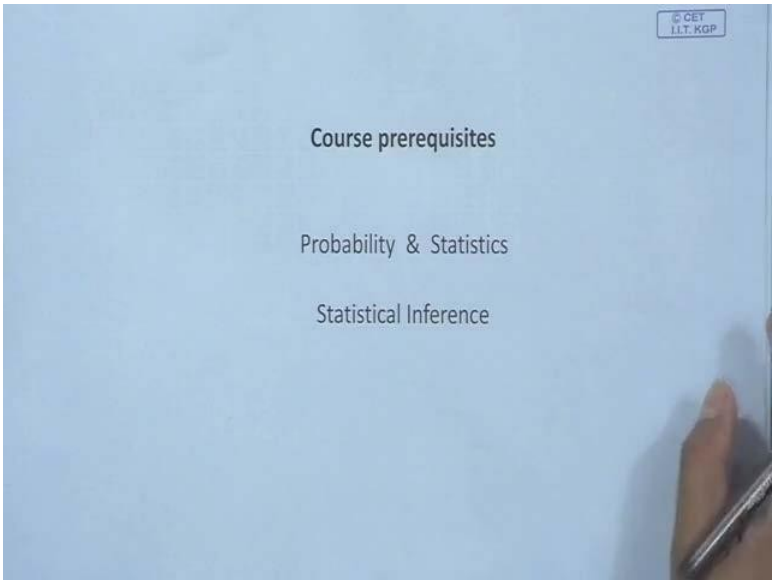
Regression Analysis
Prof. Soumen Maity
Department of Mathematics
Indian Institute of Technology, Kharagpur

Lecture - 1
Simple Linear Regression

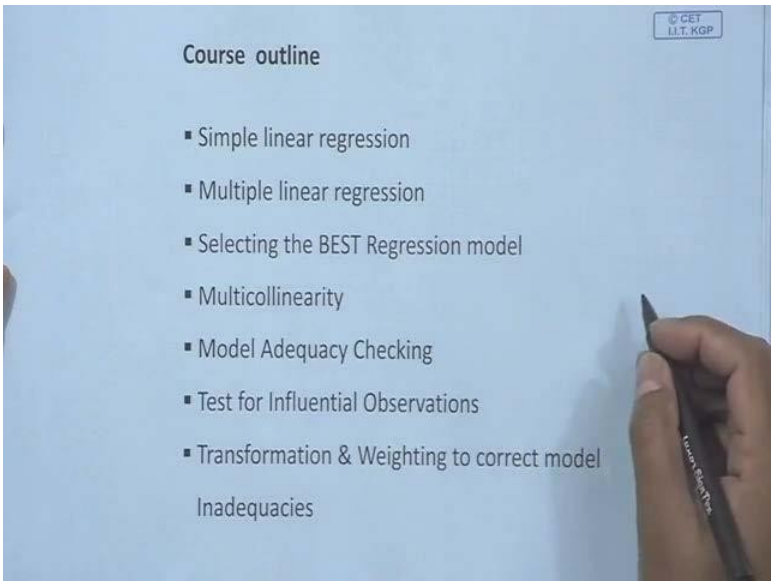
Hi, this is my 1st lecture on Regression Analysis, I would like to introduce myself as Dr. Soumen Maity, I did my B.Sc and M.Sc in statistics and **received a** Ph.D degree from Indian Statistical Institute, KolKata. Currently, I am faculty at Indian Institute of Technology, Kharagpur and Indian Institute of Science Education and Research, Pune. I am grateful to both the institutes for giving me this opportunity to work on NPTEL project.

(Refer Slide Time: 01:23)



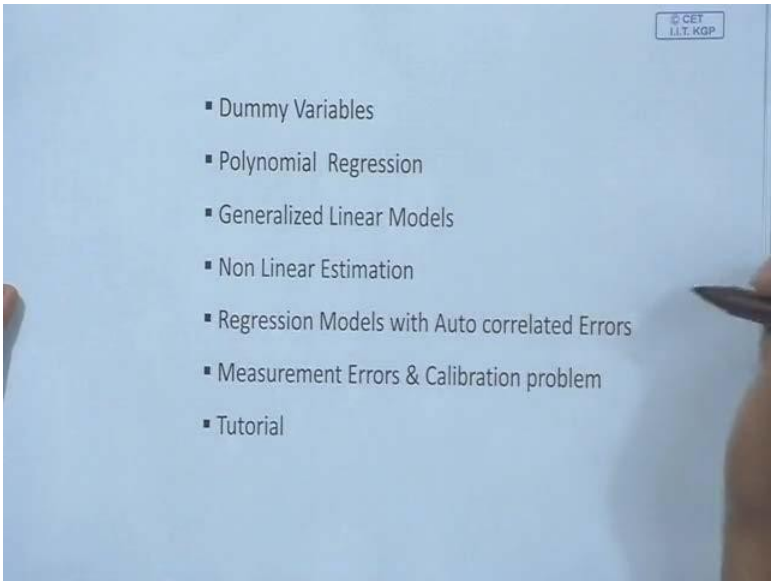
So, here is the course prerequisites, so I would expect the viewers, especially the students to know basics of probability and statistics, and statistical inference. So, more precisely now, I would like the viewer to know the discrete probability and also continuous **probability** distributions and say, point estimation, interval estimation and also testing of hypothesis. So, this course is divided into several topics or module, here are the topics.

(Refer Slide Time: 02:21)



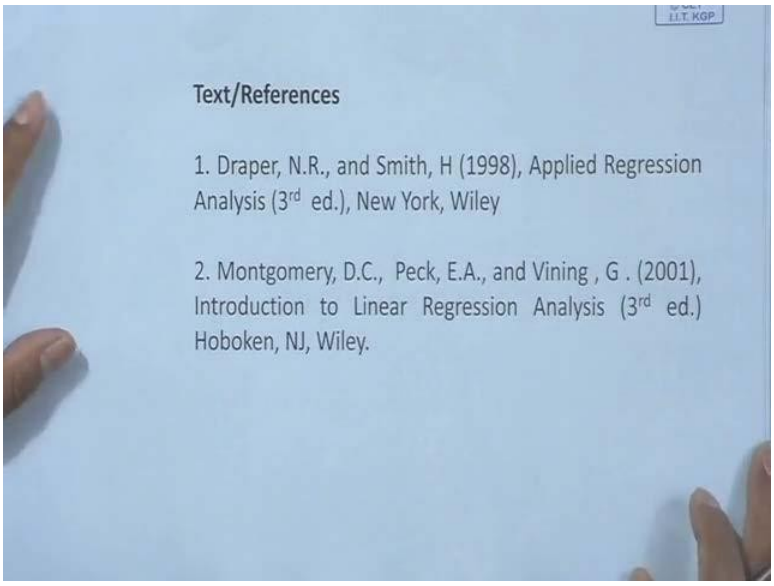
Simple linear regression, multiple linear regression, selecting the best regression model, multicollinearity, model adequacy checking, test for influential observations and then, transformation and weighting to correct model inadequacies.

