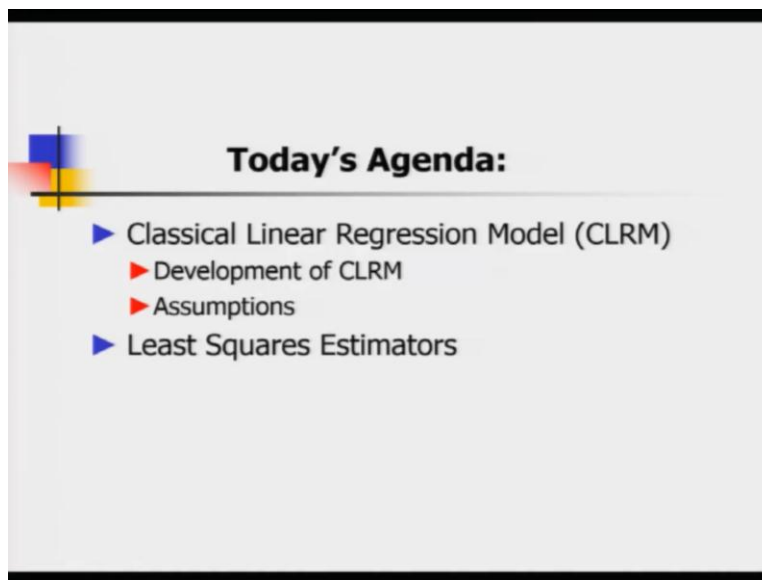


Applied Statistics and Econometrics
Professor Deep Mukherjee
Department of Economic Sciences
Indian Institute of Technology Kanpur
Lecture 25
Classical Linear Regression Model (Part-I)

Hello friends. Welcome back to the lecture series on Applied Statistics and Econometrics. So, today we are going to start the discussion on part 2 of the course which is econometrics. So, here in this course the focus is going to be the linear regression models and we are going to start from the very beginning although we have done a little bit of least squares methods in the part 1 applied statistics but I will provide you a brief recap.

And then gradually we will continue with the linear regression models and develop higher level models. So, before we start again with linear regression models in the context of econometrics, let us have a look at the agenda items for today.

(Refer Slide Time: 01:01)



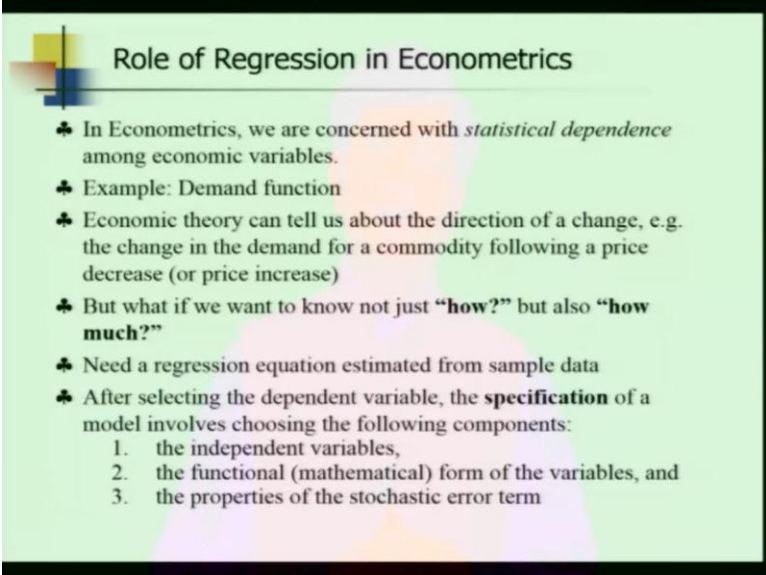
So, in today's lecture I am going to introduce what do we mean by classical linear regression model and abbreviation is also very popular in the field and that is known as CLRM. And here in today's lecture I am going to show you how the CLRM is actually developed and then I will talk about the assumptions and then finally I will end by having us brief discussion on least squares estimators.

So, in econometrics what we would like to do? So, here everything starts from an economic behavior or phenomenon that we see around us and then economic theorists have developed certain models. And these are all philosophical models mostly expressed in words and then later these models are also converted in mathematical symbols, So, we get robust mathematical models.

But then from these mathematical models we cannot guess how a particular variable is going to behave if there is a change in one particular variable. Well let me correct myself here, of course economic theory gives indication to some variables, how that variable is going to behave if there is a change in a related variable but if you quantify how much is the change then economic theory is silent about that.

So, for that you have to collect real life data and then you have to quantify certain relationships between these economic variables on which you are interested. And then an equation finally will come up from econometric estimations and that equation will be able to help you to tell that okay if there is a change in one variable X then how a related variable Y is going to change and not only that but by how much it is going to change.

