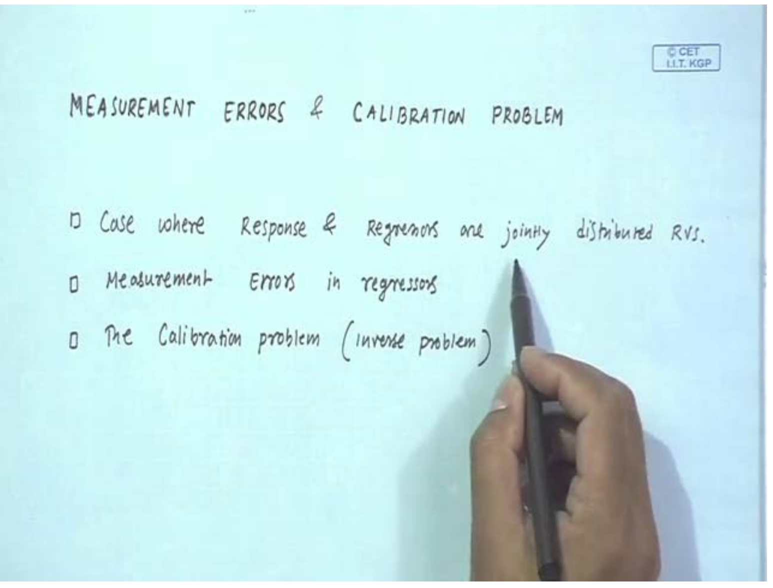
Regression Analysis
Prof. Soumen Maity
Department of Mathematics
Indian Institute of Technology, Kharagpur

Lecture - 35
Measurement Errors and Calibration Problem

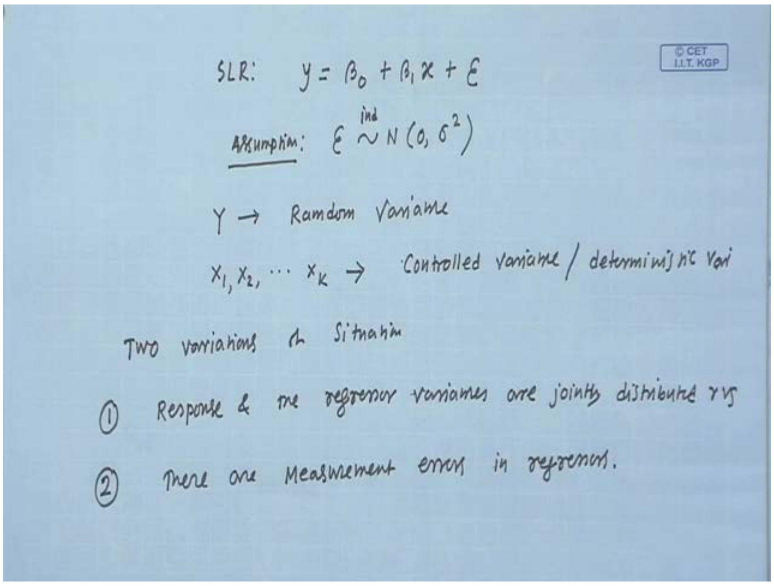
Hi, so today we will be talking on a new topic called measurement errors and calibration problem and here is the content of this topic.

(Refer Slide Time: 00:35)



So, case where the response and regressors are jointly distributed random variable and measurement error in regressors and also we will be talking about the calibration problem which is called inverse problem. Let me talk about the objective of this topic, in almost all regression model we assume that the response variable is a random variable, where the regressor variable like x 1 x 2 x k, they are called controlled variables, they are not random variable. So, what we will do in this topic is that we will talk about two variations of this situation and has I told you know that y is a random variable and x is controlled variable, which is not a random variable. Let me just recall why I say y is a random variable.

(Refer Slide Time: 02:08)

