Artificial Intelligence (AI) for Investments
Prof. Abhinava Tripathi
Department of Industrial and Management Engineering
Indian Institute of Technology, Kanpur

Lecture- 41

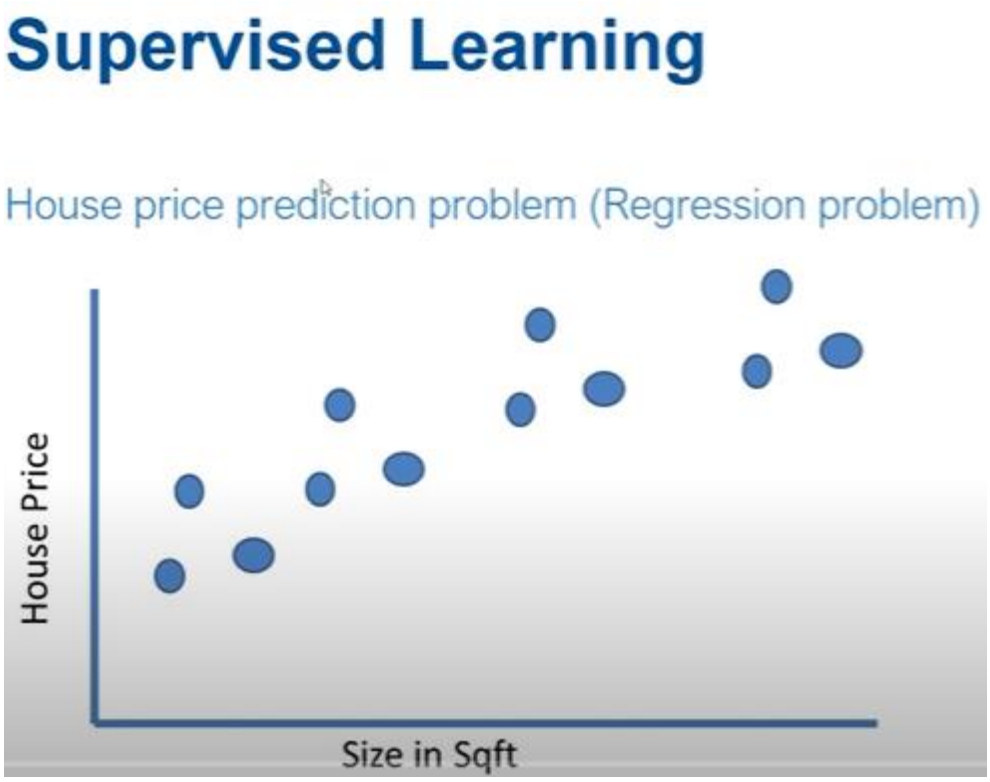
In this lesson, we will discuss the theoretical underpinnings behind the regression algorithm. We will start the discussion with the background and motivation to regression modeling technique. Next, we will discuss the types of data employed in the regression modeling. We will also discuss in great detail the cost function that is ordinary least square method, which is widely employed to fit the regression line. Next, we will discuss the simple and multiple regression modeling techniques. Next, we will examine the key assumptions behind the classical linear regression model and implications if these assumptions are violated.

We will also discuss the blue that is best linear unbiased estimator properties associated with the OLS estimators. Lastly, we will briefly cover the role of normal distribution in hypothesis testing associated with regression modeling. We will also discuss some non-linear functional forms that can be suitably modified to linear functional forms. In this video, we will provide the background and motivation for the application of machine learning algorithms in business applications.

Many times on Amazon and Netflix, we get recommendations for movies and TV series as if some human person has examined and carefully vetted our history of TV or movie watching and provided these recommendations. They are extremely accurate. Also you would have observed the spam filtering application on your mobile text messages and email applications. If you carefully examine these filtered messages, you would find almost 95% to 99%

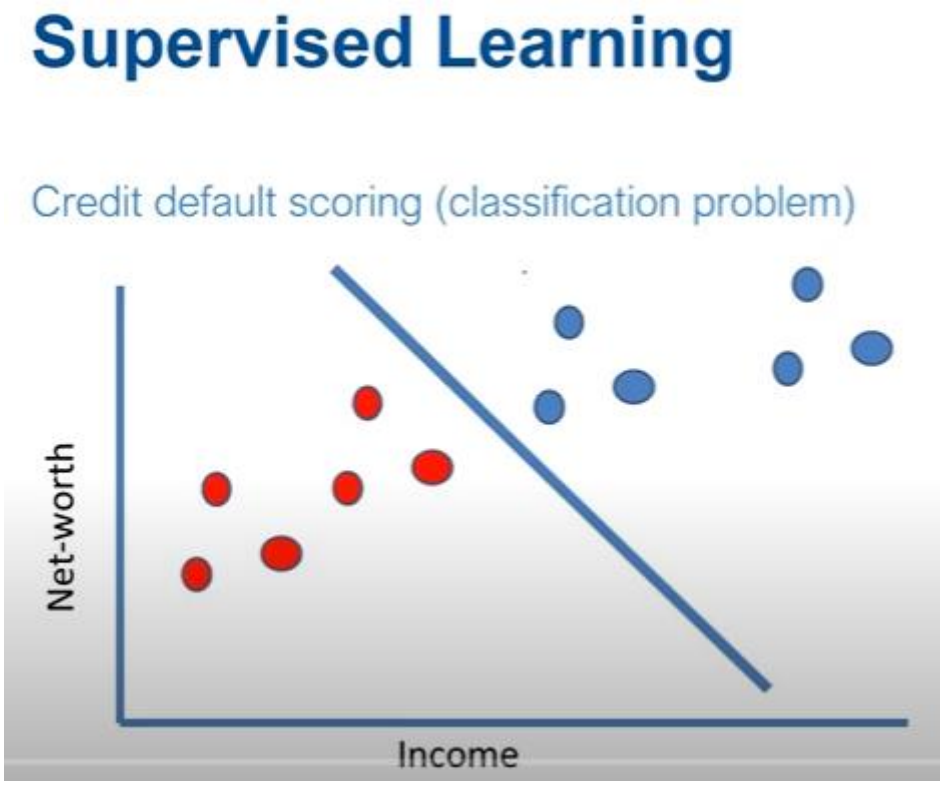
accurate messages and emails that are like phishing emails or bulk emails which are targeted to bulk audience and not very genuine emails or messages. Also if you go to website like Google news, you will find news items clustered and put together in groups as if somebody has carefully read these news items and put together similar news items, putting together these news items for our ease of reading.

To summarize this discussion, in modern day business applications, for example, Amazon and Netflix movie recommendations, filtering out spams, medical prognosis based on previous health records, algorithmic trading, credit scoring models, these applications appear or are driven by computers without being explicitly programmed. So here computers are learning without being explicitly programmed, for example, in recognizing handwriting, natural language processing and so many applications in modern day to day and business applications where computers are acting as if some human being is working behind the curtains. But obviously given the volume and nature of data, big data being handled, a human being cannot do these things on their own. And therefore we need or we clearly understand that these are computers working behind the scenes without being explicitly programmed, they are made to learn through these machine learning algorithms. In this video, we will provide a brief background to machine learning algorithms employed in business applications and finance.



Let us start with a very interesting supervised learning problem. In supervised learning, we have a set of input features which are mapped to output labels. Now, please note the critical, the important keyword here is that output is labeled. So a set of features, input features and output which is labeled, the algorithm is trained using this input output data and once the

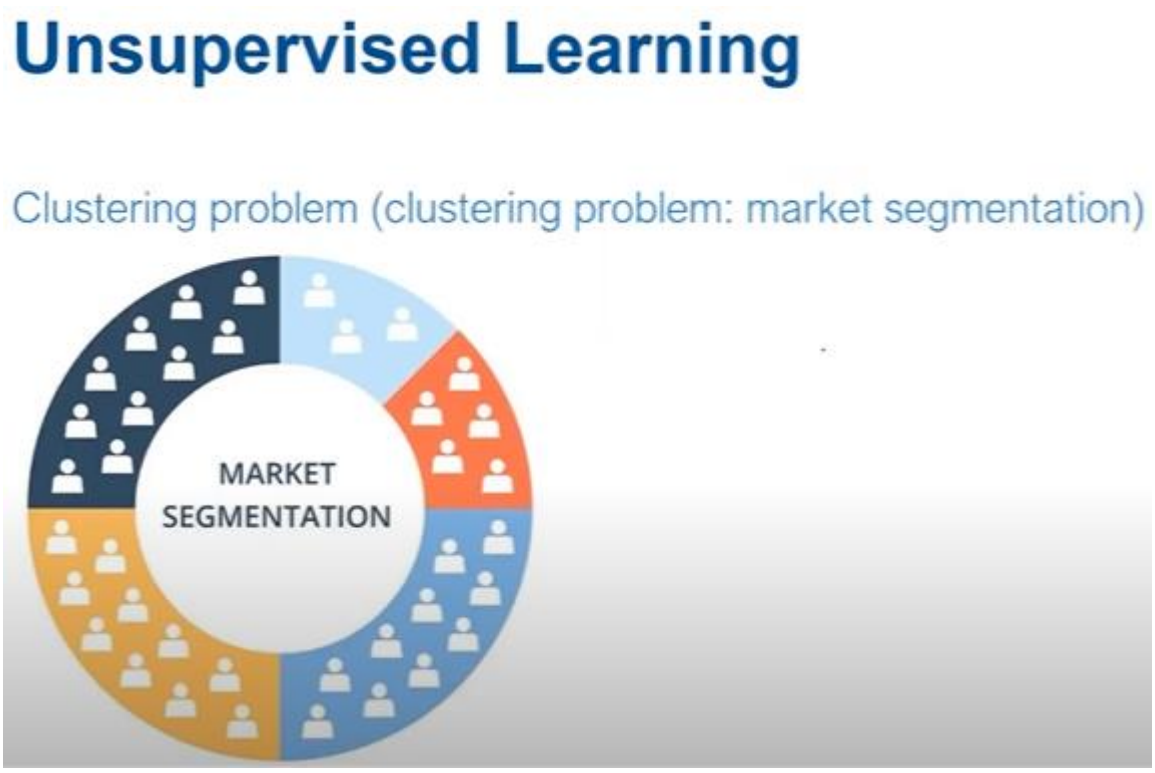
algorithm is trained, it is employed in future using these input variables or features to predict the output. A very simple example to this is house price prediction problem where a set of data that is size of house in square feet, which is one of the features and the label that is house price is given and using this data the algorithm is trained.



This is called regression problem where we try to model this using regression. We will discuss, in this lesson we will discuss the regression problem in great detail, but as a simple example using this feature, which is the size and output label which is the house price, the regression model or algorithm is developed, trained and for future when the size of a house is given, the algorithm is employed to predict its price, let us say some new points are given, then this algorithm will predict their price by fitting a regression line which we will discuss in detail in this lesson. Another interesting supervised learning problem is classification problem where a certain label data is given along with features and then algorithm is trained to classify the new observations across these labels. For example, credit default scoring problem where we try to classify observations such as loan applications into defaulters, for example here defaulters are given in red and non-defaulters or good credit borrowers are given in blue. Now we have two features, the labels are defaulted and non-defaulters while features are net worth and income.

Now using the given set of data, we train this algorithm, we train this algorithm and once we have the algorithm trained, then as the new data comes in, for example new data maybe here,

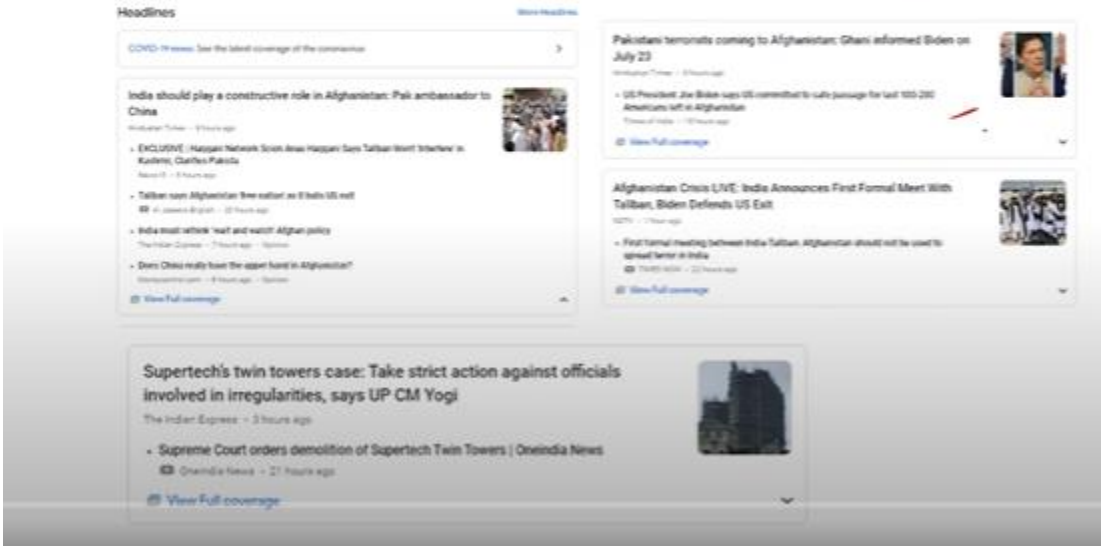
here using this classification problem, we try to classify the new observations into defaulter versus non-defaulter using this model, we try to predict that, we try to classify or segregate defaulters vis-a-vis non-defaulters using the algorithm that has been trained. In the unsupervised learning, the key difference is that the data is not labeled, we have only features. For example, you may have demographic variables such as taste, income, age and gender and based on these features, you will try to segregate individuals into different groups and this is a very useful application in market segmentation by consulting firms who employ such data in creating target segments for targeting them for their products. Such segments are based on this classification problem, clustering problem where we try to cluster individuals based on certain features such as taste, income, gender and so on. So this is a very interesting business problem, clustering problem of market segmentation.



Another very interesting clustering problem that we already discussed is clustering of news items depending upon their similarity. So natural language processing and text processing, text analytics applications and algorithms are employed to segregate news items into clusters, clusters that are similar based on certain features maybe genre of news or similarity in terms of geography or the topic of news, they are clustered together by unsupervised learning cluster algorithms that cluster the news items together. To summarize this video, we discussed supervised and unsupervised learning algorithms. In the supervised learning algorithm, we have data which is labeled along with the features and we try to develop the algorithm by mapping the relationship between input features and labeled output. Subsequently, this trained algorithm is employed to predict future observations or future units.

Unsupervised Learning

Clustering problem (clustering problem : news aggregation)



Using their features, we try to predict their output labels. In contrast, in the unsupervised learning algorithm, in the unsupervised learning algorithm, we have only features not labels and we try to map or cluster the observations or units together based on similarity in features and one of the very important and interesting application we saw was that of market segmentation. In this video, we will discuss the types of data being employed with machine learning algorithms to train and test the models in financial markets and finance domain in general. In general, there are three broad types of data that are employed in finance domain. First is observations about multiple individuals or units collected over a single period.

Types of Data

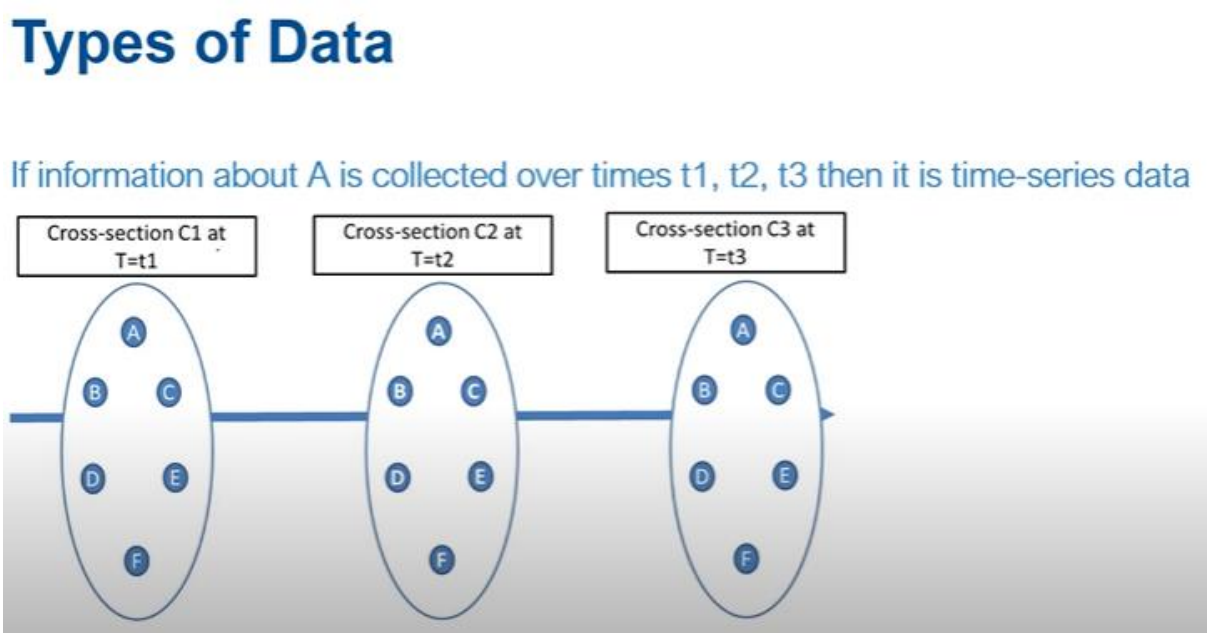
Cross-sectional data: Observations about multiple individuals (units) collected over a single period

Time-series data: Observations of a single individual (unit) collected over multiple periods

Panel or longitudinal data: Observations about multiple individuals (units) collected over various time periods

Second, observations about a single individual collected over multiple periods and lastly,

observations about multiple individuals or units collected over various time periods. We will examine this and visualize these types of data in subsequent discussion. Have a look at this diagram here. We have three cross sections C1 at time 1, C2 at time T2 and C3 at time T3. Notice if you collect only one individual data about one individual like A over all the three cross sections that is 1, T1, T2 and T3, then this kind of data is called time series data.



