

Course: Econometrics
Department: Agricultural Economics
Instructor: Bewuketu Minwuye (MSc in Agricultural and Applied Economics)

Chapter 1: What is Econometrics?

What it is all about?

1.1. Definition and Scope

The economic theories we learn in various economics courses suggest many relationships among economic variables. For instance, in microeconomics we learn demand and supply models in which the quantities demanded and supplied of a good depend on its price. In macroeconomics, we study ‘investment function’ to explain the amount of aggregate investment in the economy as the rate of interest changes; and ‘consumption function’ that relates aggregate consumption to the level of aggregate disposable income.

Each of such specifications involves a relationship among economic variables. As economists, we may be interested in questions such as: If one variable changes in a certain magnitude, by how much will another variable change? Also, given that we know the value of one variable; can we forecast or predict the corresponding value of another? The purpose of studying the relationships among economic variables and attempting to answer questions of the type raised here, is to help us understand the real economic world we live in.

However, economic theories that postulate the relationships between economic variables have to be checked against data obtained from the real world. If empirical data verify the relationship proposed by economic theory, we accept the theory as valid. If the theory is incompatible with the observed behavior, we either reject the theory or in the light of the empirical evidence of the data, modify the theory. To provide a better understanding of economic relationships and a better guidance for economic policy making we also need to know the quantitative relationships between the different economic variables. We obtain these quantitative measurements taken from the real world. *The field of knowledge which helps us to carry out such an evaluation of economic theories in empirical terms is econometrics.*

WHAT IS ECONOMETRICS?

Literally interpreted, econometrics means “economic measurement”, but the scope of econometrics is much broader as described by leading econometricians. Various econometricians

used different ways of wordings to define econometrics. But if we distill the fundamental features/concepts of all the definitions, we may obtain the following definition.

“Econometrics is the science which integrates economic theory, economic statistics, and mathematical economics to investigate the empirical support of the general schematic law established by economic theory. It is a special type of economic analysis and research in which the general economic theories, formulated in mathematical terms, is combined with empirical measurements of economic phenomena. Starting from the relationships of economic theory, we express them in mathematical terms so that they can be measured. We then use specific methods, called econometric methods in order to obtain numerical estimates of the coefficients of the economic relationships.”

Measurement is an important aspect of econometrics. However, the scope of econometrics is much broader than measurement. As D. Intriligator rightly stated the “metric” part of the word econometrics signifies ‘measurement’, and hence econometrics is basically concerned with measuring of economic relationships.

In short, econometrics may be considered as the integration of economics, mathematics, and statistics for the purpose of providing numerical values for the parameters of economic relationships and verifying economic theories.

1.2.Goals of Econometrics

There are three main goals of econometrics:

- (1) Analysis, i.e. testing of economic theory
- (2) Policy making, i.e. supplying numerical estimates of the coefficients of economic relationships, this may be then used for decision making / policy simulation.
- (3) Forecasting, i.e. using the numerical estimates of the coefficients in order to forecast the future values of the economic magnitudes. Of course, these goals are not mutually exclusive. Successful econometric applications should include some combinations of all three aims.

1.3. Methodology of Econometrics

Econometric research is concerned with the measurement of the parameters of economic relationships and with the prediction of the values of economic variables. The relationships of economic theory which can be measured with econometric techniques are relationships in which some variables are postulated as causes of the variation of other variables. Starting with the

postulated theoretical relationships among economic variables, econometric research or inquiry generally proceeds along the following lines/stages.

1. Specification the model
2. Estimation of the model
3. Evaluation of the estimates
4. Evaluation of the forecasting power of the estimated model

1. Specification of the model

In this step the econometrician has to express the relationships between economic variables in mathematical form. This step involves the determination of three important tasks:

- i) The dependent and independent (explanatory) variables which will be included in the model.
- ii) The *a priori* theoretical expectations about the size and sign of the parameters of the function.
- iii) The mathematical form of the model (number of equations, specific form of the equations, etc.)

Note: The specification of the econometric model will be based on economic theory and on any available information related to the phenomena under investigation. Thus, specification of the econometric model presupposes knowledge of economic theory and familiarity with the particular phenomenon being studied.

Specification of the model is the most important and the most difficult stage of any econometric research. It is often the weakest point of most econometric applications. In this stage there exists enormous degree of likelihood of committing errors or incorrectly specifying the model. Some of the common reasons for incorrect specification of the econometric models are:

1. The imperfections, looseness of statements in economic theories.
2. The limitation of our knowledge of the factors which are operative in any particular case.
3. The formidable obstacles presented by data requirements in the estimation of large models.

The most common errors of specification are:

- a. Omissions of some important variables from the function.
- b. The omissions of some equations (for example, in simultaneous equations model).

- c. The mistaken mathematical form of the functions.

2. Estimation of the model

This is purely a technical stage which requires knowledge of the various econometric methods, their assumptions and the economic implications for the estimates of the parameters. This stage includes the following activities.

- a. Gathering of the data on the variables included in the model.
- b. Examination of the identification conditions of the function (especially for simultaneous equations models).
- c. Examination of the aggregations problems involved in the variables of the function.
- d. Examination of the degree of correlation between the explanatory variables (i.e. examination of the problem of multi-collinearity).
- e. Choice of appropriate economic techniques for estimation, i.e. to decide a specific econometric method to be applied in estimation; such as, OLS, MLM, Logit, and Probit.

3. Evaluation of the estimates

This stage consists of deciding whether the estimates of the parameters are *theoretically meaningful and statistically satisfactory*. This stage enables the econometrician to evaluate the results of calculations and determine the reliability of the results. For this purpose we use various criteria which may be classified into three groups:

- i. *Economic a priori criteria*: These criteria are determined by economic theory and refer to the size and sign of the parameters of economic relationships.
- ii. *Statistical criteria (first-order tests)*: These are determined by statistical theory and aim at the evaluation of the statistical reliability of the estimates of the parameters of the model. Correlation coefficient test, standard error test, t-test, F-test, and R^2 -test are some of the most commonly used statistical tests.
- iii. *Econometric criteria (second-order tests)*: These are set by the theory of econometrics and aim at the investigation of whether the assumptions of the econometric method employed are satisfied or not in any particular case. The econometric criteria serve as a second order test (as test of the statistical tests) i.e. they determine the reliability of the statistical criteria; they help us establish whether the estimates have the desirable properties of unbiasedness, consistency etc.

Econometric criteria aim at the detection of the violation or validity of the assumptions of the various econometric techniques.

4) Evaluation of the forecasting power of the model:

Forecasting is one of the aims of econometric research. However, before using an estimated model for forecasting, by some way or another, the predictive power of the model should be evaluated. It is possible that the model may be economically meaningful and statistically and econometrically correct for the sample period for which the model has been estimated; yet it may not be suitable for forecasting due to various factors (reasons). Therefore, this stage involves the investigation of the stability of the estimates and their sensitivity to changes in the size of the sample. Consequently, we must establish whether the estimated function performs adequately outside the sample of data i.e. we must test an extra sample performance the model.

1.4. Elements of Econometrics

1. Data

Collecting and coding the sample data, the raw material of econometrics. Most economic data is observational or non-experimental data (as distinct from experimental data generated under controlled experimental conditions).

2. Specification

Specification of the econometric model that we think (hope) generated the sample data, that is, specification of the data generating process (or DGP).

An econometric model consists of two components:

1. An economic model: specifies the dependent or outcome variable to be explained and the independent or explanatory variables that we think are related to the dependent variable of interest.
 - Often suggested or derived from economic theory.
 - Sometimes obtained from informal intuition and observation.
2. A statistical model: specifies the statistical elements of the relationship under investigation, in particular the statistical properties of the random variables in the relationship.

3. Estimation

It consists of using the assembled sample data on the observable variables in the model to compute estimates of the numerical values of all the unknown parameters in the model.

4. Inference

Consists of using the parameter estimates computed from sample data to test hypotheses about the numerical values of the unknown population parameters that describe the behavior of the population from which the sample was selected.

1.5.Types of Econometrics

Econometrics may be divided into two broad categories: theoretical econometrics and applied econometrics.

Theoretical econometrics is concerned with the development of appropriate methods for measuring economic relationships specified by econometric models. In this aspect, econometrics leans heavily on mathematical statistics. For example, one of the methods used extensively in econometrics is least squares. Theoretical econometrics must spell out the assumptions of this method, its properties, and what happens to these properties when one or more of the assumptions of the method are not fulfilled.

In applied econometrics we use the tools of theoretical econometrics to study some special field(s) of economics and business, such as the production function, investment function, demand and supply functions, etc.

16. Types of Data

- Different types of datasets have their own issues, advantages and limitations.
- Some econometric methods may be valid (i.e., have good properties) for some types of data but not for others.
- We typically distinguish four types of datasets:
 - Cross-sectional data
 - Time series data
 - Pooled cross-sectional data
 - Panel data or longitudinal data

Cross-sectional data

- A cross-sectional dataset is a sample of individuals, or households, or firms, or cities, or states, or countries, etc, taken at a given point in time.

- We often assume that these data have been obtained by **random sampling**.
- Sometimes we do not have a random sample: sample selection problem; spatial correlation; stratified samples.

Time series data

- It consists of observations on a variable or several variables over several periods of time (days, weeks, months, years).
- A key feature of time series data is that, typically, observations are correlated across time. We do not have a random sample.
- Correlation introduces very important issues in the estimation and testing of econometric models using time series data.
- Seasonality is another common feature in many weekly, monthly or quarterly time series data.

Pooled cross sections

- Suppose that we have a **sequence of cross sections** of the same variables from the same population at years 1990, 1991, 1992, and 2012. That is called a pooled cross-sectional data.
- It is useful data to analyze the **evolution over time of the cross-sectional distribution** of variables such as individual wages, household income, firms' investments, etc.
- We should distinguish pooled cross-sections from panel data.
- In pooled cross sections we do not follow the same individuals over time. Every period we have a new random sample of individuals.

Panel or Longitudinal Data

- In panel data we have a group of individuals (or households, firms, countries, ...) who are observed at several points in time. That is, we have time series data for each individual in the sample.
- The key feature of panel data that distinguishes them from pooled cross sections is that the same individuals are followed over a given period of time.
- Using panel data we can control for time-invariant unobserved characteristics of individuals, firms, countries, ...

Chapter 2 Introduction to Linear Regression

2.1 Correlation, Causal relationship and Regression analysis

Correlation is a measure of the *strength or degree of linear relationship* (positive, negative, no relationship) between two or more variables. Correlations can be very useful research tool but they tell us nothing about the predictive power of variables. Correlation analysis does not show how the variation in the independent variables (causes) explains the change in the dependent variable (outcome). On the other hand, *causal relationship* is said to exist if the outcome is the direct result, or consequence of the action. In terms of economic variables, a given change in variable X is said to cause a change in variable Y , if the change in Y is as a result or consequence of the change in X .

Examples: 1. using fertilizer results in more crop production.