(Refer Slide Time: 03:09)



Role of Regression in Econometrics

- ♣ In Econometrics, we are concerned with *statistical dependence* among economic variables.
- ♣ Example: Demand function
- ♣ Economic theory can tell us about the direction of a change, e.g. the change in the demand for a commodity following a price decrease (or price increase)
- ♣ But what if we want to know not just “**how?**” but also “**how much?**”
- ♣ Need a regression equation estimated from sample data
- ♣ After selecting the dependent variable, the **specification** of a model involves choosing the following components:
 1. the independent variables,
 2. the functional (mathematical) form of the variables, and
 3. the properties of the stochastic error term

So, here in this slide we are going to discuss the role of regression in econometrics but already I have spoken about the 1st bullet point. So, let me start with the 2nd bullet point and that is

basically the demand function. So, here in this example I borrowed this concept from microeconomic theory and that is called demand function.

We all know about the basic idea of a demand function it says that if price of a commodity increases then quantity demanded of that commodity decreases or vice versa. But when you state this law of demand then there are many other variables or factors for which you are assuming that they are not changing.

So, at a time only price of that commodity will change and then you will see some particular behavior of the quantity demanded but what about the other factors which you are keeping fixed when you are describing the law of demand. So, there are many other factors which can also impact demand. What are these factors?

One could be income of the consumer, 2nd could be the price of a complementary good and then 3rd could be price of a substitute good. So, let me give you an example So, suppose I am interested to figure out the demand for tea. Now of course demand for tea will definitely depend primarily on the price of tea.

But what about the income of the consumer? So, if income increases then you expect the quantity demanded for tea to go up and then what about the price of sugar? Sugar is a complementary commodity to tea So, if price of sugar increases then what do you expect? Will the demand for tea increase or will it fall?

So, our economic theory says that if the price of complementary goods increase then there is a fall in the quantity of the main commodity. So, here if the price of sugar increases then we expect that quantity demanded for tea shall go down. Now what about a change in the price of a substitute commodity? So, what could be a substitute for tea? It could be coffee So, if say price of coffee increases then what will be the impact on the demand for tea?

So, if substitute goods price increase So, the coffee price increases then you expect that quantity demanded of tea shall go up because now the alternative item has become costlier. So, here you see that economic theory tells you if there is 1 unit change in the So, called explanatory or independent variables in an economic model then how my dependent variable quantity demanded for tea is going to change.

So, basically you need to estimate a regression equation from the sample data. Now in econometrics are we happy only with the estimation of regression equation? Here in this example say demand function for tea answer is no. We can make use of this demand function many ways. So, one good example could be like projecting the future demand for tea. So, the regression equation once it is estimated from the sample data can lead to the future predictions.

So, now let us come back to the general discussion. We have had enough discussion on the example. So, here after you select your dependent variable, the variable of focus then what you have to do if you want to develop an econometric model. So, you have to now look at the other specification issues of the model and there are 3 different components of a model specification except for the dependent variable Y .

So, there could be axis the independent variables or explanatory variables or in many text books you will see the term regressors and the 2nd component of the model specification is the functional mathematical form of the variable. So, when you say that your dependent variable Y is linked to a set of independent explanatory variables axis then in which way they are related.

So, there has to be an explicit functional form, a mathematical equation, So, that that your variables are going to be linked in this particular manner. And then once that is set you have to assume certain properties of the random disturbance of the stochastic error term and then these properties of this stochastic disturbance term will lead you to a particular estimator. So, at this moment let us make a distinction between estimator and estimate its very crucial.

So, what is an estimator? An estimator is a mathematical formula for which you can generate different values once you plug the values of the variables that you see in that formula. So, in this lecture and in the next lecture also we will see that when we will conduct regression our basic aim is to first find out an estimator formula which is good in many ways.

Now why we have to make the assumptions on the stochastic or random error terms? Because when we want to establish certain properties for the estimators we will require the assumptions regarding the stochastic or random error terms, hence the properties of the stochastic error term or random term is also very important in regression analysis in econometrics.

So, before we start discussion on the estimation of regression equations in econometrics, let me talk about a very simple story. And this story or example or illustration whatever you want to call

it is a motivation and this will show you how from data, from the very scratch we can think about regression problem in econometrics.