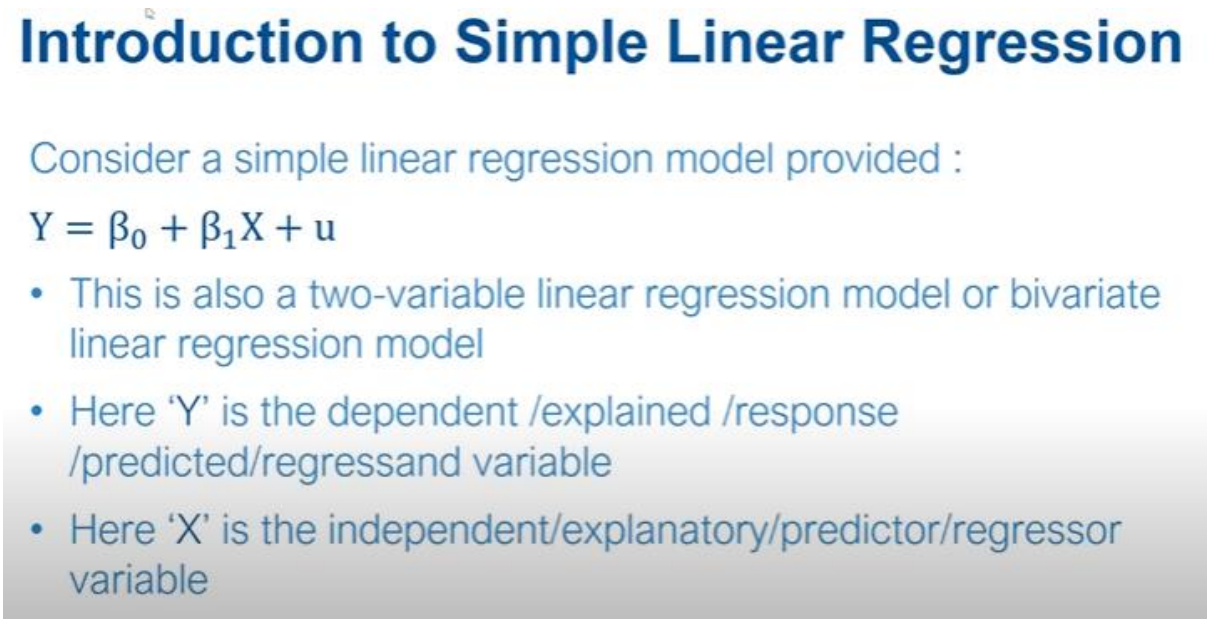
You are collecting a time series of information about individual or unit A. In contrast, if you have six individuals or more or less individuals like A, B, C, D, E and F and you collect all these individuals over a particular time let us say T1, then it is called cross sectional data because you are collecting all the information about these units for a particular time T1. So, it is a cross sectional data. Lastly, if you combine the time series and cross sectional characteristics that is you collect the information about all the units or individual of interest that is for example A, B, C, D, E and F not only for one cross section but all the desired cross sections like T equal to T1 for time T equal to T2 and time T equal to T3, then this becomes a panel or longitudinal data. That means this panel and longitudinal data comprises or involves the characteristics of time series as well as cross section.

To summarize this video, we discussed three types of data. First, time series data where we collected information about individual unit over a given time series. Next, we discussed cross sectional data where we collected and analyzed data about multiple individuals or units like A, B, C, D, E, F over a particular cross section of time and lastly, we discussed panel or longitudinal data wherein we combined the characteristics of time series and cross section. We collected the information about all the units or individuals of interest over multiple time periods and we call this panel or longitudinal data, panel or longitudinal data. In this video, we will introduce simple linear regression model.

Consider a simple linear regression model the way it is presented here, y equal to beta naught plus beta 1 x plus mu. This is often referred to as two variable linear regression model or

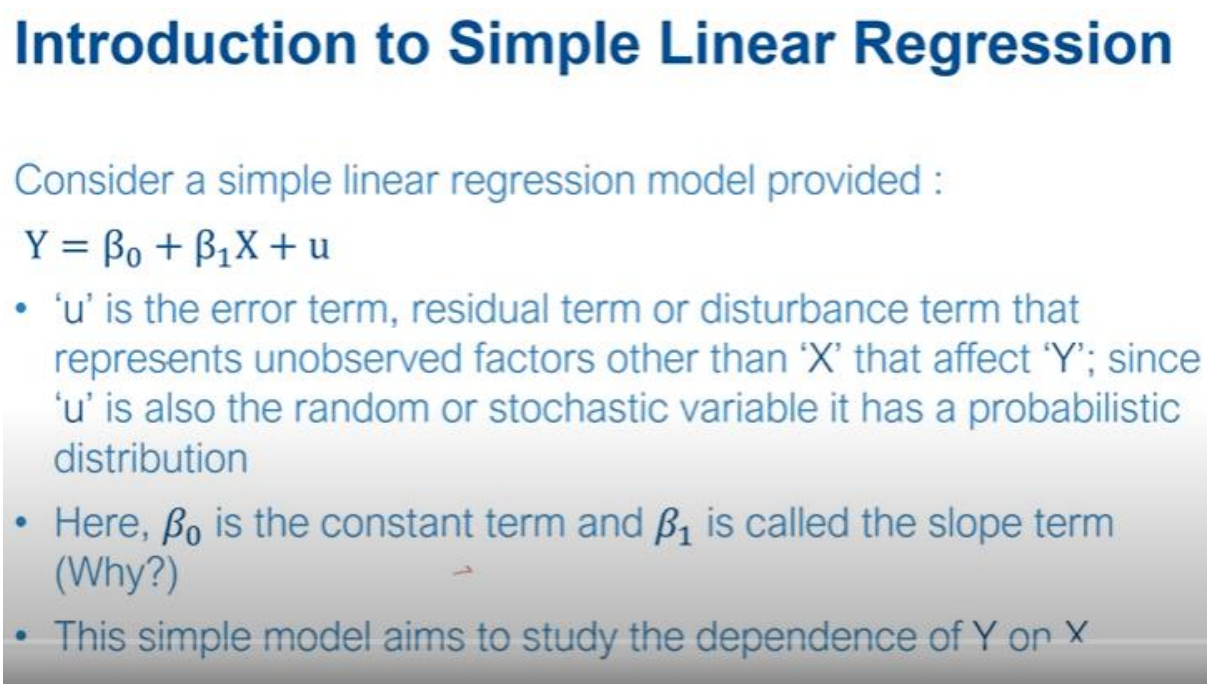
bivariate regression model because there are two variables y and x. y here is the dependent variable or explained or response or predicted or regression variable while x here is the independent explanatory predictor or regressor variable. Notice there is an error term or also called mu which is error term or residual term or disturbance term. This mu represents the aggregation of those factors that are not explained by y.

$$Y = \beta_0 + \beta_1 X + u$$

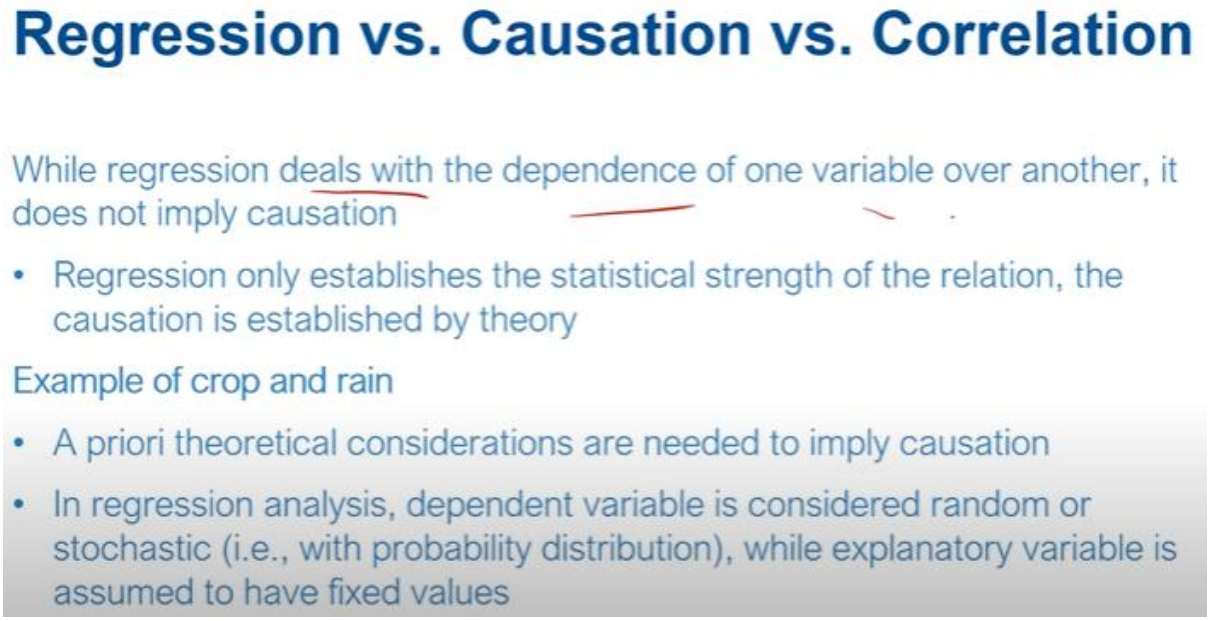


So, in this model, we are trying to explain the effect of this part, this part on y. Those factors for which we cannot explain or are not part of this model are aggregated or considered to be part of this error term. These are unobserved factors that reflect variables or factors other than x that affect y. This mu here is also a random or stochastic variable. It is considered that it has some kind of probabilistic distribution.

Often normal distribution is employed to model this error term mu. But the idea here is that this is random in nature and it assumes its values assume some kind of probabilistic distribution. Beta naught here is the constant term while beta 1 is called the slope term. This is so because this relationship, this model can be represented on a two-dimensional space xy space where this model is basically a straight line with this intercept or constant term as beta naught and slope of this line is represented as beta 1. So, this line represents in visual terms this model.



So, this is visualization of this model. This simple model aims to study the dependence of y on x that means how x affects y. Now, please note here that regression deals with the dependence of one variable over another. It does not imply simply causation. The idea is that it only establishes the statistical strength of the relation.



The causation has to be established a priori by theory. For example, if instead of y like this the dependence on y you replace an x with y and you put it like this, even then the regression model, the mathematical modeling will give you some kind of statistical relationship which obviously may be spurious in nature. Take for example crop and rain. Now a priori we already know that rain affects crop. This is established by our common understanding, common sense.

Regression vs. Causation vs. Correlation

A closely associated concept is correlation, which establishes the degree of linear relationship between the two variables

- In correlation analysis, both the variables are treated in a similar manner and considered to be random

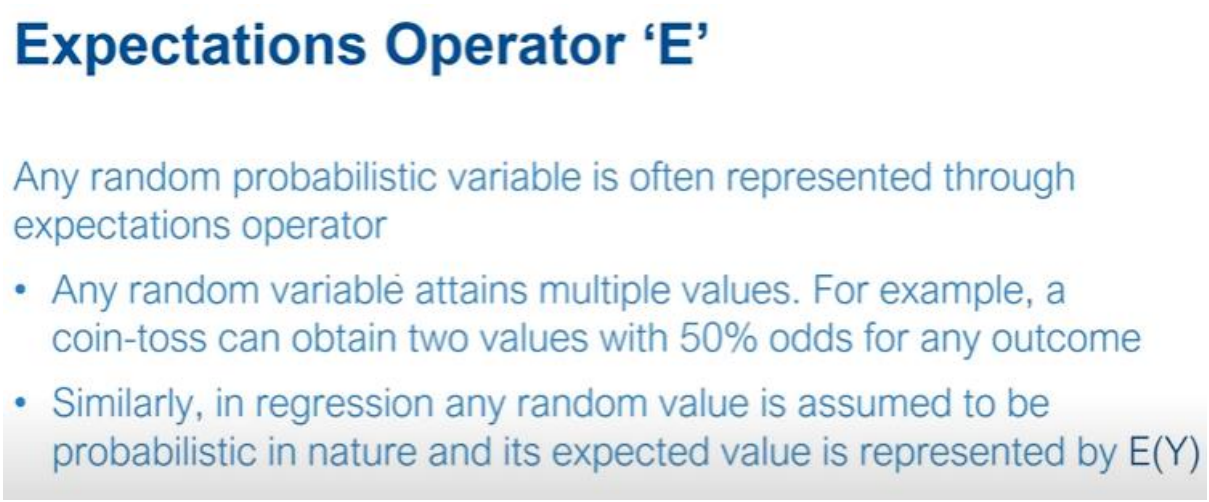
Even if you put or replace rain with crop and you try to find some kind of mathematical relationship from crop to rain, the regression model, the mathematical model will give you that relationship which is obviously spurious in nature and therefore a priori theoretical considerations are needed to imply the causation. Now in regression analysis the dependent variable which is y is considered random or stochastic in nature that means it assumes some kind of probability distribution. The reason is that it also incorporates the effect of all those unexplained or not considered factor which are aggregated in the error term mu which was stochastic in nature. So those effects or influences will also be incorporated in y and therefore y is random or stochastic in nature. In contrast, the explanatory variable which is x is assumed to have fixed value that means it is considered to be coming from outside.

While y is part of this model and therefore it is random in nature, x is fixed and taken from outside. So the x values are fixed and taken from outside in the model. A closely associated concept here is of correlation. Correlation here establishes the degree of linear relationship between the two variables let us say x1 and x2. In correlation analysis both variables are treated in a similar manner and considered to be random that means no causality from one to other is implied.

Let us say for example if variable x1 moves by 1% then x2 moves y in the same direction by let us say 0.5% then there is a 50% correlation between these two variables. However we do not employ any, we do not employ any causality here. To summarize in this video, we examined a simple linear regression model, bivariate model where there are only two variables the dependent variable y, a constant term, independent variable x and a slope or coefficient term beta1 along with the error term. y here is the dependent variable which reflects the impact of our model which is beta0 plus beta1x while those effects or influences which are not accounted for by this model are considered to be aggregated into this error term.

Now please note by definition of this model y here is random and stochastic in nature because it also incorporates the effect of these mu or unexplained part of the model. The explained or modeled part this here the variable x is considered to be fixed in nature and taken from outside while y is getting affected by this model itself which includes the effect of mu and mu being random stochastic in nature while y is also random and stochastic in nature. This relationship while it is mathematical in nature and statistical significance is

established through mathematics the theoretical underpinnings or causality from x to y has to be established a priori from theory not the mathematics. In this video we will discuss the rule of expectations operator in the context of random probabilistic variable. Any random probabilistic variable is often represented through expectations operator.



This random variable can be the error term or the dependent variable like y. Since these variables attain multiple values for example let's take an example of coin tossing game where there are two possible values and if the coin is fair then there is a 50% chance or odds for any of the outcome whether it is head or tail. Similarly in regression any random variable like the dependent variable y if it is assumed to be probabilistic in nature its expected value is represented as Ey. For example if there are n probabilities that an event has an outcome of y1, y2, y3, yn with possibilities p1, p2 these are probabilities p1, p2 and so on up till pn then the expectation of this variable y is defined as p1, y1, p2, y2 and so on that is sigma py. P is probability y is possible outcome pi, vr, i equal to 1 to n where n are the possibilities in the event.

$$E(y) = p_1 * y_1 + p_2 * y_2 + p_3 * y_3 + \cdots + p_4 * y_4$$

$$p_1 = p_2 = \cdots = p_n = \frac{1}{n}$$

$$E(y) = \frac{1}{n}(y_1 + y_2 + y_3 + \cdots + y_n)$$

Expectations Operator 'E'

For example, if there are 'n' possibilities of an event, y1, y2, y3,..,yn each with possibilities p1, p2, p3, p4…pn, then expectations operator is defined as

- E(y)= $p_1 * y_1 + p_2 * y_2 + p_3 * y_3 + \ldots\ldots + p_4 * y_4$
- This is also called probability weighted mean
- If all the probabilities are assumed to be equal then $p_1 = p_2 = \ldots\ldots = p_n = \frac{1}{n}$
- Then E(y)= $\frac{1}{n}(y_1 + y_2 + y_3 + \ldots + y_n)$, i.e., simple average of Y's

This is also called probability weighted mean. If all the probabilities are assumed to be equal that means many times we do not have any a priori knowledge about possible probabilities. So if all the probabilities are assumed to be equal then p1 equal to pn so on equal to 1 upon n. In that case the expectation is nothing but the simple average of all the possible outcomes of y's. To summarize in this video we discussed the role of expectation operator for random probabilistic variables or stochastic variables like error term or dependent variable.

We noted that expectation are nothing but simple probability weighted mean of variables. In case we do not have any a priori probability assigned to these outcomes yi's then in that case the expected value is nothing but simple average of y. In this video we will understand the basics of simple linear regression algorithm with a simple example. Consider a table shown here where family income and consumption expenditures are provided. Here a population of 60 families is divided into 10 income groups for example 80, 80 to 100, 100 to 120 and so on up till 260.

