta X_t - \beta\rho_1 X_{t-1} - \beta\rho_2 X_{t-2} + \epsilon_t \text{ .....(2)}$$

It is derived in the following manner. we have,

$$Y_t = \alpha + \beta X_t + u_t$$

$$\rho_1 Y_{t-1} = \rho_1\alpha + \beta\rho_1 X_{t-1} + \rho_1 u_{t-1}$$

$$\rho_2 Y_{t-2} = \rho_2\alpha + \beta\rho_2 X_{t-2} + \rho_2 u_{t-2}$$

So, subtracting we get,

$$\begin{aligned} Y_t - \rho_1 Y_{t-1} - \rho_2 Y_{t-2} &= \alpha(1-\rho_1-\rho_2) + \beta X_t - \beta\rho_1 X_{t-1} - \beta\rho_2 X_{t-2} + u_t - \rho_1 u_{t-1} - \rho_2 u_{t-2} \\ &= A + \beta X_t - \beta\rho_1 X_{t-1} - \beta\rho_2 X_{t-2} + \epsilon_t \end{aligned}$$

$$\text{or, } Y_t = A + \rho_1 Y_{t-1} + \rho_2 Y_{t-2} + \beta X_t - \beta\rho_1 X_{t-1} - \beta\rho_2 X_{t-2} + \epsilon_t$$

Estimating equation (2) we get the estimated values of ρ_1 and ρ_2. Using these values of ρ_1 and ρ_2 the variables are transformed as follows :

$$Y_t^* = Y_t - \hat{\rho}_1 Y_{t-1} - \hat{\rho}_2 Y_{t-2}$$

$$\text{and } X_t^* = X_t - \hat{\rho}_1 X_{t-1} - \hat{\rho}_2 X_{t-2}$$

Equation (1) is then estimated using the transformed variables Y_t^* and X_t^* to get the final estimate of the parameters.

This method is applied for solving higher order autocorrelation problems.

6.11 Some Numerical Problems

1. Consider the model : $Y_t = \alpha + \beta X_t + u_t$ with the following observations on Y and X

X	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Y	2	1	2	3	3	2	5	6	11	10	12	15	10	11	12

Test for autocorrelation.

Solution : For testing autocorrelation, we have to first estimate the regression model and then we have to estimate $e_t = Y_t - \hat{Y}_t$ where $\hat{Y}_t = \hat{\alpha} + \hat{\beta} X_t$ and finally we have to find the DW statistic

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

The following table (3.2) gives necessary calculations to estimate $\hat{\alpha}$ and $\hat{\beta}$ in the model : $Y_t = \alpha + \beta X_t + u_t$

Table 3.2

Y_t	X_t	$y_t = Y_t - \bar{Y}$	y_t^2	$x_t = X_t - \bar{X}$	x_t^2	$x_t y_t$	$\hat{Y}_t = \hat{\alpha} + \hat{\beta} X_t$	$e_t = Y_t - \hat{Y}_t$	e_t^2	$e_t - e_{t-1}$	$\Sigma(e_t - e_{t-1})^2$
2	1	–5	25	–7	49	35	0.322	1.678	2.815	—	—
1	2	–6	36	–6	36	36	1.276	–0.276	0.076	–1.954	3.818
2	3	–5	25	–5	25	25	2.230	–0.230	0.052	0.046	0.002
3	4	–4	16	–4	16	16	3.184	–0.184	0.033	0.046	0.002
3	5	–4	16	–3	9	12	4.138	–1.138	1.295	–0.954	0.910
2	6	–5	25	–2	4	10	5.092	–3.092	9.560	–1.954	3.818
5	7	–2	4	–1	1	2	6.046	–1.046	1.094	2.046	4.186
6	8	–1	1	0	0	0	7.000	–1.000	1.000	0.046	0.002
11	9	4	16	1	1	4	7.954	3.046	9.278	4.046	16.370
10	10	3	9	2	4	6	8.908	1.092	1.192	–1.954	3.818
12	11	5	25	3	9	15	9.862	2.138	4.571	1.046	1.094
15	12	8	64	4	16	32	10.816	4.184	17.505	2.046	4.186
10	13	3	9	5	25	15	11.770	–1.770	3.132	–5.594	35.450
11	14	4	16	6	36	24	12.724	–1.724	2.972	0.046	0.002
12	15	5	25	7	49	35	13.678	–1.678	2.815	0.046	0.002
$\Sigma Y_t = 105$ $\therefore \bar{Y} = \frac{\Sigma Y_t}{n} = \frac{150}{15} = 7$	$\Sigma X_t = 120$ $\bar{X} = \frac{\Sigma X_t}{n} = \frac{120}{15} = 8$	$\Sigma y_t = 0$	$\Sigma y_t^2 = 312$	$\Sigma x_t = 0$	$\Sigma x_t^2 = 280$	$\Sigma x_t y_t = 267$			$\Sigma e_t^2 = 57{,}390$		$\Sigma(e_t - e_{t-1}) = 73{,}660$