2. Increasing the price of a good leads to reduced demand of that good.

In short, Causality means a specific action (using fertilizer) leads to a specific measurable outcome (more crop production). This represents a *cause-and-effect* relationship which depicts the effect of change in one variable is the direct result of another variable. This is causal relationship. Therefore, if we cannot use correlation analysis to explain causal relationship, what analytical tools do we have to study causal relationship (explanation or predictive power of variables)? How best can we analyze (measure) the causal effect (cause-and-effect relationship)? In economics, we use the estimation process using regression techniques. But, what is regression? Regression is *an econometric tool which helps us to examine the causal relationship between two or more variables*. More formally, Regression analysis is concerned with the study of the dependence of one variable, the *dependent variable*, on one or more other variables, the *explanatory (independent) variables*, with a view to estimating and/or predicting the (population) mean or average value of the dependent variable in terms of the known or fixed values of the independent variable (s).

2.2 Stochastic and Non-stochastic Relationships

A relationship between X and Y , characterized as $Y = f(X)$ is said to be *deterministic or non-stochastic* if for each value of the independent variable (X) there is one and only one corresponding value of dependent variable (Y). On the other hand, a relationship between X and Y is said to be *stochastic or random* if for a particular value of X there is a whole probabilistic distribution of values of Y . In such a case, for any given value of X , the dependent

variable Y assumes some specific value only with some probability. Let's illustrate the distinction between stochastic and non-stochastic relationships with the help of a demand function. Assuming the demand for a certain commodity depends linearly on its price, *ceteris paribus* (other determinants assumed to be constant), which is given by:

$$Q = f(P) = \alpha - \beta P \dots\dots\dots \text{Deterministic relationship} \quad (2.1)$$

The above relationship between P and Q is such that for a particular value of P , there is only one corresponding value of Q . This is, therefore, a deterministic (non-stochastic) relationship since for each price there is always only one corresponding quantity demanded. This implies that all the variation in Y is due to solely the changes in X , and that there are no other factors affecting the dependent variable (Y). If this were true, all the points of price-quantity pairs, if plotted on a two-dimensional plane, would fall on a straight line. However, if we gather observations on the quantity actually demanded from the market at various prices and we plot them on a diagram, we see that they do not fall on a straight line. See the following diagram.

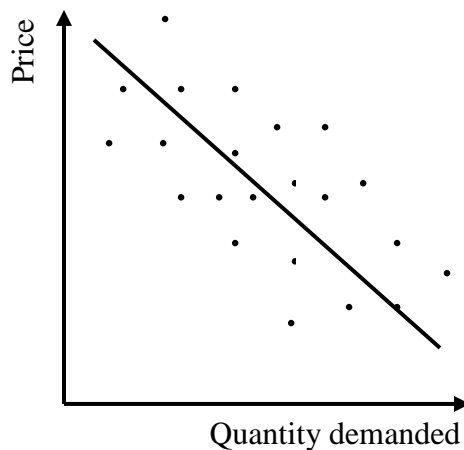


Figure 1 Scatter diagram of stochastic relationship

The deviation of the stochastic Price-Quantity demanded relationship from the deterministic relationship (line) can be caused by *omission of variables from the function, random behavior of human beings, imperfect specification of the mathematical form of the model, measurement error* and so on. Econometric models account this random relationship by introducing a random variable which is usually denoted by the lower case u or ε and is famously called *error term* or *disturbance term* or *stochastic term* of the econometric model. By introducing this random variable in the above demand function the model takes the following stochastic functional form:

$$Q_i = f(P) = \alpha - \beta P + u_i \dots\dots\dots \text{Stochastic relationship} \quad (2.2)$$

Thus, a stochastic model is a model in which the dependent variable is not only determined by the explanatory variable(s) included in the model but also by others which are not included in the model.

2.3 Basis of regression: linear regression model

The purpose of constructing an econometric relationship is usually to predict or explain the effects of one variable resulting from the changes in one or more of the explanatory variables. The basic linear regression model and the relationship between Y and X is given by:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, 2, \dots, n \quad (2.3)$$

The above model represents the linear regression model with a single regressor is called *estimator* the subscript i shows the (number of) observations. While Y represents the *dependent* variable (*explained* variable), X represents the *independent* (*explaining*) variable. The constant values β_0 and β_1 are called *coefficients* or *parameters* or *estimates*. In another way, β_0 is the *intercept* and β_1 is the *slope*. The slope β_1 is the change in Y associated with the change in X . The intercept β_0 is the value of the regression line (Y) when $X = 0$; the point at which the regression line intersects the Y axis. The parameters describe the directions and strengths of the relationship between Y and the factors used to determine Y in the model. The random component of the model that represents the other factors (variables) not included in the model but determine the dependent variable (Y) is represented by ε . Thus, the objective of an econometric analysis is in general to estimate the values of β_0, β_1 and the direction and magnitude of the effect of the independent variable X on the dependent variable Y .

Regression models should be linear in parameters in order to be linear, with out regard to the linearity of the dependent and independent variables (Say, variables can be in logarithm or radical or power form but not the parameters).

Examples:

$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$	Linear in parameters
$\ln Y_i = \beta_0 + \beta_1 \ln X_i + \varepsilon_i$	Linear in parameters
$\ln Y_i^2 = \beta_0 + \beta_1 \ln X_i^2 + \varepsilon_i$	Linear in parameters
$Y_i^2 = \beta_0 + \beta_1^2 X_i^2 + \varepsilon_i$	Non-linear in parameters
$Y_i = \sqrt{\beta_0 + \beta_1 X_i} + \varepsilon_i$	Non-linear in parameters

2.4 Curve fitting and the method of Least Squares (OLS)

Suppose we are interested in the relationship between two variables, X and Y . To describe this relationship we need a set of observations (for both X and Y) called *sample* and a specific functional form based on theory or empirics. And at this point, we assume that the relationship between X and Y is linear (straight line). Consider the following relationship between the consumption of a specific commodity and income of a given sample (8 units).

Table 1 Income- consumption relationship

Observation units (i)	1	2	3	4	5	6	7	8
Consumption (Y)	4	3	3.5	2	3	3.5	2.5	2.5
Income (X)	21	15	15	9	12	18	6	12

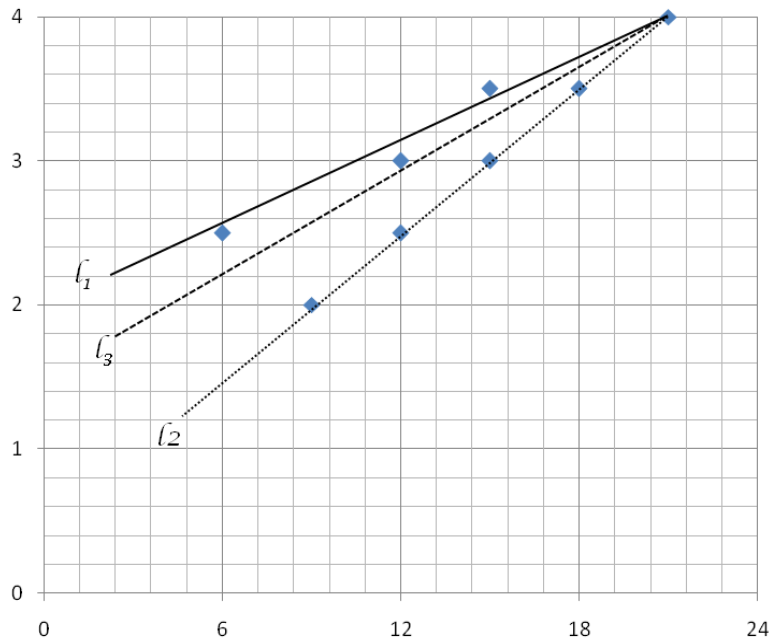


Figure 2 Representation of the scatter points by different lines

As you can see, many straight lines can be drawn (chosen) to fit the points, say in this case l_1 , l_2 and l_3 . But, if we want to represent (fit) all the points by a single line, which line fits best to the scattered points? There exists a procedure to get the line of best fit involving deviations and squared deviations. The '**line of best fit**' is the line that minimizes the sum of squared deviations of the points of the graph from the points of the straight line. The deviations are measure by the vertical distance between the straight line and the scattered points of the graph. The process of

fitting the line that fits best the scattered points using the principle of minimum sum squared deviations is called the **Ordinary Least Squares**.

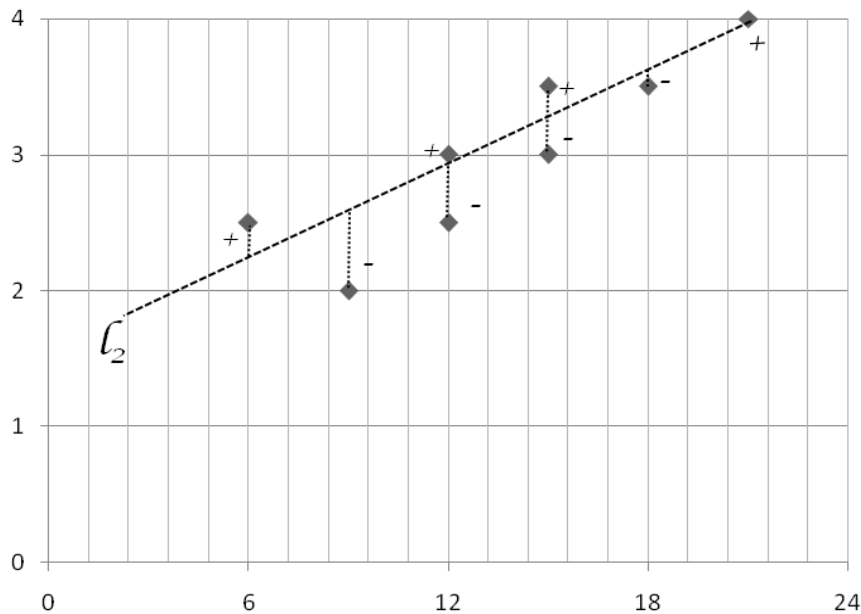


Figure 3 Deviations of the scattered points (from Line 3, ℓ_3)

In searching the best-fit line, it may be important to choose the line that minimizes the sum of squared deviations. For a given data set, the coefficients β_0 and β_1 are unknown but we can use the Least Squares procedure to estimate them. The following graph shows how actually the parameters of the linear regression model are estimated using the Least Square method. In order to do this, we must use data to the unknown slope (β_1) and intercept (β_0) of the regression line.

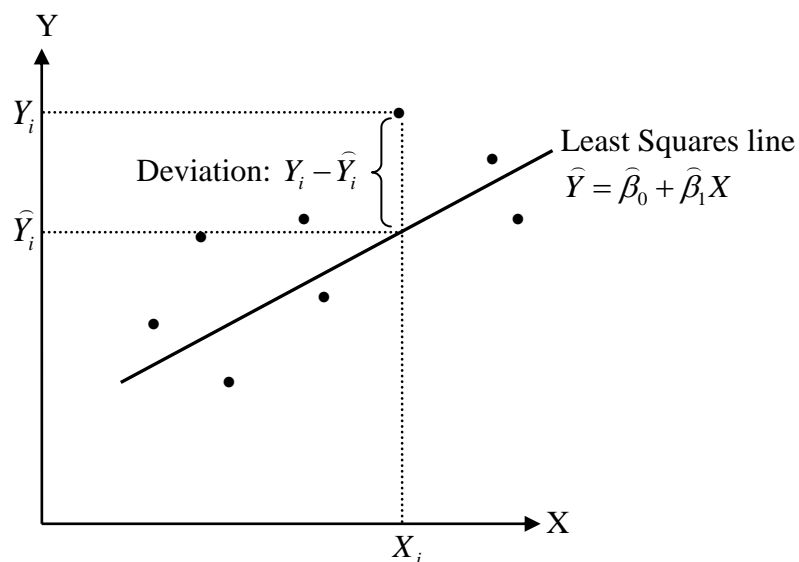


Figure 4 Constructing regression line using Least Squares

As we have said above we can plot the line that fits best the scattered points (sample) by minimizing the sum of squared deviations of the sample from the line. To obtain the Least Square formula, we have to use mathematical tools shown as follows.