A Simple Example

Consider a simple example of family income and consumption expenditure shown below

$Y_\downarrow$ / $X\rightarrow$	80	100	120	140	160	180	200	220	240	260
Weekly family consumption expenditure Y,	55	65	79	80	102	110	120	135	137	150
	60	70	84	93	107	115	136	137	145	152
	65	74	90	95	110	120	140	140	155	175
	70	80	94	103	116	130	144	152	165	178
	75	85	98	108	118	135	145	157	175	180
	–	88	–	113	125	140	–	160	189	185
	–	–	–	115	–	–	–	162	–	191
Total	325	462	445	707	678	750	685	1043	966	1211
Conditional means of Y, $E(Y\|X)$	65	77	89	101	113	125	137	149	161	173

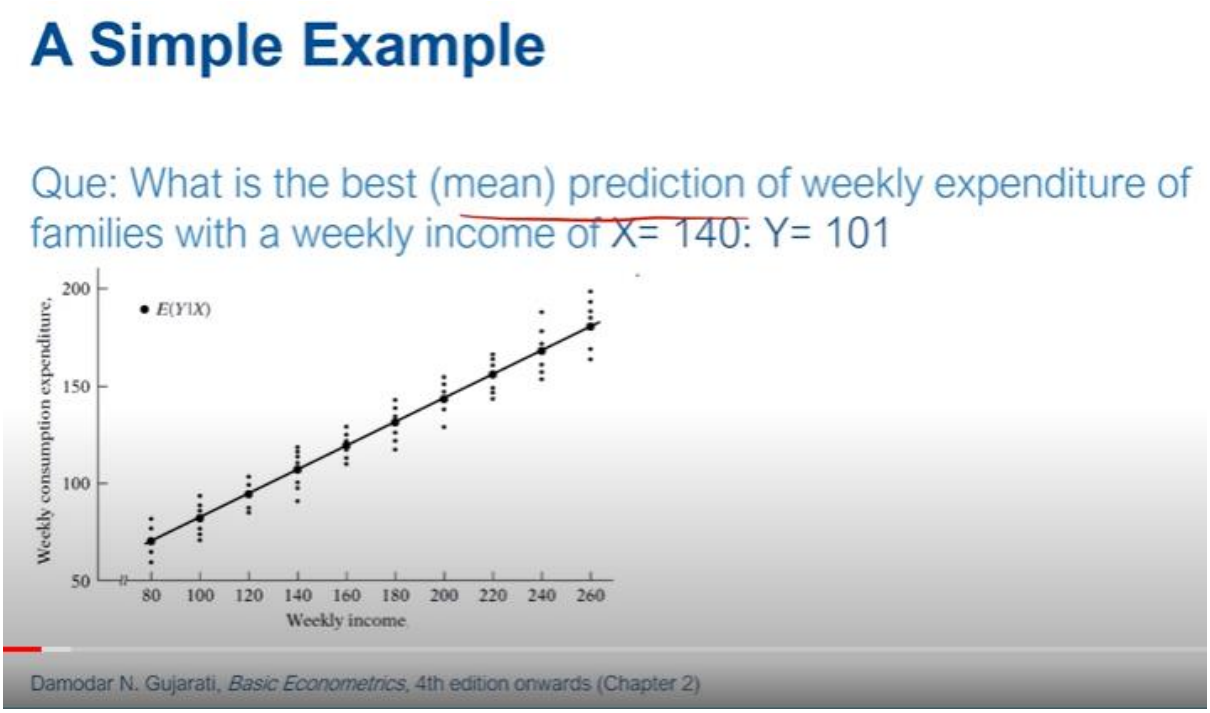
Damodar N. Gujarati, *Basic Econometrics*, 4th edition onwards (Chapter 2)

So there are 10 groups and their consumption weekly family consumption expenditure the y's are also shown. For each level of income x we try to compute the conditional mean that is expected y given x which is 65. E(Y|X) = 65. Let us see how this is computed. For example consider an income level of 80. Corresponding to this income level a set of family weekly consumption expenditure that means y's are provided.

So for this given level of x the corresponding y's are provided. The summation of these y's is 25 and their mean is 65. Now this 65 is the conditional mean of y conditioned upon a given value of x which is 80 and that is why it is called conditional mean of y expected value of y given x. Recall that in the previous video we said that if we do not have any a priori probabilities of any set of values we take simple average to compute the expected value and the same thing we did here by taking the average of these values which is 65 for x equal to 80.

Now the average of all these values is 121.2 and since this value which is the average of all the values is not conditioned upon any specific x value this will be called expected value of y which is unconditional mean or unconditional value of y because it is not conditioned on any x value it is unconditioned expected value of y. Please note that such unconditional value does not account for any income level that means any value of x and it is the prediction or expected value of y or the prediction of y when there is no knowledge of x. However one would like to improve upon the value of y by knowing the value of x if one believes that x affects y. So if x indeed affects y then the knowledge of x would definitely improve the prediction of y. So if you have the knowledge of x you can improve your prediction by computing the conditional mean of y which is this expected value of y given x and that

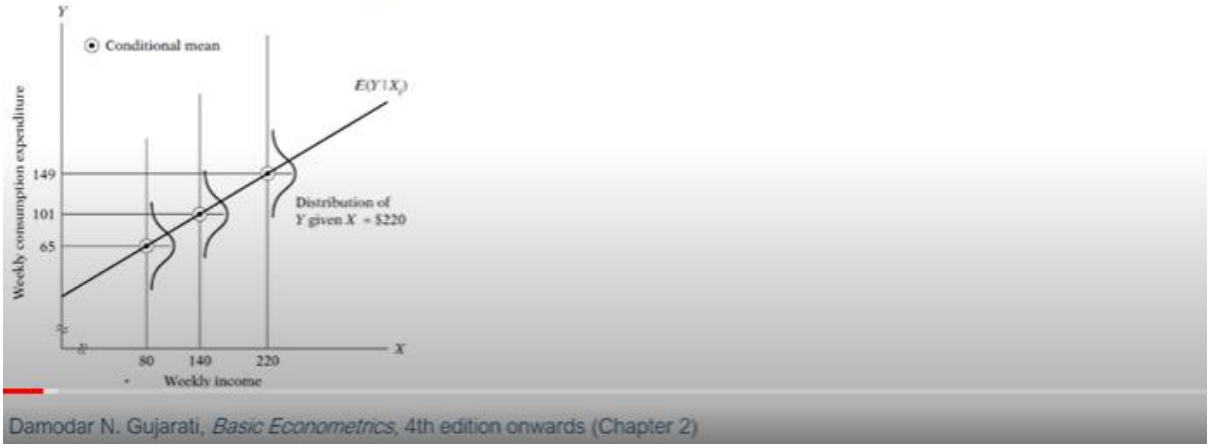
prediction will be more accurate for example the conditional mean of y given x equal to 260 will be 173 which is different from conditional mean of y when x was 80.



Now if one asks what is the best prediction or mean value of weekly expenditure for an income of x equal to 140 then you would tell him that looking at this table you will tell him a value of 101 that means the knowledge of income level that is values of x enables us to better predict the mean value of consumption. So our prediction of y which is the mean improves as compared to the situation when we do not have that knowledge. So this is the essence of regression modeling. So what we are saying here that for each level of x we have set of y values and the mean of these y values are joined together these means of y values for each level of x we have expected value of y given x and all these conditional means are joined to get the regression line. So this regression line is nothing but the conditional means conditioned upon these x values are joined together to get this fitted line.

A Simple Example

Que: What is the best (mean) prediction of weekly expenditure of families with a weekly income of X= 140: Y= 101



So for example if you have the weekly incomes given on x axis and for each set of income you have a distribution of y values this is the distribution of y values and we know that for any distribution if a variable is hazard distribution it is stochastic in nature we can compute its expected value as we have already seen using the probabilities we can get the conditional means for each given x we can compute the expected value of y that is expected value of y for each given xi where xi can be 8090 in this case 8090 and so on and once we have those conditional means we can join them to get the regression line and then in that case our estimate of y would be conditioned upon xi for any value of xi we can predict the value of y.

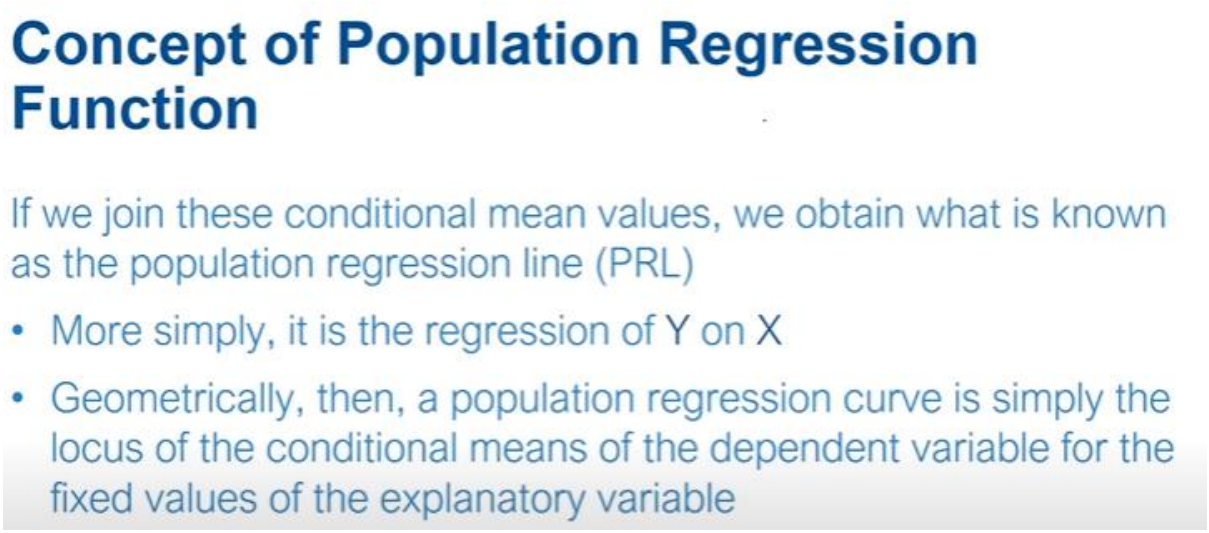
To summarize in this video we said that any regression model of this type y equal to b0 plus b1 x yi equal to b0 plus b1 xi assumes that some variable xi has some explanatory power over variable yi and therefore one can improve the prediction of yi the prediction value of yi which is yi hat can be improved by using the knowledge of xi. Thus regression modeling improves the prediction of yi by taking a set of values of xi for example if we have the knowledge of xi the corresponding values of y can be averaged and this average value of yi for a given i is a better estimate of y. Since this estimate of yi that is yi bar is for a given value which is xi and therefore often such estimate is called conditional mean of yi that means conditional value of yi given x and this is called conditional mean of y. Another type of expectation of yi which is unconditional mean which ignores the value of x which does not consider and assumes that we have no knowledge of x is expected value of y which is the average of all the values of y in respect to values of x.

$$Y_i = b_o + b_1 X_i$$

It is often considered as per regression modeling that this conditional mean is an improved estimate of y as compared to the estimate of y when we have no knowledge of xi. In this

video, we will introduce the concept of population and sample regression functions in the context of regression modeling. Recall, in the previous video we said that if we join the conditional means of y given x these were the conditional means of y given x if we join these values then we obtain what we call is a population regression line. So, we join them for each value of xi, x1, x2, x3 and so on we had a set of values on y and we join the mean of these values what we call conditional mean rather for given xi what is the mean of y when we join these conditional means what we obtained is called population regression line which is essentially the regression of y on x. Now, a population regression curve is simply the locus of these conditional means of the dependent variable which is y for the fixed or given values of explanatory variable which is x here.

So, this is the basic concept of population regression function. The population regression function discussed here can be denoted by this expected value which we have already seen expectation of y given xi this is often referred to as population regression function f xi. Here f xi is a linear function of x which is called a population regression function when we say it is a linear function it this linearity means linearity in parameters and this can be more simply written as expected value of y given xi equal to beta0 plus beta1 into xi where beta0 and beta1 are population parameters. These are population parameters because they are for the entire population assuming the population is known which is a more hypothetical construct you never have the entire population but assuming you know the population then these are the population parameters or the true values of beta0 and beta1 which reflect the relationship between y and x. Now, when we say this is the linearity in parameters that means for example if this kind of expression is there where the population regression function is beta0 plus beta1 square xi this model is non-linear in parameters and will not be estimated with linear regression modeling that we are discussing here.



Concept of Population Regression Function

Population regression function (PRF)

- $E(Y/X_i)=f(X_i)$
- In this case, $f(X_i)$ is a linear function of X
- The expression is also called population regression function

$$E\left(\frac{Y}{X_i}\right) = f(X_i)$$

However a model like expected value of y given xi which is equal to beta0 plus beta1 xi square this model is by non-linear in variables it is linear in parameters beta0 and beta1 so population parameters are linear and therefore it can be handled under linear regression models. However in real life we do not have the luxury of population we often work with much smaller set of data which is called sample and therefore what we estimate is not the population regression function but what we get is the sample regression function and it is denoted by adding a hat symbol this hat indicates that we are working with samples and therefore the estimated values of y is yi hat. So yi hat is the estimate of population value which is yi so yi is the population this yi hat is the estimate of y which is given with this regression sample regression function where beta0 hat and beta1 hat are sample parameters unlike the beta0 which was the population beta0 and beta1 which were population parameters beta0 hat and beta1 are the sample parameters or the estimators of population parameter which is beta0 or beta1. So beta0 hat here is the estimator of beta0 we will discuss some of these estimators beta1 hat is the estimator of beta1 and these estimators are applied on a given sample to estimate or proxy the population parameters using these beta0 hat and beta1 hat what we get is yi hat which is the estimate of y which is the estimator of population values based on our sample regression function. So SRF which is the sample regression function is only an estimate of population regression function based on a given sample or the sample that is available to us.

$$E\left(\frac{Y}{X_i}\right) = \beta_o + \beta_1 X_i$$

$$E\left(\frac{Y}{X_i}\right) = \beta_o + \beta_1^2 X_i$$

$$E\left(\frac{Y}{X_i}\right) = \beta_o + \beta_1 X_i^2$$

Concept of Population Regression Function

More generally, for a two variable case: $E(Y/X_i)=\beta_0+\beta_1 X_i$

- Here it is important to note that linearity means linearity in parameters
- $E(Y/X_i)=\beta_0+\beta_1^2 X_i$; this model is non-linear in parameters and will not be handled in linear regression modelling
- $\boldsymbol{E(Y/X_i)=\beta_0+\beta_1 X_i^2}$, in contrast this model is non-linear in variables and can be handled under linear regression models

Remember we do not work with population we do not have the luxury of population. Now since this is only an estimate therefore sample regression function can overestimate or underestimate the population regression function or population regression function values. For example if you are estimating yi hat and you get this kind of estimators beta0 hat and beta1 hat your sample regression function line may look like this. This is one example where actual population regression function is this. So from this point onwards you overestimate the population values of y and behind this below this point A before that you underestimate the population values.

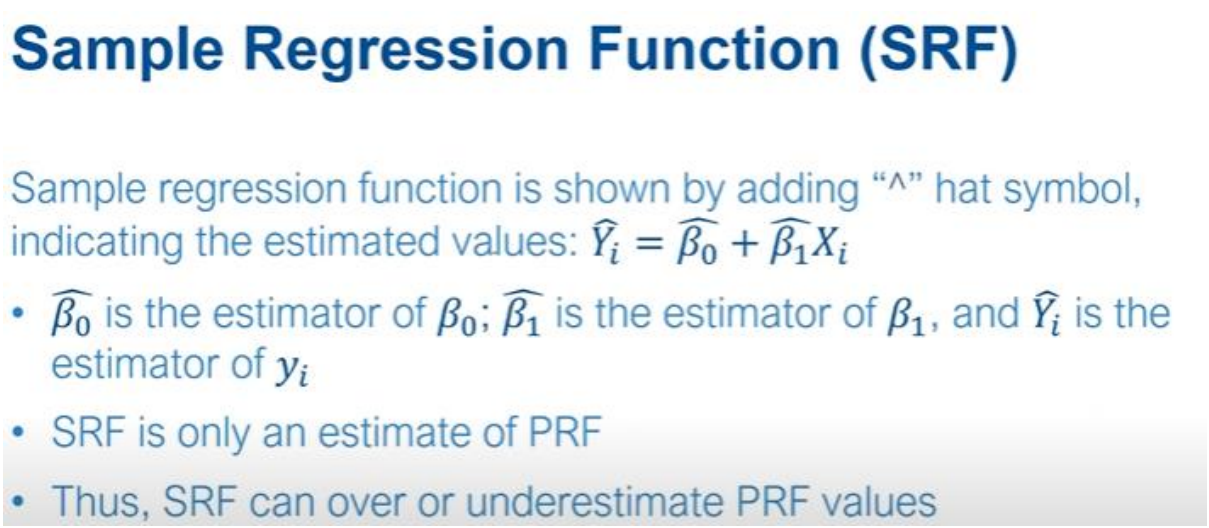
So your estimate is overestimating here underestimating here. To summarize in this video we discussed the concept of population regression function which was nothing but the values of y using population parameters beta0 plus beta1 into x. Now beta0 and beta1 here are the population parameters. Please note in real life we do not have the luxury of working with population and therefore this is not observed. This is not observed. So the set of population is not observed and we only try to map this population or estimate using a much smaller sample which is called sample.

So you work with samples in real life through which you try to make estimates or hypothesis or various predictions about the population which is much larger in size. When you work with sample you denote this with a hat symbol y hat. You do not have actual yi's and xi's you have yi hat's which are the predicted values beta0 hat plus beta1 hat into xi. Please note this yi hat is the prediction of y. This beta0 is the estimator of population parameter which was beta0.

Beta1 hat is the estimator of beta1 which was the population parameter and this is called SRF sample regression function because it is estimated on the sample which is available at hand and it tries to estimate the population values, population parameters and population yi's. However please note this yi hat is often not very accurate and it may over predict or under

predict the actual values. In this video we will discuss a very important class of estimators called ordinary least square estimators. Recall the sample regression function which we discussed was yi equal to beta0 hat plus beta1 hat xi plus mu hat. Here we have removed the hat because now we have also included the error term.

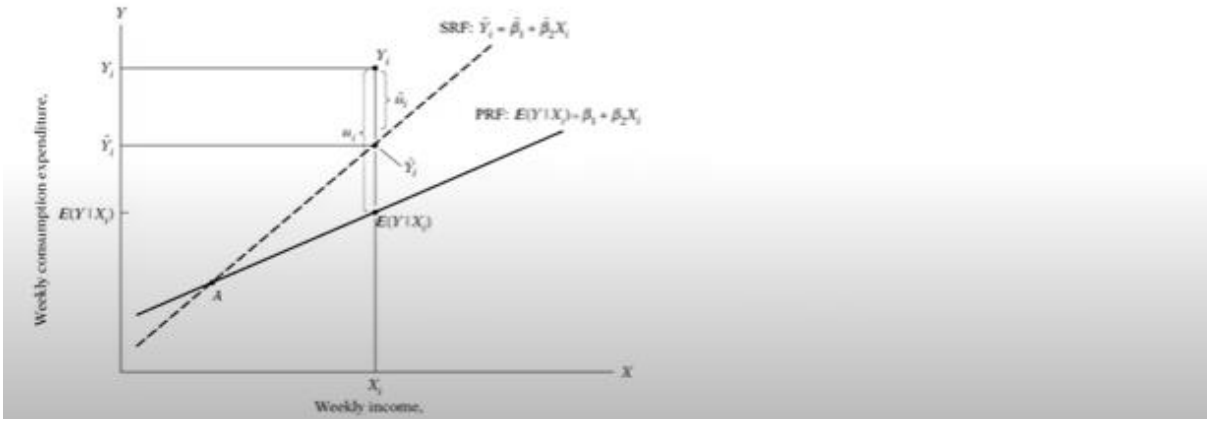
$$\hat{Y}_i = \hat{\beta}_o + \hat{\beta}_1 X_i$$



Earlier we saw this yi hat equal to beta0 hat plus beta1 hat into x hat which is nothing but the prediction of yi hat, the predicted value of y hat. Once we add the error term in the prediction it becomes the actual y hat. Now the error term this mu i hat or often also referred to as mu i is important here because this error term which is yi minus yi hat or yi minus beta0 hat minus beta1 hat xi this error term is a very important cost function and any regression fit or any estimator should minimize some function of this error. In this particular case the OLS, ordinary least square procedure as the name suggests tries to minimize the square of this function and therefore the best cost function to minimize this can be shown as follows. This is the appropriate cost function which is the sum of squares of error term which is yi minus yi hat raised to the power 2 summation.