Now, we have

$$\hat{\beta} = \frac{\sum x_t y_t}{\sum x_t^2} = \frac{267}{280} = 0.945$$

$$\hat{\alpha} = \overline{Y} - \hat{\beta}\overline{X} = 7 - 0.954 \times 8 = 7 - 7.632 = -0.632$$

So, $$\hat{\sigma}_u^2 = \frac{\sum e_t^2}{n-2} = \frac{57.390}{13} = 4.4146$$

$$\text{Var}(\hat{\beta}) = \frac{\hat{\sigma}_u^2}{\sum x_t^2} = \frac{4.4146}{280} = 0.0157$$

$$\therefore \text{ SE } (\hat{\beta}) = \sqrt{0.0157} = 0.1252$$

$$\text{Var}(\hat{\alpha}) = \hat{\sigma}_u^2 \frac{\sum X_t^2}{n\sum x_t^2}$$

$$= \frac{4.4146 \times 1240}{15 \times 280}$$

$$= \frac{5474.104}{4200} = 1.3036$$

$$\therefore \text{ SE}(\hat{\alpha}) = \sqrt{1.3036} = 1.1417$$

$$R^2 = \hat{\beta}^2 \frac{\sum x_t^2}{\sum y_t^2} = \frac{(0.954)^2 \times 280}{312} = \frac{254.8324}{312} = 0.816$$

Thus the estimated model is $\hat{Y}_t = \underset{(1.1417)}{-0.632} + \underset{(0.1252)}{0.954} X_t$, $R^2 = 0.816$

Now, *DW* statistic is $d^* = \frac{\sum (e_t - e_{t-1})^2}{\sum e_t^2} = \frac{73.660}{57.390} = 1.283$

Values of d_L and d_u at 5% level of significance with n = 15 and one explanatory variable (k' = 1) and d_L = 1.08 and d_u = 1.36. Here we see that $d_L < d^* < d_u$ and hence the test is inconclusive. In other words, on the basis of Durbin-Watson test we cannot say whether autocorrelation problem exists or not.

2. Simple consumption function is estimated from hypothetical data on income ($Y_{d,\ t}$) given in table A.

It is given that the estimated consumption functin is :

$$\hat{C}_t = \underset{(0.0055)}{3.29 + 0.906 Y_{d,t}} \quad R^2 = 0.99$$

(a) Examine whether any autocorrelation problem exists or not.

Solution :

The estimated consumption function is given

$$\hat{C}_t = \underset{(0.0055)}{3.29 + 0.906 Y_{d,t}} \quad R^2 = 0.99$$

(Calculations are given in the table A)

This equation explains almost all variations in consumption. But variance of $\hat{\beta}$ is extremely small.

Now we have to examine the error terms to see whether the evidence of autocorrelation exists or not.

The value of DW statistic is $d^* = \dfrac{\sum (e_t - e_{t-1})^2}{\sum e_t^2} = \dfrac{142.08}{143.14} = 0.9926$

(after substitution of values from the table).

At 5% level of significance for n = 19, k' = 1 (for one explanatory variable), d_L = 1.18 and d_u = 1.40. Here we see that $d^* < d_L$ (= 1.18) and hence we reject null hypothesis of no autocorrelation in favour of alternative hypothesis of positive autocorrelated disturbance terms. i.e., there is positive autocorrelation.