(Refer Slide Time: 09:41)

Development of Regression Model: An Example

- Consider a data set from randomly chosen households in a city. We observe their monthly income (in Rs. 1000, denoted by X) and monthly expenditure (in Rs. 1000, denoted by Y).

| | | X: Monthly HH Income | | | | | | | | |
|---------------------------|------|----------------------|----|------|-----|------|-------|-------|-----|-----|
| | | 40 | 60 | 80 | 100 | 120 | 140 | 160 | 180 | 200 |
| Y: Monthly HH Expenditure | 40 | 30 | 42 | 55 | 65 | 79 | 80 | 102 | 110 | 120 |
| | 60 | 32 | 44 | 60 | 70 | 84 | 93 | 107 | 115 | 130 |
| | 80 | 34 | 47 | 65 | 74 | 90 | 95 | 110 | 120 | 140 |
| | 100 | 36 | 50 | 70 | 80 | 94 | 103 | 116 | 130 | |
| | 120 | 37 | 52 | 75 | 85 | 98 | 108 | 118 | | |
| | 140 | 38 | 55 | | | | | | | |
| | 160 | | | | | | | | | |
| | 180 | | | | | | | | | |
| | 200 | | | | | | | | | |
| Mean Y | 34.5 | 48.3 | 65 | 74.8 | 89 | 95.8 | 110.6 | 118.7 | 130 | |

- What is the expected value of consumption expenditure of a household whose monthly income is Rs. 40,000? → Find conditional mean

| | | X: Monthly HH Income | | | | | | | | |
|---------------------------|------|----------------------|--------|------|-----|------|-------|-------|------|--------|
| | | 40 | 60 | 80 | 100 | 120 | 140 | 160 | 180 | 200 |
| P(Y X): Conditional prob. | 40 | 0.1667 | 0.1667 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.25 | 0.3333 |
| | 60 | 0.1667 | 0.1667 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.25 | 0.3333 |
| | 80 | 0.1667 | 0.1667 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.25 | 0.3333 |
| | 100 | 0.1667 | 0.1667 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.25 | |
| | 120 | 0.1667 | 0.1667 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | | |
| | 140 | 0.1667 | 0.1667 | | | | | | | |
| | 160 | | | | | | | | | |
| | 180 | | | | | | | | | |
| | 200 | | | | | | | | | |
| (Y X) | 34.5 | 48.3 | 65 | 74.8 | 89 | 95.8 | 110.6 | 118.7 | 130 | |

So, here we consider a data set from randomly chosen households in a city. It is a hypothetical data set of course. Now we observe their monthly income in rupees 1000 and that is denoted by X and then we also have data on their monthly expenditure again in rupees 1000 and this is denoted by Y.

Now from economic theory we have this idea that quantity demanded and consumption expenditure depends a lot on the income of the consumer. So, here we are taking household as the consumption unit. So, in that case the consumption expenditure made by household shall heavily depend on monthly income and from that economic theory or hypothesis we are developing this example.

So, suppose we have collected data on some households and I am going to now show you representation of the collected data in terms of 2 tables. So, here concentrate on the table number 1, which is the 1st table in the slide. So, note that here in the columns, I am showing monthly household income in rupees 1000.

So, all these numbers 40 60 80 100 120 these are basically monthly household income total income of the household expressed in rupees 1000. And then in the rows, I am showing you all

different numbers on the household consumption expenditures, again these are monthly totals and expressed in 1000 also.

So, now let me provide you some explanation about the data. So, you are observing some monthly income totals like 40000 60000 80000. Now this may look very crude but you may ask from where did I get this kind of exact numbers for income. In reality people do not disclose their income level.

So, here in our questionnaire we have provided the income ranges to the household and the household head or the member who is responding to the questionnaire can put a tick mark in the income box that he or she feels the best suited for his or her household. So, we started with some box like less than 30000 and then the 2nd income range or the box says that total monthly household income is between 30000 and 50000 rupees.

And again the next box or the next income range says that your total monthly household income is some number between 51000 and 70000 and this list goes on like that. And the person puts a tick in the appropriate box and then we, these are basically the class limits or different class intervals if you remember previous discussion in applied statistics part when we were discussing frequency tables.

So, we can just take the midpoint of this class. So, these are the class marks and I mean these will give you a very rough idea about the monthly household income of the particular household. So, that is how we get these discrete numbers and this is an arbitrary hypothetical data set. So, do not worry about the numbers how they are generated as such at this moment. But just assume that you see different level of household incomes.

And then you see that for each household income level, I am reporting the monthly household expenses reported by various households in my sample. So, you see for the 40000 level, I see 6 values reported then for 60000 monthly household income also I see there are 6 values of monthly household expenses. And then you see the last column 200 shows that there are only 3 numbers.

So, basically in our sample we had only 3 households who reported that high monthly household income and this table basically gives me the idea about the bivariate distribution of the X and Y. So, from here for each column now I can calculate the sample mean of Y. So, basically we can

calculate the \bar{Y} from the sample. But note that these are for all given axis. So, for example for the column 40, if you calculate the sample mean for Y then you get 34.5.