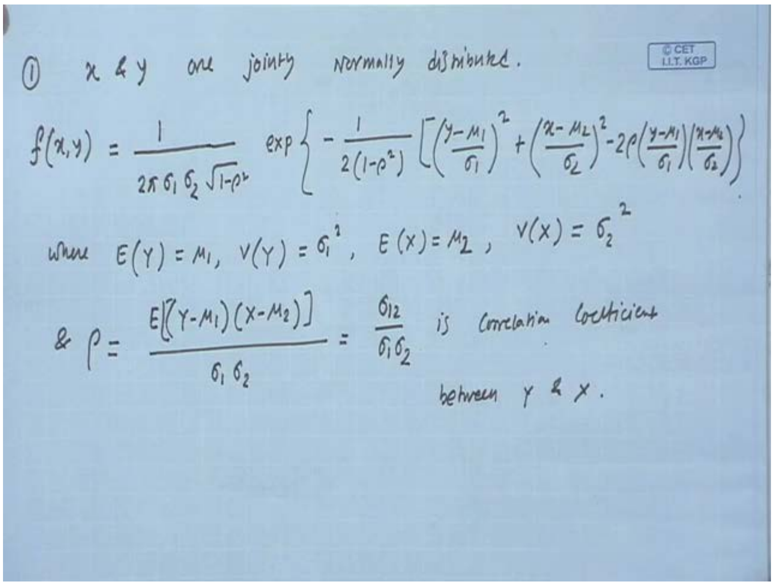


In simple linear regression model, what we consider the model is like y equal to beta naught plus beta 1 x plus epsilon. And, we assume that epsilon is random variable which follows normal distribution with mean 0 and variance sigma square. So, this is the assumption we make in simple linear regression and also in a multiple linear regression model.

Assuming, that sigma is random variable which follows normal distribution and they are independent, if you put i here, is same as assuming that y is a random variable. So, what we assume is that the response variable y is a random variable and the regressors like x 1 x 2 x k, they are not random variable, they are called controlled variable or also we call deterministic variable.

So, here we will be talking about two variation of this situation. The first one is, both response and the regressor variable are jointly distributed random variables. And the second case we will be considering the second variation is, there are measurement error in regressors. So, let me talk about the first variation where both the response and the regressor variables are jointly distributed random variable.

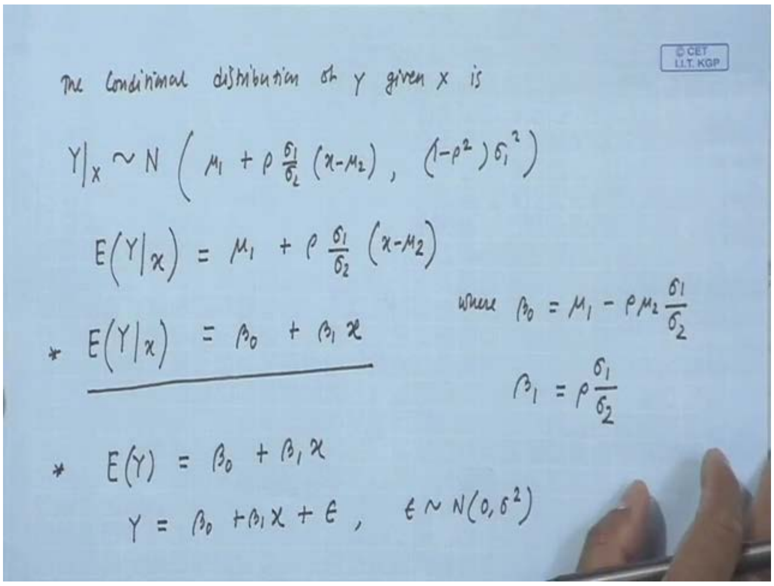
(Refer Slide Time: 05:21)



Let me assume that the first case that x and y, is the regressor and the response variable are jointly normally distributed. Usually, x is not a random variable. But, here we are considering both of them are random variable and they are jointly normal distributed. Then the joint p d f, probability density function of x and y is 1 by 2 pie sigma 1 sigma 2 1 minus rho square exponential 2 into 1 minus rho square y minus mu 1 by sigma 1 square plus x minus mu 2 by sigma 2 square minus twice rho y minus mu 1 by sigma 1 into x minus mu 2 by sigma 2.

So, this is the joint p d f of x and y when they are jointly normally distributed. Here, expectation of y is equal to mu 1, variance of y is equal to sigma 1 square, expectation of x is equal to mu 2 and variance of x is equal to sigma 2 square. And there is one more parameter that is called correlation coefficient between x and y, so the rho is the correlation coefficient which is equal to the covariance between x and y, so y minus mu 1 into x minus mu 2 by sigma 1 sigma 2. So, notation for this one is sigma 1 2 by sigma 1 sigma 2, this is the correlation coefficient between y and x. And once you have the joint distribution we know we can find the conditional distribution also.