Calculation of d-statistic for $\hat{C}_t = 3.29 + 0.908\,Y_{d,t}$

Year	Consumption Expenditure (c_t) (in Rs)	Disposable Income $Y_{d,t}$ (in Rs)	Estimated consumption $\hat{C}_t = 3.29 + 0.906\,Y_{d,t}$	e_t	e_t^2	e_{t-1}	e_{t-1}^2	$e_t - e_{t-1}$	$(e_t - e_{t-1})^2$	$e_t\, e_{t-1}$
1951	206.3	226.6	208.6	–2.3	5.29	=	=	=	=	=
1952	216.7	238.6	219.2	–2.5	6.25	–2.3	5.29	–0.2	0.04	5.75
1953	230.0	252.6	232.1	–2.1	4.41	–2.5	6.25	0.4	0.16	5.25
1954	236.5	257.4	236.5	0.00	0.00	–2.1	4.41	2.1	4.41	0.00
1955	254.4	275.3	252.7	1.7	2.89	0.0	0.00	1.7	2.89	0.00
1956	266.7	293.2	268.9	–2.2	4.84	1.7	2.89	– 3.9	15.21	–3.74
1957	281.4	308.5	282.8	–1.4	1.96	–2.2	4.84	0.8	0.64	3.08
1958	290.1	328.8	292.1	–2.0	0.04	–1.4	1.96	– 0.6	0.36	2.80
1959	311.2	337.3	308.9	2.3	5.29	–2.0	4.00	4.3	18.49	–4.60
1960	325.2	350.0	320.4	4.8	23.04	2.3	5.29	2.5	6.25	11.04
1961	335.2	364.4	333.4	1.8	3.24	4.8	23.04	– 3.0	9.00	8.64
1962	355.1	385.5	352.4	2.7	7.29	1.8	3.24	0.9	0.81	4.86
1963	375.0	404.6	369.9	5.1	26.01	2.7	7.29	2.4	5.76	13.77
1964	401.2	438.1	400.2	1.0	1.00	5.1	26.01	– 4.1	16.81	5.1
1965	432.8	437.2	432.0	0.8	0.64	1.0	1.00	–0.2	0.04	0.8

6.12 Summary

Autocorrelation (also known as serial correlation) is an econometric problem in which the current value of an error term is correlated with its past values. In CLRM it is assumed that the disturbance terms are independent to each other and if this assumption is violated then we observe autocorrelation problem.

This autocorrelation problem is mainly observed in time series data. Autocorrelation problem in time series arises due to various reasons like inertia, omission of explanatory variables, incorrect functional form, presence of lagged variable as explanatory variables, manipulation of data, transformation of data, non-stationary data etc. Presence of autocorrelation problem in any data leads of underestimation of variance of regression coefficient, underestimation of variance of disturbance term, OLS estimates of regression coefficient become inefficient and the predictions on the basis of OLS estimates are also inefficient. The autocorrelation problem in any data can be tested by various tests. The popular tests for detecting autocorrelation problem are Durbin-Waston test, Von Neumann Ratio test, Bercublut and Webb test, Wallis test, etc. These tests are applied if the explained variable is not used as an explanatory variable in lagged form. But if the lagged value of explained variable is used as an explanatory variable then Durbin's t test and Durbin's h test are popularly used.

Autocorrelation problem can be sloved by estimating first difference equation or by estimating quasi difference equation. If ρ is unknown, then Durbin's two-step procedure, Cochrane-Orcutt Iterative procedure, Grid search technique and Durbin's Higher Order Technique are used to solve the autocorrelation problem.

6.13 Exercise

Short Answer Type Questions :

(a) Choose the correct answer :

(i) In CLRM, if the assumption of independence of disturbance term is dropped, then we get the problem of _____

(a) Heteroscedasticity

(b) Multicollinearity

(c) Autocorrelation

(d) None of the above

(ii) Which of the following is not a popular test of detecting autocorrelation problem?

(a) Durbin-Watson test

(b) C N test

(c) Von Neumann Ratio test

(d) Berenblut & Webb test

(b) State whether the statements are true or false.

(i) Autocorrelation problem is generally observed in time series data unlike heteroscedasticity which is seen in cross section data also.

(ii) In the presence of autocorrelation problem, the OLS estimate of variance of regression coefficient ($\hat{\beta}$) overstates the true variance.

(c) Fill in the blanks :

(i) Two popular solutions of autocorrelation problem are _____ and _____.

(ii) If the value of DW statistic is less than the lower bound d_L then there autocorrelation is _____.

Medium Answer Type Questions :

1. Write a short note on DW test for autocorrelation.
2. Discuss in brief any one way to solve the autocorrelation problem.
3. Define autocorrelation. Describe in brief the structure of autocorrelation problem.

Long Answer Type Questions :

1. Define autocorrelation. Describe the structure of autocorrelation problem. What are the major sources of autocorrelation problem?
2. How can autocorrelation problem be detected? Describe different tests to detect autocorrelation problem.
3. How can autocorrelation problem be solved? Discuss different methods in details.

6.14 References

1. Gujarati, D (2003) : *Basic Econometrics*, McGraw Hill Higher Education
2. Kuotsoyannis, A (1996) : *Theory of Econometrics*, ELBS with Macmillan
3. Sarkhel, Jaydeb and Santosh Kumar Dutta (2020) : *An Introduction to Econometrics*, Book Syndicate Private Limited.