So, now an interesting question could be asked that, what is the expected value of consumption expenditure of a household whose monthly income is some number in the table say 40000 or 60000 or 200000 So, then how would you answer? So, to answer this particular question you have to look at the conditional mean. Now one can say that why conditional mean, So, for that let us go back to the slide again and then I will explain you through the 2nd table in the slide.

So, here you see we have table number 2 at the bottom of the slide and there I am showing you a bivariate probability distribution for X and Y. So, what do you see here? So, the X values are given the 40 60 80 numbers are given and then you see in the rows, I am not reporting the consumption figures. I am reporting some numbers in the fractions. So, these are basically the probabilities. So, these are all the conditional probabilities.

So, what do I mean by that? So, I am actually measuring probability of Y given for some particular value of X. So, if I fix the value of X to be 40 then you see in the 1st table I have 6 numbers for Y. Now these 6, I assume they are equi-probable. So, if they are equi-probable then the probability of a particular number say 32 or 37 in that column will be all 1 over 6 that is 0.1667.

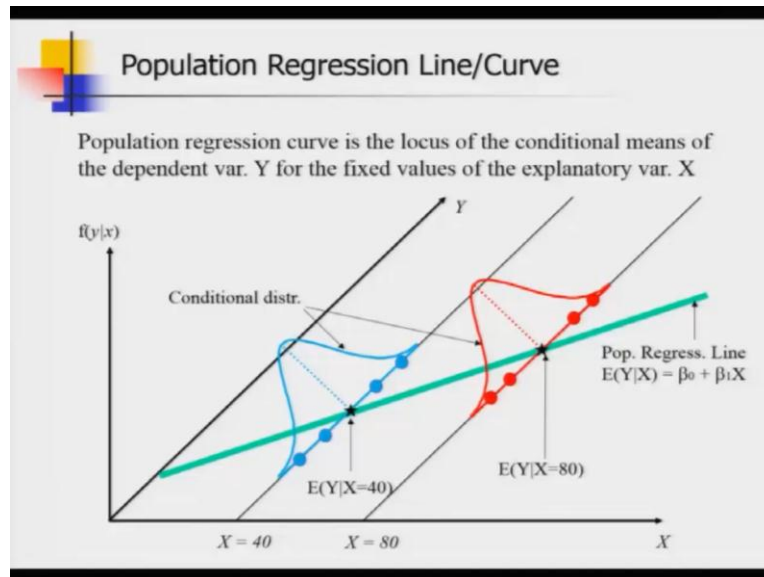
So, that is why I am showing you 0.1667 in all 6 rows under the column 40. And similarly for other columns also I have computed the probabilities. So, now you can apply the expected mean value or the mean formula for a random variable. So, now you say that my Y the consumption expenses data that actually is a random variable and it will take different values with different probabilities.

So, here in the data you see there are 6 possible values for Y for given X equal to 40 and there are probabilities associated with them and then you multiply the probability values with the realized value and then you report no our mean the expected value and that expected value is basically a conditional mean because this is computed after assuming one particular value of X.

So, this is the way you can actually find the answer. But note that this is a small sample case. If you are dealing with say very large samples a 1000 then of course you can have more number of data points and then you also have more values or levels for variables X and Y. So, these tables

will look very complicated and as the sample increases then you can say that okay you can assume existence of normal distribution, you can bring in the story of normal approximation and this conditional probabilities will not be uniform as you are shown here in this simple example.

(Refer Slide Time: 18:13)



So, in this slide we will now take the lessons from the previous slide and try to tell you the same story but through a graph. So, here we are going to introduce the concept of population regression line or population regression curve. So, here before I talk about the formal definition let us concentrate on the graph.

So, here I am showing you a 3 dimensional diagram, here along the horizontal axis I am measuring the explanatory variable X that is my income household income level and then along the vertical axis number 1 that is for Y. So, that is basically measuring my consumption level, total monthly consumption for household. And then there is another axis here and that is basically the z axis or that is named as the conditional probability f of y given x.

So, I am also measuring the conditional probabilities. Now suppose I have a very large sample it was not a small sample like I had in knowing the previous slide. So, the story from the previous slide stays in this slide also. But here I am dealing with a large data set. So, now let us see how the story can be retold in terms of the diagram. So, let us assume that we are interested to discuss the case of X equal to 40 and X equal to 80.

These are arbitrarily chosen numbers there is no specific reason why I have chosen 40 and 80. You can choose 60 and 120 also it does not matter. Anyway So, now after you have chosen 2 particular values of X then you draw some vertical lines parallel to the Y axis and these vertical lines I have shown you both of them are in black. Now you note that here suppose you concentrate on X equal to 40.

