Microeconomics: Theory & Applications
Prof. Deep Mukherjee
Department of Economic Sciences
Indian Institute of Technology, Kanpur

Lecture – 58
Linear Regression (Part-1)

Hi, welcome back to the lecture series on Microeconomics. Today we are going to start our discussion on an empirical tool called Linear Regression analysis. Linear regression analysis is one of the tools in the vast field of econometric methods. Now, what does econometrics do? So, microeconomic theory or for that matter any economic theory actually states or make some hypotheses to explain the real life economic phenomenon, but these are all qualitative in nature.

So, for that matter take the example of the law of demand that we have seen in the theory of consumer behavior. What does it say? It says that if price of a commodity increases then quantity demanded of that particular commodity will decrease given other things remaining the same. So, this is a theoretical result. But if I now ask you that suppose, we have a commodity whose price has gone down by 1 rupees or 1 dollar. Can you tell me by how much the quantity demand is going to change? The answer is no.

From this theoretical result, you cannot quantify the magnitude of the change in the quantity demand due to a price change. Econometrician we will collect data and try to fit some kind of statistical relationship between price and quantity so, that you will be able to give this answer, you will be able to deliver the answer to this very question. Now, one may ask that we have also gone through the concepts like price elasticity right. So, if price changes by 1 percent by what percent this quantity demand will change? That elasticity can also be computed only when a functional form a specific functional form for demand function is known to you.

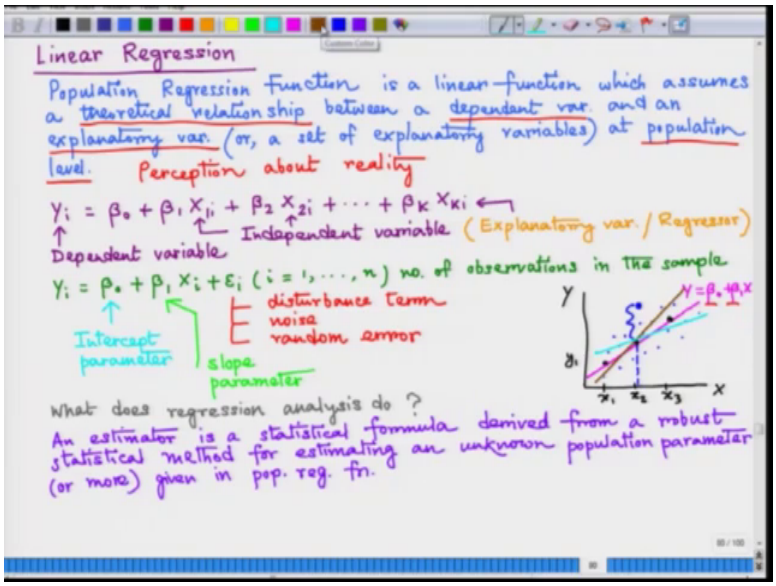
So, to compute the elasticity also we need to quantify the demand function. And simple linear regression analysis will help us to establish some kind of statistical associationship between statistical association between the price and quantity in this case and in general between any 2 or more economic variables.

So, actually the economists will start with some kind of economic theory which we will talk about some true relationship between some variables. For that matter we can assume that we are interested only on 2 variables. So, there will be some kind of true relationship which we will prevail at the population level between these 2 variables.

Now, there may be a true relationship between 2 variables, but in the real life when we observe real life data then basically there could be some error or noise because of many many issues; one factor could be omitted variables. So, as we know that economy is a very complex thing and one variable gets influenced by not only one particular variable, but many other variables. So, in a simple theoretical model, when we incorporate 1 or 2 or 3 variables in a functional form in a mathematical functional form, we tend to rule out many other factors which are probably unforeseen to an economist.

So, that is why when we observe real life data on 2 or more economic variables, we actually do not observe a curve or a straight line which we are used to draw in microeconomics rather we actually observe a scatter. So, a true relationship which can be also called a hypothesis can be represented through the notion of population regression function. So, let us have a definition for this concept, ok.

(Refer Slide Time: 04:20)



So, let us start with definition of population regression function. So, here we are talking about a theoretical relationship between 2 or more variables. The theoretical relationship can come from micro economic theory in our case and you know they had basically

describes some kind of relationship between a dependent variable and one or more explanatory variables. So, just if you remember our previous discussion that we just had on the law of demand, in that case the dependent variable would be the quantity demanded of a particular commodity and the explanatory variable could be the price of that particular commodity and income level of the consumer etcetera, etcetera.

So, this is basically conceived at the population level. So, this is some kind of a relationship which is kind of perception about reality ok. Now, whether this perception about reality is actually the case or not that needs to be tested and econometrics does so. So, here we are going to study linear regression analysis and the purpose of linear regression analysis is to take a purely theoretical equation which is basically a population regression function, something like this; Y i equal to beta naught plus beta 1 X 1 and this is basically my population regression function 1 i ok.