(Refer Slide Time: 02:49)



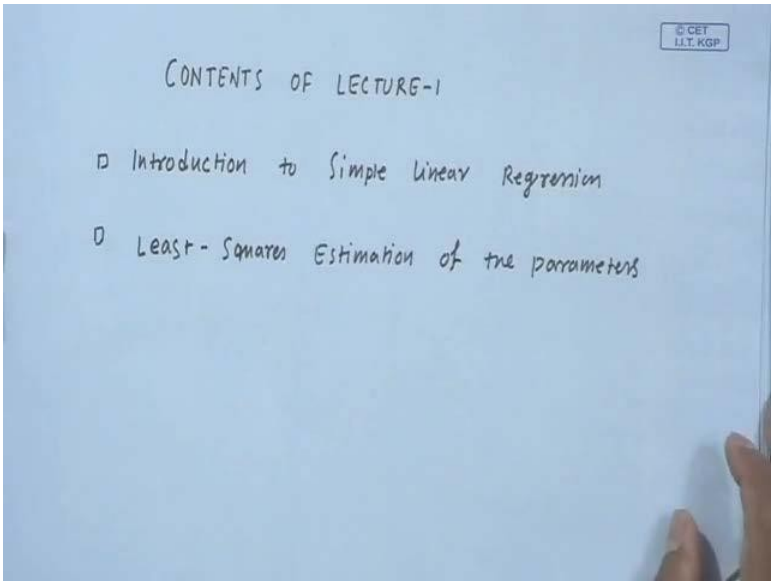
Dummy variables, polynomial regression, generalized linear models, non linear estimation and regression models with auto correlated errors, measurement errors and calibration problem. And finally, will be solving some problems, so I will have some sort of tutorial classes.

(Refer Slide Time: 03:26)



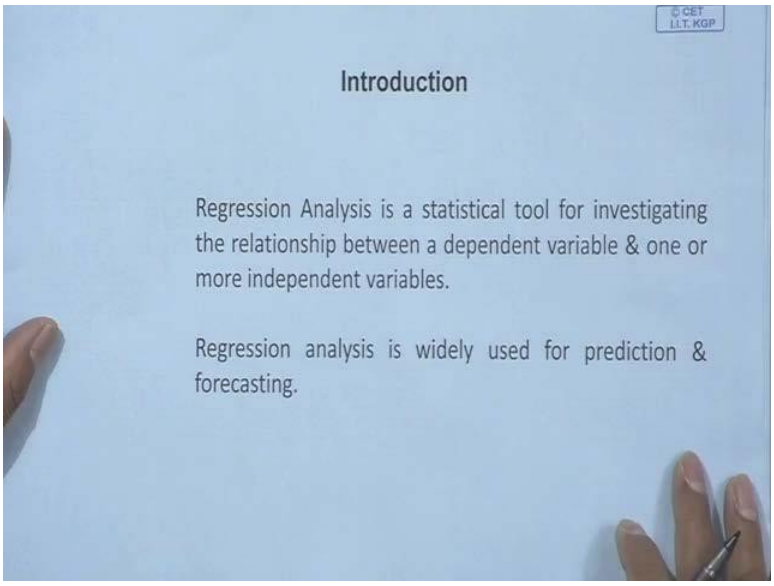
And I would be basically following these two books, the first one is applied regression analysis by Draper and Smith, and the second one is introduction to linear regression analysis by Montgomery, Peck and Vining.

(Refer Slide Time: 03:54)



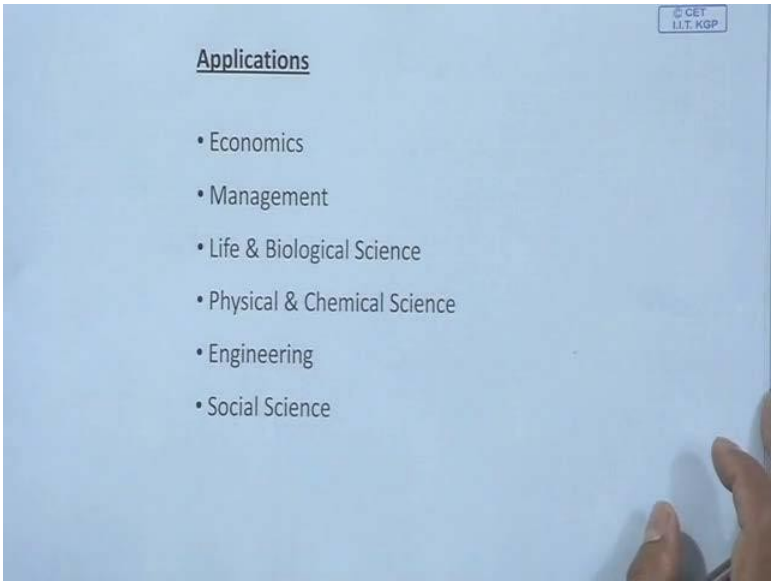
So, here is the content of today's lecture, so today basically I will introduce what is a regression analysis. And then, I will be talking about simple linear regression and least square estimation of the parameters, that means, the regression coefficients. So, let me talk about, what is regression analysis.