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad \text{Population regression function} \quad (2.4)$$

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \quad \text{Estimated regression line} \quad (2.5)$$

$$\varepsilon_i = Y_i - \hat{Y}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i) \quad \text{Deviations (residual)} \quad (2.6)$$

$$[Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)]^2 = (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 \quad \text{Squared deviations} \quad (2.7)$$

$$\sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 \quad \text{Sum of squared deviations} \quad (2.8)$$

Where $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ represents the equation of the straight line with intercept $\hat{\beta}_0$ and slope $\hat{\beta}_1$. In those notations, Y_i is the actual value for observation i and corresponds to the value of X for that observation, while n is the number of observations. \hat{Y}_i , called the *fitted* or *predicted* or *estimated* value is the value of Y on the straight line associated with observation X_i . The deviation (residual) is calculated by subtracting the fitted value of Y_i (which is \hat{Y}_i) from the actual value (Y_i). Thus, for each observation on X there is a corresponding deviation of the fitted value from the actual value of Y . So, our objective is to minimize the sum of squared deviations where we can use elementary calculus to do so. We use procedures of partial derivatives with respect to $\hat{\beta}_0$ and $\hat{\beta}_1$, and setting each equal to zero to solve their values using simultaneous equations.

$$\text{Minimize } \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 \quad (2.9)$$

$$\begin{aligned} \frac{\partial}{\partial \hat{\beta}_0} \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 &= -2 \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) \\ \frac{\partial}{\partial \hat{\beta}_1} \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 &= -2 \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) X_i \\ -2 \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) &= 0 \end{aligned} \quad (2.10)$$

$$-2 \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) X_i = 0 \quad (2.11)$$

By rearranging equations (2.10) and (2.11) we obtain a pair of simultaneous equations, called *Normal Equations*, given below in equations (2.12) and (2.13) .

$$\sum_{i=1}^n Y_i = \hat{\beta}_0 n + \hat{\beta}_1 \sum_{i=1}^n X_i \quad (2.12)$$

$$\sum_{i=1}^n X_i Y_i = \hat{\beta}_0 \sum_{i=1}^n X_i + \hat{\beta}_1 \sum_{i=1}^n X_i^2 \quad (2.13)$$

Now we can solve for $\hat{\beta}_0$ and $\hat{\beta}_1$ simultaneously by multiplying equation (2.12) by $\sum_{i=1}^n X_i$ and equation (2.13) by n .

$$\sum_{i=1}^n X_i \sum_{i=1}^n Y_i = \hat{\beta}_0 n \sum_{i=1}^n X_i + \hat{\beta}_1 \sum_{i=1}^n X_i \sum_{i=1}^n X_i \quad (2.14)$$

$$n \sum_{i=1}^n X_i Y_i = \hat{\beta}_0 n \sum_{i=1}^n X_i + \hat{\beta}_1 n \sum_{i=1}^n X_i^2 \quad (2.15)$$

Solving the system simultaneously and subtracting equation (2.14) from equation (2.15), we get

$$n \sum_{i=1}^n X_i Y_i - \sum_{i=1}^n X_i \sum_{i=1}^n Y_i = \hat{\beta}_1 \left[n \sum_{i=1}^n X_i^2 - \left(\sum_{i=1}^n X_i \right)^2 \right]$$

From which it follows that

$$\hat{\beta}_1 = \frac{n \sum_{i=1}^n X_i Y_i - \sum_{i=1}^n X_i \sum_{i=1}^n Y_i}{n \sum_{i=1}^n X_i^2 - \left(\sum_{i=1}^n X_i \right)^2} \quad \text{and} \quad \hat{\beta}_0 = \frac{\sum_{i=1}^n Y_i - \hat{\beta}_1 \sum_{i=1}^n X_i}{n} = \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \quad (2.16)$$

How would the values of $\hat{\beta}_0$ and $\hat{\beta}_1$ change if we use the deviation values [$x_i = X_i - \bar{X}$] and [$y_i = Y_i - \bar{Y}$] rather than the actual values X_i and Y_i ?

Having obtained the coefficients of the model ($\hat{\beta}_0$ and $\hat{\beta}_1$), they need to be incorporated into the model and the standard errors of the respective coefficients have to be as well fitted within the model as follows:

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i, \text{ where } s.e \text{ represents the standard errors of the model} \quad (2.17)$$

We need to calculate first the estimated population variance (s^2)¹ in order to calculate the standard errors ($s.e$) of the regression line. The estimate of the population variance is calculated as follows

$$s^2 = \frac{\sum \hat{\varepsilon}_i^2}{n-2} = \frac{\sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2}{n-2} \quad (2.18)$$

Note that the degrees of freedom ($n-2$) is constrained by the number of coefficients to be estimated. In this case, we have two coefficients ($\hat{\beta}_0$ and $\hat{\beta}_1$) to be estimated and hence the $n-2$.

It follows that the standard error of the slope of the regression line ($\hat{\beta}_1$) and the intercept ($\hat{\beta}_0$) is given by

$$s.e_{\hat{\beta}_1} = \sqrt{s^2 / \sum x_i^2} \quad \text{and} \quad s.e_{\hat{\beta}_0} = \sqrt{s^2 \left(\frac{\sum X_i^2}{n \sum x_i^2} \right)} \quad (2.19)$$

Notice that the standard errors of the coefficients ($s.e_{\hat{\beta}_1}$ and $s.e_{\hat{\beta}_0}$) measure the dispersion of the estimates about their means. How do they differ from $s = \sqrt{s^2}$? Remember s is the standard error of the regression, which measures the dispersion of the error term associated with the regression line.

¹ Notice that $\sqrt{s^2}$ is known as **Root MSE** in STATA's ANOVA table.

Example 2.1

Observations (n)	Y_i	X_i	\bar{Y}	\bar{X}	$Y_i - \bar{Y}$	$X_i - \bar{X}$	X_i^2	$Y_i X_i$
1	4	21	3	13.5	1	7.5	441	84
2	3	15	3	13.5	0	1.5	225	45
3	3.5	15	3	13.5	0.5	1.5	225	52.5
4	2	9	3	13.5	-1	-4.5	81	18
5	3	12	3	13.5	0	-1.5	144	36
6	3.5	18	3	13.5	0.5	4.5	324	63
7	2.5	6	3	13.5	-0.5	-7.5	36	15
8	2.5	12	3	13.5	-0.5	-1.5	144	30
	24	108	24	108	0	0	1620	343.5
	$\sum_{i=1}^n Y_i$	$\sum_{i=1}^n X_i$	$\sum \bar{Y}$	$\sum \bar{X}$	$\sum (Y_i - \bar{Y})$	$\sum (X_i - \bar{X})$	$\sum_{i=1}^n X_i^2$	$\sum_{i=1}^n X_i Y_i$

$$1. \quad \hat{\beta}_1 = \frac{n \sum_{i=1}^n X_i Y_i - \sum_{i=1}^n X_i \sum_{i=1}^n Y_i}{n \sum_{i=1}^n X_i^2 - \left(\sum_{i=1}^n X_i \right)^2} = \frac{8 \times (343.5) - (24 \times 108)}{8 \times (1620) - (108)^2} = \frac{2748 - 2592}{12960 - 11664} = \frac{156}{1296} = 0.12$$

$$2. \quad \hat{\beta}_0 = \frac{\sum_{i=1}^n Y_i - \hat{\beta}_1 \sum_{i=1}^n X_i}{n} = \frac{24 - 0.12 \times 108}{8} = \frac{24 - 12.96}{8} = \frac{11.04}{8} = 1.38$$

So, the fitted or estimated regression line of the above data is normally given as $\hat{Y} = 1.38 + 0.12 X$, the numbers in parentheses are the standard errors of the estimated coefficients. That is normally how a fitted regression line is sketched. How do you think the Least Square formulae would change if we use the deviation forms of the variables?

The formulae will be less complicated if we write the Least Square estimates in terms of variables that are expressed as deviations from their respective sample means. Hence, we transform the data to *deviations form* by expressing each observation of X and Y in terms of deviations from their respective means. The deviations form are represented by lower cases of X and Y .

$$x_i = X_i - \bar{X} \quad \text{and} \quad y_i = Y_i - \bar{Y} \quad (2.20)$$

Summing the population regression function, $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ over all n observations and dividing it by n , we find that

$$\sum_{i=1}^n Y_i = \sum_{i=1}^n \beta_0 + \beta_1 \sum_{i=1}^n X_i + \sum_{i=1}^n \varepsilon_i \Rightarrow \frac{\sum_{i=1}^n Y_i}{n} = \frac{\sum_{i=1}^n \beta_0}{n} + \frac{\beta_1 \sum_{i=1}^n X_i}{n} + \frac{\sum_{i=1}^n \varepsilon_i}{n} = \bar{Y} = \beta_0 + \beta_1 \bar{X} + \bar{\varepsilon} \quad (2.21)$$

Subtracting equation (6) from the population regression function and combining like terms gives

$$Y_i - \bar{Y} = \beta_1 (X_i - \bar{X}) + (\varepsilon_i - \bar{\varepsilon}) \Rightarrow y_i = \beta_1 x_i + \varepsilon_i \quad (2.22)$$

Equation (7) represents the regression function in *deviations form* and notice that the intercept, β_0 dissolves out. From equation (7), the estimated slope of the regression line is given by

$$\hat{\beta}_1 = \frac{\sum x_i y_i}{\sum x_i^2} \quad (2.23)$$

If you calculate the value of $\hat{\beta}_1$ in **Example 1.1** using this formula, the result would be undoubtedly the same.

2.5 Assumptions of Least Square

A stochastic model of the following type

$$Y_i = \underbrace{\beta_0 + \beta_1 X_i}_{\text{Deterministic component}} + \underbrace{\varepsilon_i}_{\text{random component}} \quad \forall i = 1, 2, 3, \dots, n \quad (2.24)$$

is known as the **Classical Linear Regression Model** if it satisfies the following assumptions.

1. The model is linear in parameters, without regard to the linearity of the dependent and independent variables. The relationship between X and Y is linear as given by Equation (9).
2. ε is a random (real number) variable

- i. With zero mean or The expected value of the error term has mean zero given any value of the explanatory variable, X .

$$E(\varepsilon_i) = 0 \Rightarrow E(\varepsilon_i | X) = 0 \quad \forall i = 1, 2, 3, \dots, n$$

Thus, observing a high or a low value of X does not imply a high or a low value of ε . This effectively means X and ε are uncorrelated. The implication is that changes in X are not associated with changes in ε in any particular direction - Hence the associated changes in Y can be attributed to the impact of X . This assumption allows us to interpret the estimated coefficients as reflecting causal impacts of X on Y .