Now see you have 4 circles in light blue colors those are plotted on this vertical line. What are these circles? So, these circles are basically the observed numbers of total monthly household consumption expenditure for some households who has reported total monthly household income of rupees 40000. Now these 4 are basically representative numbers. There could be 40 more.

So, basically if you have a large number of y's for X equal to 40 then actually you can assume the normal distribution, you can bring in the normal approximation because you have large sample. So, then you can assume that there is some kind of distribution here approximately normal bell shaped and that is given in light blue color as well. So, if this distribution you can assume from the observed data then of course there will be a mean for this distribution is not it?

So, that distribution mean is also marked in light blue. So, I have this broken line which is connecting this peak of the distribution and the asterisk or a star point here in black color. So, that star or this black asterisk point expresses or represents the conditional mean and this conditional mean is basically represented mathematically by this expression expected value of Y given X equal to 40.

So, now let me move on to the case of X equal to 80, same story is applicable here also. So, if you follow the same steps or the same logic that I told you in the case of X equal to 40 you get another asterisk point or black star here. And that is again the conditional mean and mathematically that is expressed as expected value of Y given X equal to 80. So, now you have got 2 points on X and Y plane.

So, we all know that you can draw a straight line if you have 2 data points. So, basically I am drawing a straight line here that will pass through these 2 conditional mean points in, these are represented by black stars. So, note that I have a thick green line or a straight line that I have drawn and this is going through the 2 black star points. So, this thick green line is called the population regression line.

So, now I go with the formal definition that I am showing here at the top of the slide. So, a population regression curve is the locus of the conditional means of the dependent variable Y for the fixed values of the explanatory variable X. So, note that in that slide we have written line slash curve. Why I have written both? Because if you are assuming the linear relationship between 2 variables Y and X then you draw a line.

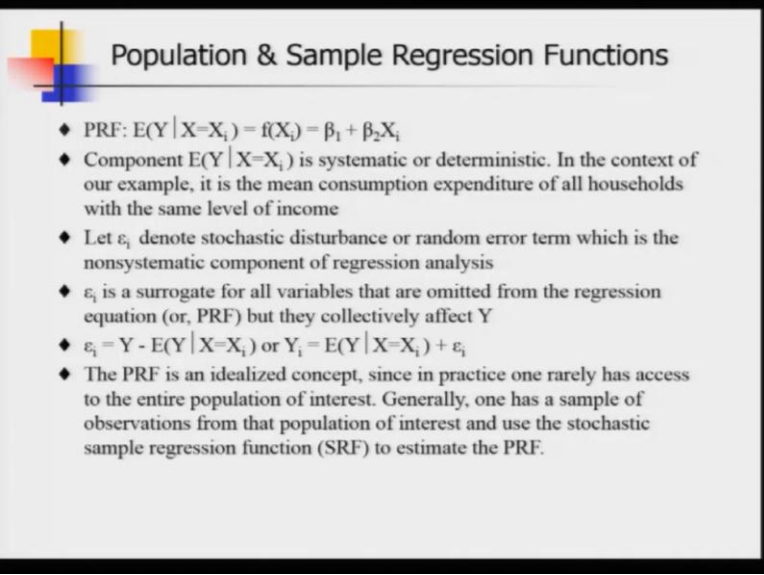
But if you assume that there is non-linearity between Y and X and indeed if you see that after plotting 4 or 5 conditional mean data points on the X Y coordinate plane. If you see that there is sufficient non-linearity, then you can also connect these points by a curve. So, that is why this concept can also be straight line, it can also be a curve. So, that is why population regression can be expressed in both ways either in terms of a straight line or in terms of a curve.

So, now if I take you back to the diagram you see that population regression line, as it is a straight line we can assume a functional form for that or a mathematical expression for that and that is what is written here, expected value of Y given X equals to beta naught plus beta 1 times X. So, here this beta naught and beta 1 are the regression coefficients and you know beta naught is my intercept coefficient and beta 1 is my slope coefficient from the knowledge of mathematics.

We have also discussed this issue in the case of curve fitting in the part 1 of the course as well. So, this is not new to you. So, we got a relationship called population regression function and from there now we will begin our journey of estimation of this relationship between 2 economic variables. Population regression function actually is unknown to you. Why? Because you know that, that is the locus of the conditional mean of Y given the axis.

But in your data set you may not be lucky to have these data points. So, basically what you observe you have raw data set and you have 100 or 1000 observations but you do not know which point actually is this black star point as per the slide that I have shown you last. So, you have to define some kind of a statistical technique So, that you can estimate some straight line you can get a very close proxy for that straight line that population regression line.