Sample Regression Function (SRF)

Sample regression function (SRF): $\widehat{Y_i} = \widehat{\beta_0} + \widehat{\beta_1} X_i$



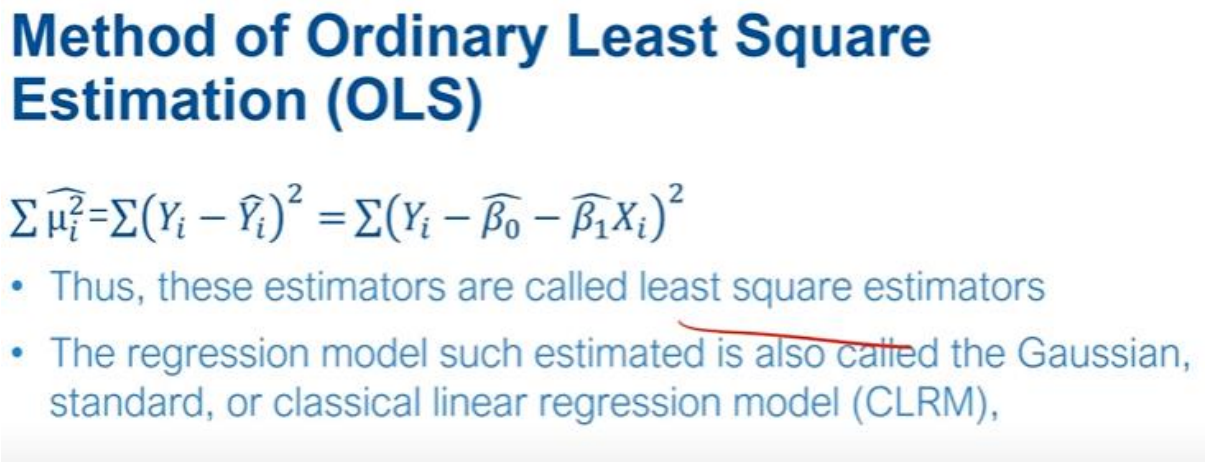
$$\hat{Y}_i = \hat{\beta}_o + \hat{\beta}_1 X_i$$

This is our cost function which is minimized here. Now we can also replace this yi hat using this expression here so we get yi minus beta0 hat minus beta1 hat into xi raised to the power 2 summation is our sum of squares. Now we minimize these squared residuals. One is why not minimize just the residuals or absolute residuals. Now talking about simple residuals it fails to recognize the fact that large positive and negative errors for example plus 6 and minus 6 will tend to cancel each other out. So if there are large deviations large error terms mu1, i and muj which are very large but on the summation they tend to cancel each other out simple summation of residuals will ignore that.

If we are minimizing the absolute residuals to an extent that is an improvement because it will recognize that plus 6 and minus 6 or positive and negative residuals should not cancel each other and therefore large errors whether positive or negative they are penalized. However in case of squared residuals it penalizes large errors more strongly so the penalty on large errors is much more stringent in case of squares as compared to simple absolute residuals. To further illustrate and visualize this argument let us have a look at set of observations plotted on this x-y diagram. These are the actual values that we have rather not population but a sample that we have obtained. Now we try to fit a sample regression function or sample regression the line corresponding to sample regression function which is represented as yi hat equal to beta1 hat plus beta2 hat.

Now obviously because we are working with samples these are estimators since these are OLS estimators we try to minimize the sum of these squared residuals. The difference between the actual values and this fitted line so this is our sample regression function the line

corresponding to sample regression function this distance is my error in prediction. Now this error the larger this error the problematic it is for me and therefore a function which minimizes the square of these errors gives higher penalty to large errors as compared to small errors. If I would have added simply the absolute residuals it would have definitely penalized errors but a square of residuals penalizes large errors much more than simple absolute summation. Now in this scheme our final function our final function or squared residual summation is this which we can also write as summation of my square is a function obviously a function of beta0 hat and beta1 because it is written like this so it must be a function of beta0 hat and beta1 hat which are the key parameters or estimators of importance.



So these are sample parameters that we are interested in knowing. Now a very regular sample scheme to minimize this we want to minimize the squared residuals we can set the differential of this partial differential with respect to our betas equal to 0 we can set this differential with respect to parameters different parameter estimates beta0 hat and beta1 hat equal to 0 and then set the condition of the double differential to positive for minima condition to obtain the estimates of beta0 and beta1 hat as we will shortly see and this is the reason why these estimators are called least square estimators because we minimize the square residuals and such scheme or regression model obtained from such scheme is also called Gaussian standard or classical linear regression model by minimizing these least square errors. Now if you recall we said that my square can be written as this function which is yi minus beta0 hat minus beta1 hat xi square summation we can take the partial differential with respect to beta0 which is since it is square term this is minus 2 times this term which is same and then differential of beta0 is nothing but a minus sign so we get this minus 2 times mu i hat why because this term itself is yi hat and yi minus yi hat is nothing but mu i hat or often referred to as mu i itself the error we are interchangeably using mu i hat or mu i here so this is nothing but minus 2 times mu i which is the partial differential with respect to beta0 hat similarly we can take partial differential with respect to beta1 hat which will be very similar only that now because beta1 hat has a multiple of xi we have an extra xi here so we have minus 2 times mu i hat into xi summation this is the partial differential with respect to beta1 hat. Now we need to set this differential equal to 0 and solve the equation a closed form solution can be obtained which is shown here this closed

form solution can be obtained only if certain assumptions are met now in this lesson in a set of series of videos we will discuss some of these assumptions that are needed and if these assumptions are met then we can obtain a closed form solution like this that means we can precisely estimate the values of beta1 hat and beta0 hat so we can obtain these beta1 hat and beta0 hat estimators and these estimators are often referred to as OLS ordinary least square estimators because of the minimization of these residual square scheme these are called ordinary least square estimators beta0 hat and beta1 hat and once applied to a given sample we obtain the estimates of population parameter so what the estimates that we obtain the beta0 hat and beta1 hat based on these formulas these are the estimates from the sample. To summarize in this video we discussed a very important class of estimators called ordinary least square estimators the ordinary least square estimators are obtained by minimizing the sum of squares of residuals shown by a formula here to visualize this scheme we plotted all the observations from our sample like this then we try to fit a line which we often called the line corresponding to the sample regression function because we are working with samples so the sample regression function line can be written as yi hat because these are only estimate not the population values beta1 hat plus beta2 hat xi and then we compute these error terms that means the difference from these observed values to the predicted values which is yi minus yi hat this is the gap which is called mu i or we are referring to as mu i hat or mu i interchangeably and then the square of these mu i's are minimized to obtain the estimators beta1 hat and beta2 hat which are called the OLS estimators.

$$SRF\ Function: Y_i = \hat{\beta}_o + \hat{\beta}_1 X_i + \hat{\mu}_i$$
$$where\ \hat{Y}_i = \hat{\beta}_o + \hat{\beta}_1 X_i$$
$$\hat{\mu}_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_o + \hat{\beta}_1 X_i$$

Method of Ordinary Least Square Estimation (OLS)

Recall the SRF function: $Y_i = \widehat{\beta_0} + \widehat{\beta_1} X_i + \widehat{\mu_i}$; where $\hat{Y}_i = \widehat{\beta_0} + \widehat{\beta_1} X_i$

- Here, $\widehat{\mu_i} = Y_i - \hat{Y}_i = Y_i - \widehat{\beta_0} - \widehat{\beta_1} X_i$
- The line fit should aim to minimize the square this error $\widehat{\mu_i}$
- Concept of OLS suggests that the best cost function to minimize is as follows

$$SRF\ Function: Y_i = \hat{\beta}_o + \hat{\beta}_1 X_i + \hat{\mu}_i$$
$$where\ \hat{Y}_i = \hat{\beta}_o + \hat{\beta}_1 X_i$$

$$\sum\widehat{\mu^2}_i = \sum(Y_i - \hat{Y}_i)^2 = \sum(Y_i - \hat{\beta}_o - \hat{\beta}_1 X_i)^2$$

Method of Ordinary Least Square Estimation (OLS)

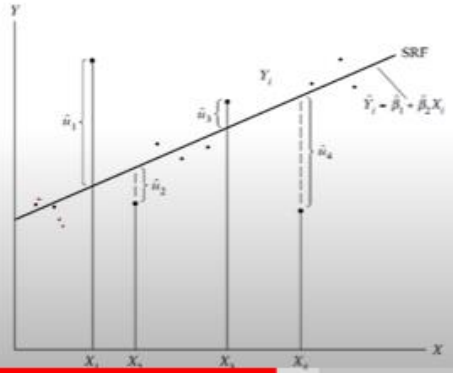
Recall the SRF function: $Y_i = \widehat{\beta_0} + \widehat{\beta_1} X_i + \widehat{\mu_i}$; where $\hat{Y}_i = \widehat{\beta_0} + \widehat{\beta_1} X_i$

- $\sum \widehat{\mu_i^2} = \sum(Y_i - \hat{Y}_i)^2 = \sum(Y_i - \widehat{\beta_0} - \widehat{\beta_1} X_i)^2$
- That is we minimize squared residuals (why not just residuals or absolute residuals)

$$\sum\widehat{\mu^2}_i = \sum(Y_i - \hat{Y}_i)^2 = \sum(Y_i - \hat{\beta}_o - \hat{\beta}_1 X_i)^2$$

Method of Ordinary Least Square Estimation (OLS)

- $\sum \widehat{\mu_i^2} = \sum(Y_i - \hat{Y}_i)^2 = \sum(Y_i - \widehat{\beta_0} - \widehat{\beta_1} X_i)^2$: Minimize these squared residuals



$$\sum\widehat{\mu^2}_i = \sum(Y_i - \hat{Y}_i)^2 = \sum(Y_i - \hat{\beta}_o - \hat{\beta}_1 X_i)^2$$

Method of Ordinary Least Square Estimation (OLS)

$$\sum \widehat{\mu_i^2} = \sum (Y_i - \widehat{Y_i})^2 = \sum (Y_i - \widehat{\beta_0} - \widehat{\beta_1} X_i)^2$$

- Obvious to note here that $\sum \widehat{\mu_i^2} = f(\widehat{\beta_0}, \widehat{\beta_1})$
- Setting differential of $\sum \widehat{\mu_i^2} = 0$ that satisfies and double differential to positive for minima condition, one obtains the estimates that is, $\widehat{\beta_0}$ and $\widehat{\beta_1}$

Method of Ordinary Least Square Estimation (OLS)

$$\sum \widehat{\mu_i^2} = \sum (Y_i - \widehat{Y_i})^2 = \sum (Y_i - \widehat{\beta_0} - \widehat{\beta_1} X_i)^2$$

- Thus, these estimators are called least square estimators
- The regression model such estimated is also called the Gaussian, standard, or classical linear regression model (CLRM),

$$\sum \widehat{\mu^2}_i = \sum (Y_i - \hat{Y}_i)^2 = \sum (Y_i - \hat{\beta}_o - \hat{\beta}_1 X_i)^2$$

$$\frac{\partial (\sum \widehat{\mu^2}_i)}{\partial \hat{\beta}_o} = -2\sum (Y_i - \hat{\beta}_o + \hat{\beta}_1 X_i) = -2\sum \hat{\mu}_i$$

$$\frac{\partial (\sum \widehat{\mu^2}_i)}{\partial \hat{\beta}_1} = -2\sum (Y_i - \hat{\beta}_o + \hat{\beta}_1 X_i) X_i = -2\sum \hat{\mu}_i X_i$$

Method of Ordinary Least Square Estimation (OLS)

$$\sum \widehat{\mu_i^2} = \sum (Y_i - \widehat{Y_i})^2 = \sum (Y_i - \widehat{\beta_0} - \widehat{\beta_1} X_i)^2$$

- $\frac{\partial (\sum \widehat{\mu_i^2})}{\partial \widehat{\beta_0}} = -2 \sum (Y_i - \widehat{\beta_0} + \widehat{\beta_1} X_i) = -2 \sum \widehat{\mu_i}$ (partial differential w.r.t. to $\widehat{\beta_0}$)
- $\frac{\partial (\sum \widehat{\mu_i^2})}{\partial \widehat{\beta_1}} = -2 \sum (Y_i - \widehat{\beta_0} + \widehat{\beta_1} X_i) X_i = -2 \sum \widehat{\mu_i} X_i$ (partial differential w.r.t. to $\widehat{\beta_1}$)

$$\sum \widehat{\mu^2}_i = \sum (Y_i - \hat{Y}_i)^2 = \sum (Y_i - \hat{\beta}_o - \hat{\beta}_1 X_i)^2$$

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

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

Method of Ordinary Least Square Estimation (OLS)

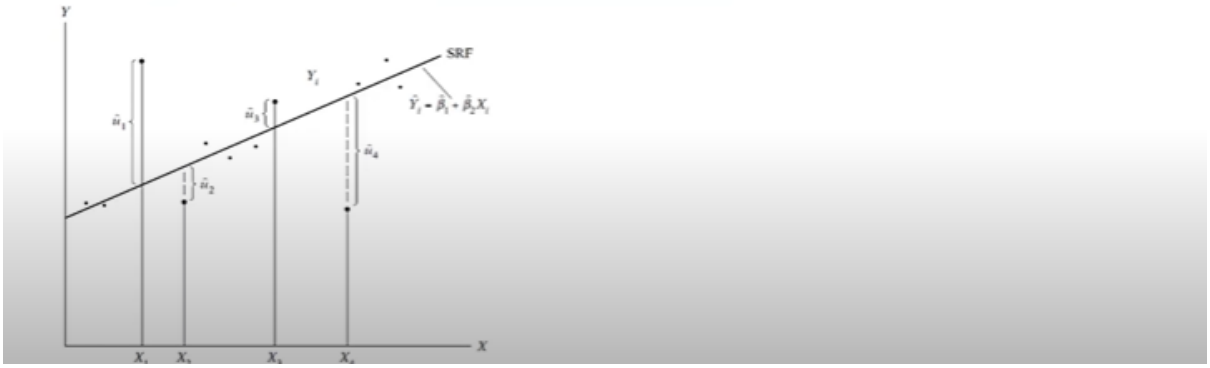
$$\sum \widehat{\mu_i^2} = \sum (Y_i - \widehat{Y_i})^2 = \sum (Y_i - \widehat{\beta_0} - \widehat{\beta_1} X_i)^2$$

- $\widehat{\boldsymbol{\beta_1}} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{(X_i - \bar{X})^2}$ and $\widehat{\boldsymbol{\beta_0}} = \bar{Y} - \widehat{\boldsymbol{\beta_1}} \bar{X}$
- However, to achieve this closed form solution CLRM-OLS makes certain assumptions

$$\sum \widehat{\mu^2}_i = \sum (Y_i - \hat{Y}_i)^2 = \sum (Y_i - \hat{\beta}_o - \hat{\beta}_1 X_i)^2$$

Summary

- $\sum \widehat{\mu_i^2} = \sum (Y_i - \widehat{Y_i})^2 = \sum (Y_i - \widehat{\beta_0} - \widehat{\beta_1} X_i)^2$: Minimize these squared residuals



In this video we will introduce multiple linear regression algorithm. Recall the simple linear regression problem where there was only one dependent and one independent variable which was of the following form b0 plus b1 into x plus error term here there was only one independent and one dependent variable x and y now all the discussion that we had about the simple two variable bivariate linear regression model we can extend this to a multiple linear regression very simply as shown here where y equal to beta0 plus beta1 x1 plus beta2 x2 and so on up till beta1 xn plus there are term. Here xi's represent the explanatory or independent variables and their coefficients beta1 beta2 beta2 or called partial regression coefficients all that we discussed about the simple bivariate linear regression problem apply to this model as well. All the discussion and properties of simple linear regression model and various other aspects of the regression remain the same including the properties of error term that is this error term is the random stochastic variable with the probability distribution. Now there are some very important properties of this error term that we need to recall first and foremost the expectation of this error term conditioned upon all the independent variables xi equal to 0 that means the mean of this error term given all the xi's is equal to 0 for all the i's. For example if you remember for our income and expenditure example where given xi there were a number of y values the corresponding error terms all the mu i's here this mean of this mu i the mean of this mu i that is expectation of this mu i and since it is for given xi the condition upon xi is equal to 0.

Introduction to Multiple Linear Regression

We can generalize the two variable problem into multiple linear regression as $\boldsymbol{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \ldots . + \beta_n X_n + u}$

- $\boldsymbol{X_i's}$ represent the explanatory variables
- Here the coefficients $\boldsymbol{\beta_1, \beta_2, \ldots, \beta_n}$ are called the partial regression coefficients

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n n + u$$

Introduction to Multiple Linear Regression

Multiple linear regression $\boldsymbol{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \ldots . + \beta_n X_n + u}$

- Other aspects of the regression remain the same, including the properties of the error term, that is, $\boldsymbol{u}$
- Zero conditional mean of error: $E(u_i|X_{1i}, X_{2i}, \ldots, X_{ni}) = 0$ for each 'i'
- No serial correlation: $cov(u_i, u_j) = 0$; Homoscedasticity: $var(u_i) = \sigma^2$

$$E\ (u_i|X_{1i}, X_{2i}, \ldots, X_{ni}\ ) = 0$$

Similarly for all the xi's x1 x2 xn i the conditional mean of error term is assumed to be 0. Next no serial correlation or covariance between mu i's that means for xi a set of mu i's are there then for xj a set of mu j's are there then the correlation between mu i and mu j's equal to 0 that means these mu i and mu j's are not correlated that is second very important assumption here and the third the variance of mu i's that is sigma square is assumed to be constant that means whatever the variance for this mu i mu j or if there was x1 2 and so on xk's the error terms are mu k's the variance of these mu i mu j or mu k is same as sigma square that is the third assumption. Next another very important assumption is that mu i is independent of x's that means all these independent variable x1 x2 and xn and so on these are not correlated with error terms that means correlation or covariance between mu i xi mu i x2 and so on mu i xn equal to 0. And please also note the assumption is that model is correctly specified that means all the relevant influences that is independent variables xi's are incorporated in the model there is no important or variable or factor that is influencing y is left out. Just one caveat correlation and covariance represent similar property the normalized version of covariance is correlation itself.