APPENDIX
STATISTICAL TABLES
TABLE I
ORDINATES AND AREA OF THE DISTRIBUTION OF STANDARD NORMAL VARIABLE*

τ	$\phi(\tau)$	$\Phi(\tau)$	τ	$\phi(\tau)$	$\Phi(\tau)$	τ	$\phi(\tau)$	$\Phi(\tau)$
.00	.3989423	.5000000						
.01	.3989223	.5039894	.51	.3502919	.6949743	1.01	.2395511	.8437524
.02	.3988625	.5079783	.52	.3484925	.6984682	1.02	.2371320	.8461358
.03	.3987628	.5119665	.53	.3466677	.7019440	1.03	.2347138	.8484950
.04	.3986233	.5159534	.54	.3448180	.7054015	1.04	.2322970	.8508300
.05	.3984439	.5199388	.55	.3429439	.7088403	1.05	.2298821	.8531409
.06	.3982248	.5239222	.56	.3410458	.7122603	1.06	.2274696	.8554277
.07	.3979661	.5279032	.57	.3391243	.7156612	1.07	.2250599	.8576903
.08	.3976677	.5318814	.58	.3371799	.7190427	1.08	.2226535	.8599289
.09	.3973298	.5358564	.59	.3352132	.7224047	1.09	.2202508	.8621434
.10	.3969525	.5398278	.60	.3332246	.7257469	1.10	.2178522	.8643339
.11	.2965360	.5437953	.61	.3312147	.7290691	1.11	.2154582	.8665005
.12	.3960802	.5477584	.62	.3291840	.7323711	1.12	.2130691	.8686431
.13	.3955854	.5517168	.63	.3271330	.7356527	1.13	.2106856	.8707619
.14	.3950517	.5556700	.64	.3250623	.7389137	1.14	.2083078	.8728568
.15	.3944793	.5596177	.65	.3229724	.7421539	1.15	.2059363	.8749281
.16	.3938684	.5635595	.66	.3208638	.7453731	1.16	.2035714	.8769756
.17	.3932190	.5674949	.67	.3187371	.7485711	1.17	.2012135	.8789995
.18	.3925315	.5714237	.68	.3165929	.7517478	1.18	.1988631	.8809999
.19	.3918060	.5753454	.69	.3144317	.7549029	1.19	.1965205	.8829768
.20	.3910427	.5792597	.70	.3122539	.7580363	1.20	.1941861	.8849303
.21	.3902419	.5831662	.71	.3100603	.7611479	1.21	.1918602	.8868606
.22	.3894038	.5870644	.72	.3078513	.7642375	1.22	.1895432	.8887676
.23	.3885286	.5909541	.73	.3056274	.7673049	1.23	.1872354	.8906514
.24	.3876166	.5948349	.74	.3033893	.7703500	1.24	.1849373	.8925123
.25	.3866681	.5987063	.75	.3011374	.7733726	1.25	.1826491	.8943502
.26	.3856834	.6025681	.76	.2988724	.7763727	1.26	.1803712	.8961653
.27	.3846627	.6064199	.77	.2965948	.7793501	1.27	.1781038	.8979577
.28	.3836063	.6102612	.78	.2943050	.7823046	1.28	.1758474	.8997274
.29	.3825146	.6140919	.79	.2920038	.7852361	1.29	.1736022	.9014747
.30	.3813878	.6179114	.80	.2896916	.7881446	1.30	.1713686	.9031995

TABLE I (Contd.)