(Refer Slide Time: 08:50)

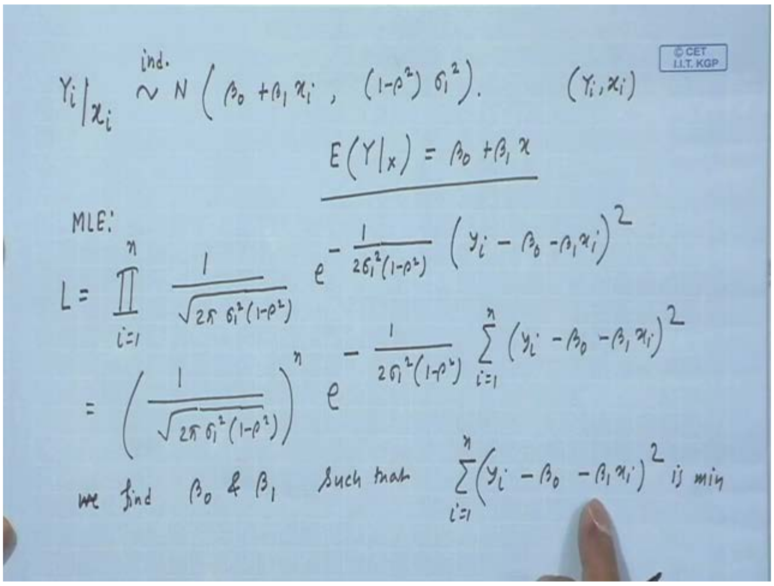


Let, me write down the conditional distribution of y given x follows normal distribution with mean mu 1 plus rho sigma 1 sigma 2 x minus mu 2 and variance 1 minus rho square sigma 1 square. So, this is the conditional distribution of x of y given x, which is normally distributed with some mean and variance. So, what we can write here is that the expectation of y given x is equal to mu 1 plus rho sigma 1 by sigma 2 x minus mu 2.And, this I can write as beta naught plus beta 1 x, where my beta naught is equal to mu 1 minus rho mu 2 sigma 1 by sigma 2 and beta 1 is equal to rho sigma 1 by sigma 2.

So, why I derived all these things because just to say that here you can see the conditional expectation of y or the expectation of y given x is this one. So, this is the model we need to consider when both x and y are random variable. And, what we do when x is not a random variable, when x is a controlled variable or deterministic variable, what we do is that we fit the model expectation of y which is equal to beta naught plus beta 1 x. This is the model we fit in the case when y is a random variable and x is a deterministic variable, it is not a random variable.

So, writing this is same as y equal to beta naught plus beta 1 x plus epsilon, where epsilon follows normal 0 sigma square. Similarly, this is the difference only, this is the model we need to consider here whereas when both at x and y are random variable. And this is the model we consider when y is a random variable but, x is not a random variable. That is, in almost all cases this is the situation that y is the random variable but, x is not random variable, it is a controlled variable. Now, how to fit this model because we are talking about the case when both x and y are random variable, so from here we know the conditional distribution of y given x.

(Refer Slide Time: 13:02)



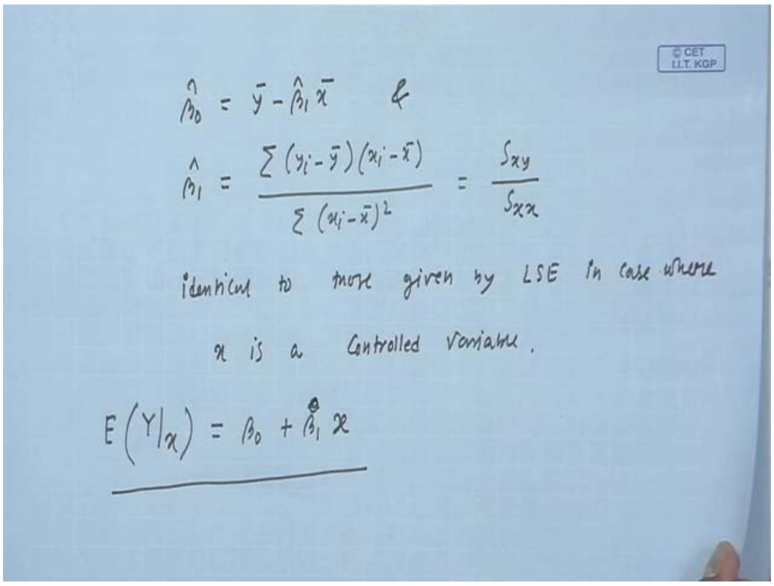
From, here we can say that this y i given x i this follows normal distribution with mean beta naught plus beta 1 x i and the variance is 1 minus rho square sigma 1 square and they are independent. So, I am given the observations y i x i both are random variable and I know that this is true, that is the random variable y given x is or random variable y i given x i are independent random variable. And they follow normal distribution with mean beta naught plus beta 1 x i and variance, a constant variance.