(Refer Slide Time: 04:29)



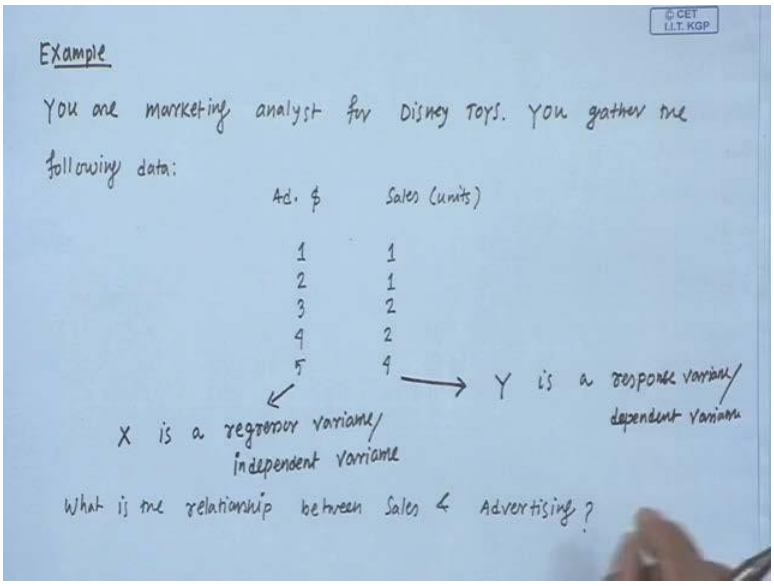
So, regression analysis is a statistical tool for investigating the relationship between a dependent variable and one or more independent variables. Just now, I will give an example to explain, what we mean by dependent variable and independent variables, and regression analysis is widely used for prediction and forecasting.

(Refer Slide Time: 05:16)



And it has application in different fields like economics, management, life and biological science, physical and chemical science, engineering and social sciences.

(Refer Slide Time: 05:43)



So, here is the example, I told that, I will give an example to explain, what I mean by independent and dependent variable. So, I said that, regression analysis is a statistical tool for investigating the relationship between one dependent variable and one or more independent variables. So, consider this example, suppose you are marketing analyst for Disney toys and you gather the following data, the first column is advertising cost and the second one is sales amount.

So, here what we want is that, what is the relationship between the sales and advertising cost. So, here you can see that, the amount of money that we want to spend, that is sort of controlled variable, you can decide how much amount of money you want to spent for advertising. But, the sales amount is not a controlled variable, you cannot control the sales amount, so the sales amount is a dependent variable, it depends on advertisement cost, this is one factor, but it depends on the amount of money spent on advertisement.

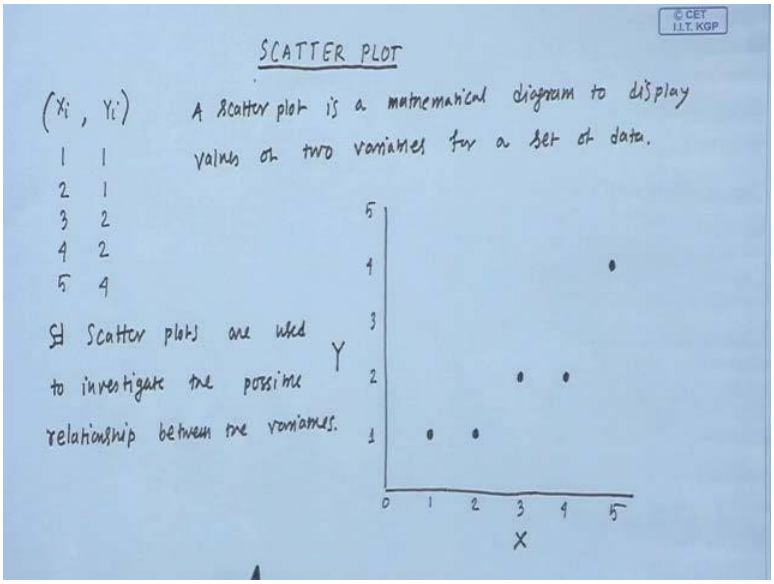
So, this is the dependent variable, but the advertising cost is an independent variable, also we call it controlled variable, you can control it. So usually, I hope that you understood the difference between the independent and dependent variable. So, usually this variable, which is independent, this is denoted by X and X is a regressor

variable or also we call it independent variable. Whereas the sales amount, we do not have any control on sales

amount and this type of variable is denoted by Y and Y is a response variable and also we call it dependent variable.

So, as I told, regression analysis is statistical tool for investigating the relationship between one dependent variable and one or more independent variables. So, here I mean, the whole objective of this course is to find the relationship between the variables, one response variable and several independent variables, let me talk about scatter plot next.

(Refer Slide Time: 10:14)



So, here these are same observation, you know, I have been given a set of observation X i Y i. So, X i stands for regression variable and Y i stands for response variable and i have five observations like X 1 Y 1, X 2 Y 2, X 3 Y 3, X 4 Y 4, X 5 Y 5 and the scatter plot is basically obtained by plotting this data on X Y plane. Formally we can see that, scatter plot is mathematical diagram to display values of two variables for a set of data. So now, I will explain scatter plot for this Disney toy data, so the first observation X 1 Y 1 is plotted here, the second observation is X 2 Y 2, so that is plotted here.

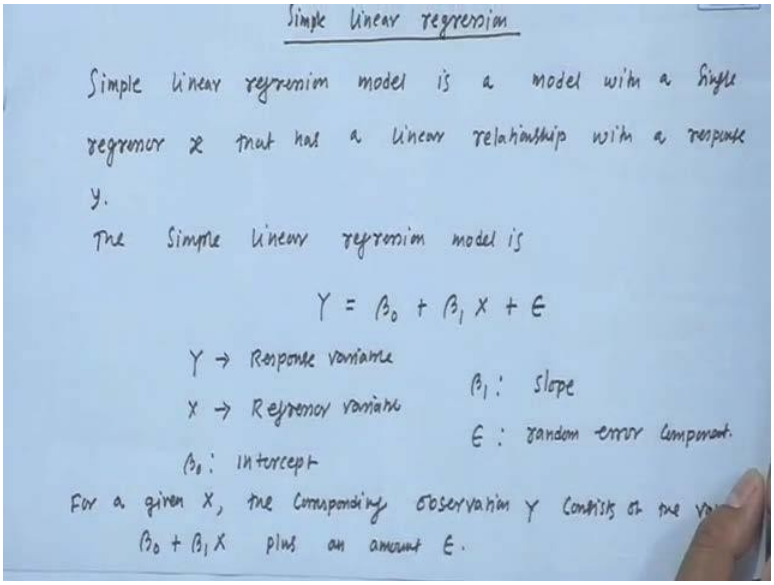
I should say that, usually in the regression analysis, the regression variable is plotted along the X axis and the response variable is plotted along the Y axis. So, these two points corresponds to this two data point and then, 3 2 is here and I have next 4 2 here and then, 5 4 here. So, this is the scatter plot corresponds to the data for Disney toy problem and this scatter plots are used to investigate the possible relationship between

two variables. So, this scatter plots are used to investigate the possible relationship between two variables.