So, here I define Y as my dependent variable. So, dependent variable is basically the variable whose behavior is explained through a model and this X 1 in this case is basically my independent variable or explanatory variable and needless to say that you can add more explanatory variables. If you assume that there are more number variables which are affecting the variation in this dependent variable Y. So, you can have as many as you want in terms of the variables.

So, this is basically the notion of the population regression function. So, after laying out the fundamental population regression function, let us now, go and study the regression analysis in deeper details. So, in a regression analysis, we take a pure theoretical relation and let me now simplify the analysis by assuming only one variable, one explanatory variable, but needless to say that this can be generalized to more than one variables anytime right. So, only one explanatory variables X and what is i? i is basically 1 to n; these are basically the number of observations right ok.

So, now let us go back to the statistical relationship. We have already told you that in theory we can have a straight line or a simple curve, but in reality there can be many random noises in the data and for that reason we introduce an extra term in this regression equation. And, this is called disturbance term or some people call it noise, sometimes it is also called random error ok.

Now, quickly let us note down the interpretations for these parameters beta naught and beta 1. So, beta naught is basically called an intercept parameter and this beta 1 is called slope parameter. So, these are the relationships defined at the population level. So, this is the true relationship. Now, what are the interpretations for this? Let us first talk about the intercept parameter.

So, intercept parameter actually gives a constant variation in the; so, beta naught, the intercept parameter actually is a naive regressor or explanatory variable which tries to explain the variation in Y, but you know this is a constant term basically. So, later we will see that this is nothing but the sample mean of the dependent variable ok.

Now, let us move on to beta 1. So, beta 1 is basically the slope parameter and that basically comes with the explanatory variable. So, if we throw 1 or more explanatory variables in the regression equation, these variables are expected to explain some variation in the main variable or dependent variable Y right. So, what would be the interpretation for beta 1? It is simply say is that, it simply says that if there is 1 unit change in the explanatory variable or the independent variable, then by how much my dependent variable Y is going to change.

So, now let me try to address what does regression analysis do. So, as I told you that the population regression function actually talks about the true population relationship or its coming from some theory or hypothesis., But, the problem is this, we really do not know the exact values of beta naught and beta 1 and beta k, if there are k number explanatory variables. We need to know some numbers for these parameters in the regression equation.

So, how to find some proxy values for these true population parameters? Our regression analysis actually tries to find out good proxies for this population parameters beta naught, beta 1 to beta k. And there are many methods to obtain these proxies. We are going to study only one such proxy and the name of the method is ordinary least squares. But, before we get into the details of the ordinary least squares method, let us have a look at 2 more definitions.

So, any regression method actually leads to an estimator and here we are writing the definition of an estimator. So, an estimator is a statistical formula derived from ok so, now, let us going to have another definition and this time it is for fine, no definition. So,

now, let me have a simple diagram to explain the philosophy behind regression analysis. I hope this simple diagram we will help you to understand this tool better.

So, let me have this explanatory variable plotted along the horizontal axis and the dependent variable or the variable of focus Y along the horizontal axis. So, the population regression function says that there is some kind of a linear relationship between these 2 variables and if we plot that then we can say that say for some value of X say X 1, we observe some value Y 1 and so on so forth.

So, let me just you know have 2 more data points, there can be many. So, let me assume that there is some positive relationship between Y and X and this comes from some kind of theory or hypothesis. So, basically the population regression function is basically straight line which joins these points and this is the true relationship between Y and X. So, basically one can write Y equal to beta naught plus beta 1 X ok.

But due to the presence of noise in the economy and of course, as a result in the data, we actually observe lots of points. In this Y explain, if an econometrician collects some data from reality then basically this diagram is called a scatter diagram and now the task at hand is to find out why there is a deviation from this pink line or the population regression line.

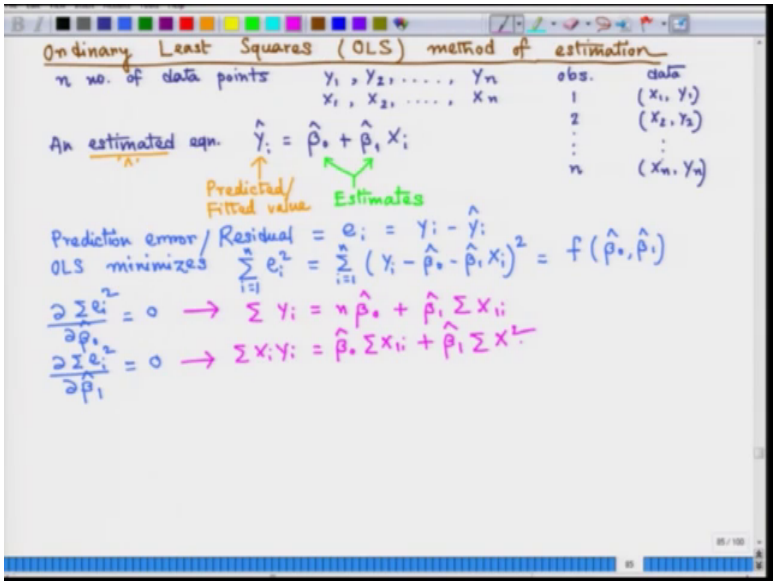
So, suppose let me take a particular value of X say X 2, then I expect that my population regression function as it is giving me value of Y. In reality, I should see that Y value only for X 2, but I am observing some data in reality, here this big large blue dot. So, what is causing this difference? So, can we explain this variation in Y? So, that is the question with which regression analysis. So, this is the problem with which regression analysis is concerned.