- ii. With constant (equal) variance (homoskedastic or homoscedastic distribution)

$$Var(\varepsilon_i) = E(\varepsilon_i - E(\varepsilon_i))^2 = E(\varepsilon_i^2) = \sigma^2$$

3. The error terms from any two observations are uncorrelated with each other, which implies there is no autocorrelation (no serial correlation). When the observations are drawn sequentially over time (time series data) we say that there is no serial correlation or no autocorrelation. When the observations are cross sectional (survey data) we say that we have no spatial correlation.

$$Cov(\varepsilon_i, \varepsilon_j) = 0 \quad \forall i, j \quad i \neq j$$

4. The error terms (ε) are independent of X 's. It follows that there is no correlation between the random variable (ε) and the explanatory variable (X). If two variables are unrelated, then their covariance is zero.

$$Cov(\varepsilon_i, X) = 0 \quad \forall i$$

5. The variance of the independent variable X must be non-zero.

$$Var(X_i) > 0$$

This is a crucial requirement. To identify the impact of X on Y , it must be that we observe situations with different values of X . In the absence of such variability there is no information about the impact of X on Y . It means that the values of the independent variable (X) should not be constant.

6. The error term has a normal distribution with mean zero ($E(\varepsilon_i) = 0$) and constant variance (σ^2).

$$\varepsilon \sim N(0, \sigma^2)$$

If conditions 1-6 hold, we have the Best Linear Unbiased Estimator (BLUE). Unbiased: Over repeated samples, the estimator will give us the true population parameter. Efficient (Best): Unbiased as well as smallest variance for the $\hat{\beta}$ estimate. Consistent: What we find is that over many different samples, the Ordinary Least Square (OLS) estimates will be close to the population estimates.

2.6 Desirable Properties of the OLS

A good starting point for this section is to notice that the Least Square estimates result from a specific sample of observations of dependent and independent variables. It follows that the estimates may vary from sample to sample. Remember also that the estimates of Least Square ($\hat{\beta}_0$ and $\hat{\beta}_1$) refer not only to regression estimates from a specific *sample* but are also used to make inferences about the population from which the sample is taken (i.e. the estimator or formula which is also used to compute the estimates from many different samples).

i. Unbiasedness

We want our estimator to be unbiased. Remember that there actually exist **true** values of the coefficients population regression function, which of course we do not know. These reflect the true underlying relationship between Y and X . We want to use a technique to estimate these true coefficients. Our results will only be approximations to reality. **An unbiased estimator is such that the average of the estimates, across an infinite set of different samples of the same size n , is equal to the TRUE value.** This means that on average the estimator $\hat{\beta}$ is correct, even though any single estimate of $\hat{\beta}$ for a particular sample of data may not equal β .

$$E(\hat{\beta}) = \beta \Leftrightarrow E(\hat{\beta}) - \beta = 0, \text{ i.e. Average or expected value of } \hat{\beta} \text{ is equal to the true value of } \beta.$$

ii. Efficiency (Minimum variance plus unbiasedness)

An estimator is efficient if within the set of assumptions that we make, it provides the most precise estimates in the sense that the variance is the lowest possible in the class of estimators we are considering. How do we choose between the OLS estimator and any other unbiased estimator? Our criterion is efficiency.

$$Var(\hat{\beta}) \leq Var(\tilde{\beta})$$

The **variance of an estimator** is an **inverse measure** of its **statistical precision**, i.e., of its dispersion or spread around its mean. The **smaller the variance** of an estimator, the **more**

statistically precise it is. A **minimum variance estimator** is therefore the statistically **most precise estimator** of an unknown population parameter, although it may be biased or unbiased. Among all the linear unbiased estimators, which one has the smallest variance? It is OLS. Thus, efficiency which includes unbiasedness and minimum variance characteristics is also another desirable property of an estimator.

iii. Consistency

An estimator is said to be consistent, if $\hat{\beta}$ approaches its true value, β when the sample size gets larger and larger (approaches infinity). More formally, $\hat{\beta}$ is a consistent estimator of β if the probability limit of $\hat{\beta}$ is β . Given Assumptions 1-6, the Ordinary Least Squares Estimator (OLS) is the Best Linear Unbiased Estimator (BLUE). This means that the OLS estimator is the most efficient (least variance) estimator in the class of linear unbiased estimators. This is known as the Gauss-Markov Theorem.

2.7 Goodness of fit

It is basically essential to scrutinize how the model we estimated fits well the data. We can perform a number of goodness of fit criteria in order to examine the quality of the estimated model. After fitting the model (estimating the parameters), the model has to be subjected to a number of goodness of fit tests in order to make sure that the model is a good explanation for the relationship among economic variables in consideration. A good regression equation is one that helps explain a large proportion of the variation in Y .

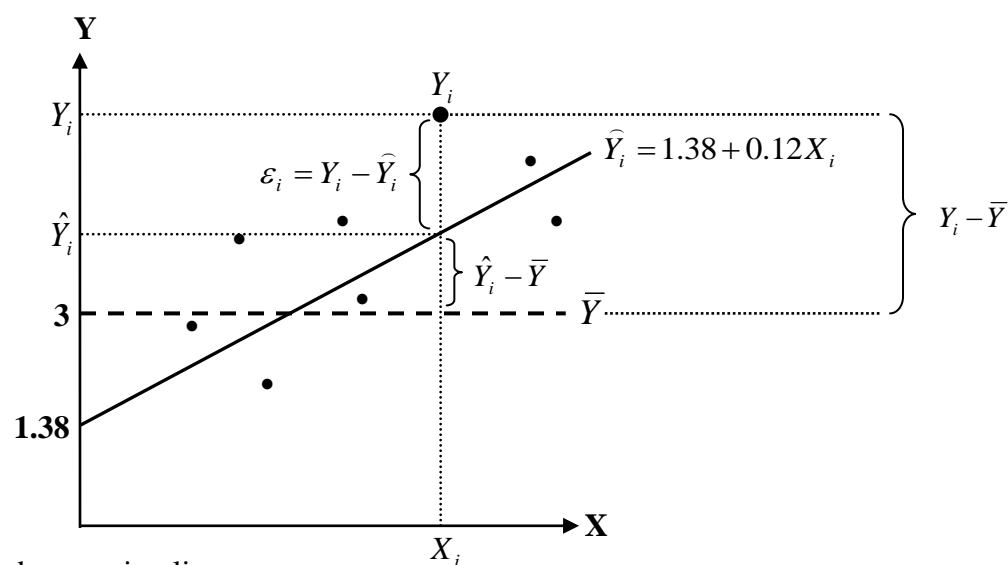


Figure 5 a fitted regression line

Remember that the variation of the dependent variable of the population regression function ($Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$) is attributed to (1) the change in the deterministic part, $\beta_0 + \beta_1 X_i$ (as a result of the change in X) and, (2) the unexplained portion of random component, ε_i . At this point we would like to separate the variation effects (causes) in to two parts. In order to do so, consider the following identity.

$$Y_i - \bar{Y} = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y}) \quad (2.25)$$

In the above identity (equation (2.25)), $Y_i - \bar{Y}$ measures the deviation of the observed (actual) value from its mean (i.e. difference between the observed value of Y and its mean). On the right hand side of the identity, $Y_i - \hat{Y}_i$ denotes the residual value, ε_i . And, the second right-hand term, $\hat{Y}_i - \bar{Y}$ gives the difference between the predicted (estimated value) of Y and the mean of Y (see figure 5).

It follows that in order to measure the variation; we square both sides of equation (2.25) and then sum over all observations: $1, 2, 3, \dots, n$.

$$\sum (Y_i - \bar{Y})^2 = \sum (Y_i - \hat{Y}_i)^2 + \sum (\hat{Y}_i - \bar{Y})^2 + 2 \sum (Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) \quad (2.26)$$

The last term in equation (2.26) can be shown to be equal to zero using the properties (assumption) of the Least Square residuals, $\sum \hat{\varepsilon}_i = 0$ and $\sum \hat{\varepsilon}_i X_i = 0$. It follows that

$$\sum (Y_i - \bar{Y})^2 = \sum (Y_i - \hat{Y}_i)^2 + \sum (\hat{Y}_i - \bar{Y})^2 \quad (2.27)$$

Total variation of Y Residual variation of Y Explained variation of Y
(Total Sum Squarè) (Residual Sum Squarè) (Explained Sum Squarè)

Or in short, $TSS = RSS + ESS$. When we divide both sides of this identity by TSS , we get

$$1 = \frac{RSS}{TSS} + \frac{ESS}{TSS} \Rightarrow 1 - \frac{RSS}{TSS} = \frac{ESS}{TSS} = R^2 \quad (2.28)$$

The ratio $\frac{ESS}{TSS} (= 1 - \frac{RSS}{TSS})$ measures the variation of Y explained by the model, X (regression line). It represents the proportion of the change in Y caused by the variables included in the model, X . This is usually represented by the Coefficient of Determination (R^2). It is a measure of how much of the variance of Y is explained by the regressor X . The computed R^2 following an OLS regression is always between 0 and 1. A low R^2 is not necessarily an indication that the model is wrong - just that the included X has low explanatory power. The key to whether the

results are interpretable as **causal impacts** is whether the explanatory variable is uncorrelated with the error term.

We can also get the value of R^2 from the parameter estimates of the regression line shown as follows. Considering the estimated regression line using deviations forms of the variables, $x_i = (X_i - \bar{X})$ and $y_i = (Y_i - \bar{Y})$, the regression estimator can be given as follows

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \Rightarrow (\hat{Y}_i - \bar{\hat{Y}}) = (\hat{\beta}_0 - \bar{\beta}_0) + \hat{\beta}_1 (X_i - \bar{X}) \Rightarrow \hat{y}_i = \hat{\beta}_1 x_i$$

And, for each observation the following holds true

$$y_i = \hat{y}_i + \hat{\varepsilon}_i \quad (2.29)$$

Squaring both sides of equation (2.29) and summing over all observations, we get

$$\begin{aligned} \sum y_i^2 &= \sum \hat{y}_i^2 + \sum \hat{\varepsilon}_i^2 \\ &= \hat{\beta}_1^2 \sum x_i^2 + \sum \hat{\varepsilon}_i^2 \end{aligned}$$

From which it follows that

$$R^2 = \frac{ESS}{TSS} = \frac{\sum \hat{y}_i^2}{\sum y_i^2} = \hat{\beta}^2 \frac{\sum x_i^2}{\sum y_i^2} \quad (2.30)$$

Notice that the value of R^2 shows the explanation power of the variables included in the model (X_i 's) to explain the change (variation) in the dependent variable (Y). A high value of R^2 is associated with good fit and low value of R^2 with poor fit. We must realize however that a good fit model can still result in relatively low R^2 . Low R^2 does not mean just poor model. It just shows the power of X variables in explaining the variation in Y . Notice also that R^2 usually increases with each added explanatory variable (X) despite the added X -variable is totally unrelated with the dependent variable. It follows that we must have some theoretical or empirical (observation as well) justifications to include a specific variable as an explanatory variable. Suppose the estimated regression line provides an $R^2 = 0.47$. The implication is that the estimated regression curve explains 47% of the total variation of the dependent variable (Y -value). The remaining 53% of the total variation in Y is unaccounted for by the regression line and is attributed to the factors included in the random component of the model (ε_i).