If you compute the correlation if you want to compute the correlation between two variables to compute their covariance let us say covariance of x1 and covariance of x2 if you divide them by standard deviation of x1 and standard deviation of x2 you get the correlation between x1 and x2 so this is just a simple information. Another very important condition for this model to work is of collinearity that is all these xi's let us take in together x1 x2 and so on xn none of xi's can be explained by remaining independent variables for example if I take xi the remaining ones x1 to xi-1 and xi plus 2 to xn they cannot explain this xi completely that means there cannot be any function which is of this form alpha 1 x1 alpha 2 x2 and so on till alpha n xn which is equal to 0 that means any one of these variable cannot be expressed linearly as a linear combination of other independent variables. If that were to happen this is called perfect multicollinearity that means if there is a solution to this problem this equation this represents multicollinearity that means one of these xi's can be explained by the remaining independent variables and in that case of perfect multicollinearity the model will not run that is the equation this equation will become inter-terminate it cannot be determined there is no solution. Let us look at the example of collinearity visually so in this diagram notice that the independent variables x2 and x3 they have some relationship with y which is represented by these common areas but there is no common area between x2 and x3 that means there is no collinearity. However, if there were to exist some common area between x2 and x3 it represents low collinearity.

$$cov(u_i, X_1) = cov(u_i, X_2) = cov(u_i, X_{ni}) = 0$$

Introduction to Multiple Linear Regression

Multiple linear regression $\boldsymbol{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \ldots . + \beta_n X_n + u}$

- Zero correlation (or covariance) between u_i and X:
  $cov(u_i, X_1)=cov(u_i, X_2)=cov(u_i, X_n)=0$
- The model is correctly specified

Introduction to Multiple Linear Regression

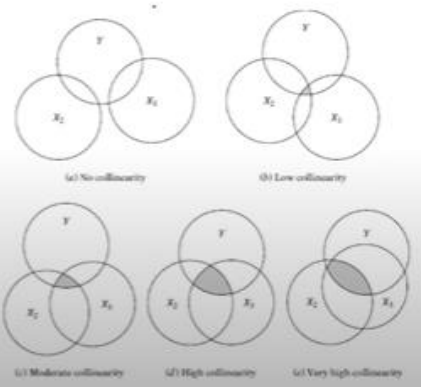
Multiple linear regression $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \ldots . + \beta_n X_n + u$

- Lastly, one more condition is added; that is no exact linear relationship between X_i and X_js (X_1, $X_2 \ldots X_n$): $\alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \ldots \alpha_n X_n \neq 0$
- If such a relationship exists, the model will be affected by the problem of perfect multicollinearity, and will not run (i.e., indeterminate)

Introduction to Multiple Linear regression

However, there may be instances of less than perfect collinearity across variables and can affect the estimation

- If the multicollinearity is not perfect but high, then the estimators have large variances (standard errors of estimate)



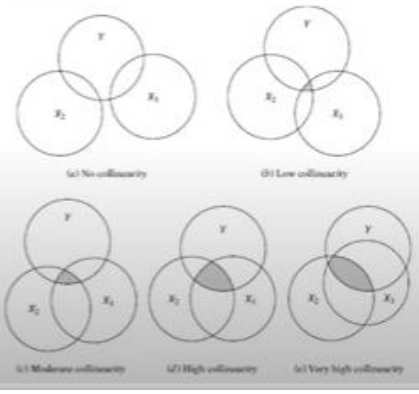
Depending upon this common area if it increases a lot like here or here or here this represents very high levels of collinearity and this is not good for the estimation. Particularly the fact that if the multicollinearity is not perfect the previous case that we just discussed but high then model will run the model is not inter-terminable the model can be determined it is not indeterminate. However then the problem is that the estimators for example beta0 estimator beta1 estimator these will have very large variances or what we call standard errors of estimate. Also recall that the significance of these betas is calculated in the form of t statistics and the corresponding p values the t values are simply calculated as coefficient divided by the standard error of estimate which is basically variance of these estimators and if these standard error of estimates are very high which go into denominator this value is very low that means your resulting t values are very low. This means your t values are very low and therefore the lower the t value the lower is the are the corresponding p values which

increases the chances of failure to reject the null hypothesis that means you will have rather poor confidence interval more wider confidence interval you need and therefore even though your r square may be high which indicates very good fit in the regression but still the power of your test will be low and because of low t values and you will be failing to reject the null that means even the coefficient may be significant it may be significant but still you will not be able to say that by rejecting the null.

Introduction to Multiple Linear regression

However, there may be instances of less than perfect collinearity across variables and can affect the estimation

- This makes the 't-values' low and high chances of failure to reject the null-hypothesis (wider confidence intervals), even though the R^2 may be high



Interpreting the Multiple Linear Regression

Similar to the two variable regression, the following expression

$$E(Y_i|X_1 \ldots . X_n) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \ldots . + \beta_n X_n$$

- Represents the conditional mean or expected value of Y given the fixed values of all the $X_i's$
- The partial coefficient β_1 is the effect of X_1 on Y, net of any effect from other explanatory variables ($X_i's$), or in other words, keeping all the $X_i's$ constant

To summarize in this video we discussed an induced multiple linear regression model we noted that all the properties of simple linear regression model apply to this multiple linear regression as well we also highlighted some important properties of this multiple linear

regression model for example the conditional mean of error term is 0 which is expectation of mu i given or conditioned upon all the xi 0 we also noted that this should be no correlation between mu i's and mu j's or set of error terms we also noted that the variance of these error terms should be constant which is a very important property called homoscedasticity we also said that there should be no correlation or covariance between the error term and independent variables that is xi's and it should be equal to 0 the model should be correctly specified that means all the xi's that are there in the model there should be no important influence or factor should be left out lastly we also noted the property of multicollinearity that means none of the independent variables should be explained completely by the remaining independent variable or all the linear all the independent variable should not be expressed as a linear combination which is equal to 0 this leads to perfect multicollinearity and it makes the model indeterminate in this video we will understand the interpretation of various aspects of multiple linear regression algorithm similar to our understanding of two variable or bivariate regression model this expression here represents the conditional mean or expected value of y given the values of x1 x2 and xn and so on so this is very similar to the expression that we saw here expected value of y given x here instead of single independent variable we are considering all the independent variables from x1 to x1 and therefore this represents the expected value of yi given all the xi's this is called the conditional mean or expected value of y given the fixed values of xi or conditioned upon the fixed values of xi notice the coefficients beta 1 beta 2 and so on up till beta n here these are called partial coefficients for example this beta 1 is the partial coefficient and it measures the effect of x1 on this variable y net of any effect from any other explanatory variables that is xi or in other words keeping all the other xi's constant if all the xi's are held constant except x1 then beta 1 will represent the impact of x1 on y let us introduce a very important parameter which is r square this r square represents the part of variation of variable y which is explained by this model I repeat this r square represents the variation or variance in the variable y which is the dependent variable which can be explained by the model this all the xi's and coefficients to put together the computation is very simple basically this is explained sum of squares divided by total sum of square which is equal to 1 minus residual sum of squares upon total sum of squares let me give you a little bit background on this recall that the values of y on xy axis are provided like this and this is your fitted line so this is the total let us say this is yi then this is the total gap sum of squares of these total yi's will be total sum of squares however this much part is explained this much part is explained part so this squares of these distances would be explained sum of squares and this is unexplained part these are the error terms so this is unexplained or residual sum of squares so here explained sum of squares that means these sum of squares divide by total sum of squares which are these total term total sum of squares is the explained part ratio of explained part this can also be computed as 1 minus residual sum of squares upon total sum of squares where the sum of squares of these distances is called residuals. This remains same whether we are talking about bivariate regression or multivariate regression as we are seeing here a more improved version is also computed in the form of adjusted R square where a small adjustment is made here instead of using RSS and TSS we divide these residual sum of squares and total sum squares by their respective degrees of freedom. For RSS the degrees of freedom are n minus k here n is the total number of variables here n is the total number of observations in the sample k is the number of

coefficients excluding beta naught so all the beta 1 to beta n all the coefficients excluding beta naught divide by n minus 1 n minus 1 is the degrees of freedom for TSS so I repeat RSS upon n minus k where residual sum of squares has n minus k degrees of freedom where n is the total number of observations in the sample and k is the coefficients from beta 1 to beta n and TSS divide by n minus 1 degrees of freedom where n is the total observations. This is often referred to as mean sum of squares of RSS mean sum of squares of RSS that is residual sum of squares and mean sum of square of TSS that is total sum of squares.

$$E(Y|X_1 \ldots . X_n) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n$$

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

$$Adjusted - R^2 = 1 - \frac{\frac{RSS}{n-k}}{\frac{TSS}{n-1}} = 1 - \frac{(MSS\ of\ RSS)}{(MSS\ of\ TSS)}$$

Interpreting the Multiple Linear Regression

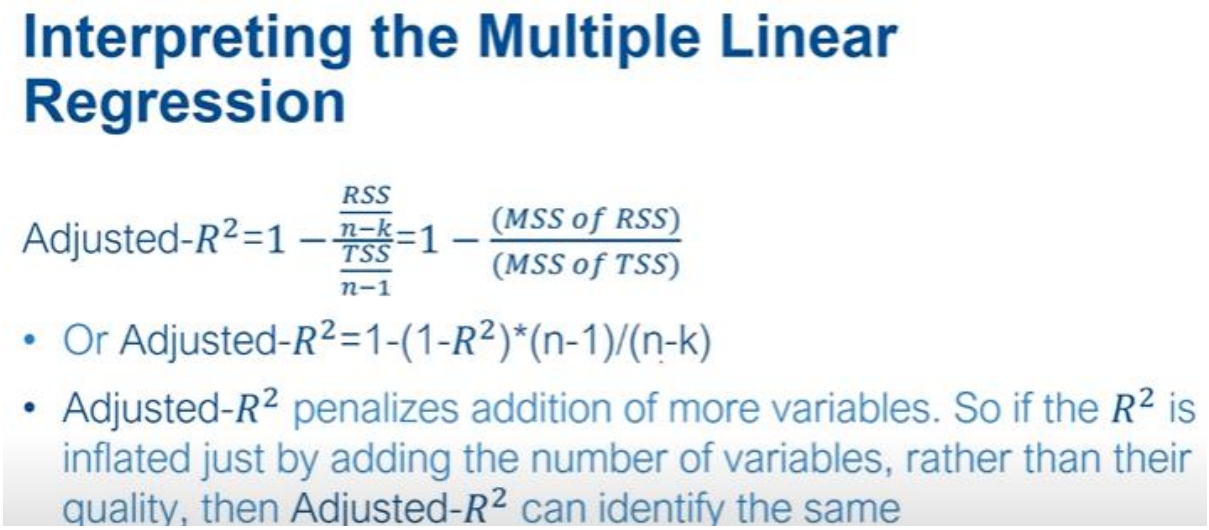
Similar to the two variable regression, the following expression

$E(Y|X_1 \ldots . X_n) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \ldots . + \beta_n X_n$

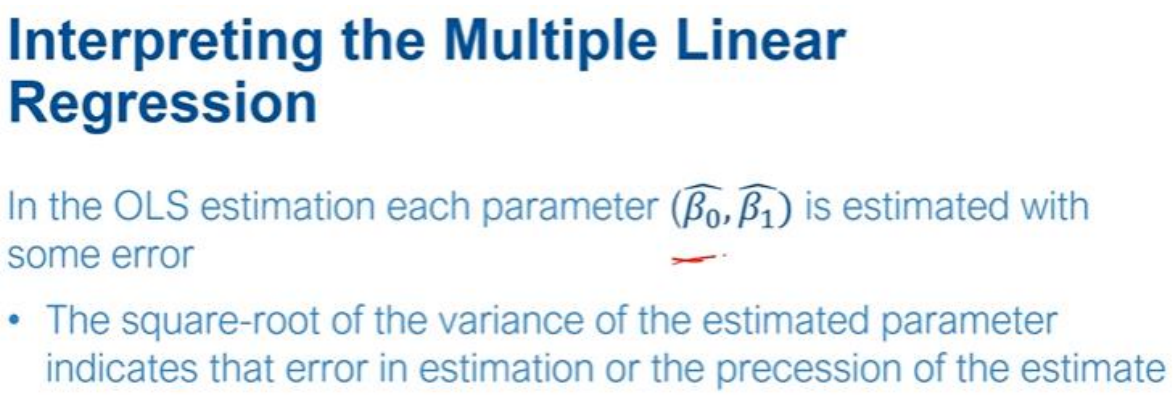
- The definition of $R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$, which is same as earlier
- One also calculates adjusted-$R^2 = 1 - \frac{\frac{RSS}{n-k}}{\frac{TSS}{n-1}} = 1 - \frac{(MSS\ of\ RSS)}{(MSS\ of\ TSS)}$; remember the dfs?

Now these n minus k and n minus 1 are called degrees of freedom. Please notice this adjusted R square is very interesting entity here. If we rearrange the expression for R square and just R square we get something like this. Please notice in R square in this expression we are putting a certain additional penalty in the form of these n minus 1 upon n minus k that means the more the number of variables the more the number of these k's or variables in the model this R square is penalized in the computation of adjusted R square that means the more the number of k's the lower the value of adjusted R square so there is certain penalty that this expression puts on R square and therefore this adjusted R square which is after adjusting for these degrees of freedom it puts a certain penalty for addition of variables. The reason being the normal R square can be easily inflated just by adding the number of variables, variables small small variables where the contribution to the explanatory power is so low that they are not very important but still if their numbers are too large the R square

value is inflated. So this penal term this penalty is inflicted on this computation so that there is a tradeoff of for adding more variables which in the form of explanatory power to the model as well as this penalty and therefore the resulting expression adjusted R square would identify whether the addition of variable is important or not.'



$$Adjusted - R^2 = 1 - (1 - R^2) * (n-1)/(n-k)$$

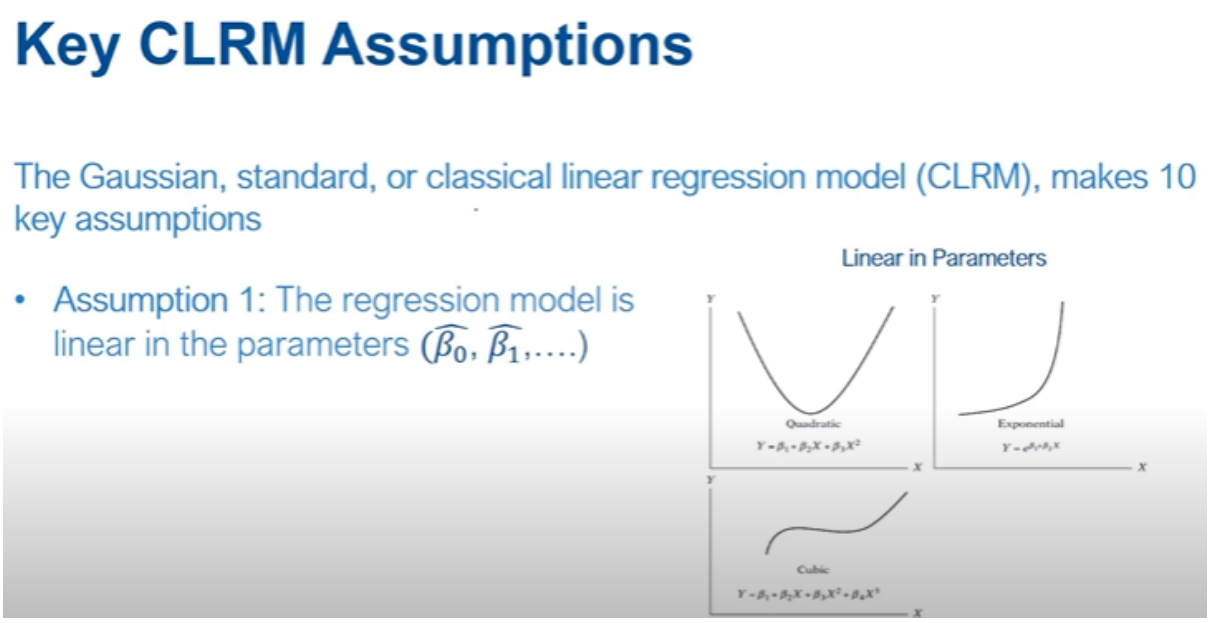


Lastly another very important aspect of this model is the coefficients or estimates of the coefficients beta 0 hat and beta 1 hat these are simply the estimates of the coefficients using OLS estimators based on OLS estimators these estimates are made these estimates have some variance so in repeated samples these estimates have variances and the square root of the variance of these estimated parameters indicates the error in estimation or standard error of the estimate. So if you have repeated samples these beta 0 will have a distribution estimates of beta 0 will have a distribution and higher this the standard error of this estimate or the variance of the estimate over the estimator. Remember we said that t statistic which was very important is coefficient upon standard error of estimate the higher the standard error of estimate the lower its t value and lower its significance in the regression model. So for us the better model or better estimator where SE is lower standard is lower given certain assumptions that we made earlier about heterosoxicity, autocorrelation, multicollinearity and so on OLS estimators are supposedly the best in class estimators available with us and that is

why we discussed OLS estimators in great detail. To summarize in this video we discussed the interpretation of multiple linear regression we noted that the expression such as conditional expectation of the dependent variable remains similar as bivariate regression model.

The interpretation of coefficient is also very simple these are simply partial coefficients which reflect the impact of a corresponding variable on the dependent variable keeping all the variables or all the variables constant. We also discussed the computation of R square measure and adjusted R square measure. We noted that adjusted R square measure provides slightly improved version of R square where we adjust for addition of more variables and therefore it recognizes the fact that addition of more variables simply just to increase explanatory power is not desirable and therefore it puts a penalty or a penalizing term penalizing adjustment to reflect that variables should contribute significantly and not just in small small explanatory power. We also noted that it is desirable that the estimators of the coefficient like beta naught hat and beta 1 hat these estimators should have lower variance that means their standard error of the estimate should be lower which increases the power of the estimate or power of the regression so that the t values and therefore the ability to reject the null and bring out its impact on the model more efficiently.