Now, if the scatter plot indicates sort of linear relationship between the variables, so in that case, we need to go for linear model. But, if the scatter plot indicates sort of non linear relationship between X and Y then, we need to go for like, maybe quadratic fit or a cubic fit or the higher **order** polynomial fit. And looking at this scatter plot, I feel that, this scatter plot indicates sort of linear relationship between the response variable and the regression variable.

So, for this Disney toy data, we would go for linear model between X and Y, and the objective of this module is to study, how to fit linear relationship, more specifically simple linear regression for given a response variable and one regression variable.

(Refer Slide Time: 15:59)



So now, we will talk about simple linear regression, so simple linear model, regression model is a model with a single regressor X, that has a linear relationship with a response Y. So, the simple linear regression model is Y equal to beta naught plus beta 1 X plus epsilon, I will explain it. So, here you know that, Y is response

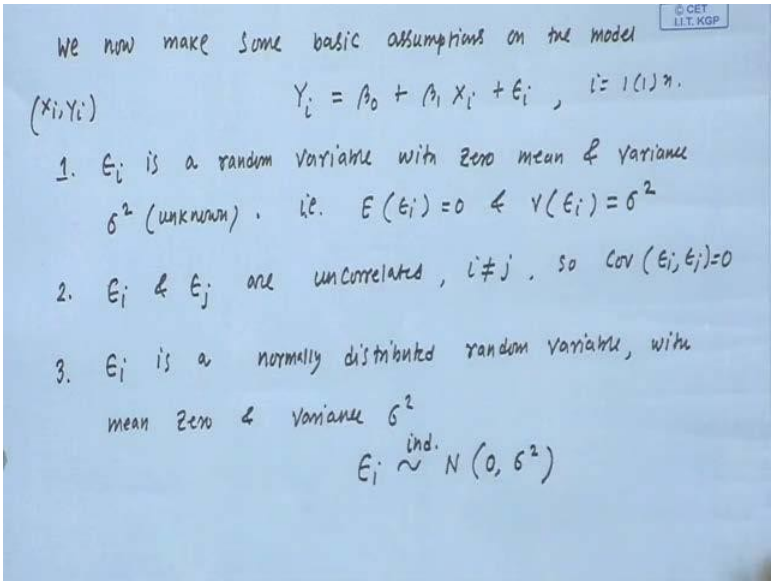
variable, X is regressor variable, beta naught is called intercept, beta 1 is called slope and epsilon is a random error component.

Before going into the detail of this one, I want to mention one more thing like, just recall the Disney toy example, there we have one variable that is, the advertising cost and X stands for advertising cost, the other one is sales amount. So, I told that, X is the controlled variable, so you can decide how much money you want to spend for advertising. So, X is not a random variable, whereas Y is dependent variable, you cannot control the sales amount, so Y is dependent variable, it depends on regressor variable and it cannot be controlled.

So, Y is a random variable and X is not a random variable, it is a controlled variable, you can say it is a deterministic variable or mathematical variable, but X is not a random variable, Y is a random variable. So, come back to this simple linear regression model, Y equal to beta naught plus beta 1 X plus epsilon, so what is the meaning of this one is that, for a given X that means, given advertising cost, the corresponding observation Y that means, corresponding sales amount, consist of the value beta naught plus beta 1 X plus an amount epsilon.

So, it says that, given the advertising cost, the corresponding sales amount consist of the value beta naught plus beta 1 X plus some error component I mean, the variable component. So, next we will make some basic assumptions on simple linear model.

(Refer Slide Time: 21:53)



We now make some basic assumption on the model, the model is Y i equal to beta naught plus beta 1 X i plus epsilon i, for i equal to 1 to n. So, before I wrote Y equal to

beta naught plus beta 1 X plus epsilon, now I am writing the same model for the i th observation. And here I said that, this is a random error components, so what we assume that, the first assumption is that, epsilon i is a random variable with zero mean and variance sigma square, which is unknown.

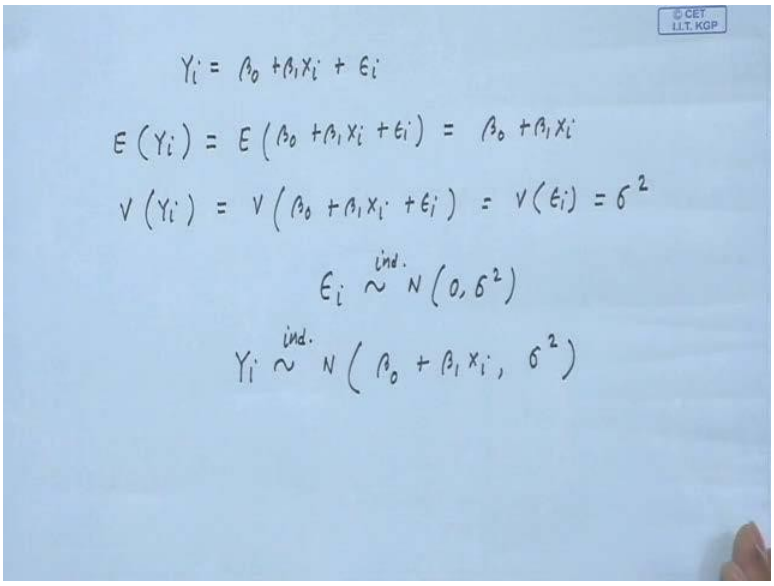
So, what you are given is that, you have just given a set of observation X i Y i, for i equal to 1 to no, that is all. And from the scatter plot, if you see that, the relationship is linear then, you are going to fit the simple linear regression model and you are making some assumption on the model. So, the epsilon i, the error term is a random variable with 0 mean and variable sigma square, which is unknown. So that means, expectation of epsilon i is equal to 0 and variance of epsilon i is equal to sigma square.