Now, to estimate this parameter and do we want to fit the model, y given x is equal to beta naught plus beta 1 x. Fitting this model means we need to estimate the coefficients beta naught and beta 1. So, what we will do is that we will go for maximum likelihood estimator, because here we know the distribution of conditional distribution of y given x i.

So, the likelihood function is, hence they are independent the likelihood function of y 1 y 2 y n given x 1 x 2 x n is just the product of marginal density, so that is nothing but, 1 by root over of 2 pie sigma 1 square 1 minus rho square and e to the power of minus 1 by 2 sigma 1 square 1 minus rho square y i minus beta naught minus beta 1 x i square, i equal to 1 to n. This is the likelihood function and this can be written as 1 by root over of 2 pie sigma 1 square 1 minus rho square to the power of n e to the power of minus 1 by 2 sigma 1 square 1 minus rho square summation y i minus beta naught minus beta 1 x i whole square.

So, what the maximum likelihood estimate technique suggest is that, you construct the likelihood function and here you can go for and find the parameter beta naught and beta 1 in such a way that the likelihood function is maximum. Maximizing this likelihood is same as minimizing this thing, so we find beta naught and beta 1 such that this is minimum. Find beta naught and beta 1 such that y i minus beta naught minus beta 1 x i whole square, i equal to 1 to n, this is minimum. This is nothing but, the least squared function we consider while estimating beta naught and beta 1 using the least squared technique in simple linear regression model. So, we know what is beta naught hat and beta 1 hat.

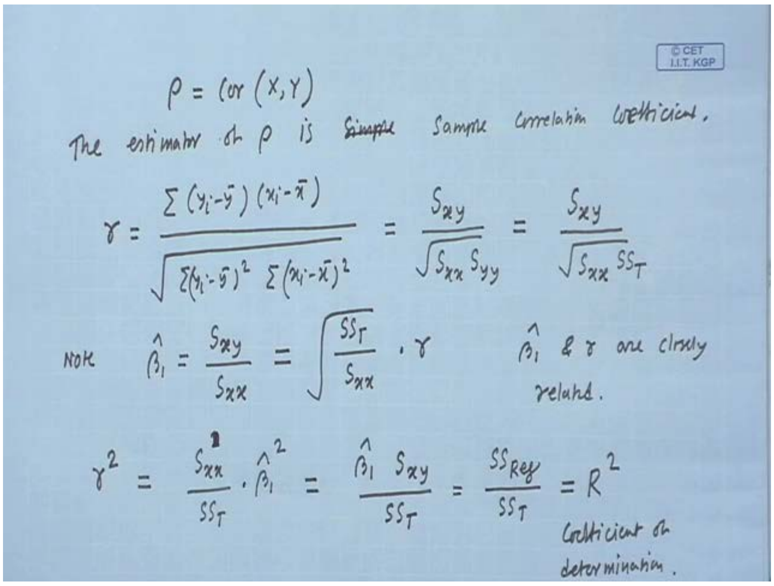
(Refer Slide Time: 18:00)



Beta naught hat is equal to y bar minus beta 1 hat x bar and beta 1 hat is equal to summation y i minus y bar x i minus x bar summation x i minus x bar whole square. This is nothing but, S x y by S x x. So, these are identical to those given by least square estimate, because we are minimizing the same function here we are minimizing the least square function here also, in case where x is a controlled variable.

So, here we are trying to estimate the model expectation of y given x which is equal to beta naught plus beta 1 hat, sorry beta naught plus beta 1 x. And we found that the estimate, the maximum likelihood estimate for beta naught and beta 1 are the same as obtained by least squared technique in case when x is a controlled variable.

(Refer Slide Time: 20:04)



Here, we have new parameter called the correlation coefficient rho. This is the correlation coefficient between x and y. What you want is that, we want to draw some inference about this correlation coefficient. First, we find an estimator for this one. The estimator of row which is a correlation coefficient, is the sample correlation coefficient that is equal to r.