(Refer Slide Time: 26:17)



Population & Sample Regression Functions

- ◆ PRF: $E(Y | X=X_i) = f(X_i) = \beta_1 + \beta_2 X_i$
- ◆ Component $E(Y | X=X_i)$ is systematic or deterministic. In the context of our example, it is the mean consumption expenditure of all households with the same level of income
- ◆ Let ε_i denote stochastic disturbance or random error term which is the nonsystematic component of regression analysis
- ◆ ε_i is a surrogate for all variables that are omitted from the regression equation (or, PRF) but they collectively affect Y
- ◆ $\varepsilon_i = Y - E(Y | X=X_i)$ or $Y_i = E(Y | X=X_i) + \varepsilon_i$
- ◆ The PRF is an idealized concept, since in practice one rarely has access to the entire population of interest. Generally, one has a sample of observations from that population of interest and use the stochastic sample regression function (SRF) to estimate the PRF.

So, here in this slide we are going to talk about the population regression function and the sample regression function. So, the 1st bullet point is basically repeating the population regression function concept from the previous slide. So, we have already discussed about it. Now you see that this component that I have written in the 1st bullet point which is the population regression function is a systematic or it is a deterministic component, there is no stochasticity involved in it.

Because mean once it is computed is a constant number. So, there is no fluctuation there. Now in the context of our sample or example it is the mean consumption expenditure of all households with the same level of income. But real life is not this simple like text book or classroom. So, here we are trying to explain the variation in monthly household consumption expenditure through one single explanatory variable that is monthly household income, of course total for all household members.

Now in reality there could be many other factors which are also affecting the household consumption expenditure but you are not accommodating them in your simple linear regression model. So, what could be the other factors? It could be the number of children in the household, number of aged in the household, number of college going students in the, school going students in the household. So, for aged people of course there could be medical expenses and also in overall expenses for a household could be a bit on the higher side.

For the school and college going children or students you have to spend money on tuition, then text books and uniform and these and that. So, of course in on an average the expenditure will be little bit higher then there could be another criteria or factor like the socio economic status. So, the factors I have spoken about which will impact the Y are also observed. So, if you collect data on these factors then you may think that I have collected every variable that I need to model Y. And then there should not be any need for any random component or any stochastic component. So, even a deterministic model is good enough do not you think so?

But the answer is no, unfortunately because still there could be other factors which may have impacted the households consumption expenses that you do not know at the time of survey probably you forgot to ask some questions like you may not have asked questions that okay in the last month or in last 3 months did you have some family gathering or some ceremony in your place like you have, have you married your daughter or son or was there any invitation or have you been to some tourist place for tourism.

So, you may not have asked these questions to the household and if one of these things actually happen then it would have an impact on the household consumption expenditure and unfortunately your model does not take care of that. So, to be on the safer side in all regression models you should have a random noise component.

So, here let us assume ϵ_i denote the stochastic disturbance or the random error term which is the non-systematic component of the regression model you are trying to develop. So, what is this i subscript doing? So, i is denoting the i th household. So, if you have n number of data points i will range from 1 to n . Now this ϵ_i is a surrogate for all the variables that are omitted from the regression equation or the population regression function.

But they are collectively can affect Y So, that is basically the idea of bringing in this ϵ_i in the model. So, I write ϵ_i as the difference between the actual observed value of dependent variable Y and the conditional mean that is expectation of Y given X equal to X_i or that can be simplified by this expression Y_i . So, the original observation Y_i the value of the dependent variable Y for the i th individual now can be broken into 2 components.

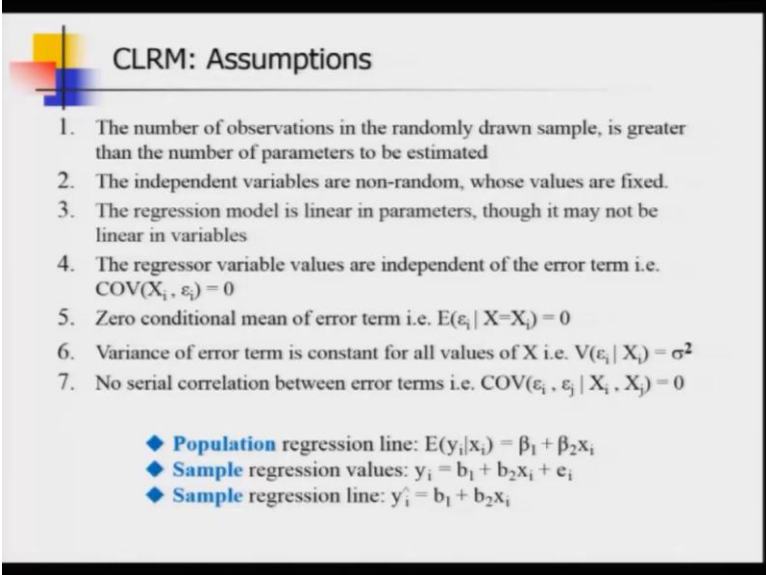
First the deterministic component which is coming from the population regression function and that has this $\beta_1 + \beta_2 X_i$ and the 2nd component is basically the random error

component which is the epsilon. So, now let us discuss about the distinction between population regression function and the sample regression function. So, what is the sample regression function?