τ	$\phi(\tau)$	$\Phi(\tau)$	τ	$\phi(\tau)$	$\Phi(\tau)$	τ	$\phi(\tau)$	$\Phi(\tau)$
.31	.3802264	.6217195	.81	.2873689	.7910299	1.31	.1691468	.9049021
.32	.3790305	.6255158	.82	.2850364	.7938919	1.32	.1669370	.9065825
.33	.3778007	.6293000	.83	.2826945	.7967306	1.33	.1647397	.9082409
.34	.3765372	.6330717	.84	.2803438	.7995458	1.34	.1625551	.9098773
.35	.3752403	.6368307	.85	.2779849	.8023375	1.35	.1603833	.9114920
.36	.3739106	.6405764	.86	.2756182	.8051055	1.36	.1582248	.9130850
.37	.3725483	.6443088	.87	.2732444	.8078498	1.37	.1560797	.9146565
.38	.3711539	.6480273	.88	.2708640	.8105703	1.38	.1539483	.9162067
.39	.3697277	.6517317	.89	.2684774	.8132671	1.39	.1518308	.9177356
.40	.3682701	.6554217	.90	.2660852	.8159399	1.40	.1497275	.9192433
.41	.3667817	.6590970	.91	.2636880	.8185887	1.41	.1476385	.9207302
.42	.3652627	.6627573	.92	.2612863	.8212136	1.42	.1455641	.9221962
.43	.3637136	.6664022	.93	.2588805	.8238145	1.43	.1435046	.9236415
.44	.3621349	.6700314	.94	.2564713	.8263912	1.44	.1414600	.9250663
.45	.3605270	.6736448	.95	.2540591	.8289439	1.45	.1394306	.9264707
.46	.3588903	.6772419	.96	.2516443	.8314724	1.46	.1374165	.9278550
.47	.3572253	.6808225	.97	.2492277	.8339768	1.47	.1354181	.9292191
.48	.3555325	.6843863	.98	.2468095	.8364569	1.48	.1334353	.9305634
.49	.3538124	.6879331	.99	.2443904	.8389129	1.49	.1314684	.9318879
.50	.3520653	.6914625	1.00	.2419707	.8413447	1.50	.1295176	.9331928
1.51	.1275830	.9344783	2.01	.0529192	.9777844	2.51	0.170947	.9939634
1.52	.1256646	.9357445	2.02	.0518636	.9783083	2.52	.0166701	.9941323
1.53	.1237628	.9369916	2.04	.0508239	.9788217	2.53	.0162545	.9942969
1.54	.1218775	.9382198	2.04	.0498001	.9793248	2.54	.0158476	.9944574
1.55	.1200090	.9394292	2.05	.0487920	.9798178	2.55	.0154493	.9946139
1.56	.1181573	.9406201	2.06	.0477996	.9803007	2.56	.0150596	.9947664
1.57	.1163225	.9417924	2.07	.0468226	.9807738	2.57	.0146782	.9949151
1.58	.1145048	.9429466	2.08	.0458611	.9812372	2.58	.0143051	.9950600
1.59	.1127042	.9440826	2.09	.0449148	.9816911	2.59	.0139401	.9952012
1.60	.1109208	.9452007	2.10	.0439836	.9821356	2.60	.0135830	.9953388
1.61	.1091548	.9463011	2.11	.0430674	.9825708	2.61	.0132337	.9954729
1.62	.1074061	.9473839	2.12	.0421661	.9829970	2.62	.0128921	.9956035
1.63	.1056748	.9484493	2.13	.0412795	.9834142	2.63	.0125581	.9957308
1.64	.1039611	.9494974	2.14	.0404076	.9838226	2.64	.0122315	.9958547
1.65	.1022649	.9505285	2.15	.0395500	.9842224	2.65	.0119122	.9959754

TABLE I (Contd.)

τ	ϕ(τ)	Φ(τ)	τ	ϕ(τ)	Φ(τ)	τ	ϕ(τ)	Φ(τ)
1.66	.1005864	.9515428	2.16	.0387069	.9846137	2.66	.0116001	.9960930
1.67	.0989255	.9525403	2.17	.0378779	.9849966	2.67	.0112951	.9962074
1.68	.0972823	.09535213	2.18	.0370629	.9853713	2.68	.0109969	.9963189
1.69	.0956568	.9544860	2.19	.0362619	.9857379	2.69	.0107056	.9964274
1.70	.0940491	.9554345	2.20	.0354746	.9860966	2.70	.0104209	.9965330
1.71	.0924591	.9563671	2.21	.0347009	.9864474	2.71	.0101428	.9966358
1.72	.0908870	.9572838	2.22	.0339408	.9867906	2.72	.0098712	.9967359
1.73	.0893326	.9581849	2.23	.0331939	.9871263	2.73	.0096058	.9968333
1.74	.0877961	.9590705	2.24	.0324603	.9874545	2.74	.0093466	.9969280
1.75	.0862773	.9599408	2.25	.0317397	.9877755	2.75	.0090936	.9970202
1.76	.0847764	.9607961	2.26	.0310319	.9880894	2.76	.0088465	.9971099
1.77	.0832932	.9616364	2.27	.0303370	.9883962	2.77	.0086052	.9971972
1.78	.0818278	.9624620	2.28	.0296546	.9886962	2.78	.0083697	.9972821
1.79	.0803801	.9632730	2.29	.0289847	.9889393	2.79	.0081398	.9973646
1.80	.0789502	.9640697	2.30	.0283270	.9892759	2.80	.0079155	.9974449
1.81	.0775379	.9648521	2.31	.0276816	.9895559	2.81	.0076965	.9975229
1.82	.0761433	.9656205	2.32	.0270481	.9898296	2.82	.0074829	.9975988
1.83	.0747663	.9663750	2.33	.0264265	.9900969	2.83	.0072744	.9976726
1.84	.0734068	.9671159	2.34	.0258166	.9903581	2.84	.0070711	.9977443
1.85	.0720649	.9678432	2.35	.0252182	.9906133	2.85	.0068728	.9978140
1.86	.0707404	.9685572	2.36	.0246313	.9908625	2.86	.0066793	.9978818
1.87	.0694333	.9692581	2.37	.0240556	.9911060	2.87	.0064907	.9979476
1.88	.0681436	.9699460	2.38	.0234910	.9913437	2.88	.0063067	.9980116
1.89	.0668711	.9706210	2.39	.0229374	.9915758	2.89	.0061274	.9980738
1.90	.0656158	.9712834	2.40	.0223945	.9918025	2.90	.0059525	.9981342
1.91	.0643777	.9719334	2.41	.0218624	.9920237	2.91	.0057821	.9981929
1.92	.0631566	.9725711	2.42	.0213407	.9922397	2.92	.0056160	.9982498
1.93	.0619524	.9731966	2.43	.0208294	.9924506	2.93	.0054541	.9983052
1.94	.0607652	.9738102	2.44	.0203284	.9926564	2.94	.0052963	.9983589
1.95	.0595947	.9744119	2.45	.0198374	.9928572	2.95	.0051426	.9984111
1.96	.0584409	.9750021	2.46	.0193563	.9930531	2.96	.0049929	.9984618
1.97	.0573038	.9755808	2.47	.0188850	.9932443	2.97	.0048470	.9985110
1.98	.0561831	.9761482	2.48	.0184233	.9934309	2.98	.0047050	.9985588
1.99	.0550789	.9767045	2.49	.0179711	.9936128	2.99	.0045666	.9986051
2.00	.0539910	.9772499	2.50	.0175283	.9937903	3.00	.0044318	.9986501