So, for that we need estimators with lower standard errors. In this backdrop OLS estimators are considered to be the best estimators as they offer the lowest variance in the entire class of linear estimators. In this video we will briefly review the key assumptions behind the CLRM that is classical linear regression model. The Gaussian standard or classical linear regression model as it is called makes 10 very important assumptions. The first assumption is that the linear regression model is linear in the sense it is linear in parameters.



For example, let us look at these three diagrams. This one y equal to beta 1 plus beta 2x plus beta 3x square is linear in parameters but non-linear in variables which is x squared term

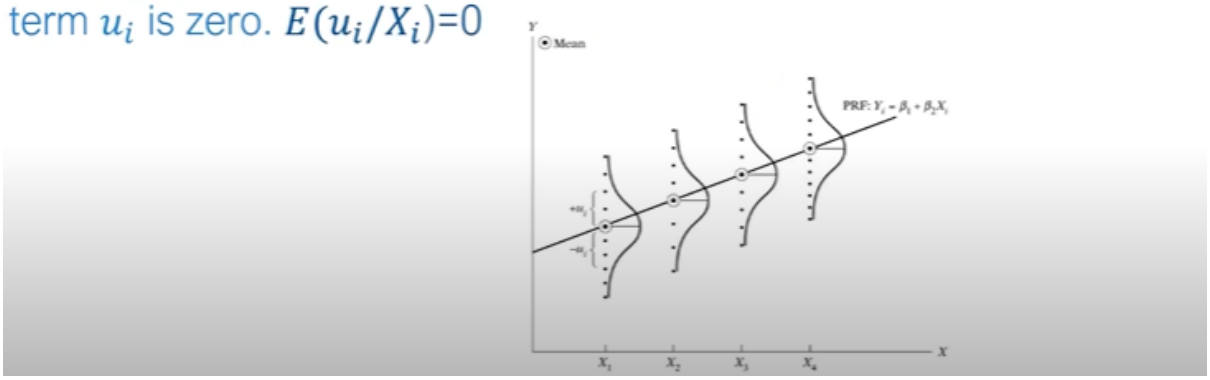
however it can be estimated with the CLRM because it is linear in parameters. On the contrary this one y equal to e to the power beta 1 plus beta 2x is non-linear in parameters. Although there are transformations as we will see that can make this expression linear but as of now it is non-linear in parameters. Look at this expression earlier we had a quadratic expression where x squared term was there here we have x cubed term but still it is linear in parameters and therefore directly estimated it can be directly estimated through classical linear regression model. Second assumption here is that the values of x in repeated samples are fixed that means they are not stochastic and they do not have any random component so they are fixed in nature unlike y or error term they are not stochastic they do not have any probabilistic distribution they are fixed and estimated with certainty there is no stochastic aspects of x.

Key CLRM Assumptions

Assumption 2: Values taken by the regressor X are considered fixed in repeated samples. More technically, X is assumed to be non-stochastic

Key CLRM Assumptions

Assumption 3: Zero conditional mean of disturbance (u_i): given the value of X, the mean, or expected, value of the random disturbance term u_i is zero. $E(u_i/X_i)$=0



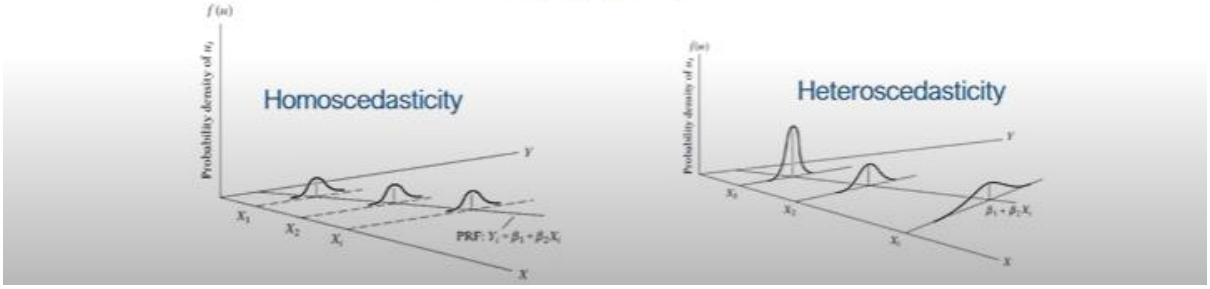
Next and very important assumption is that zero conditional mean of the disturbance term ui. Or the expected value of disturbance term mu i or error term is zero conditioned upon each xi. For example let us say we have x1 variable corresponding to which there are number of y's and for all these y's error terms mu 1 i's would be estimated mu 1 all these mu i's if we take the mean of these mu i's since it is for only x1 then mu i x1 equal to 0. Similarly for x2 all the mu i's mu 1 i mu 1 2 mu 2 2 and so on all these mu i's will be zero. So conditional

mean of all these error terms for given xi should be equal to zero. The next assumption, assumption number four is of homose elasticity or equal variance of these error terms.

Key CLRM Assumptions

Assumption 4: Homoscedasticity or equal variance of u_i. Given the value of X, the variance of u_i is the same for all observations. That is, the conditional variances of u_i are identical. $var(u_i/x_i)=E[u_i - E(u_i|X_i)]^2 = constant = \sigma^2$

- Heteroscedastic variance= $var(u_i/x_i)=\sigma_i^2$



$$var\left(\frac{u_i}{x_i}\right) = E\,[u_i - E(u_i|X_i)]^2 = constant = \sigma^2$$

$$Heteroscedastic\ variance = var\left(\frac{u_i}{x_i}\right) = \sigma^2$$

So like we said earlier for a given value of x the variance of mu i is same for all the xi's that is conditional variances of mu i are identical that is expected value of mu i minus expected value of mu i xi raised to the power 2 is constant and equal to sigma square. If they are not constant that means they are varying with xi's for each xi the variance is sigma square then this is called heterostatic variance and it creates problems that means the solution that we obtain for beta 0 hat the closed form solution that we obtain for beta 0 hat and beta 1 hat and so on so forth for other coefficients will not remain valid. Look at this diagram here, here with the assumption of homose elasticity the variance of error term for each x are same. Notice for each xi the variance of error terms is same. However, here the variances are different for different xi's and therefore it has heterostatic variance of error term.

In this particular case the solution the closed form solution for coefficients beta i hat that we obtained earlier we saw earlier for the simple linear regression will not remain valid. Another very important assumption, assumption number 5 is the no auto correlation between the disturbances or error terms. So, for any two values of xi and xj where i equal to 0 j there is the correlation between mu i's which are corresponding to xi and mu j's which are corresponding to xj should be 0. In symbolic terms mu i and mu j the covariance between mu i and mu j given xi xj which is also equal to expectations of this expression into this

expression whole square or expectation of mu i given xi into mu i mu j given xj because this expectation is anyway 0 expectation of mu i 0 and mu j is also 0. So, this only simplifies to mu i given xi and mu j given xj and this equal to 0 which is also a way of saying that correlation between mu i mu j given xi xj is 0.

$$Cov\left(u_i, u_j | X_i, X_j\right) = E\left[[u_i - E(u_i)|X_i][u_j - E(u_j)|X_j]\right]^2 = E[(u_j|X_i)(u_j|X_j)] = 0$$

Key CLRM Assumptions

Assumption 5: No autocorrelation between the disturbances. Given any two X values, Xi and Xj (i ≠ j), the correlation between any two u_i and u_j (i ≠ j) is zero.

Symbolically, $Cov(u_i, u_j|X_i, X_j) = E\left[[u_i - E(u_i)|X_i][u_j - E(u_j)|X_j]\right]^2$

$= E[(u_i|X_i)(u_j|X_j)] = 0$

So, this can also be said because correlation is nothing but normalized or standardized form of covariance. So, we can say that covariance between mu i mu j is 0 given xi xj or correlation between mu i mu j is 0 given xi xj. If there is some correlation then that can also be examined visually before going to the mathematics of it. For example, if there is no auto correlation then the error terms will be plotted randomly without any pattern. If there is positive auto correlation then error terms will plot like this with each other it indicates positive correlation and if it is plotted like this then it indicates a negative correlation between error terms that is mu i and mu j.

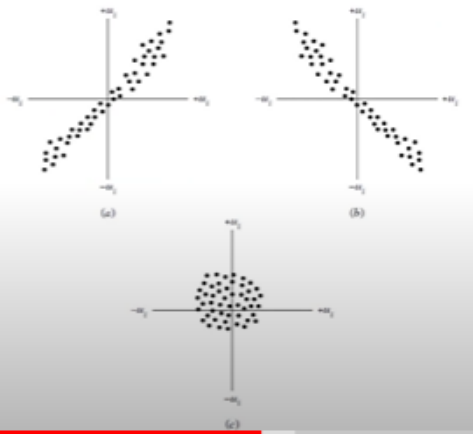
Key CLRM Assumptions

Assumption 5: No autocorrelation between the disturbances

(a) Positive autocorrelation

(b) negative autocorrelation

(c) No autocorrelation



Assumption number 6 another very important assumption is that mu i the error terms and xi are not correlated. This is very intuitive assumption because we said xi is fixed it is not stochastic it is taken from outside and is not getting affected by the process. So, mu i which if the model is correctly specified all the relevance influences are taken with xi's mu i should not be correlated with xi which is to suggest that covariance between mu and xi which is represented as this term expected value of mu into expected mu i and xi minus expected value of xi this should be equal to 0. Notice here we already know that the expected value of mu i this is 0 already. So, this expression simply becomes covariance of mu i xi simply becomes expectations of mu xi which is this term which is 0 which in simple terms says that correlation between mu i and xi is 0 because like we said earlier correlation is nothing but standardized form of covariance.

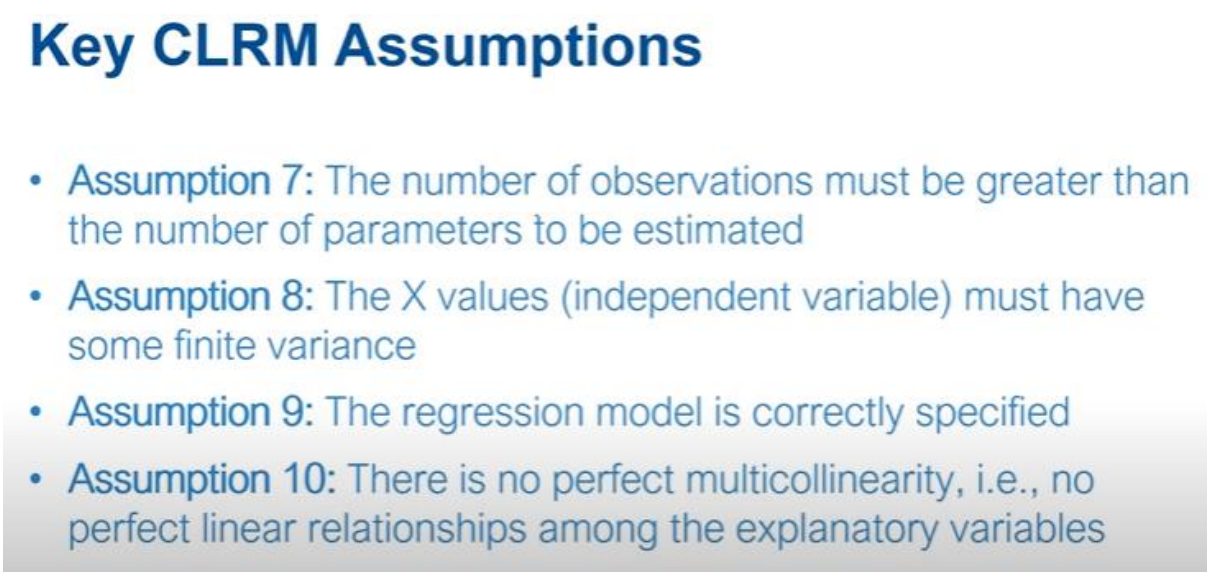
$$Cov(u_i, X_i) = E[(u_i - E(u_i))(X_i - E(X_i)]$$

Key CLRM Assumptions

Assumption 6: Zero covariance between u_i and X_i, or $E(u_i X_i) = 0$.

- $Cov(u_i, X_i) = E[(u_i - E(u_i))(X_i - E(X_i)]$
- By definition: $E(u_i)=0$ and $E(X_i)=X_i$
- $Cov(u_i, X_i) = E(u_i X_i) = 0$
- That is, u_i and X_i are not correlated

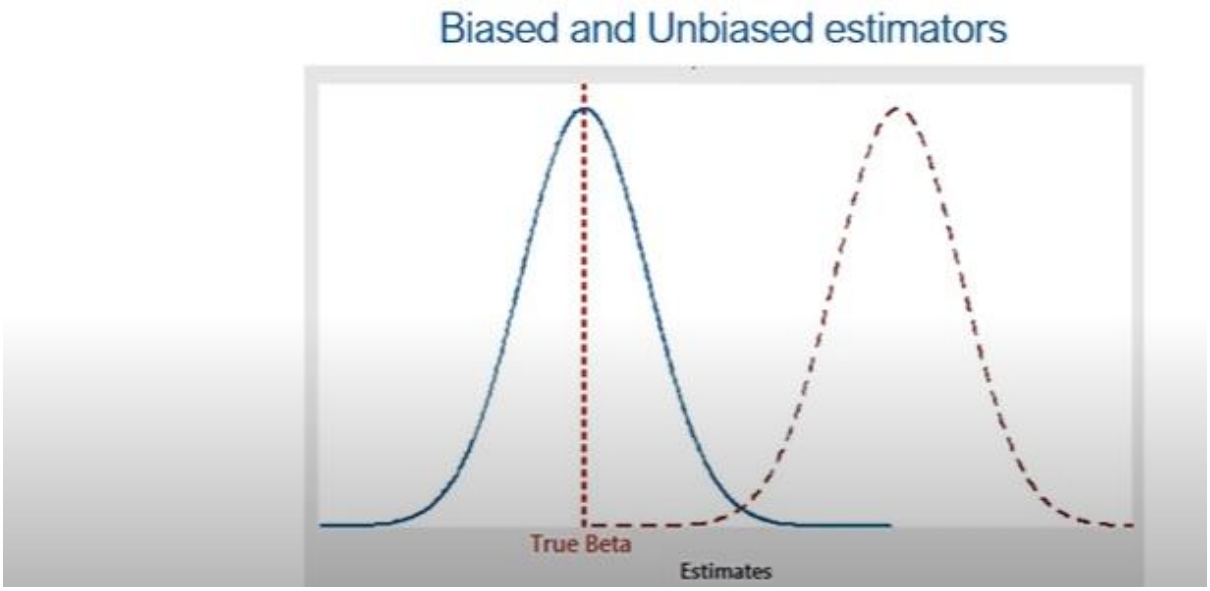
So, mu i and xi are not correlated which is our assumption 6. Then the remaining assumptions are quite intuitive as well. These remaining four assumptions are assumption 7, 8, 9, 10. The 7th being the number of observations must be greater than the number of parameters to be estimated. In fact, for a better estimation the number of observations should be much much larger than the parameters to be estimated. A good thumb rule should be that number of observations should be at least five times the number of parameters.



The x values the dependent variable must have some finite variance. For example, if the variance is very low or if all the x values are almost same then the model cannot be estimated or even if estimated the estimates are very poor. The regression model is correctly specified which means all the relevant influences that is the x signs are taken in the model and those that are left or not considered should be very small that even if they are mixed with the error term because we are not accounting for them specifically in the model they should not create any trouble in estimation. Lastly, there is no perfect multicollinearity that means there should be no perfect relationship between the independent variables of this form. So, they if it is of this form this indicates that one of the independent variables can be perfectly explained by the remaining variable and in this case the model will be indeterminable, it cannot be determined. To summarize in this video we reviewed and examined all the key 10 very important assumptions of classical linear regression model.

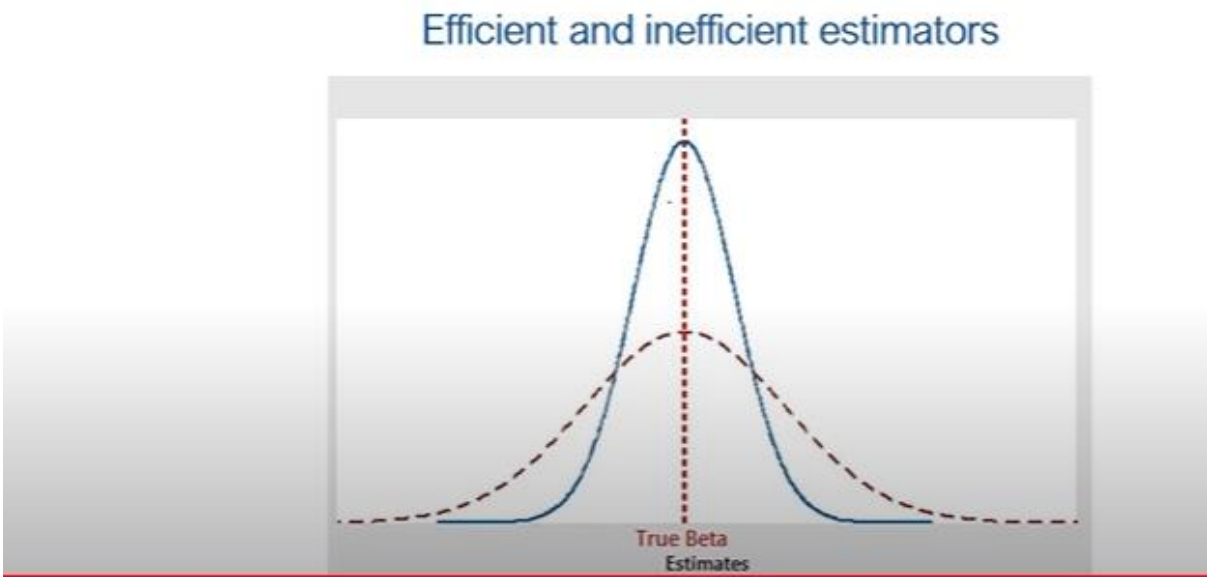
In this video we will introduce blue properties of OLS estimators. OLS estimators are often considered to be as best linear, unbiased and efficient estimators. These are called blue properties. We already know that OLS estimators are linear in parameters. In addition, there are two very important properties that is unbiased and efficient.