So, we have already gone through the case of population regression function but that is basically in hypothetical world. You cannot get the exact values for beta 1 and beta 2 because these are the unknown true population parameter values. So, you have to get some proxies for them. How to get the proxies? You get the sample and then you follow some method So, that you can get some proxy values for beta 1 and beta 2.

So, once you have say sample 1, you can get some proxies like beta 1 hat and beta 2 hat and then you get a sample 2 So, you will get another set of beta 1 hat and beta 2 hat. So, these all possible combinations of beta 1 hat and beta 2 hat will give you a range of regression equations and these are called the sample regression functions.

(Refer Slide Time: 32:42)



CLRM: Assumptions

1. The number of observations in the randomly drawn sample, is greater than the number of parameters to be estimated
2. The independent variables are non-random, whose values are fixed.
3. The regression model is linear in parameters, though it may not be linear in variables
4. The regressor variable values are independent of the error term i.e. $\text{COV}(X_i, \varepsilon_i) = 0$
5. Zero conditional mean of error term i.e. $E(\varepsilon_i | X=X_i) = 0$
6. Variance of error term is constant for all values of X i.e. $V(\varepsilon_i | X_i) = \sigma^2$
7. No serial correlation between error terms i.e. $\text{COV}(\varepsilon_i, \varepsilon_j | X_i, X_j) = 0$

◆ **Population** regression line: $E(y_i|x_i) = \beta_1 + \beta_2x_i$
◆ **Sample** regression values: $y_i = b_1 + b_2x_i + e_i$
◆ **Sample** regression line: $y_i = b_1 + b_2x_i$

So, now we are going to talk about the assumptions that you must make in order to estimate a classical linear regression model. So, we will start with a very simple assumption and it says that the number of observations in the randomly drawn sample should be greater than the number of parameters to be estimated.

So, although we know this looks a bit silly but if you have a simple model that I have shown you previously with one intercept parameter and one slope parameter then you must have at least 3

observations to estimate the slope and the intercept parameter. So, that is basically in the nutshell we are talking about here.

Now in bullet point 2 we are saying the independent variables are non-random whose values are fixed. So, that is a very important assumption. Third assumption says that the regression model is linear in parameters though it may not be linear in variable. So, what do we mean by that it may not be linear in variables?

So, if there is non-linearity between 2 variables Y and X our regression technique the same linear regression model technique is well equipped to capture the non-linearity in the model. So, for capturing the non-linearity between 2 variables you can still adopt a linear regression analysis but you have to make the assumption that there could be non-linear relationship between 2 variables. And that you can take care of by taking square term or the logarithmic value of the explanatory variables in the right hand side. But the parameters of course they will be all linear. So, you should not have like some term like beta square or log of beta.

Now the 4th assumption says that the regressor variable values are independent of the error term. So, in statistical notation we write covariance between X_i and ϵ_i must be equal to 0. This is also famously known as the exogenous condition. Now we move on and the 5th point talks about the 0 conditional mean of the error term and that is expected value of ϵ_i given X equal to X_i and that should be equal to 0 and this is a very useful property we will see why later on when we are going to talk about the properties of the least squares estimators.

Now the 6th assumption says that the variance of error term is constant for all values of X_i and we call this constant variance assumption as the assumption of homoscedasticity and σ^2 is the common variance term that is assumed for all ϵ_i . Now the last one says that no serial correlation between the error terms it implies the covariance between ϵ_i and ϵ_j where i is not equal to j given X_i and X_j is equal to 0.

So, here what do we mean? So, if we have n number of observations we are saying that the error term value for say individual 1 and individual 17 and say individual 127 they are not linked to each other or you can also say that the error term between individual number 2 and 3 and then 3 and 4 they are also not linked in any way. So, we are done with the assumptions for the classical linear regression models.

Now you note the functional forms that I am showing you here in 3 bullet points at the bottom and we start with the population regression line So, you note that there is no error term So, it is a deterministic component of your regression analysis. Then the 2nd bullet point shows you the sample regression values. So, here you see that when you are talking about the sample regression values these will not involve the beta 1 and beta 2 neither it will involve the expectation.

Because expectation is basically coming from the probability distribution but here when you are dealing with sample you have the raw data with you So, you write Y_i that is basically the raw observation, the number that you see for the dependent variable Y for an individual element i in your sample and you see that beta 1 and beta 2 are now replaced by small b_1 and small b_2 . So, these are basically the estimates for the beta 1 and beta 2.