TABLE I (Contd.)

τ	$\phi(\tau)$	$\Phi(\tau)$	τ	$\phi(\tau)$	$\Phi(\tau)$	τ	$\phi(\tau)$	$\Phi(\tau)$
3.01	.0043007	.9986938	3.21	.0023089	.9993363	3.41	.0011910	.9996752
3.02	.0041729	.9987361	3.22	.0022358	.9993590	3.42	.0011510	.9996869
3.03	.0040486	.9987772	3.23	.0021649	.9993810	3.43	.0011122	.9996982
3.04	.0039276	.9988171	3.24	.0020960	.9994024	3.44	.0010747	.9997091
3.05	.0038098	.9988558	3.25	.0020290	.9994230	3.46	.0010383	.9997197
3.06	.0036951	.9988933	3.26	.0019641	.9994429	3.46	.0010030	.9997299
3.07	.0035836	.9989297	3.27	.0019010	.9994623	3.47	.0009689	.9997398
3.08	.0034751	.9989650	3.28	.0018397	.9994810	3.48	.0009358	.9997493
3.09	.0033695	.9989992	3.29	.0017803	.9994991	3.49	.0009037	.9997585
3.10	.0032668	.9990324	3.30	.0017226	.9995166	3.50	.0008727	.9997674
3.11	.0031669	.9990646	3.31	.0016666	.9995335	3.51	.0008436	.9997759
3.12	.0030698	.9990957	3.32	.0016122	.9995499	3.52	.0008135	.9997842
3.13	.0029754	.9992360	3.33	.0015595	.9995658	3.53	.0007853	.9997922
3.14	.0028835	.9991553	3.34	.0015084	.9995811	3.54	.0007581	.9997999
3.15	.0027943	.9991836	3.35	.0014587	.9995959	3.55	.0007317	.9998146
3.16	.0027075	.9992112	3.36	.0014106	.9996103	3.56	.0007001	.9998146
3.17	.0026231	.9992378	3.37	.0013639	.9996242	3.57	.0006814	.9998146
3.18	.0025412	.9992636	3.38	.0013187	.9996376	3.58	.0006575	.9998282
3.19	.0024615	.9992886	3.39	.0012748	.9996505	3.59	.0006343	.9998347
3.20	.0023841	.9993129	3.40	.0012322	.9996631	3.60	.0006119	.9998409

*Abridged from Table 1 of *Biometrika Tables for Statisticians*, vol. I, with the kind permission of the Biometrika Trustees.