The second assumption is that, this is very important part, the second one is, the epsilon i and epsilon j are uncorrelated, i not equal to j, that means, so the covariance between epsilon i and epsilon j is equal to 0. The third one is that, epsilon i is a normally distributed random variable with a mean 0 and variance sigma square. That means, we are assuming that, epsilon i follows normal distribution with mean 0 and variance sigma square.

Now, what you can see that, this epsilon i's are uncorrelated and they are normally distributed. So, under this normality assumption, now this epsilon i's are not only uncorrelated, they independent also, so these are independent. So, what is the consequence I mean, of this one, in terms of the response variable Y i. So, what we are basically assuming is that, let me write down, see I said that, Y is a random variable and X is controlled variable, it is a deterministic variable, it is not a random variable.

So we made several assumption on epsilon i, now what is the consequence of this assumptions on Y, in terms of say, Y.

(Refer Slide Time: 26:57)



So, Y i is equal to beta naught plus beta 1 X i plus epsilon i, from here I can write, expectation of Y i is equal to expectation of beta naught plus beta 1 X i plus epsilon i and this is equal to beta naught plus beta 1 X i just, plus expectation of epsilon i, which is equal to 0. And what is the variance of Y i, variance of Y i is equal to variance of beta naught plus beta 1 X i plus epsilon i, which is equal to variance of epsilon i, because these are not random variable, so which is equal to sigma square.

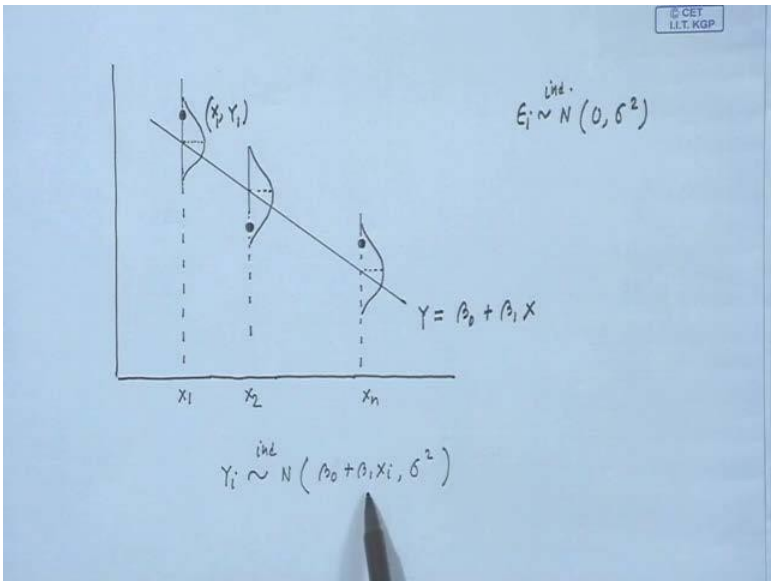
And also finally, we assume that, epsilon i follows normal distribution with mean 0 and variance sigma square and they are independent. And the consequence of this one in terms of the response variable Y is, so Y i follows normal distribution with mean beta naught plus beta 1 X i and variance sigma square and they are also independent. So, the assumptions on the error term like epsilon i having expectation 0, variance sigma square and follows, they are uncorrelated and epsilon i follows normal distribution.

So, finally, epsilon i is following normal distribution with mean 0, variance sigma square and they are independent. So, the consequence of that in terms of response variable is that, Y i follows normal distribution with mean beta naught plus beta 1 X i and variance constant variance sigma square. So, we are assuming that, the i th

observation is from normal distribution with mean beta naught plus beta 1 X i and the constant variance sigma square.

So, given a set of data, you need to be very careful about, whether your data set satisfy this basic assumption or not. But, if the dataset is not satisfying the basic assumptions then, you cannot go for the usual least square fit, I will be talking about those things may be in this class only. So, there will be topic called model adequacy checking, so that talks about, given a dataset, while fitting a simple linear regression model, how to check, whether the basic assumption are true or not, so we have to wait for that model adequacy checking topic.

(Refer Slide Time: 31:05)



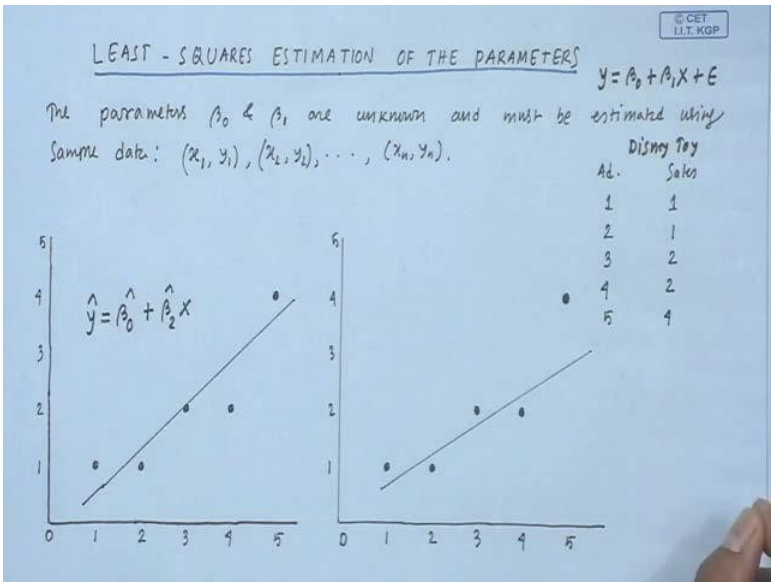
So, let me once again graphically say, how this situation I mean, how this assumption is illustrated in this figure. So, we made the assumption like epsilon i follows normal distribution with 0 mean and variance sigma square and they are independent and the consequence of that, in terms of response variable is that, Y i follows this. So, what you are assuming is that, this is my X 1 Y 1 data, this is X 2 Y 2, this is X n Y n and this line is Y equal to beta naught plus beta 1 X.