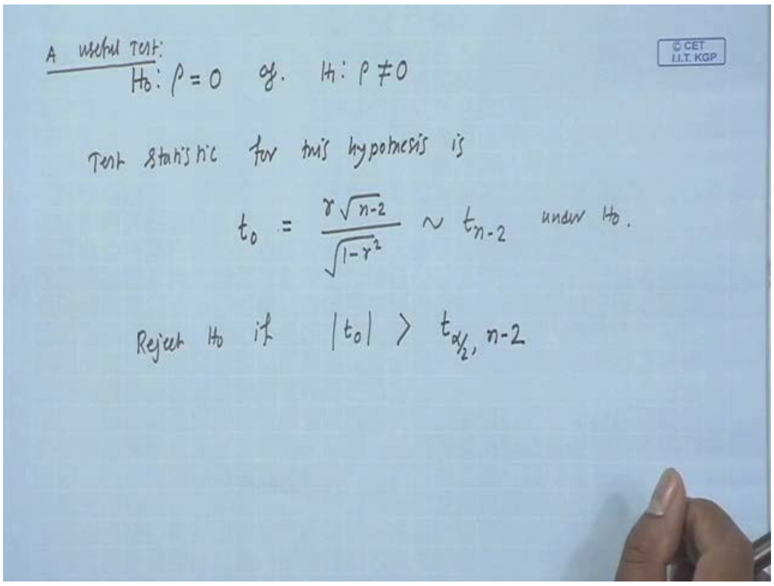
So, r is the sample correlation coefficient which is y i minus y bar into x i minus x bar by square root of summation y i minus y bar whole square into x i minus x bar whole square. This is the sample correlation coefficient and this is the estimator for population correlation coefficient rho.

And this can be written as, you need a standard notation s x y by square root of S x x S y y. So, S y y is this one and we know that this one is also called S S square, sorry S S total. So, I will write this as S x y by square root of S x x and then S S total. Now, note that whether there is a relation between this sample correlation coefficient, which is an estimator for rho and regression coefficient beta 1. So, we know that beta 1 hat is equal to S x y by S x x which can be written in terms of r. This can be written as S S total by S x x square root of this into r, you can check that, just plug r here you will get back this one. So, this says that beta 1 hat and r are closely related.

Also, what we will do is, we will see what is r square here. So, from this expression r square is equal to S x x by S S total into beta 1 hat square. And this one can be written as, we know that beta 1 hat is S x y by S x x. This can be written as, beta 1 hat into S x y by S S total. You can check that, you know just take out 1 beta 1 hat and plug this value here. We know that this one is nothing but, S S regression. This is interesting so, S S regression by S S total. And you know what is this quantity this is called the coefficient of determination, that is a capital r square. This is called coefficient of determination.

So, what does this r square do is that it measures the proportion of variability in the response variable that is explained by the regression model. And what does r do is that it measures the linear association between x and y, here we observe that you know r square is equal to the small r square which is the sample correlation coefficient, is equal to the capital r square which is the coefficient of determination. So, what you do is that we have a new parameter rho which is the correlation coefficient between x and y. And we learned how to test the significance of this, I mean whether this correlation coefficient is significant or not by testing the hypothesis.

(Refer Slide Time: 26:53)

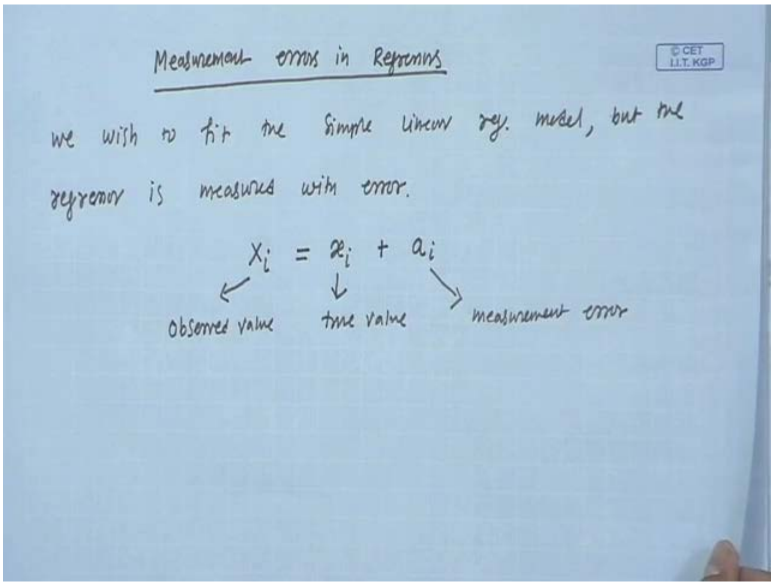


That h naught is rho equal to 0 against h 1 that rho is not equal to 0. So, this is a useful test. And the tester statistic for this hypothesis is t naught which is equal to r root of r of n minus 2 by 1 minus r square, square root of this thing. And this follows t n minus 2 degree of freedom under h naught.