TABLE II
DISTRIBUTION OF STANDARD NORMAL VARIABLE
Values of τ_α

α	0.05	0.025	0.01	0.005
τ_α	1.645	1.960	2.326	2.576

TABLE III
χ^2 – DISTRIBUTION*
VALUES OF $\chi^2_{\alpha,\upsilon}$

υ \ α	0.995	0.99	0.975	0.95	0.05	0.025	0.01	0.005
1	0.000	0.000	0.001	0.004	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	5.991	7.378	0.210	10.597
3	0.072	0.115	0.216	0.352	7.815	9.348	11.345	12.828
4	0.207	0.297	0.484	0.711	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	11.070	12.832	15.086	16.750
6	0.676	0.872	1.237	1.635	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	28.869	31.526	34.805	37.156
19	6.844	7.833	8.907	10.117	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	33.924	36.781	40.289	42.796
23	9.260	10.196	11.688	13.091	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	36.415	39.364	42.980	45.558
25	10.520	11.524	13.120	14.611	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	40.113	43.194	46.963	49.645
28	12.461	13.565	15.308	16.928	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	43.773	46.979	50.892	53.672
40	20.706	22.164	24.433	26.509	55.759	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	67.505	71.420	76.154	79.490
60	35.535	37.485	40.482	43.188	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	90531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	124.342	129.561	135.807	140.169

For larger values of ν, the quantity $\sqrt{2\chi^2} - \sqrt{2\nu - 1}$ may be used as a standard normal variable.

*Abridged from Table 8 of *Biometrika Tables for Statisticians,* vol. I, with the kind permission of the Biometrika Trustees.

TABLE IV
t-DISTRIBUTION*
Values of $t_{\alpha,\upsilon}$

υ \ α	0.05	0.025	0.01	0.005
1	6.314	12.706	31.821	63.657
2	2.920	4.303	6.965	9.925
3	2.353	3182	4.541	5.841
4	2.132	2.776	3.747	4.604
5	2.015	2.571	3.365	4.032
6	1.943	2.447	3.143	3.707
7	1.895	2.365	2.998	3.499
8	1.860	2.306	2.896	3.355
9	1.833	2.262	2.821	3.250
10	1.812	2.228	2.764	3.169
11	1.796	2.201	2.718	3.106
12	1.782	2.179	2.681	3.055
13	1.771	2.160	2.650	3.012
14	1.761	2.145	2.624	2.977
15	1.753	2.131	2.602	2.947
16	1.746	2.120	2.583	2.921
17	1.740	2.110	2.567	2.898
18	1.734	2.101	2.552	2.878
19	1.729	2.093	2.539	2.861
20	1.725	2.086	2.528	2.845
21	1.721	2.080	2.518	2.831
22	1.717	2.074	2.508	2.819
23	1.714	2.069	2.500	2.807
24	1.711	2.064	2.492	2.797
25	1.708	2.060	2.485	2.787
26	1.706	2.056	2.479	2.779
27	1.703	2.052	2.473	2.771
28	1.701	2.048	2.467	2.763
29	1.699	2.045	2.462	2.756
30	1.697	2.042	2.457	2.750
40	1.684	2.021	2.423	2.704
60	1.671	2.000	2.390	2.660
120	1.658	1.980	2.358	2.617
∞	1.645	1.960	2.326	2.576

*Abridged from Table 1 of *Biometrika Tables for Statisticians*, vol. I, with the kind permission of the Biometrika Trustees.

TABLE V
F-DISTRIBUTION*
Values of F0.05; v1,v2

υ_2 \ υ_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∝
1	161.4	199.5	215.7	224.6	230.2	234.2	236.8	238.9	240.5	241.9	243.9	245.9	248.0	249.1	250.1	251.1	252.2	253.3	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84

TABLE V (Contd.)

υ_2 \ υ_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∝
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
∝	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

For other values of v_1 and v_2, one may use linear interpolation, taking l/v_1 and l/v_2 as the independent variables.

TABLE V (Contd.)

Values of $F_{0.01, v_1, v_2}$

v_2 \ v_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∝
1	4052	4999.5	5403	5625	5764	5859	5928	5982	6022	6056	6106	6157	6209	6235	6261	6287	6313	6339	6366
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.42	99.43	99.45	99.46	99.47	99.47	99.48	99.49	99.50
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	27.05	26.87	26.69	26.60	26.50	26.41	26.32	26.22	26.13
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.37	14.20	14.02	13.93	13.84	13.75	13.65	13.56	13.46
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.88
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.86
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.91
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.17
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.00
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.75
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.65
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.57
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.49

Table V (Contd.)

v_2 \ v_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∝
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.31
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	2.96	2.81	2.66	2.58	2.50	2.42	2.33	2.23	2.13
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.90	2.75	2.60	2.52	2.44	2.35	2.26	2.17	2.06
30	7.56	5.34	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.80
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.60
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.38
∝	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	1.00

For other values of v_1 and v_2, one may use linear interpolation, taking $1/v_1$ and $1/v_2$ as the independent variables.