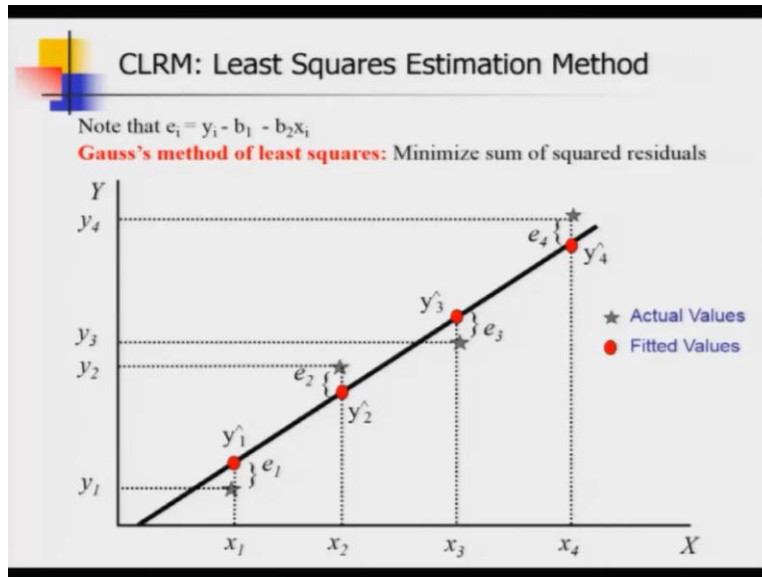
So, now how to get this estimates that we were going to discuss next. But note that as we are talking about the sample and we do not know the true parameter values beta 1 and beta 2 there could be uncertainty and that uncertainty is captured by the random disturbance of the stochastic disturbance term e_i here. And the last in the list is sample regression line and that is basically again a deterministic expression. It will not involve any stochasticity.

So, why because once you have estimated the proxies for beta 1 and beta 2 namely b_1 and b_2 then you have one mathematical equation and the only unknown thing is X . So, if you plug values of X in that equation you will get the predicted or the fitted value for Y and this is denoted by \hat{Y}_i . So, once you provide a prediction then there is no error.

Because it is coming from a model So, that is why here I am not in accommodating any stochastic or random disturbance term. Now the question of million dollar is, how do we get the proxies for beta 1 and beta 2? Beta 1 beta 2 are unknown we will never be able to know their true values that we have accepted but can we get some very good proxies about beta 1 and beta 2? So, that is basically we are interested in and that is what we will be doing in this 2nd part of the course.

We will try to get best possible proxies for beta 1 and beta 2. So, here we are going to talk about the method that we have already actually gone through in the 1st part of the course and that is the least squares approach as proposed by the German mathematician and statistician Frederick Gauss who also invented the normal distribution.

(Refer Slide Time: 39:50)



So, note that we can write the regression residual as e_i equals to y_i minus b_1 minus b_2 times X_i . So, basically this is basically my regression residual and Gauss suggested that we have to minimize the sum of the squared residuals. So, we have done this least squares approach estimation method when we were talking about curve fitting and all in the part 1. So, I will not talk about this in great details but probably I will just spend a minute here on the diagram. This one also you have seen earlier.

So, here Y is dependent variable X is my independent variable and the star values in grey or you can also call them as asterisk. So, you these are basically the actual observations. So, these are basically $x_1 y_1$, $x_2 y_2$, $x_3 y_3$ and $x_4 y_4$ pairs. Note that it is basically illustration So, its heavily simplified So, there are only 4 observations. And now I am saying that I have to find out the least squares estimator for β_1 and β_2 , So, how to proceed?

So, suppose I have the b_1 and b_2 values someone has given me. Now plug the value of x_1 x_2 x_3 and x_4 one by one in that expression that I am showing you are at the top. So, this is basically the regression residual expression and then I can generate what? I can generate e_1 e_2 e_3 and e_4 . Why because this y_i is basically my observed value for the dependent variable for the i th individual in the sample and that is given the by this star.

And once I plug different values for axis and if I am told previously that I know the values of b_1 and b_2 then actually I know from the 2nd part of this expression I can get the fitted values which

are denoted as \hat{y}_1 \hat{y}_2 \hat{y}_3 and \hat{y}_4 . Now there is difference, So, these fitted values are all falling on the straight line as expected because they are coming from the equation of a straight line. So, the fitted value shall lie on the straight line but there is a gap and this gap is basically the regression residual.

So, some residual is positive sum is negative. So, basically you have to take square and then you need to sum and then finally you need to minimize the sum of squared residuals. So, we stop our discussion here. In the next lecture I am going to start with the calculus of least squares principle and then I am going to talk about the properties of the wireless estimators. See you then, thank you.