So, this you can put also i anyway, so the first situation like I mean, the assumption in terms of Y is graphically illustrate here. So, it says that, the i th observation or the i th response variable Y i, that is coming from a normal distribution with mean beta naught plus beta 1 X i and variance sigma square. So, this is the normal, so this is the

data X 1 Y 1, so Y 1 is from normal distribution with mean beta naught plus beta 1 X 1 and variance sigma square, so this is from this distribution, this is normal distribution.

And Y 2 is again from normal distribution with a different mean, with mean beta naught plus beta 1 X 2 and constant variance sigma square. It is necessary to understand this part or the basic assumption we made. So, assuming this means, you are assuming that, the response variable follows normal distribution and the i th observation is coming from the normal distribution with mean beta naught plus beta 1 X i and constant variance sigma square.

(Refer Slide Time: 34:00)



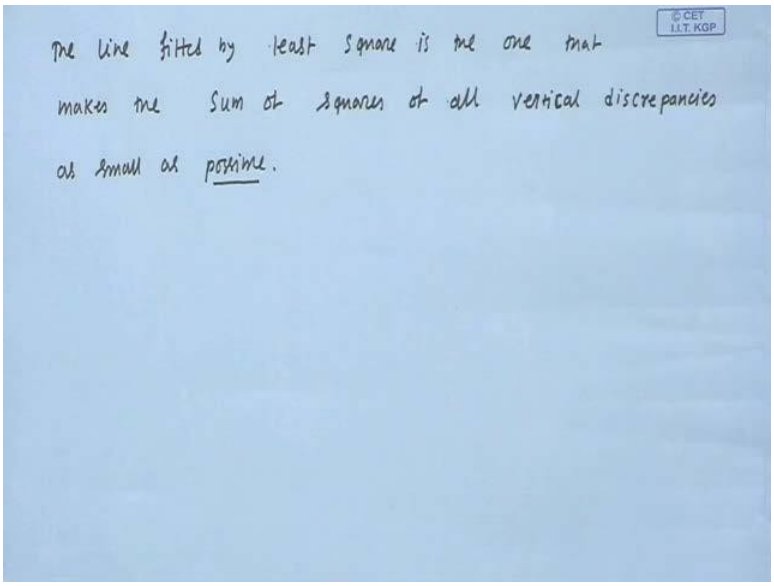
So, next we will move for least squares estimation of the parameters, so we talked about, we know what is a simple linear regression model, Y equal to beta naught plus beta 1 X plus epsilon. So, least squares estimation of the parameters means, estimating the regression coefficients beta naught and beta 1, so this is called intercept and this is called slope and fitting the simple linear regression is nothing but, estimating this regression coefficients. So, it says that, the parameter beta naught and beta 1 are unknown and must be estimated using the data.

So, what you are given is that, you are just given a set of observations and you have to fit, if the scatter plot indicates that, there is linear relationship, you can go for simple linear regression fit and also in the regression analysis, the starting point is generally fitting linear model. So, suppose this is the scatter plot for the Disney toy

data and we have to fit, so we have to estimate the regression coefficients that means, we have to fit straight line for the given data.

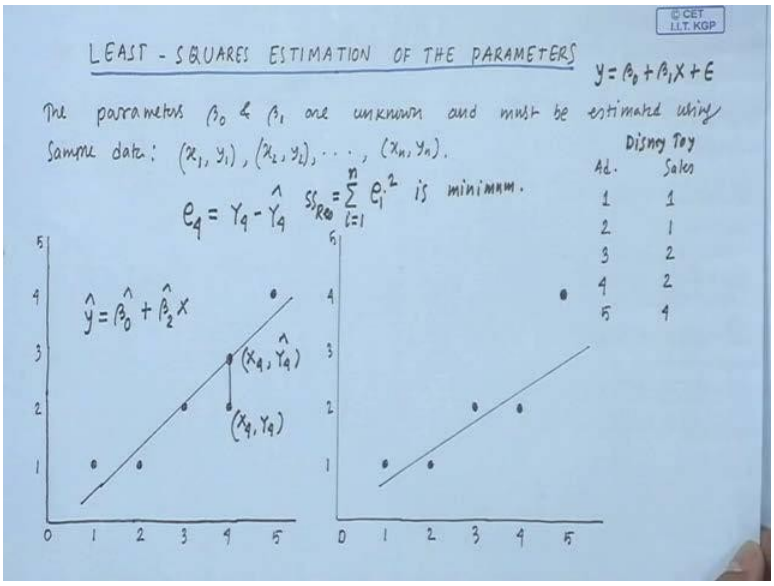
Suppose, the fitted model is Y hat, which is equal to beta naught hat plus beta 2 hat into X, so this is the fitted line. And you can see that, I have drawn two lines, the same scatter plot, this is one straight line, say suppose this is my fitted model for this scatter plot or for this data and this is another fit. Now, which one is better, whether this one better or this one is better. So, I will come back to this slide again, let me write one important thing.

(Refer Slide Time: 37:28)



The line fitted by least square is the one, that a makes the sum of squares of all vertical discrepancies as small as possible. So, this is the main idea behind the least square fit, the line fitted by least square technique is the one, that makes the sum of square of all vertical discrepancies as small as possible.

(Refer Slide Time: 38:52)



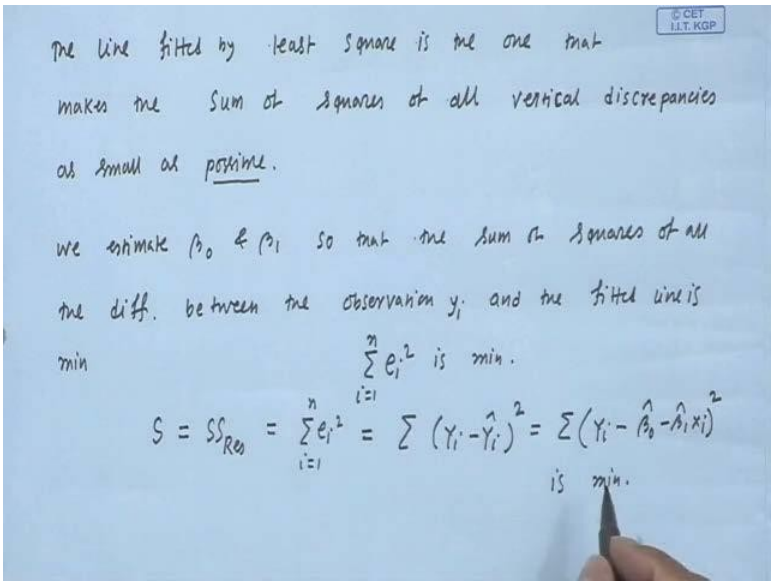
So, what is the meaning of that, so what the least square technique does is that, it fits a line such that, what I mean by this vertical discrepancy, this is the vertical discrepancy for the fourth observation. So, for the fourth observation, basically this is (X 4 Y 4), X 4 Y 4 is equal to 4, 2. And suppose, this is a fit, this is the fitted line Y hat is equal to beta naught hat plus beta 1 hat X and then, this point is nothing but, X 4 Y 4.