2.7.1 Hypothesis testing

Another important thing to do in examining the goodness of fit of the regression model is hypothesis testing. In cases where the true population variance is unknown (in cases of OLS estimates), we use the t-distribution and *t-test*.

$$t_{\hat{\beta}_i, n-K} = \frac{\hat{\beta}_i - \beta}{se(\hat{\beta}_i)}, \text{ where } \hat{\beta}_i = \hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_n \quad (2.31)$$

The *t-test* for a particular coefficient $\hat{\beta}_i$ is given by the formula in equation (2.31) with $n-K$ degrees of freedom where K is the number of coefficients to be estimated, including the intercept term. The following steps may be followed in performing *t-test*.

Steps of Hypothesis testing

1. Formulate null and alternative hypotheses: alternative depends on 1- or 2-tailed test

$$H_0 : \beta_i = 0, H_1 : \beta_i \neq 0 \text{ (2-tailed)} \quad H_0 : \beta_i = 0, H_1 : \beta_i > 0 \text{ (1-tailed)}$$

2. Specify test statistic and appropriate distribution

$$t^* = t_{\alpha/2, n-K} = \frac{\hat{\beta}_i - \beta}{se(\hat{\beta}_i)} \dots\dots \text{ (2-tailed), and } t^* = t_{\alpha, n-K} = \frac{\hat{\beta}_i - \beta}{se(\hat{\beta}_i)} \dots\dots \text{ (1-tailed)}$$

3. Choose rejection region: α (example, $\alpha = 0.05$)

4. Calculate the test statistic (t^*) for sample

5. Reject or do not reject the null hypothesis

If $P(|t| > t_{\alpha/2})$ reject null hypothesis (2-tailed)

If $P(|t| > t_{\alpha})$ reject null hypothesis (1-tailed)

6. State Conclusion (take inference)

Notice that if the calculated test statistic (t^*) is larger than the critical value (t_c), reject the null hypothesis, which implies the coefficient is statistically significant. If the calculated test statistic (t^*) is smaller than the critical value (t_c), do not reject (accept) the null hypothesis, which implies the coefficient is statistically not significant.

The t -test analyzes the significance of each parameters. Apart from t -test , we can also test if the overall model (the model that postulates the existence a linear relationship between X and Y) has/has not a significant impact in explaining the variation in Y (if the coefficients of the model are together equal to zero). The F -test enables us to perform and examine this statistical relationship. The following formula shows the F -test .

$$F_{K-1, n-K} = \frac{\text{Explained Variance}}{\text{Unexplained Variance}} = \frac{ESS/K-1}{RSS/n-K} = \frac{\hat{\beta}_1^2 \sum x_i^2}{s^2} = \frac{R^2/K-1}{(1-R^2)/n-K} \quad (2.32)$$

The F -test follows F-distribution with $K-1$ and $n-K$ degrees of freedom for the explained and unexplained variation respectively. Notice that K represents the number of parameters (coefficients) estimated from the model. Other things being equal, a large ratio of explained to unexplained variance shows the existence of strong statistical relationship between X and Y . And we can reject or accept the null hypothesis of no relationship between X and Y at the 5 percent significance level by looking up the appropriate critical level of the F-distribution with $K-1$ and $n-K$ degrees of freedom. If the value $F_{K-1, n-K}$ calculated from the regression is larger than the critical value, we reject the null hypothesis at the 5 percent level. If the value of $F_{K-1, n-K}$ is lower than the critical value, we cannot reject the null hypothesis.

Chapter 3 Multiple Regression, Interpretation and Comparison

3.1 Introduction

In simple regression we study the relationship between a dependent variable and a single explanatory (independent variable). But it is rarely the case that economic relationships involve just two variables. Rather a dependent variable Y can depend on a whole series of explanatory variables or regressors. For instance, in demand studies we study the relationship between quantity demanded of a good and price of the good, price of substitute goods and the consumer's income. The model we assume is:

$$Y_i = \beta_0 + \beta_1 P_1 + \beta_2 P_2 + \beta_3 X_i + u_i \text{-----} (3.1)$$

Where Y_i = quantity demanded, P_1 is price of the good, P_2 is price of substitute goods, X_i is consumer's income, and β 's are unknown parameters and u_i is the disturbance.

Equation (3.1) is a multiple regression with three explanatory variables. In general for K -explanatory variable we can write the model as follows:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \dots + \beta_k X_{ki} + u_i \text{-----} (3.2)$$

Where $X_{ki} = (i = 1, 2, 3, \dots, K)$ are explanatory variables, Y_i is the dependent variable and $\beta_j (j = 0, 1, 2, \dots, (k+1))$ are unknown parameters and u_i is the disturbance term. The disturbance term is of similar nature to that in simple regression, reflecting:

- the basic random nature of human responses
- errors of aggregation
- errors of measurement
- errors in specification of the mathematical form of the model

and any other (minor) factors, other than x_i that might influence Y .

In this chapter we will first start our discussion with the assumptions of the multiple regressions and we will proceed our analysis with the case of two explanatory variables and then we will generalize the multiple regression model in the case of k -explanatory variables using matrix algebra.

3.2 Assumptions of Multiple Regression Model

In order to specify our multiple linear regression model and proceed our analysis with regard to this model, some assumptions are compulsory. But these assumptions are the same as in the single explanatory variable model developed earlier except the assumption of no perfect multicollinearity. These assumptions are:

1. *Randomness of the error term:* The variable u is a real random variable.
2. *Zero mean of the error term:* $E(u_i) = 0$
3. *Homoscedasticity:* The variance of each u_i is the same for all the x_i values.

$$\text{i.e. } E(u_i^2) = \sigma_u^2 \text{ (constant)}$$

4. *Normality of u :* The values of each u_i are normally distributed.

$$\text{i.e. } U_i \sim N(0, \sigma^2)$$

5. *No auto or serial correlation:* The values of u_i (corresponding to X_i) are independent from the values of any other u_i (corresponding to X_j) for $i \neq j$.

$$\text{i.e. } E(u_i u_j) = 0 \text{ for } x_i \neq j$$

6. *Independence of u_i and X_i :* Every disturbance term u_i is independent of the explanatory variables. i.e. $E(u_i X_{1i}) = E(u_i X_{2i}) = 0$

This condition is automatically fulfilled if we assume that the values of the X 's are a set of fixed numbers in all (hypothetical) samples.

7. *No perfect multicollinearity:* The explanatory variables are not perfectly linearly correlated.

We can't exclusively list all the assumptions but the above assumptions are some of the basic assumptions that enable us to precede our analysis.

3.3 A Model With Two Explanatory Variables

In order to understand the nature of multiple regression model easily, we start our analysis with the case of two explanatory variables, then extend this to the case of k -explanatory variables.

3.3.1 Estimation of parameters of two-explanatory variables model

The model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + U_i \dots\dots\dots(3.3)$

is multiple regression with two explanatory variables. The expected value of the above model is called population regression equation i.e.

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2, \text{ Since } E(U_i) = 0. \dots\dots\dots(3.4)$$

Where β_i is the population parameters. β_0 is referred to as the intercept and β_1 and β_2 are also sometimes known as regression slopes of the regression. Note that, β_2 for example measures the effect on $E(Y)$ of a unit change in X_2 when X_1 is held constant.

Since the population regression equation is unknown to any investigator, it has to be estimated from sample data. Let us suppose that the sample data has been used to estimate the population regression equation. We leave the method of estimation unspecified for the present and merely assume that equation (3.4) has been estimated by sample regression equation, which we write as:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 \dots\dots\dots(3.5)$$

Where $\hat{\beta}_j$ are estimates of the β_j and \hat{Y} is known as the predicted value of Y.

Now it is time to state how (3.3) is estimated. Given sample observation on Y, X_1 & X_2 , we estimate (3.3) using the method of least square (OLS).

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + e_i \dots\dots\dots(3.6)$$

is sample relation between Y, X_1 & X_2 .

$$e_i = Y_i - \hat{Y} = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i} \dots\dots\dots(3.7)$$

To obtain expressions for the least square estimators, we partially differentiate $\sum e_i^2$ with respect to $\hat{\beta}_0, \hat{\beta}_1$ and $\hat{\beta}_2$ and set the partial derivatives equal to zero.

$$\frac{\partial [\sum e_i^2]}{\partial \hat{\beta}_0} = -2 \sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i}) = 0 \dots\dots\dots(3.8)$$

$$\frac{\partial [\sum e_i^2]}{\partial \hat{\beta}_1} = -2 \sum X_{1i} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i}) = 0 \dots\dots\dots(3.9)$$

$$\frac{\partial [\sum e_i^2]}{\partial \hat{\beta}_2} = -2 \sum X_{2i} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i}) = 0 \dots\dots\dots(3.10)$$

Summing from 1 to n, the multiple regression equation produces three **Normal Equations**:

$$\sum Y = n\hat{\beta}_0 + \hat{\beta}_1 \sum X_{1i} + \hat{\beta}_2 \sum X_{2i} \dots\dots\dots(3.11)$$

$$\sum X_{2i} Y_i = \hat{\beta}_0 \sum X_{2i} + \hat{\beta}_1 \sum X_{1i} X_{2i} + \hat{\beta}_2 \sum X_{2i}^2 \dots\dots\dots(3.12)$$

$$\sum X_{2i}Y_i = \hat{\beta}_0 \sum X_{2i} + \hat{\beta}_1 \sum X_{1i}X_{2i} + \hat{\beta}_2 \sum X_{2i}^2 \dots\dots\dots(3.13)$$

From (3.11) we obtain $\hat{\beta}_0$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}_1 - \hat{\beta}_2 \bar{X}_2 \dots\dots\dots(3.14)$$

Substituting (3.14) in (3.12) , we get:

$$\begin{aligned} \sum X_{1i}Y_i &= (\bar{Y} - \hat{\beta}_1 \bar{X}_1 - \hat{\beta}_2 \bar{X}_2) \sum X_{1i} + \hat{\beta}_1 \sum X_{1i}^2 + \hat{\beta}_2 \sum X_{2i} \\ \Rightarrow \sum X_{1i}Y_i - \hat{Y} \sum X_{1i} &= \hat{\beta}_1 (\sum X_{1i}^2 - \bar{X}_1 \sum X_{2i}) + \hat{\beta}_2 (\sum X_{1i}X_{2i} - \bar{X}_2 \sum X_{2i}) \\ \Rightarrow \sum X_{1i}Y_i - n\bar{Y} \bar{X}_{1i} &= \hat{\beta}_2 (\sum X_{1i}^2 - n\bar{X}_{1i}^2) + \hat{\beta}_2 (\sum X_{1i}X_{2i} - n\bar{X}_1 \bar{X}_2) \dots\dots\dots(3.15) \end{aligned}$$

We know that

$$\begin{aligned} \sum (X_i - Y_i)^2 &= (\sum X_i Y_i - n\bar{X}_i \bar{Y}_i) = \sum x_i y_i \\ \sum (X_i - \bar{X}_i)^2 &= (\sum X_i^2 - n\bar{X}_i^2) = \sum x_i^2 \end{aligned}$$

Substituting the above equations in equation (3.14), the normal equation (3.12) can be written in deviation form as follows:

$$\sum x_1 y = \hat{\beta}_1 \sum x_1^2 + \hat{\beta}_2 \sum x_1 x_2 \dots\dots\dots(3.16)$$

Using the above procedure if we substitute (3.14) in (3.13), we get

$$\sum x_2 y = \hat{\beta}_1 \sum x_1 x_2 + \hat{\beta}_2 \sum x_2^2 \dots\dots\dots(3.17)$$