Now, as we really do not know the values for beta naught and beta 1, then we can take help of statistical methods to come up with some kind of sample regression function. So, we know we can devise some estimator which will now be a fitted line amongst these data points in the scatter plot. So, we are basically saying that we do not know the true relation between the we do not know the true parameter values. But, note we can adopt some statistical method to come up with you know some estimator, so that we can feed some kind of a straight line through the data.

Now, note that there could be many estimators and hence there is no unique straight line. So, if you draw this brown straight line then it will represent one particular combination of beta naught and beta 1. And if I now draw this sky blue straight line, this also passes through the same scatter plot. So, if I now plot the sky blue straight line which also passes through the data that I have or the scatter diagram that I have, but it represents a different combination of beta naught and beta 1.

So, these 2 lines that I have drawn these are called the sample regression functions right. The task at hand is to find the best possible sample regression function which fits the data ok. There are many methods to find proxy values for beta parameters. We are going to now study the simplest possible method which is called ordinary least squares.

(Refer Slide Time: 22:27)



Let us talk about some estimated equation from the real life data. So, suppose we are given some data points like there are n number of data points in the scatter plot. Note that, I am going to restrict my discussion only on one explanatory variables; in the next class we are going to expand the number of explanatory variables. So, whatever I am going to discuss now is quite general. So, it will hold for key number of explanatory variables as well.

So, let us start with n number of data points. So, we observe some values of the dependent variable like Y 1, Y 2, up to Y n corresponding to the explanatory variable or the regressor variable values X 1, X 2, to X n right. So, basically we are observing the

data as a pair. So, basically this is the way you can think of the data set. So, observation number 1, number 2 dot dot number n and here is basically my data and data is basically X 1, Y 1 a pair of observed pair of variables values; X 2, Y 2 and finally, X n and Y n. So, these are basically the points in the scatter diagram ok. So, now an estimated equation using this data would look like right. So, here note that I have introduced this new notation hat. So, hat basically represents this term estimated ok. So, this is basically represented by this carrot or hat sign ok. So, this Y i Y hat i is called the fitted value or the predicted value ok.

Now, then this beta naught hat and beta 1 hat are basically called the estimates of the population parameter from the data. Now, the question is how to find these estimates. So, the first thing that we would like to do is to write down another expression for prediction error or residual. Of course, we can understand that that as we do not know the true values for beta naught and beta 1, given a dataset we can only produce some proxy numbers and they may be close to the original or true population parameter value or they may not.

So, whatever be the case, you know in most cases you know they are not equal to the population parameter values; these estimates I am talking about. Hence, basically the Y hat i, the predicted or fitted value that we obtain from our estimated equation or regression equation will not match the actual data which is Y i. So, there will be some difference and then this difference is called prediction error or residual of the regression model; this is given by e i and that is basically defined as Y i minus Y hat i; so, the actual value minus the fitted value.

Now, this method OLS, minimizes the sums sum of squared residuals ok. Now, why you know we need to minimize this sum squared of total? First of all, our target is 2 ok; why we have to minimize this sum squared of residuals? First of all as an econometrician or a statistician, I would like to minimize the error. So, that is why I have to minimize. But why square term?