The vertical discrepancy is nothing but, let me write that as e 4, that is called a residual for the fourth observation. So, e 4 is equal to this distance that is, Y 4 minus Y 4 hat, so this is what we mean by vertical discrepancy and what the least square estimation or least square technique does is that, it fits a model such that, this e i square for i equal to 1 to n in general, but here it is 1 to 5, this is minimum. So, in order to say which fit is good, whether this is good or this one is good, so what you do is that, you compute this e i square, this is called residual sum of square, this is SS residual.

You compute SS residual for this fit, you compute SS residual for this fit and you see, which one is smaller, that one is better than the other one. And what the least squares estimation does is that, it provides a fit, which has minimum SS residual. So,

I hope that, you understood the basic and very nice and natural idea behind the least square estimation.

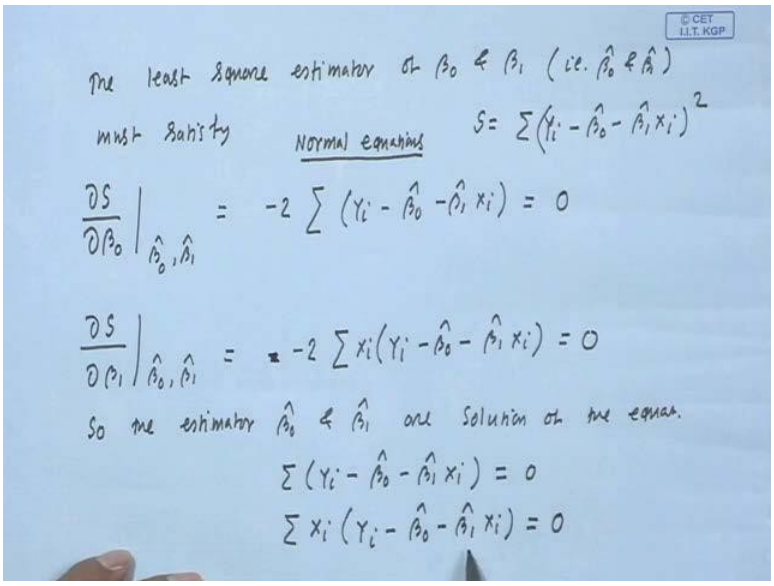
(Refer Slide Time: 41:45)



So, we estimate beta naught and beta 1 so that, the sum of square of all the differences between the observation Y i and the fitted line is minimum. So, the minimum of this one is that, compute all the residuals e 1, e 2, e 3, e n and then, this beta naught and beta 1 are estimated so that, this summation e i square, i equal to 1 to n is minimum. I will write this, so estimate beta naught and beta 1 so that, the sum of square of all the difference between the observation Y i and the fitted line is minimum.

That means S, which is nothing but, SS residual, sum of square residual, which is equal to e i square, i equal to 1 to n, which is nothing but, Y i minus Y i hat square, which is nothing but, Y i minus beta naught hat minus beta 1 hat X i square is minimum. So, you have to estimate, you have to find this beta naught hat beta 1 hat, which is beta naught hat is the estimate of beta naught and beta 1 hat is the estimate of beta 1 such that, this is minimum.

(Refer Slide Time: 44:25)



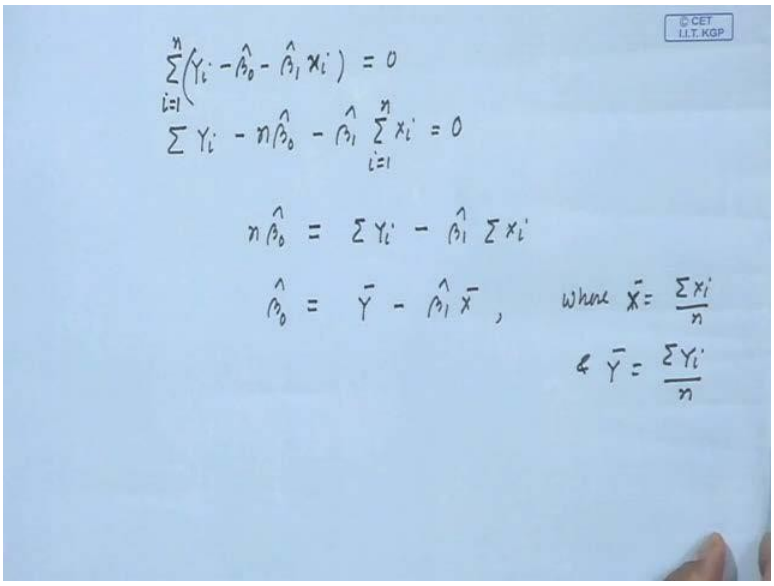
So, the least square estimator of beta naught and beta 1 that is, a beta naught hat and beta 1 hat, they must satisfy the following two equations, you differentiate S with respect to beta naught and at the point beta naught hat beta 1 hat. So, let me just write down what is S, S is equal to summation Y i minus beta naught hat minus beta 1 hat X i square, so this is what the S is, so you find beta naught and beta 1 such that, this is minimum. So, this one is equal to partial derivative of this one with respect to beta naught is minus 2 Y i minus beta naught hat minus beta 1 hat X i equal to 0, so this is one equation.

The other one is partial derivative of S with respect to beta 1 at the point beta naught hat beta 1 hat. So now, we are differentiating with respect to beta 1, so that is equal to minus 2 summation Y i minus beta naught hat beta 1 hat X i into X i, so this is equal to 0. So, these two equations are called normal equations, since there are two unknown parameter, you will get two normal equations and you can see that, this normal equations are independent.