Let's bring (2.17) and (2.18) together

$$\sum x_1 y = \hat{\beta}_1 \sum x_1^2 + \hat{\beta}_2 \sum x_1 x_2 \dots\dots\dots(3.18)$$

$$\sum x_2 y = \hat{\beta}_1 \sum x_1 x_2 + \hat{\beta}_2 \sum x_2^2 \dots\dots\dots(3.19)$$

$\hat{\beta}_1$ and $\hat{\beta}_2$ can easily be solved using matrix

We can rewrite the above two equations in matrix form as follows.

$$\begin{pmatrix} \sum x_1^2 & \sum x_1 x_2 \\ \sum x_1 x_2 & \sum x_2^2 \end{pmatrix} \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} \sum x_2 y \\ \sum x_3 y \end{pmatrix} \dots\dots\dots(3.20)$$

If we use Cramer's rule to solve the above matrix we obtain

$$\hat{\beta}_1 = \frac{\sum x_1 y \cdot \sum x_2^2 - \sum x_1 x_2 \cdot \sum x_2 y}{\sum x_1^2 \cdot \sum x_2^2 - (\sum x_1 x_2)^2} \dots\dots\dots(3.21)$$

$$\hat{\beta}_2 = \frac{\sum x_2 y \cdot \sum x_1^2 - \sum x_1 x_2 \cdot \sum x_1 y}{\sum x_1^2 \cdot \sum x_2^2 - (\sum x_1 x_2)^2} \dots\dots\dots (3.22)$$

We can also express $\hat{\beta}_1$ and $\hat{\beta}_2$ in terms of covariance and variances of Y , X_1 and X_2

$$\hat{\beta}_1 = \frac{\text{Cov}(X_1, Y) \cdot \text{Var}(X_2) - \text{Cov}(X_1, X_2) \cdot \text{Cov}(X_2, Y)}{\text{Var}(X_1) \cdot \text{Var}(X_2) - [\text{Cov}(X_1, X_2)]^2} \dots\dots\dots (3.23)$$

$$\hat{\beta}_2 = \frac{\text{Cov}(X_2, Y) \cdot \text{Var}(X_1) - \text{Cov}(X_1, X_2) \cdot \text{Cov}(X_1, Y)}{\text{Var}(X_1) \cdot \text{Var}(X_2) - [\text{Cov}(X_1, X_2)]^2} \dots\dots\dots (3.24)$$

3.3.2 The coefficient of determination (R^2): two explanatory variables case

In the simple regression model, we introduced R^2 as a measure of the proportion of variation in the dependent variable that is explained by variation in the explanatory variable. In multiple regression model the same measure is relevant, and the same formulas are valid but now we talk of the proportion of variation in the dependent variable explained by all explanatory variables included in the model. The coefficient of determination is:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum e_i^2}{\sum y_i^2} \dots\dots\dots (3.25)$$

In the present model of two explanatory variables:

$$\begin{aligned} \sum e_i^2 &= \sum (y_i - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i})^2 \\ &= \sum e_i (y_i - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i}) \\ &= \sum e_i y - \hat{\beta}_1 \sum x_{1i} e_i - \hat{\beta}_2 \sum x_{2i} e_i \\ &= \sum e_i y_i \text{ since } \sum e_i x_{1i} = \sum e_i x_{2i} = 0 \\ &= \sum y_i (y_i - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i}) \\ \text{i.e } \sum e_i^2 &= \sum y^2 - \hat{\beta}_1 \sum x_{1i} y_i - \hat{\beta}_2 \sum x_{2i} y_i \\ \Rightarrow \underbrace{\sum y^2}_{\text{Total sum of squares (Total variation)}} &= \underbrace{\hat{\beta}_1 \sum x_{1i} y_i + \hat{\beta}_2 \sum x_{2i} y_i}_{\text{Explained sum of squares (Explained variation)}} + \underbrace{\sum e_i^2}_{\text{Residual sum of squares (unexplained variation)}} \dots\dots\dots (3.26) \end{aligned}$$

$$\therefore R^2 = \frac{ESS}{TSS} = \frac{\hat{\beta}_1 \sum x_{1i} y_i + \hat{\beta}_2 \sum x_{2i} y_i}{\sum y^2} \dots\dots\dots (3.27)$$

As in simple regression, R^2 is also viewed as a measure of the prediction ability of the model over the sample period, or as a measure of how well the estimated regression fits the data. The

value of R^2 is also equal to the squared sample correlation coefficient between \hat{Y}_t & Y_t . Since the sample correlation coefficient measures the linear association between two variables, if R^2 is high, that means there is a close association between the values of Y_t and the values of predicted by the model, \hat{Y}_t . In this case, the model is said to “fit” the data well. If R^2 is low, there is no association between the values of Y_t and the values predicted by the model, \hat{Y}_t and the model does not fit the data well.

3.3.3 Adjusted Coefficient of Determination (\bar{R}^2)

One difficulty with R^2 is that it can be made large by adding more and more variables, even if the variables added have no economic justification. Algebraically, it is the fact that as the variables are added the sum of squared errors (RSS) goes down (it can remain unchanged, but this is rare) and thus R^2 goes up. If the model contains $n-1$ variables then $R^2=1$. The manipulation of model just to obtain a high R^2 is not wise. An alternative measure of goodness of fit, called the adjusted R^2 and often symbolized as \bar{R}^2 , is usually reported by regression programs. It is computed as:

$$\bar{R}^2 = 1 - \frac{\sum e_i^2 / n - k}{\sum y^2 / n - 1} = 1 - (1 - R^2) \left(\frac{n-1}{n-k} \right) \text{-----}(3.28)$$

This measure does not always goes up when a variable is added because of the degree of freedom term $n-k$ is the numerator. As the number of variables k increases, RSS goes down, but so does $n-k$. The effect on \bar{R}^2 depends on the amount by which R^2 falls. While solving one problem, this corrected measure of goodness of fit unfortunately introduces another one. It losses its interpretation; \bar{R}^2 is no longer the percent of variation explained. This modified \bar{R}^2 is sometimes used and misused as a device for selecting the appropriate set of explanatory variables.

3.4. Hypothesis Testing in Multiple Regression Model

In multiple regression models we will undertake two tests of significance. One is significance of individual parameters of the model. This test of significance is the same as the tests discussed in simple regression model. The second test is overall significance of the model.

3.4.1. Tests of individual significance

If we invoke the assumption that $U_i \sim N(0, \sigma^2)$, then we can use either the t-test or standard error test to test a hypothesis about any individual partial regression coefficient. To illustrate consider the following example.

$$\text{Let } Y = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + e_i \dots\dots\dots (3.51)$$

$$\text{A. } H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

$$\text{B. } H_0 : \beta_2 = 0$$

$$H_1 : \beta_2 \neq 0$$

The null hypothesis (A) states that, holding X_2 constant X_1 has no (linear) influence on Y . Similarly hypothesis (B) states that holding X_1 constant, X_2 has no influence on the dependent variable Y_i . To test these null hypothesis we will use the following tests:

- i- **Standard error test:** under this and the following testing methods we test only for $\hat{\beta}_1$. The test for $\hat{\beta}_2$ will be done in the same way.

$$SE(\hat{\beta}_1) = \sqrt{\text{var}(\hat{\beta}_1)} = \sqrt{\frac{\hat{\sigma}^2 \sum x_{2i}^2}{\sum x_{1i}^2 \sum x_{2i}^2 - (\sum x_1 x_2)^2}} ; \text{ where } \hat{\sigma}^2 = \frac{\sum e_i^2}{n-3}$$

- If $SE(\hat{\beta}_1) > \frac{1}{2} \hat{\beta}_1$, we accept the null hypothesis that is, we can conclude that the estimate β_i is not statistically significant.
- If $SE(\hat{\beta}_1) < \frac{1}{2} \hat{\beta}_1$, we reject the null hypothesis that is, we can conclude that the estimate β_i is statistically significant.

Note: The smaller the standard errors, the stronger the evidence that the estimates are statistically reliable.

- ii. **The student's t-test:** We compute the t-ratio for each $\hat{\beta}_i$

$$t^* = \frac{\hat{\beta}_i - \beta}{SE(\hat{\beta}_i)} \sim t_{n-k} , \text{ where } n \text{ is number of observation and } k \text{ is number of parameters. If}$$

we have 3 parameters, the degree of freedom will be $n-3$. So;

$$t^* = \frac{\hat{\beta}_2 - \beta_2}{SE(\hat{\beta}_2)} ; \text{ with } n-3 \text{ degree of freedom}$$

In our null hypothesis $\beta_2 = 0$, the t^* becomes:

$$t^* = \frac{\hat{\beta}_2}{SE(\hat{\beta}_2)}$$

- If $t^* < t$ (tabulated), we accept the null hypothesis, i.e. we can conclude that $\hat{\beta}_2$ is not significant and hence the regressor does not appear to contribute to the explanation of the variations in Y.
- If $t^* > t$ (tabulated), we reject the null hypothesis and we accept the alternative one; $\hat{\beta}_2$ is statistically significant. Thus, the greater the value of t^* the stronger the evidence that β_i is statistically significant.

3.4.2 Test of Overall Significance

Throughout the previous section we were concerned with testing the significance of the estimated partial regression coefficients individually, i.e. under the separate hypothesis that each of the true population partial regression coefficient was zero.

In this section we extend this idea to joint test of the relevance of all the included explanatory variables. Now consider the following:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + U_i$$

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \dots = \beta_k = 0$$

$$H_1 : \text{at least one of the } \beta_k \text{ is non-zero}$$

This null hypothesis is a joint hypothesis that $\beta_1, \beta_2, \dots, \beta_k$ are jointly or simultaneously equal to zero. A test of such a hypothesis is called **a test of overall significance** of the observed or estimated regression line, that is, whether Y is linearly related to X_1, X_2, \dots, X_k .

Can the joint hypothesis be tested by testing the significance of individual significance of $\hat{\beta}_i$'s as the above? The answer is no, and the reasoning is as follows.

In testing the individual significance of an observed partial regression coefficient, we assumed implicitly that each test of significance was based on different (i.e. independent) sample. Thus, in testing the significance of $\hat{\beta}_2$ under the hypothesis that $\beta_2 = 0$, it was assumed tacitly that the testing was based on different sample from the one used in testing the significance of $\hat{\beta}_3$ under the null hypothesis that $\beta_3 = 0$. But to test the joint hypothesis of the above, we shall be violating the assumption underlying the test procedure.

“.....testing a series of single (individual) hypothesis is not equivalent to testing those same hypothesis. The institutive reason for this is that in a joint test of several hypotheses any single hypothesis is affected by the information in the other hypothesis.”²

The test procedure for any set of hypothesis can be based on a *comparison of the sum of squared errors from the original, the unrestricted multiple regression model to the sum of squared errors from a regression model in which the null hypothesis is assumed to be true.*

When a null hypothesis is assumed to be true, we in effect place conditions or constraints, on the values that the parameters can take, and the sum of squared errors increases. The idea of the test is that if these sum of squared errors are substantially different, then the assumption that the joint null hypothesis is true has significantly reduced the ability of the model to fit the data, and the data do not support the null hypothesis.

If the null hypothesis is true, we expect that the data are compliable with the conditions placed on the parameters. Thus, there would be little change in the sum of squared errors when the null hypothesis is assumed to be true.