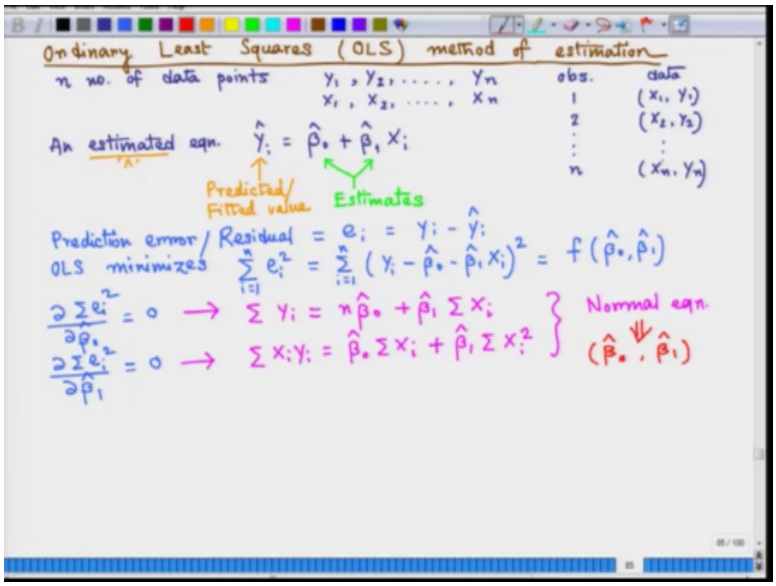
Because note from that diagram that we had which depicted the population regression function and the sample regression equations that you know error could be of either positive or negative type. So, if you have only some of errors then you know positive and negative numbers will cancel out altogether. That is why we you are squaring the error

term, so that we deal with only the positive numbers and we minimize the sum of squared residuals.

So, basically you see this sum of squared residuals. Now, let me expand this. So, we can now write ok. So, basically what we see? The sum squared of residuals is a function of the unknown entities like beta naught hat and beta 1 hat. So, the target is to find beta naught hat and beta 1 hat in such a way that the sum of squared residuals is minimized. So, we know how to minimize any function. So, we have to basically take partial derivatives and set the m 2 equal to 0. So, there will be 2 such derivatives ok.

So, note that these expression leads to an equation like this. Here, the sum ranges from i equal to 1 to n; just to just to have a less cluttered expression, I am avoiding the sum range ok. So, this is one equation and this first order condition leads to another equation.

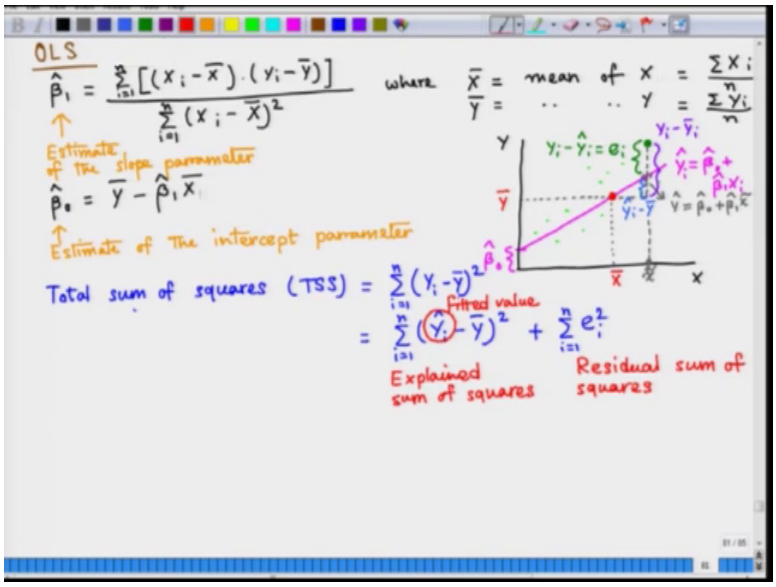
(Refer Slide Time: 31:38)



So, the second fast order condition also leads to an equation which looks like; oops I mistakenly written 1 here. There is only one explanatory variable. So, let me just remove that 1 suffix, sorry subscript ok. So, these are called the normal equations and if you solve these 2 normal equations, then you will get beta naught hat and beta 1 hat such that as a pair they minimize the sum of squared residuals.

Let us now look at a graph which talks about the decomposition of the variation in Y. Once we have estimated the sample linear regression function, that means, that we know our estimates for the population parameters beta naught and beta 1.

(Refer Slide Time: 33:47)



Plot Y bar X bar here in the graph. Let me assume any particular value of X, say X tilde and for this value this point is basically Y hat i not i, Y hat equals beta naught hat plus beta 1 X tilde right. Suppose, the original data says that the corresponding Y value, the Y value corresponding to the X tilde value is this green dot here right.

So, in that case basically, this gap is my residual right; e i which is basically the difference between the actual value minus the fitted value Y i right ok. And then for the diagrams completeness sake, let me extend this Y bar. So, now, we are going to see some important concepts which emerges from the OLS method. The first concept is of Total Sum of Squares or abbreviated as TSS. So, now this could be broken down into 2 pieces ok. So, the first component is known as the explained sum of squares. Why this is called explained sum of squares? Because of this fact that here the Y hat i this component is actually given by the sample regression function. So, here we see the role of the explanatory variables because, you know this Y hat i is the fitted value right.

So, given the values of X's and the relationship between these variables that we have assumed; this is the fitted value and this is basically some kind of value which comes

from within the model. So, the other component is known as the residual sum of squares ok, fine ok.

So, we will continue with this discussion on linear regression analysis in the next lecture.