BLUE Properties of OLS Estimators



Let us start with the unbiased estimators. Any estimator that has its expected value same as the true population parameter for example if we are trying to estimate beta1 using some estimator beta1 hat like OLS estimator and if the expected value of this beta1 hat in large samples in large repeated samples if in repeated samples the expected value of beta1 hat is same as the true population then it is called unbiased estimator. However, if the expected value is different for example this curve notice this if the beta distribution the distribution of the coefficient here and expected value is somewhere here which is different from the true beta then this is unbiased estimator. In contrast, the blue one this one is called unbiased because its expected value is somewhere around true beta and such estimators are often called unbiased estimators. It is considered that if certain assumptions that we discussed regarding multicolinearity, autocorrelation, heteroscedasticity the 10 classical assumptions if they are held then OLS estimators are considered as unbiased estimators.

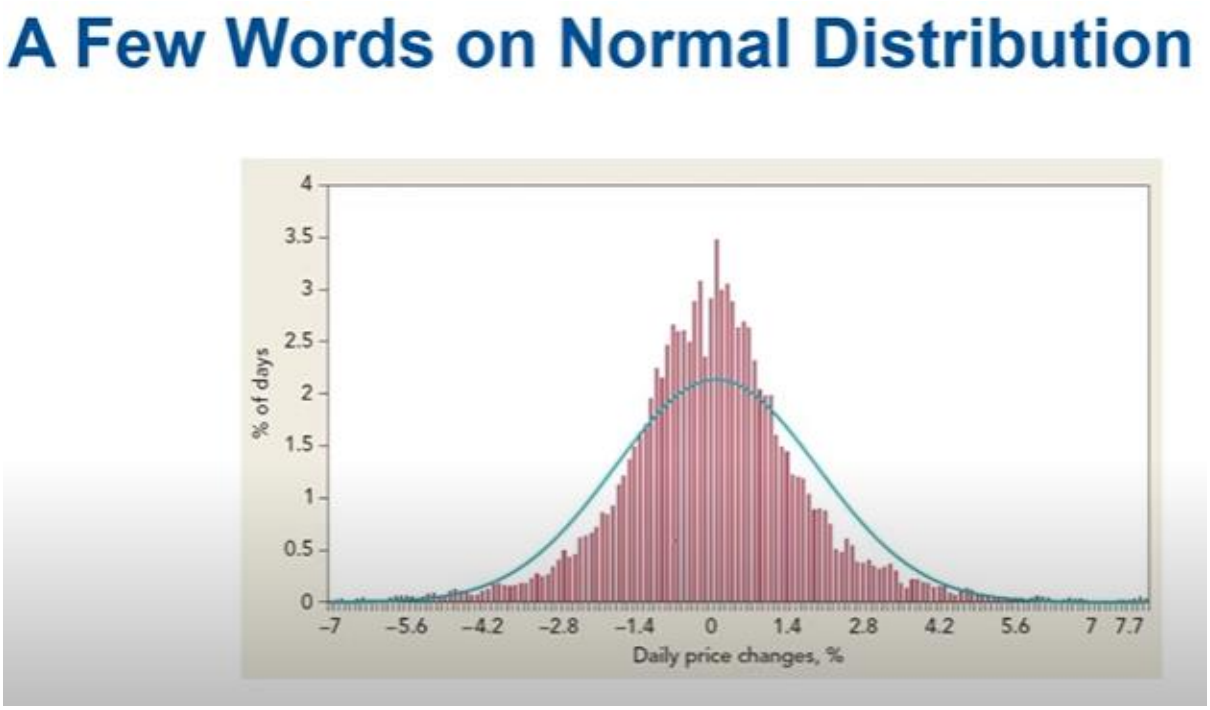
BLUE Properties of OLS Estimators



Next, we have efficient versus inefficient estimators. Estimators with very low variance or standard error, standard is nothing but the variance of estimator. So, if the variance of estimator is low it is called relatively efficient. In this figure this estimator has a lower variance lower standard error as compared to the dotted one, the solid blue line has lower variance as compared to dotted one and therefore it would be considered as more efficient as compared to this. Across different classes of estimators OLS estimators are said to be best in terms of their efficiency that means they have the lowest standard error of the estimate. If these two properties of unbiased and efficiency are combined it is said that for OLS estimators in large samples they are more their expected value converges closer and closer to the true beta and they become more and more efficient that means their variance becomes lower and lower and therefore the combined property of efficiency and unbiasedness is often referred to as consistency.

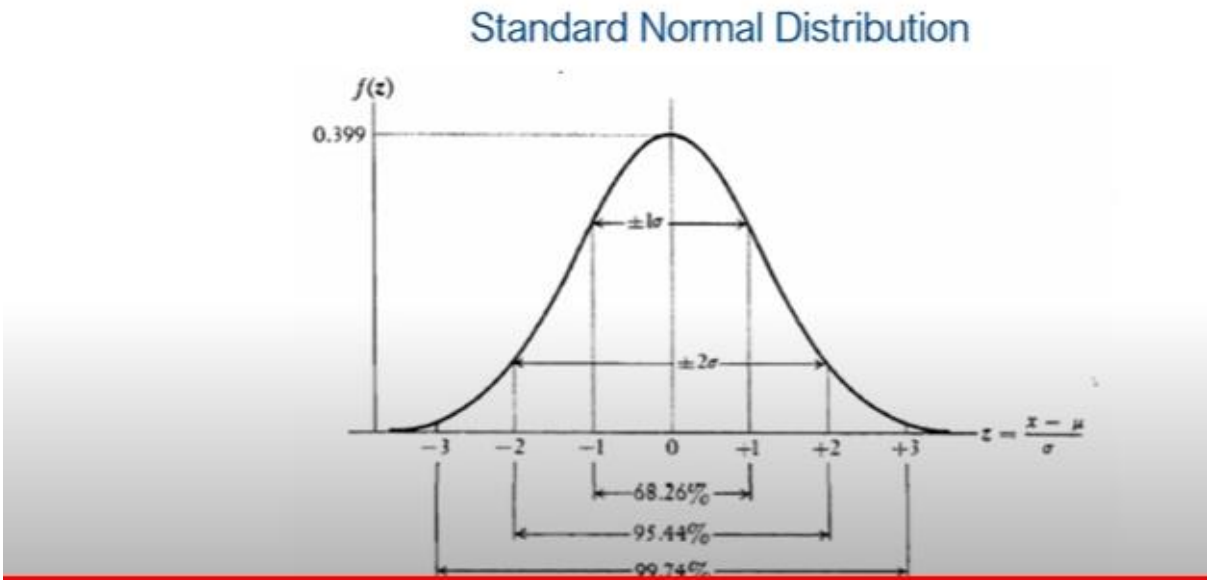
What is consistency? Here consistency is that in large samples the expected value and the very expected value and variance converge to true beta that means the estimate converges to true beta its expected value becomes closer to true beta and its variance comes lower and lower so essentially it converges to true beta in large samples which is called the consistency. Combining these properties OLS estimators are best in class and therefore they are often referred to as best linear unbiased and efficient estimates. To summarize in this video we discussed the blue property of OLS estimators we noted that across all the samples across all the estimators OLS estimators are best linear and biased estimators and combining the property of unbiasedness and efficiency they are consistent estimators as well that is in large samples the sample estimates converge to population parameters. Classical linear regression model CNLRM and hypothesis testing part 1. In this video we will introduce the

normal distribution and its application in classical linear regression model for hypothesis testing and estimating the significance of coefficients or parameter estimates.



Recall the discussion on normal distribution. A simple normal distribution on returns have daily price changes or returns on the x-axis here on the horizontal x-axis and on y-axis we may have frequency of days or percentage of days. A fitted normal distribution would look something like this green bell shaped curve while the bars are actual returns the continuous shape normal distribution can be superimposed on this. The important aspect of normal distribution is that it can be simply defined by two parameters one is the mean which is mu and its variance sigma square where sigma is the standard deviation. Using just these two parameters we can define a normal distribution and as we will see in this video the application of normal distribution considerably improves the quality of analysis or regression analysis with hypothesis testing and estimation of significance of parameter estimates.

A Few Words on Normal Distribution



A version of normal distribution which is most often employed is called standard normal distribution. Most of the properties of normal distribution in fact all the properties remain similar except that on x-axis instead of the actual values say returns we use the scaled or normalized or standardized return that is x which is the value of return minus mean upon standard deviation so it is normalized so that its mean is 0 mean of 0 this distribution will have a mean of 0 and standard deviation of 1 because of this transformation. On the y-axis instead of frequency or percentage frequency we use something called probabilities or rather probability densities. So, for example, each point will represent a particular standardized return in z form and on y-axis we will have the corresponding probability which is essentially a probability density the probability corresponding to this point which is essentially nothing but probability density so it would make more sense to have a segment of strip on the x-axis so the area will represent the probability of observing return in this window which is represented by a z interval from x1 to x2 where x1 and x2 can be mapped to certain return. For example, from z or minus 1 to plus 1 that is in terms of standard deviation this essentially represents the standard deviation form of returns so minus 1 to plus 1 standard deviation we have 68.

Classical Normal Linear Regression Model (CNLRM)

The estimation of sample parameters is not complete without hypothesis testing ($\widehat{\beta_0}, \widehat{\beta_1}$)

- It is important to draw inferences about population parameters using sample estimates, more clearly, we would like the estimated parameters to be as close as possible to population parameters
- It must be noted that the randomness in the beta (coefficient) estimates is introduced by μ_i (error term): How?
- Thus, these sample coefficient estimates also have a probability distribution [as one take different samples from population, one gets different estimates]

26 percent of return plus minus 2 standard deviation area that is 95.44 percent of the probability will lap within plus minus 2 standard deviation and within plus minus 3 standard deviation we have 99.74 percent and this is a respective of different returns because this is now in standardized form so we can simply say that within plus if it follows normal distribution then as per standard normal distribution irrespective of mean the mean and standard deviation can be anything if it follows normal distribution then for standard normal distribution these properties will be held no matter what are the mean and standard deviation that means within plus minus 3 standard deviation you have 99.74 percent of the probability within plus minus 2 percent you have 95.44 percent probability and within plus minus 1 percent there is 60.

286 percent probability that observations will lie within this. Now with this introduction of normal distribution we can apply the hypothesis testing concept on our parameter estimates beta naught hat and beta 1 hat because hypothesis testing and inference inferential statistics are required for sample parameters we need to draw inferences about population parameters using these sample estimates of beta naught hat and beta 1 hat and more clearly to be more precise we would like the estimated parameters these beta naught hat and beta 1 hat to be as close as possible to the population parameters that is beta 1 beta naught. Now please remember why these beta naught hat and beta 1 hat are different from beta naught beta 1 because in repeated samples these beta naught hat and beta 1 hat will be different because there is certain randomness introduced because this error term. So, this error term introduces randomness in the model and because of that the estimates are not same as population parameters every time you change the sample you will obtain certain different beta naught hat and beta 1 hat and other population parameters. Please note because this randomness is introduced by this error term the sample coefficient estimates that is beta naught hat and beta 1 hat will also have a probability distribution and this probability distribution will be similar to the error term because this randomness is introduced by this error term itself and therefore

every time you change the sample you will obtain different estimates of beta naught hat and beta 1 hat and therefore if you want to make any inference about beta naught hat and beta 1 hat you need to know their distribution which is essentially the distribution of the error term itself because this is the one causes causing randomness in beta naught hat and beta 1. So, we assume we make the following assumption about the error term in the model remember the simple model y equal to beta this model which we extended to multiple linear regression where this mu was the error term.

The Normality Assumption of the Error Term μ_i

To make any inference about the probability distribution of the estimate, we need to make some assumption about the distribution of the error term μ_i

- The CNLRM assumes that μ_i is distributed normally with the following:
- $Mean = E(u_i) = 0$
- $Variance = E[u_i - E(u_i)]^2 = E(u_i)^2 = \sigma^2$
- $Covariance = E\left[[u_i - E(u_i)][u_j - E(u_j)]\right] = E(u_i, u_j) = 0;\ i \neq j$
- These assumptions are summarised as $\boldsymbol{u_i \sim N(0, \sigma^2)}$

$$Mean = E(u_i) = 0$$
$$Variance = E[u_i - E(u_i)]^2 = E(u_i)^2 = \sigma^2$$
$$Covariance = E\left[[u_i - E(u_i)][u_j - E(u_j)]\right] = E(u_i, u_j) = 0$$

So, the following assumptions are made about this error term first the expectations of this error term or the mean of this error term is 0 first. Second we assume that the variance that is expected value of mu i minus expected mu i square or this because this is 0 is same as sigma square. So, this is homoscedasticity the variance of error term is constant. Second the covariance or correlation between mu i and mu j because this is 0 it converges to this expression is also 0 that means mu i and mu j across observations i and j are not correlated. Combined the assumption can be written as in this form that is mu i is distributed normally with a mean of 0 and variance of sigma square with the 0 autocorrelation.

Please note that normal distributions like we said are very easily defined just by two parameters one is the mean mu and variance sigma square. Now, as we have made the normality assumption and previous other classical linear regression model assumptions that we discussed 10 assumptions plus normality. OLS estimates are blue estimates that is their best linear unbiased efficient and consistent estimates as we discussed in the previous video and these are therefore since they are consistent these estimates converge to their true

population values as sample size increases. The model as we have already seen y i equal to beta naught hat plus beta 1 hat x i plus mu i hat here the predicted values of y i are written as y i hat which is beta naught hat plus beta 1 hat into x i. Now, please note the mean of it is assumed that the mean of beta 1 hat or expected value of beta 1 hat is same as the population parameter because if these assumptions are held then this expected value of beta 1 hat is same as beta 1 and variance of beta 1 is sigma beta 1 square and to summarize these two definitions we have beta 1 hat distributed normally.

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\mu}_i; where\ \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$
$$Mean: E(\hat{\beta}_1) = \beta_1; Variance = var(\hat{\beta}_1) = \sigma^2_{\hat{\beta}_1}$$

Properties of OLS Estimators under Normality

Normal distributions are very easily defined with just two parameters, i.e., mean and variance of the population

- Under the normality assumption, OLS estimates are unbiased, efficient, consistent (estimates converge to their to population values as sample size increases) $Y_i = \widehat{\beta_0} + \widehat{\beta_1} X_i + \widehat{\mu_i}$; where $\widehat{Y_i} = \widehat{\beta_0} + \widehat{\beta_1} X_i$
- Mean: $E(\widehat{\beta_1}) = \beta_1$; Variance = $var(\widehat{\beta_1}) = \sigma^2_{\widehat{\beta_1}}$; then $\widehat{\boldsymbol{\beta_1}} \sim \boldsymbol{N(\beta_1, \sigma^2_{\widehat{\beta_1}})}$

So, this represents that beta 1 hat is distributed normally and its mean is beta 1 and its variance is sigma square beta 1 hat. So, this is the combined expression for the assumptions. Now, we can also translate these assumptions that we discussed in the previous slide in the standard normal distribution format. Recall that this z or the standard normal version is the value itself beta 1 hat minus its mean which is the estimate of population this is expected to be same as population parameter. So, beta 1 hat minus its mean divided by standard deviation of beta 1 is the standardized version which is z and like we said this z is distributed normally with the 0 mean or expected value and a variance which is 1 because it is the standardized version. So, it is normally distributed with a mean of 0 and standard deviation of 1 and it helps us in doing any kind of hypothesis testing and inferential statistics.

Properties of OLS Estimators under Normality

Normal distributions are very easily defined with just two parameters, i.e., mean and variance of the population

- By the properties of standard normal distribution $Z = \frac{\widehat{\beta_1} - \beta_1}{\sigma_{\widehat{\beta_1}}}$
- Where $\boldsymbol{Z \sim N(0,1)}$: Z is normally distributed with mean of 0, and SD=1

$$Z = \frac{\hat{\beta}_1 - \beta_1}{\sigma_{\hat{\beta}_1}}$$

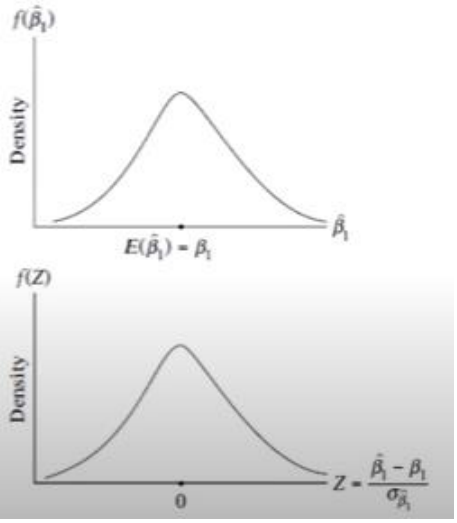
To summarize in this video we introduced the normality assumption with error term since the error terms are the ones introducing randomness in the population parameter estimates or sample estimates. Therefore, the sample estimate itself is normally distributed the beta 1 hat beta 0 hat sample estimates. Normal distribution because normal distributions are very easily defined with just two parameters that is mean and variance and if the assumptions are held then the expected value of beta 1 hat is same as beta 1 which is the population parameter. So, the sample estimate converges to population parameter estimate if the assumptions are held as we discussed the classical normal linear regression model assumptions are held. Its variance is variance of beta 1 hat which we called as sigma square beta 1 hat and which is this sigma square beta 1 hat and therefore we said that beta 1 hat is normally distributed with the mean of beta 1 and variance of sigma square beta 1 hat.

$$Mean: E(\hat{\beta}_1) = \beta_1; Variance = var(\hat{\beta}_1) = \sigma^2_{\hat{\beta}_1}$$

Properties of OLS Estimators under Normality: Summary

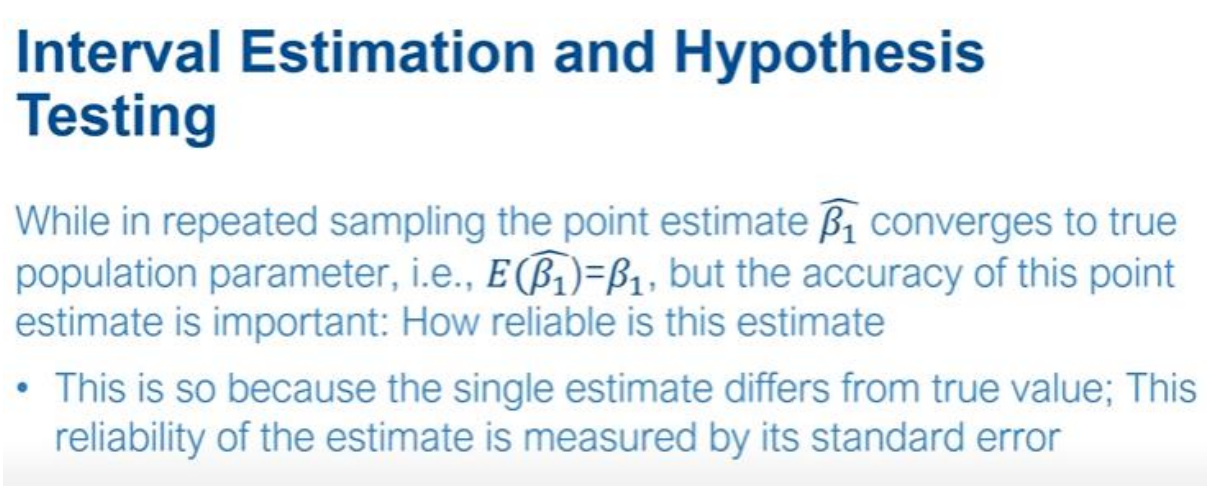
Normal distributions are very easily defined with just two parameters, i.e., mean and variance of the population

Mean: $\boldsymbol{E(\widehat{\beta_1}) = \beta_1}$; Variance = $\boldsymbol{var(\widehat{\beta_1})}$ $\boldsymbol{= \sigma^2_{\widehat{\beta_1}}}$; then $\boldsymbol{\widehat{\beta_1} \sim N(\beta_1, \sigma^2_{\widehat{\beta_1}})}$



We also saw that in standardized form this is represented by the statistic Tz which is standardized with this formula it has a mean of 0 and a variance of 1 and distributed normally. In the next video we will see the application of this normal distribution and conclude its role in hypothesis testing and making inferences about the coefficient. In this video we will conclude our discussion on hypothesis testing with classical normal linear regression model. Please remember in repeated sampling the point estimate that is beta 1 hat converges to population parameter which is beta 1.