So, you can uniquely fit beta naught and beta 1, so the estimator beta naught hat and beta 1 hat are solution of the equations, summation Y i minus beta naught hat minus beta 1 hat X i equal to 0 and X i into Y i minus beta naught hat minus beta 1 hat X i

equal to 0. So, you have two independent normal equations and from here, you can estimate beta naught hat and beta 1 hat, so you will be doing that, let me start with this one.

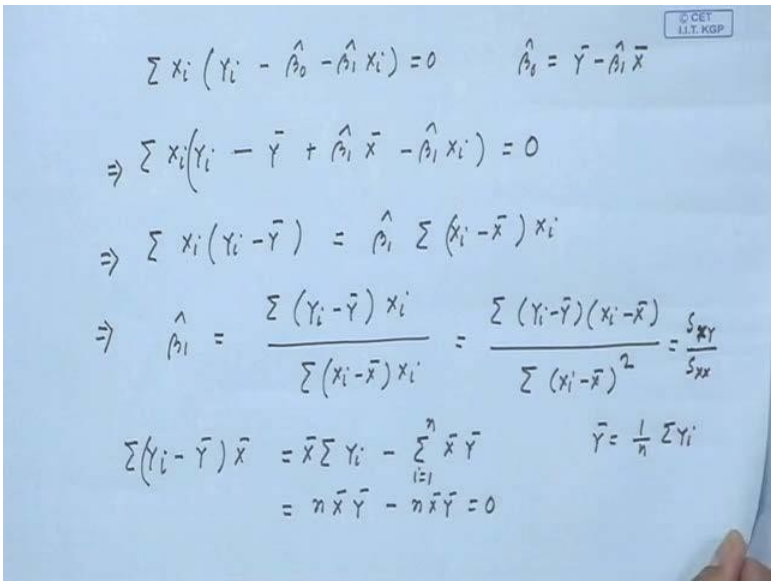
(Refer Slide Time: 48:24)



So, what the first equation is, summation Y i minus beta naught hat minus beta 1 hat X i equal to 0. So, from here, I can write that, summation Y i minus n beta naught hat, because this sum is over from 1 to n, minus beta 1 hat sum X i, i is from 1 to n, this is equal to 0. So then, n beta naught hat is equal to summation over Y i minus beta 1 hat summation X i.

And from here, I can write that, beta naught hat equal to Y bar minus beta 1 hat X bar, of course where Y bar is equal to summation X i, X bar is equal to summation X i by n and Y bar is equal to summation Y i by n. So, this involves beta 1 hat, so you need to estimate beta 1 hat also I mean, you need to find beta 1 hat also.

(Refer Slide Time: 50:23)

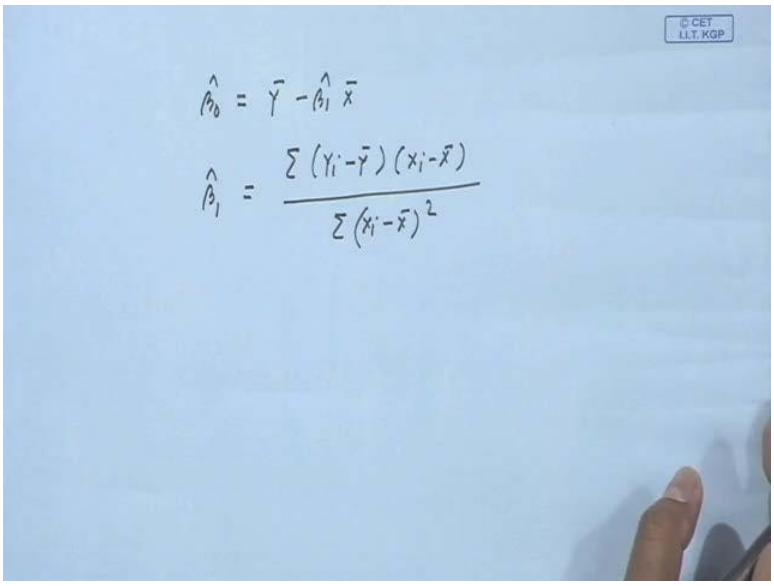


So, let me start with the second normal equation, that was summation X i into Y i minus beta naught hat minus beta 1 hat X i equal to 0 and just now what we obtained is that, beta naught hat is equal to Y bar minus beta 1 hat X bar. So, I can plug this one here, so what I will get a that, X i of Y i, let me write one more line, minus Y bar plus beta 1 hat X bar minus beta 1 hat X i is equal to 0. So, from here, I can write that, X i into Y i minus Y bar is equal to beta 1 hat sum over X i minus X bar, I hope you understand this one.

So, from here, I can write that, my beta 1 hat is equal to sum over Y i minus Y bar into X i by, I missed one X i here, by sum over X i minus X bar into X i. This can be written as sum over Y i minus Y bar into X i minus X bar by summation X i minus X bar into X i minus X bar, so this is X i minus X bar square. So, what I have added is that, I have added a term here, I can prove that, see because of the fact that I can prove that, X i Y bar into X bar is 0. So, let me just prove that, this is equal to sum over Y i X bar minus summation X bar Y bar and if I write Y bar is equal to 1 by n summation Y i.

So, summation Y i, I can write as n into Y bar that is, n into X bar into Y bar minus, this sum is from 1 to n and it is independent of i, so n into X bar and Y bar. So, this is 0 and also you will use annotation that, this is equal to S X Y by S X X.

(Refer Slide Time: 53:44)



So, what we got is that, finally we got that, beta naught hat is equal to Y bar minus beta 1 hat X bar and also we got that, beta 1 hat is equal to summation Y i minus Y bar into X i minus X bar by summation X i minus X bar whole square. So, we have learned, how to fit a simple linear regression model, given a set of observations X i Y i, for i equal to 1. We know how to fit simple linear regression model like, Y i is equal to beta naught plus beta 1 X i plus epsilon and here are the least square estimators, beta naught hat and beta 1 hat.

And in the next class, we will be talking about several properties of this least square estimators and it can be proved that, these are the best linear unbiased estimators using Gauss Markov theorem.

Thank you.