*Abridged from Table 18 of Biometrika Tables for Statisticians, vol. I, with the kind permission of the Biometrika Trustees.

TABLE VI
THE DURBIN-WATSON d-STATISTIC
SIGNIFICANCE POINTS OF *dt.* AND *dv* : 5%

	$k' = 1$		$k' = 2$		$k' = 3$		$k' = 4$		$k' = 5$	
n	d_L	*du*	d_L	*du*	d_L	*du*	d_L	*du*	d_L	*du*
15	1.08	1.36	0.95	1.54	0.82	1.75	0.69	1.97	0.56	2.21
16	1.10	1.37	0.98	1.54	0.86	1.73	0.74	1.93	0.62	2.15
17	1.13	1.38	1.02	1.54	0.90	1.71	0.78	1.90	0.67	2.10
18	1.16	1.39	1.05	1.53	0.93	1.69	0.82	1.87	0.71	2.06
19	1.18	1.40	1.08	1.53	0.97	1.68	0.88	1.85	0.75	2.02
20	1.20	1.41	1.10	1.54	1.00	1.68	0.90	1.83	0.79	1.99
21	1.22	1.42	1.13	1.54	1.03	1.67	0.93	1.81	0.83	1.96
22	1.24	1.43	1.15	1.54	1.05	1.66	0.96	1.80	0.86	1.94
23	1.26	1.44	1.17	1.54	1.08	1.66	0.99	1.79	0.90	1.92
24	1.27	1.45	1.19	1.55	1.10	1.66	1.01	1.78	0.93	1.90
25	1.29	1.45	1.21	1.55	1.12	1.66	1.04	1.77	0.95	1.89
26	1.30	1.46	1.22	1.55	1.14	1.65	1.60	1.76	0.98	1.88
27	1.32	1.47	1.24	1.56	1.16	1.65	1.08	1.76	1.01	1.86
28	1.33	1.48	1.26	1.56	1.18	1.65	1.10	1.75	1.03	1.85
29	1.34	1.48	1.27	1.56	1.20	1.65	1.12	1.74	1.05	1.84
30	1.35	1.49	1.28	1.57	1.21	1.65	1.14	1.74	1.07	1.83
31	1.36	1.50	1.30	1.57	1.23	1.65	1.16	1.74	1.09	1.83
32	1.37	1.50	1.31	1.57	1.24	1.65	1.18	1.73	1.11	1.82
33	1.38	1.51	1.32	1.58	1.26	1.65	1.19	1.73	1.13	1.81
34	1.39	1.51	1.33	1.58	1.27	1.65	1.21	1.73	1.15	1.81
35	1.40	1.52	1.34	1.58	1.28	1.65	1.22	1.73	1.16	1.80
36	1.41	1.52	1.35	1.59	1.29	1.65	1.24	1.73	1.18	1.80
37	1.42	1.53	1.36	1.59	1.31	1.66	1.25	1.72	1.19	1.80
38	1.43	1.54	1.37	1.59	1.32	1.66	1.26	1.72	1.21	1.79
39	1.43	1.54	1.38	1.60	1.33	1.66	1.27	1.72	1.22	1.79
40	1.44	1.54	1.39	1.60	1.34	1.66	1.29	1.72	1.23	1.79
45	1.41	1.57	1.43	1.62	1.38	1.67	1.34	1.72	1.29	1.78
50	1.50	1.59	1.46	1.63	1.42	1.67	1.38	1.72	1.34	1.77
55	1.53	1.60	1.49	1.64	1.45	1.68	1.41	1.72	1.38	1.77
60	1.55	1.62	1.51	1.65	1.48	1.69	1.44	1.73	1.41	1.77
65	1.57	1.63	1.54	1.66	1.50	1.70	1.47	1.73	1.44	1.77
70	1.58	1.64	1.55	1.67	1.52	1.70	1.49	1.74	1.46	1.77
75	1.60	1.S5	1.57	1.68	1.54	1.71	1.51	1.74	1.49	1.77
80	1.61	1.66	1.59	1.69	1.56	1.72	1.53	1.74	1.51	1.77
85	1.62	1.67	1.60	1.70	1.57	1.72	1.55	1.75	1.52	1.77
90	1.63	1.68	1.61	1.70	1.59	1.73	1.57	1.75	1.54	1.78
95	1.64	1.69	1.62	1.71	1.60	1.73	1.58	1.75	1.56	1.78
100	1.65	1.69	1.63	1.72	1.61	1.74	1.59	1.76	1.57	1.78

Note : k' = Number of explanatory variables excluding the constant.

NOTES