Here, is the rejection criteria. We reject h naught if the modulus value of this one is greater than t alpha by 2 n minus 2. So, here we learned about how to test the correlation coefficient is equal to 0 against the alternative hypothesis that the correlation coefficient is not equal to 0. Well so, we talked about one case here where both the regressor variable x and the response variable y, both of them are random variables. And we observed that the linear model we need to fit here is very similar to the case when y is a random variable and x is a controlled variable, that is the situation in almost all cases.

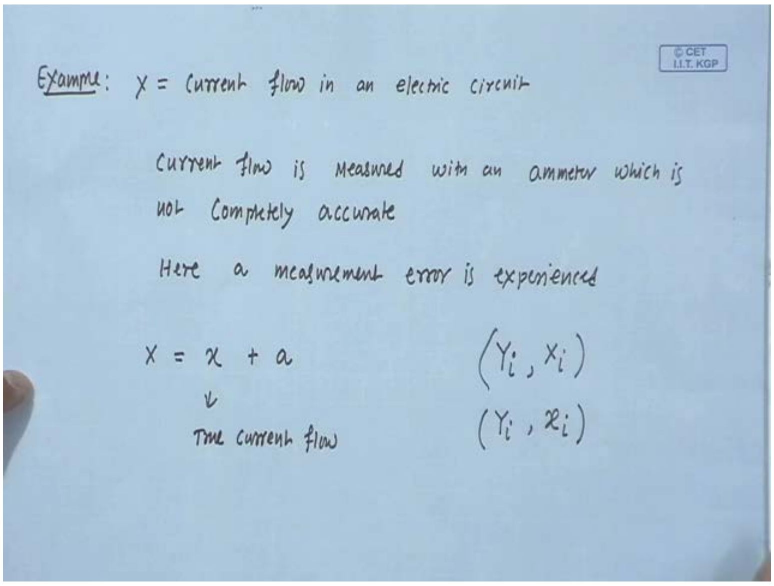
Here, we assume that the random variable x and y they jointly follow normal distribution by variant normal distribution. And we have a new parameter called rho here, which is the correlation coefficient between this two random variable and if you see we learned how to test the hypothesis that rho is equal to 0 against rho is not equal to 0. If you see from this testing you know rho is equal to 0, that means there is no linear relationship between the regressor variable and response variable, which is same as testing the beta 1 is equal to 0 against beta 1 is not equal to 0. Also, we say that if beta naught sorry if the beta 1 is equal to 0, then there is no relationship between x and y.

(Refer Slide Time: 30:45)



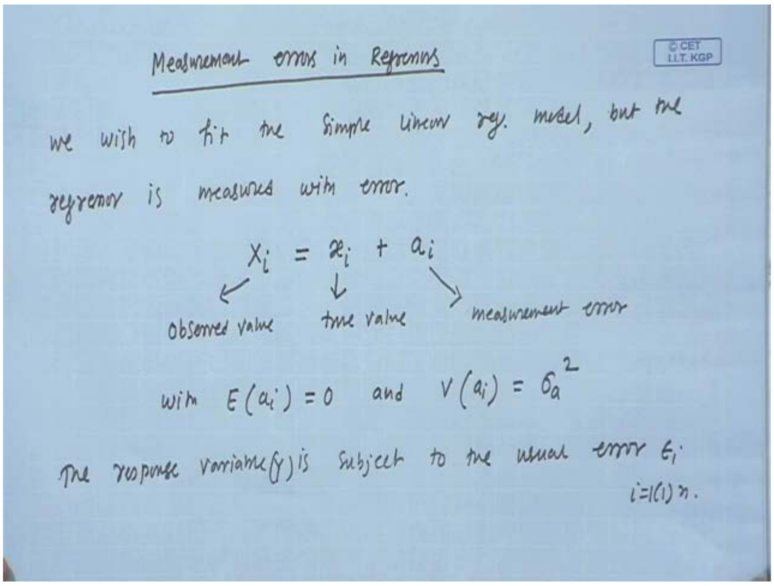
Next, we will be talking about another deviation that is measurement errors in regressors. Here, we wish to fit the simple linear regression model but, the problem here is that, the regressor is measured with error. So, what I mean by this, here suppose x i is the observed value of the regressor, so this is the observed value. In usual case you know we consider that y is sorry x is controlled variable and then there is no error while measuring the value of x I, which is equal to small x i but, here the regressor is measured with error, so x i is equal to small x i plus a i. So, what is this small x i, small x i is the true value and this a i is called the measurement error. I will give an example to illustrate this situation.