Let the Restricted Residual Sum of Square (**RRSS**) be the sum of squared errors in the model obtained by assuming that the null hypothesis is true and **URSS** be the sum of the squared error of the original unrestricted model i.e. unrestricted residual sum of square (URSS). It is always true that $RRSS - URSS \geq 0$.

Consider $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \dots + \hat{\beta}_k X_k + e_i$.

This model is called unrestricted. The test of joint hypothesis is that:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \dots = \beta_k = 0$$

H_1 : at least one of the β_k is different from zero.

We know that: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \dots + \hat{\beta}_k X_{ki}$

$$Y_i = \hat{Y} + e$$

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

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

This sum of squared error is called unrestricted residual sum of square (URSS). This is the case when the null hypothesis is not true. If the null hypothesis is assumed to be true, i.e. when all the slope coefficients are zero.

$$Y = \hat{\beta}_0 + e_i$$

$$\hat{\beta}_0 = \frac{\sum Y_i}{n} = \bar{Y} \rightarrow \quad (\text{applying OLS}) \dots \dots \dots (3.52)$$

$$e = Y - \hat{\beta}_0 \quad \text{but} \quad \hat{\beta}_0 = \bar{Y}$$

$$e = Y - \bar{Y}$$

$$\sum e_i^2 = \sum (Y_i - \hat{Y}_i)^2 = \sum y^2 = TSS$$

The sum of squared error when the null hypothesis is assumed to be true is called Restricted Residual Sum of Square (RRSS) and this is equal to the total sum of square (TSS).

The ratio: $\frac{RRSS - URSS / K - 1}{URSS / n - K} \sim F_{(k-1, n-k)} \dots \dots \dots (3.53);$ (has an F-distribution with k-1 and n-k degrees of freedom for the numerator and denominator respectively)

$$RRSS = TSS$$

$$URSS = \sum e_i^2 = \sum y^2 - \hat{\beta}_1 \sum yx_1 - \hat{\beta}_2 \sum yx_2 + \dots \dots \dots \hat{\beta}_k \sum yx_k = RSS$$

$$F = \frac{(TSS - RSS) / k - 1}{RSS / n - k}$$

$$F = \frac{ESS / k - 1}{RSS / n - k} \dots\dots\dots (3.54)$$

If we divide the above numerator and denominator by $\Sigma y^2 = TSS$ then:

$$F = \frac{\frac{ESS}{TSS} / k - 1}{\frac{RSS}{TSS} / n - k} \dots\dots\dots (3.55)$$

This implies the computed value of F can be calculated either as a ratio of ESS & TSS or R^2 & $1 - R^2$. If the null hypothesis is not true, then the difference between RRSS and URSS (TSS & RSS) becomes large, implying that the constraints placed on the model by the null hypothesis have large effect on the ability of the model to fit the data, and the value of F tends to be large. Thus, we reject the null hypothesis if the F test static becomes too large. This value is compared with the critical value of F which leaves the probability of α in the upper tail of the F-distribution with k-1 and n-k degree of freedom.

If the computed value of F is greater than the critical value of F (k-1, n-k), then the parameters of the model are jointly significant or the dependent variable Y is linearly related to the independent variables included in the model.

Topic 5: Limited Dependent Variable Models

Introduction

Many different types of linear models have been discussed in the course so far. But in all the models considered, the response variable has been a quantitative variable, which has been assumed to be normally distributed. In this Subtopic, we consider situations where the response variable is a categorical random variable, attaining only two possible outcomes. Examples of this type of data are very common. For example, the response can be whether or not a farmer has adopted a technology, whether or not an item in a manufacturing process passes the quality control, whether or not the farmer has credit access, etc. Since the response variables are dichotomous (that is, they have only two possible outcomes), it is inappropriate to assume that they are normally distributed—thus the data cannot be analyzed using the methods discussed so far in the course. The most common method to use for analyzing data with dichotomous response variables is logit and probit models.

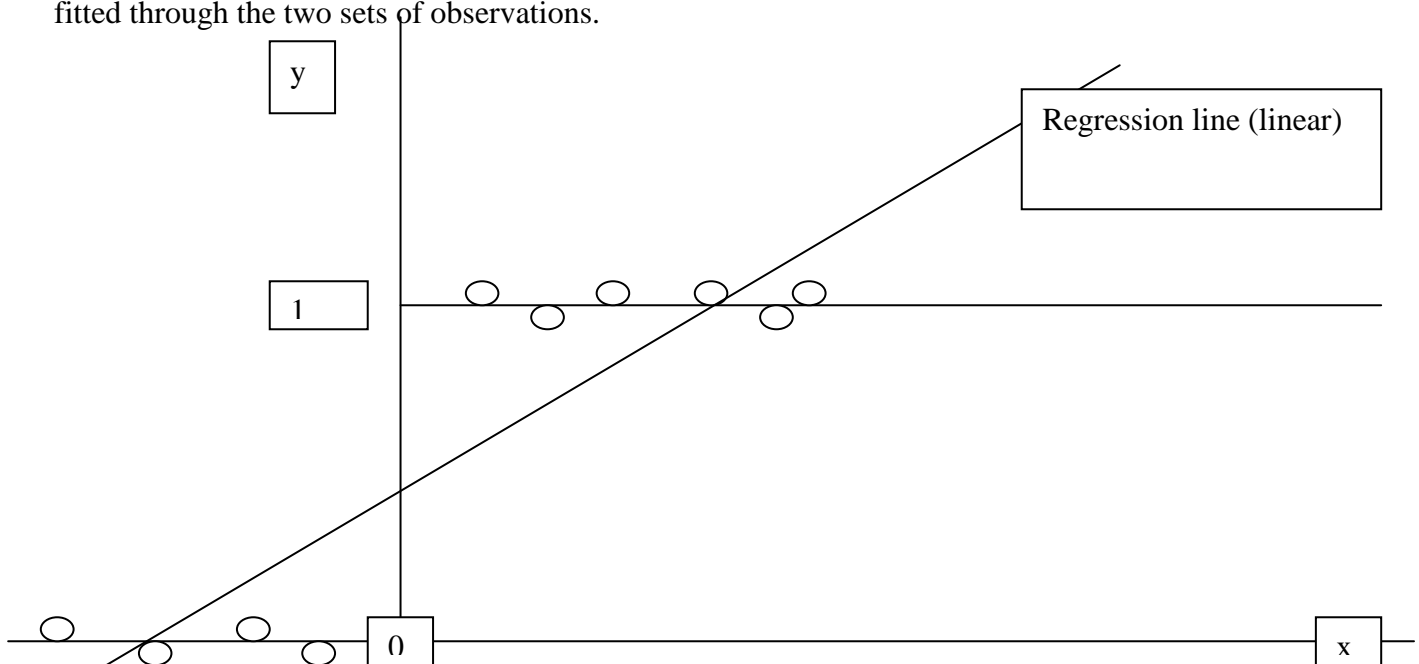
5.1 Dummy Dependent Variables

When the response variable is dichotomous, it is convenient to denote one of the outcomes as success and the other as failure. For example, if a farmer adopted a technology, the response is ‘success’, if not, then the response is ‘failure’; if an item passes the quality control, the response is ‘success’, if not, then the response is ‘failure’; if a has credit access, the response is ‘success’, if not the response is ‘failure’. It is standard to let the dependent variable Y be a binary variable, which attains the value 1, if the outcome is ‘success’, and 0 if the outcome is ‘failure’. In a regression situation, each response variable is associated with given values of a set of explanatory variables X_1, X_2, \dots, X_k . For example, whether or not a farmer adopted a technology may depend on the educational status, farm size, age, gender, etc.; whether or not an item in a manufacturing process passes the quality control may depend on various conditions regarding the production process, such as temperature, quality of raw material, time since last service of the machinery, etc.

When examining the dummy dependent variables we need to ensure there are sufficient numbers of 0s and 1s. If we were assessing technology adoptions, we would need a sample of both farmers that have adopted a technology and those that have not adopted.

5.1.1 Linear Probability Model (LPM)

The Linear Probability Model uses OLS to estimate the model, the coefficients and t-statistics etc are then interpreted in the usual way. This produces the usual linear regression line, which is fitted through the two sets of observations.



5.1.1.1 Features of the LPM

1. The dependent variable has two values, the value 1 has a probability of p and the value 0 has a probability of $(1-p)$.
2. This is known as the Bernoulli probability distribution. In this case the expected value of a random variable following a Bernoulli distribution is the probability the variable equals 1.
3. Since the probability of p must lie between 0 and 1, then the expected value of the dependent variable must also lie between 0 and 1.

5.1.1.2 Problems with LPM

1. The error term is not normally distributed, it also follows the Bernoulli distribution.
2. The variance of the error term is heteroskedastic. The variance for the Bernoulli distribution is $p(1-p)$, where p is the probability of a success.
3. The value of the R-squared statistic is limited, given the distribution of the LPMs.
4. Possibly the most problematic aspect of the LPM is the non-fulfilment of the requirement that the estimated value of the dependent variable y lies between 0 and 1.
5. One way around the problem is to assume that all values below 0 and above 1 are actually 0 or 1 respectively

6. An alternative and much better remedy to the problem is to use an alternative technique such as the Logit or Probit models.
7. The final problem with the LPM is that it is a linear model and assumes that the probability of the dependent variable equalling 1 is linearly related to the explanatory variable.

For example if we have a model where the dependent variable takes the value of 1 if a farmer has extension contact and 0 otherwise, regressed on the farmers education level. The probability of contacting an extension agent will rise as education level rises.

5.1.1.3 LPM model example

The following model of technology adoption (TA) was estimated, with extension visit (EV) and education (ED) as the explanatory variables. Regression using OLS gives the following result.

$$\hat{t}_i = 2.79 + 0.76e_i - 0.12d_i$$

(2.10) (0.06) (0.04)

$$R^2 = 0.15, DW = 1.78$$

$$t = \begin{cases} 1 - \text{Adopted} \\ 0 - \text{Not adopted} \end{cases}$$

The coefficients are interpreted as in the usual OLS models, i.e. a 1% rise in extension contact, gives a 0.76% increase in the probability of technology adoption.

The R-squared statistic is low, but this is probably due to the LPM approach, so we would usually ignore it. The t-statistics are interpreted in the usual way.

5.1.2 The Logit Model

The main way around the problems mentioned earlier is to use a different distribution to the Bernoulli distribution, where the relationship between x and p is non-linear and the p is always between 0 and 1. This requires the use of a 's' shaped curve, which resembles the cumulative distribution function (CDF) of a random variable. The CDFs used to represent a discrete variable are the logistic (Logit model) and normal (Probit model).

If we assume we have the following basic model, we can express the probability that $y=1$ as a cumulative logistic distribution function.

$$y_i = \alpha_0 + \alpha_1 x_i + u_i$$

$$p_i = E(y = 1 / x_i) = \alpha_0 + \alpha_1 x_i$$

The cumulative Logistic distributive function can then be written as:

$$p_i = \frac{1}{1 + e^{-z_i}}$$

$$\text{Where : } z_i = \alpha_0 + \alpha_1 x_i$$

There is a problem with non-linearity in the previous expression, but this can be solved by creating the odds ratio:

$$1 - p_i = \frac{1}{1 + e^{z_i}}$$

$$\frac{p_i}{1 - p_i} = \frac{1 + e^{z_i}}{1 + e^{-z_i}} = e^{z_i}$$

$$L_i = \ln\left(\frac{p_i}{1 - p_i}\right) = z_i = \alpha_0 + \alpha_1 x_i$$

Note that L is the log of the odds ratio and is linear in the parameters. The odds ratio can be interpreted as the probability of something happening to the probability it won't happen. i.e. the odds ratio of getting a mortgage is the probability of getting a mortgage to the probability they will not get one. If p is 0.8, the odds are 4 to 1 that the person will get a mortgage.