This is to suggest that if you have number of samples the mean of beta 1 hat will converge to two population parameter beta 1. However, we do not have the luxury of working with many samples and often we are left with one sample even that is a very small sample as compared to the overall population and therefore the accuracy of this point estimate of this beta 1 hat is important. How reliable is this estimate? Because this single estimate definitely differs from the true population parameter and therefore the reliability of this estimate is measured by standard error. Please recall that in OLS estimation of regression model each parameter estimate that is for example beta 0 hat or beta 1 hat is estimated with some error. The square root of the variance of this estimated parameter indicates that error in estimation or essentially the precision of that estimate. So the variance of this parameter estimate we take the square root of that which we take the square root of that which becomes the standard error of this estimate.



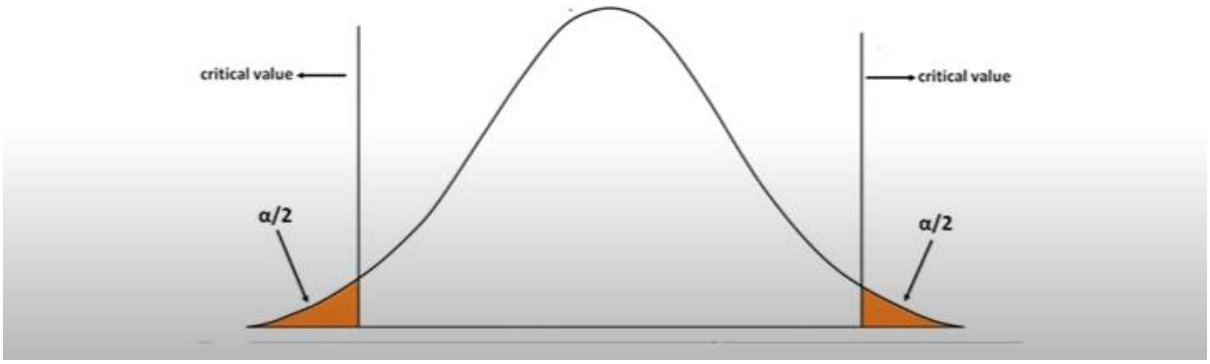
Interval Estimation and Hypothesis Testing

In the OLS estimation each parameter $(\widehat{\beta_0}, \widehat{\beta_1})$ is estimated with some error

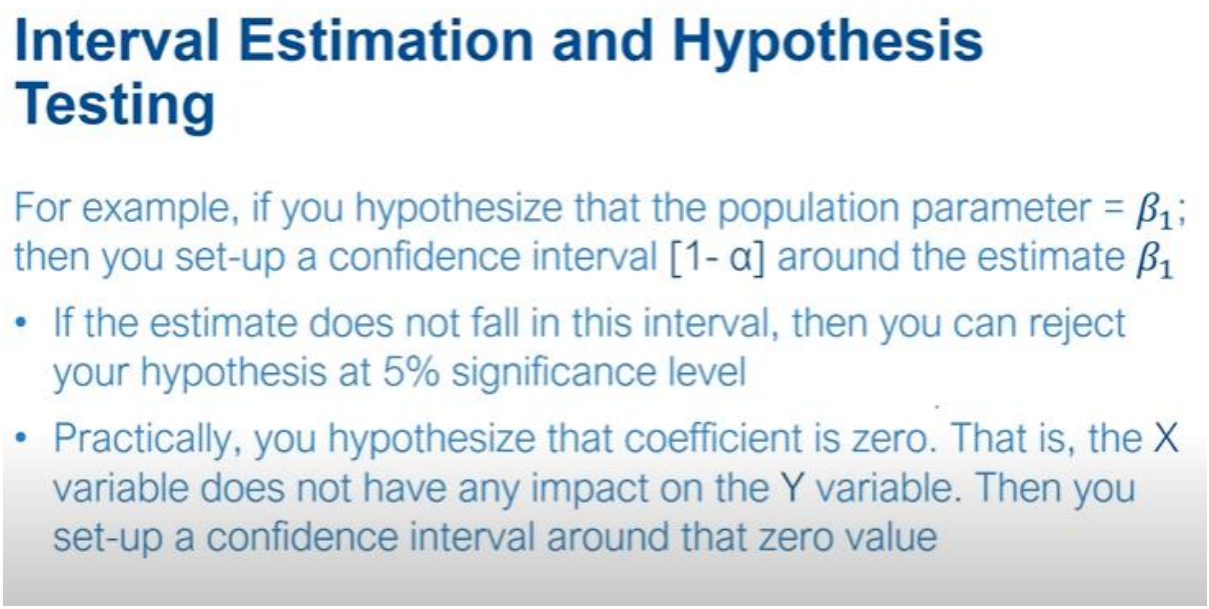
- The square-root of the variance of the estimated parameter indicates that error in estimation or the precession of the estimate

Interval Estimation and Hypothesis Testing

In statistics we configure the confidence interval around the estimate



Let us examine how this standard error plays a very important role in hypothesis testing and confidence interval estimation around this coefficient. In the hypothesis testing procedure you set up a confidence interval for example let's say you are you want to be 95 percent confident in your estimate of that coefficient or parameter estimate then there is a five percent possibility or five percent chance you are willing to take where you may be wrong and therefore if you have taken this 95 percent region of confidence you are considering then the remaining five percent which is distributed in two half regions five percent divided by two and five percent divided by two which is 2.5 percent is called critical region or region of significance where you may be mistaken.



So this is the region which you are willing to take a chance where you may be mistaken 2.5 percent on the right and 2.5 percent on the left which is overall five percent because you have taken a 90 you are willing to take a 95 percent confidence. Let's see how this works in the case of regression. Let us say you hypothesize the population parameter as beta1 which means if it is a simple linear regression model you are estimating a regression like this beta0 plus beta1 into x and you estimate a coefficient of beta1. Now the way the convention suggests you consider this beta1 or hypothesize it to be 0 which has a practical implication that if beta1 is 0 x our variable of interest has no impact on y. This is how is the convention works so you assume that you start with a hypothesis that beta1 is the 0 or your true population parameter is 0 which means x does not affect y.

Now you want to set up a 95 percent confidence interval and therefore if you want to set up a 95 percent confidence interval which is 1 minus alpha your alpha or significance level is 5 percent so you set up a 95 percent confidence interval around your population estimate beta1 which is 0 you set up this confidence interval and you are willing to take a 5 percent chance 2.5 percent on the left and 2.5 percent on the right that you may be wrong as well. Practically you are assuming that beta1 is 0 that means x does not affect y so this is your essentially your null hypothesis and if this null is rejected you consider a null that beta1 is 0 and if your estimate falls in this region you tend to not reject your null however if it your estimate falls in this region which is the significance region or critical region then you reject the null and say that beta1 is not 0 it is significant and the implication is that x indeed affects y so there is a relationship between x and y if you are able to reject the null. Now in practice along with 95 percent we also check the significance with 90 percent and 99 percent that means this value in case of 90 percent this value is 5 percent on the left and 5 percent on the right in case of 99 percent it is 0.

5 percent on the left and 0.5 percent on the right that means 1 percent significance. Now that we have seen standard normal distribution we also know that if this is a normal distribution

then for each value of probability whether it is 5 percent 2.5 percent we have a z value so we depending upon the area in this region we can find a relevant z value corresponding to each point that we have already seen. In practice we have a small we have small samples and often population variances are not known so instead of using z parameter we make use of students t distribution which is very similar to standard normal distribution only minor difference is that students t distribution has relatively fatter tails it has fat tails which means for any hypothesis testing rejection of null requires more evidence so it needs more evidence to reject the null and therefore we because we work with small samples and we do not want to make any major mistakes we tend to work with students t distribution however the procedure remains exactly similar whether you use normal standard normal distribution or students t distribution the procedure exactly similar however in practice we tend to do hypothesis testing with regression coefficients using students t distribution as we'll see in our practice example because it requires more evidence to reject the null and therefore if you are rejecting the null we are more confident so we try to find the corresponding t value corresponding to this alpha value we try to find this t value which is on this side left side it is minus t alpha by t by 2 or plus t alpha by 2 alpha by 2 indicates that this t value is corresponding to alpha by 2 so in the t table t statistic table we take the values which are corresponding to plus alpha by 2 and minus alpha by 2 they are same because of symmetry in the distribution they are same now the way it is done you compute the relevant t values you compute the relevant t values or critical t values corresponding to these alpha by 2 for example if alpha is 5 then you compute the t value corresponding to 2.5 level on the left side and 97.5 on the positive side the right side so you compute t values corresponding t values these are called critical t values which is plus minus t alpha by 2 1 for 2.

5 on this side critical value and 1 for 97.5 this is in the case of 95 confidence interval you can find the relevant values similarly for the 90 percent or 99 percent level of confidence once you compute these critical or significant t values you compare them with the t value that is obtained in the sample so you compute a sample t value which is there for your sample and you compute if these t values is larger than this plus minus t alpha by 2 that means it falls on the right side either on the right side in the critical region or on the left side of the critical region then you reject the null that means you say that beta estimate of beta 1 hat or beta 2 hat or whatever is estimate is under consideration it is significantly different from 0 it is not equal to 0 if it is on this side that means it is negatively negative and if it is on this side right side it is positive but it is significantly different from 0 that means it is not 0 and therefore we are rejecting our null and saying with that level let's say in this case 95 confidence or 5 significance that beta 1 is significant and x the variable of interest indeed affects the dependent variable y so just to summarize the discussion you assume a null hypothesis which is called h0 that the true population parameter beta 1 which is assumed to be 0 in most of the cases and therefore your alternate hypothesis h1 is that true population parameter beta 1 is not equal to 0 or the variable of interest which is xi affects indeed affects the variable y the decision rule here goes like this you construct a confidence level of 1 minus alpha where alpha is 5 percent the level of significance so your confidence interval is 95 percent for the population parameter beta 1 then you compute the relevant t statistic corresponding to this critical value of alpha by 2 on this side and here alpha by 2 on this side so you compute the

critical t value so let's call on this side this is minus t alpha by 2 this will be negative so for example for 95 percent here the relevant t statistic will be corresponding to a 2.5 percent value and on the right side it will be corresponding to a 97.5 percent value if your computed t statistic for the coefficient if the computed t statistic which is beta 1 estimate estimate of beta 1 minus 0 upon its standard deviation of beta 1 hat this number this is the t statistic that you compute for the sample if it is greater than or falls in this region greater than this value or lower than this value and fall in this region then you reject the null h nWW you don't say that i accept the null but you reject the null you reject the null or fail to reject the null you don't say that you accept the null either you say i fail to reject the null or you reject the null so if you reject the null then you say that x this variable affects y if you fail to reject the null that means your parameters that means your parameters fall in this confidence region then you fail to reject the null and say x does not affect y. In this video we'll discuss some of the other functional forms of OLS regression that are non-linear in nature and we'll also discuss how to transform them for suitable estimation with OLS regression procedure. To begin with we'll have a look at some of the log linear or log log formations for example have a look at this expression y i equal to beta one times x i to the power beta two into e to the power mu i now in order to estimate this notice the parameters are of non-linear form and therefore if you take the natural log transformation of this expression on both sides left hand side and the right hand side on the left hand side we have ln y i on the right hand side we have ln of beta one plus beta two times ln of x i plus mu i which can be further transformed into y i dash equal to alpha plus beta two times x i dash plus mu i.

$$Log - linear\ or\ log - \log model: Y_i = \beta_1 X_i^{\beta_2} * e^{u_i}$$

$$\ln(Y_i) = \ln(\beta_1) + \beta_2 \ln(X_1) + u_i$$

$$Y_i' = \alpha + \beta_2 X_i' + \mu_i$$

Other Functional Forms and Non-linear Transformations

- Log-linear or log-log model: $Y_i = \beta_1 X_i^{\beta_2} * e^{u_i}$; take natural log and transform the model as below
- $\ln(Y_i) = \ln(\beta_1) + \beta_2 \ln(X_i) + u_i$ or alternatively
- $Y_i' = \alpha + \beta_2 X_i' + u_i$: The model is now linear in parameters α and β_2
- The interpretation goes as follows: β_2 measures percentage change in Y_i for a given percentage change in X_i

Now this final transformed model notice it is linear in parameters alpha and beta two and therefore if we estimate this model through OLS regression procedure then repetition of beta two here is important which is of coefficient of interest which is beta two measures the percentage change in y i for a percentage change in x i because this is log log transformation that is why it is percentage change in y i for a given percentage change in x i but for all times to come remember that the original nature of equation has transformed so if you wanted to estimate this beta two the interpretation with respect to original x i and this x i dash is different nonetheless this expression can be estimated with the OLS regression procedure. Now have a look at this another expression for which we are going to use log-lin transformation y t equal to y naught plus one plus r raised to the power t if we take the natural log and transform the model as below here natural log of y t equal to beta one plus t times beta two because it was one plus r to the power t the logarithm transformation has made it in this form. Now please note this is a semi-log model because the dependent variable y t is in log form while the independent variable t is not in log form and therefore the interpretation of beta two is as follows beta two measures percentage of proportional change in y t for a given absolute change in t however if your coefficient of interest were y naught and one plus r the interpretation is not exactly identical to them nonetheless this model can be transformed with the help of OLS procedure. Also this was a log-lin transformation a vice versa interpretation for lin log model as we can see here that is given change absolute change in y t for a percentage change in x t will be applicable to a model of this kind where x t the dependent variable is in the log form while y t the independent variable is not in log form and therefore the interpretation of beta two will be like we said absolute change in y t for a percentage or relative change in x t. To summarize this video we noted that there are certain expressions where which are not exactly linear in parameters however with some transformations such as taking log on both sides can transform them into different or resulting format which is linear in parameters and therefore the resulting format or the transformed expression can be estimated using OLS regression procedure.

$$Log - lin\ model: Y_t = Y_0(1+r)^t$$
$$\ln(Y_t) = \beta_1 + t\beta_2$$
$$Y_t = \beta_1 + \ln(X_t)\beta_2$$

Other Functional Forms and Non-linear Transformations

- Log-lin model: $Y_t = Y_0(1+r)^t$; take natural log and transform the model as below
- $\ln(Y_t) = \beta_1 + t\beta_2$
- This is a semi-log model, and β_2 measures proportional change in Y_t for a given absolute change in t
- Vice-versa interpretation goes for Lin-log model below (absolute change in Y_t for a % or relative change in X_t.
- $Y_t = \beta_1 + ln(X_t)\,\beta_2$

However the interpretation of coefficient from this transformed equation may not be identical to the original equation nonetheless it is still useful for estimating the coefficients and making policy decisions and from such perspectives. Among supervised learning algorithms regression algorithm is a very important tool employed in finance domain for applications such as forecasting security prices or create scoring. Regression algorithms can be run with only two variables one independent and one dependent which is known as simple linear regression or with more than two variables which is called as multiple linear regression. The key variables in a regression model include a dependent variable, one or more independent variables, coefficients of these variables and an error term.

The error term accounts for the variation in the dependent variable that cannot be explained by the model. While regression analysis can provide the statistical significance of the relationship the direction of causality should come a priori from the theoretical underpinnings. Refer to the rain versus crop example in this lesson. Ordering least square or OLS is the most often employed method to estimate a regression model which involves minimizing residual sum of squares. OLS estimation or ordinary least square estimation of regression involves 10 key assumptions. The most important assumptions here include linearity in parameters, exogeneity of independent variables, zero conditional mean of the error or residual term, homoscedasticity of error variances, absence of multicollinearity, no autocorrelation across error terms, no correlation between error and dependent variables.

If these assumptions are held then OLS estimators are referred to as BLUE that is best linear unbiased and efficient estimates. The statistical significance of OLS estimators is determined through hypothesis testing of coefficients individually. This requires normality assumption of the error that is residuals. Very often the model is not linear and may require some kind of transformation to make it linear which can be subsequently estimated through OLS. However, the interpretation of coefficients also change with such transformations. Thank you.