(Refer Slide Time: 33:13)



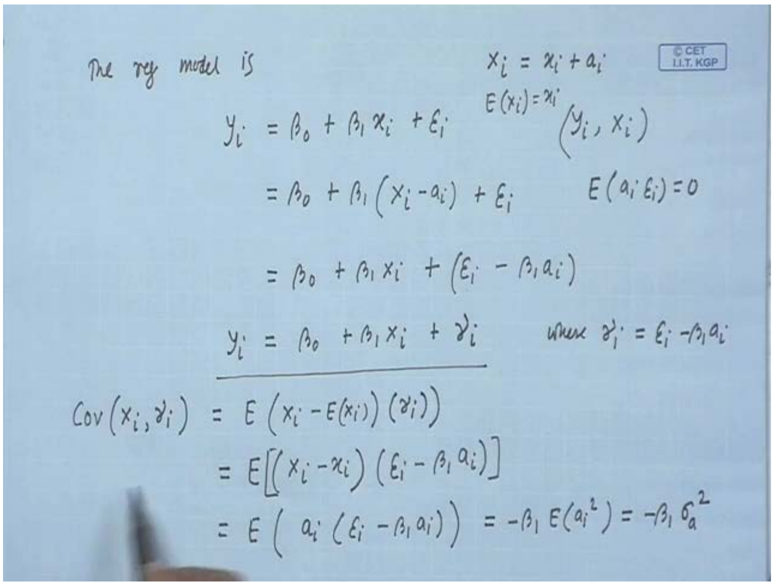
Let, me consider this example. Suppose x is a regressor variable which stands for the current flow in an electric circuit. And the current flow is measured with an ammeter, which is not completely accurate. So, here a measurement error is experienced. So this is the observed current flow capital x and the small x is the true current flow and this one is the measurement error. What we are given is that we are given the observation say y i that is a response variable and we are given x i capital x i we are given the observed value, but we want to find linear relationship between y i and the true value of the regressor variable. So, we will be talking about how to deal with such situation. Now, let me go to my previous slide.

(Refer Slide Time: 34:55)



Here, is the observed value you understood this is the observed current flow, this is the true current flow and here is the measurement error. And here we make the assumption that, with this expected value of measurement error is equal to 0 and variance of a i the constant variance sigma a square. Of course, the response variable is a random variable. The response variable is subject to the usual error epsilon i for i equal to 1 to n.

(Refer Slide Time: 36:17)



Now, what we want is that we want to find relationship between the response variable y and the true value of x. So, we want to fit we want to consider the model. The regression model is y i is equal to beta naught plus beta 1 x i plus epsilon. So, here only the problem is that we do not have small x I, this is the true value. But, what we have is that we are given y i the response variable and the capital x i that is the measured value of the regressor variable.

This can be written as beta naught plus beta 1 in terms of capital x i. I can write this as, x i minus a i, because of the fact that we assume that the measured value is equal to the true value plus the measurement error plus epsilon, because we are given y i and capital x i so we have to convert this model in terms of capital x i, that is quite clear.

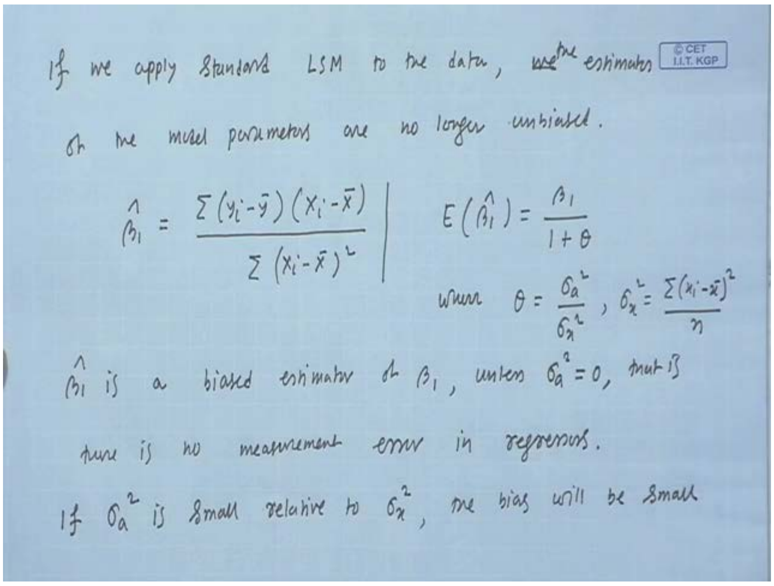
So, this is equal to beta naught plus beta 1 x i plus epsilon i minus beta 1 a i. This is equal to beta naught plus beta 1 capital x i. I should write, I mean I should not mix here capital x i and small x I, I used to do that before, because both of them are same but, here I have to be careful. So, this is equal to gamma i, where this gamma i is equal to epsilon i minus beta 1 a i. Now, this appears to be you know now we have the model in terms of capital x i, y i is equal to beta naught plus beta 1 x i plus some error term.