5.1.2.1 Logit model features

1. Although L is linear in the parameters, the probabilities are non-linear.
2. The Logit model can be used in multiple regression tests.
3. If L is positive, as the value of the explanatory variables increase, the odds that the dependent variable equals 1 increase.
4. The slope coefficient measures the change in the log-odds ratio for a unit change in the explanatory variable.
5. These models are usually estimated using Maximum Likelihood techniques.
6. The R-squared statistic is not suitable for measuring the goodness of fit in discrete dependent variable models, instead we compute the count R-squared statistic.

If we assume any probability greater than 0.5 counts as a 1 and any probability less than 0.5 counts as a 0, then we count the number of correct predictions. This is defined as:

$$\text{Count } R^2 = \frac{\text{number of correct predictions}}{\text{Total number of observations}}$$

The Logit model can be interpreted in a similar way to the LPM, given the following model, where the dependent variable is granting of a mortgage (1) or not (0). The explanatory variable is a customer's income:

The coefficient on y suggests that a 1% increase in income (y) produces a 0.32% rise in the log of the odds of getting a mortgage. This is difficult to interpret, so the coefficient is often ignored, the z -statistic (same as t -statistic) and sign on the coefficient is however used for the interpretation of the results. We could include a specific value for the income of a customer and then find the probability of getting a mortgage.

5.1.2.2 Logit model result

If we have a customer with 0.5 units of income, we can estimate a value for the Logit of $0.56 + 0.32 \cdot 0.5 = 0.72$. We can use this estimated Logit value to find the estimated probability of getting a mortgage. By including it in the formula given earlier for the Logit Model we get:

$$p_i = \frac{1}{(1 + e^{-(0.72)})} = \frac{1}{1.49} = 0.67$$

Given that this estimated probability is bigger than 0.5, we assume it is nearer 1, therefore we predict this customer would be given a mortgage. With the Logit model we tend to report the sign of the variable and its z -statistic which is the same as the t -statistic in large samples.

5.1.3 Probit Model

An alternative CDF to that used in the Logit Model is the normal CDF, when this is used we refer to it as the Probit Model. In many respects this is very similar to the Logit model. The Probit model has also been interpreted as a 'latent variable' model. This has implications for how we explain the dependent variable. i.e. we tend to interpret it as a desire or ability to achieve something.

5.1.4 The models compared

1. The coefficient estimates from all three models are related.
2. According to Amemiya, if you multiply the coefficients from a Logit model by 0.625, they are approximately the same as the Probit model.
3. If the coefficients from the LPM are multiplied by 2.5 (also 1.25 needs to be subtracted from the constant term) they are approximately the same as those produced by a Probit model.

5.2 The Tobit model

Researchers sometimes encounter dependent variables that have a mixture of discrete and continuous properties. The problem is that for some values of the outcome variable, the response has discrete properties; for other values, it is continuous.

5.2.1 Variables with discrete and continuous responses

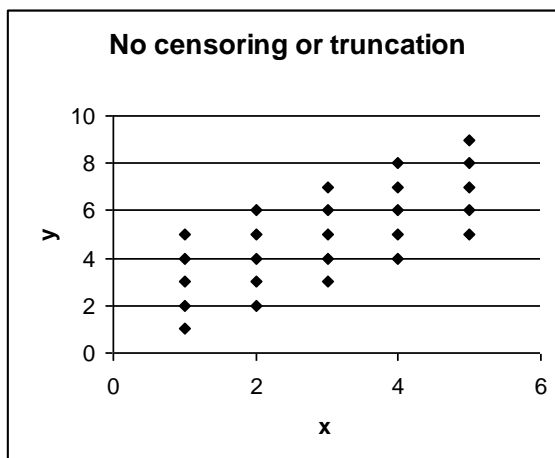
Sometimes the mixture of discrete and continuous values is a result of surveys that only gather partial information. For example, income categories 0–4999, 5000–9999, 10000–19999, 20000–29999, 30000+. Sometimes the true responses are discrete across a certain range and continuous across another range. Examples are days spent in the hospital last year or money spent on clothing last year

5.2.2 Some terms and definitions

Y is “censored” when we observe X for all observations, but we only know the true value of Y for a restricted range of observations. If $Y = k$ or $Y > k$ for all Y, then Y is “censored from below”. If $Y = k$ or $Y < k$ for all Y, then Y is “censored from above”.

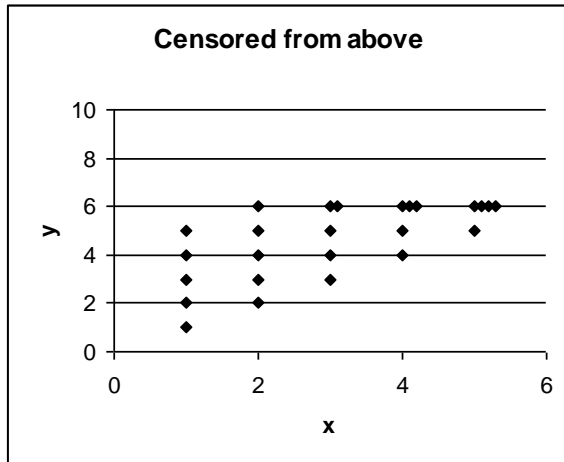
Y is “truncated” when we only observe X for observations where Y would not be censored.

Example 5.1: Non-censoring but truncation



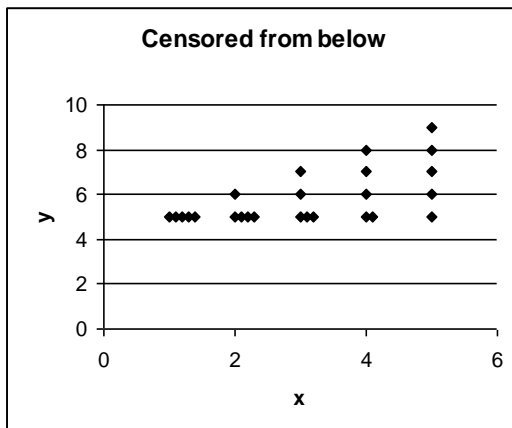
We observe the full range of Y and the full range of X

Example 5.2: Censoring from above



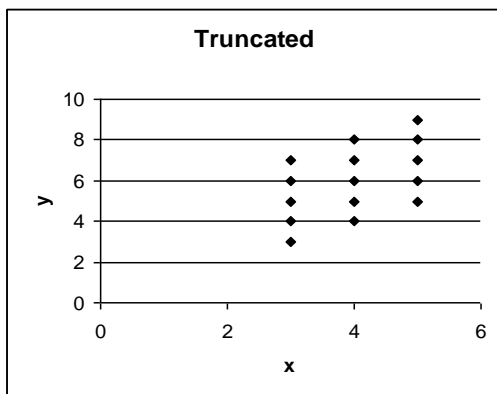
Here if $Y \geq 6$, we do not know its exact value

Example 5.3: Censoring from below



Here, if $Y \leq 5$, we do not know its exact value.

Example 5.4: Truncation



Here if $X < 3$, we do not know the value of Y .

5.2.3 Conceptualizing censored data

What do we make of a variable like “Days spent in the hospital in the last year”? For all the respondents with 0 days, we think of those cases as “left censored from below”. Think of a latent variable for sickness that underlies “days spent in the hospital in the past year”. Extremely healthy individuals would have a latent level of sickness far below zero if that were possible.

Possible solutions for Censored Data

Assume that Y is censored from below at 0. Then we have the following options:

1) Do a logit or probit for $Y = 0$ vs. $Y > 0$.

You should always try this solution to check your results. However, this approach omits much of the information about Y .

2) Do an OLS regression for truncated ranges of X where all $Y > 0$. This is another valuable double-check. However, you lose all information about Y for wide ranges of X .

3) Do OLS on observations where $Y > 0$. This is bad. It leads to censoring bias, and tends to underestimate the true relationship between X and Y .

4) Do OLS on all cases. This is usually an implausible model. By averaging the flat part and the sloped part, you come up with an overall prediction line that fits poorly for all values of X .

5.2.4 The selection part of the Tobit Model

To understand selection, we need equations to identify cases that are not censored.

$$y_i > 0 \quad \text{implies} \quad \beta x_i + \varepsilon_i > 0$$

$$\text{So } \text{pr}(y > 0 \mid x) = \text{pr}(\beta x + \varepsilon) > 0$$

So $\text{pr}(y > 0 \mid x)$ is the probability associated with a z-score $z = \beta x / \varepsilon$

Hence, if we can estimate β and ε for noncensored cases, we can estimate the probability that a case will be noncensored!

5.2.5 The regression part of the Tobit Model

To understand regression, we need equations to identify the predicted value for cases that are not censored.

$$E(y) \mid y > 0 = \beta x + \varepsilon$$

where β is the slope of the latent regression line and where ε is the standard deviation of y , conditional on x .

5.2.6 A warning about the regression part of the Tobit Model

It is important to note that the slope β of the latent regression line will not be the observed slope for the uncensored cases! This is because $E(\epsilon) > 0$ for the uncensored cases!

For a given value of x , the censored cases will be the ones with the most negative ϵ . The more censored cases there are at a given value of x , the higher the $E(\epsilon)$ for the few uncensored cases. This pattern tends to flatten the observed regression line for uncensored cases.

5.2.7 The catch of the Tobit model

To estimate the probit part (the probability of being uncensored), one needs to estimate β and ϵ from the regression part. To estimate the regression part (β and ϵ), one needs to estimate the probability of being uncensored from the probit part. The solution is obtained through repeated (iterative) guesses by maximum likelihood estimation.

5.2.8 OLS regression without censoring or truncation

Why is the slope too shallow in the censored model? Think about the two cases where $x = 0$ and $y > 3$. In those cases $E(\epsilon) \neq 0$, because all cases where the error is negative or near zero have been censored from the population. The regression model is reading cases with a strongly positive error, and it is assuming that the average error is in fact zero. As a result the model assumes that the true value of Y is too high when X is near zero. This makes the regression line too flat.

5.2.9 Why does the Tobit work where the OLS failed?

When a Tobit model “looks” at a value of Y , it does not assume that the error is zero. Instead, it estimates a value for the error based on the number of censored cases for other observations of Y for comparable values of X . Actually, STATA does not “look” at observations one at a time. It simply finds the maximum likelihood for the whole matrix, including censored and noncensored cases.

5.2.10 The grave weakness of the Tobit model

The Tobit model makes the same assumptions about error distributions as the OLS model, but it is much more vulnerable to violations of those assumptions. The examples I will show you involve violations of the assumption of *homoskedasticity*. In an OLS model with heteroskedastic errors, the estimated standard errors can be too small. In a Tobit model with heteroskedastic errors, the computer uses a bad estimate of the error distribution to determine the chance that a case would be censored, and the coefficient is badly biased.