You may think that we are done, because this is the model we know how to fit this model before also but, the problem here is that see this capital x i is random variable and this gamma i is also random variable, now we need to check whether before when x was controlled variable there was no correlation between this two right. Now, here we need to check whether they are correlated or they are independent. So, what we will do is that we will compute the covariance of x i and gamma i which is nothing but expectation of x i minus e x i into gamma i. Because expectation of gamma is of course, equal to 0 because both expectation of epsilon i is equal to the expectation of a i that is 0.

So, here I can write that this is equal to capital x i, now the expected value of capital x i is equal to small x i, that is small x i and this one is equal to epsilon i minus beta 1 a i. Now, this one is equal to expectation of what is this capital x i minus small x i is equal to a i, so a i into epsilon i minus beta 1 a i. And we assume that, you know perhaps I forgot to mention this that here while I was talking about this model, here we assume that expect this two are independent, a i and epsilon i they are independent this is equal to 0.

Then this one is equal to minus beta 1 expectation of a i square. So, expectation of a i square is nothing but, the variance of a i square which is equal to minus beta 1 sigma a square. So, here as you see that you know this observed value of x i of the regressor and the model error they are correlated, so you cannot apply the standard or the ordinary least square technique to estimate the parameter beta naught and beta 1 here.

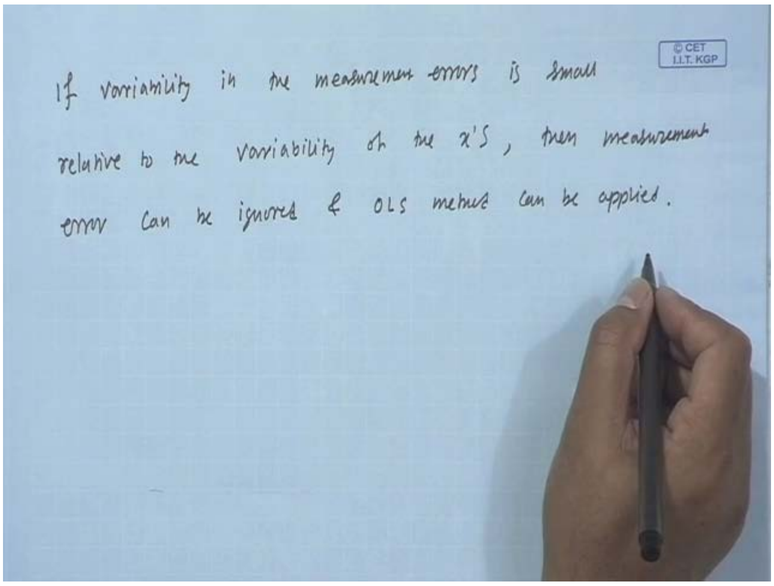
(Refer Slide Time: 42:18)



So, it says that if we apply standard least square method to the data, the estimates of the model parameters are no longer unbiased. If you apply just simple or ordinary least square technique what we will get is that, beta 1 hat is equal to summation y i minus y bar into capital x i minus x bar, because this is the observed regressor value, by summation x i minus x bar whole square. But, you can check that the expected value of this beta 1 hat is equal to beta 1 by 1 plus theta, that is the beta 1 hat we got here is not an unbiased estimator of beta. Where theta is equal to sigma a square by sigma x square, again the sigma x square is equal to you need to check you know this one sigma x square is this.

So, what this indicates is that if beta 1 hat is a biased estimator of beta 1, unless this theta is equal to 0 that means unless sigma a square is equal to 0, sigma a square is the measurement error variance. This will be 0 that is there is no measurement error in regressor. Also, you know if this sigma a square is very small relative to sigma x square, the biased will be then theta will be small, so the bias will be I mean once theta is small this quantity is almost close to 1 then the bias will be small.

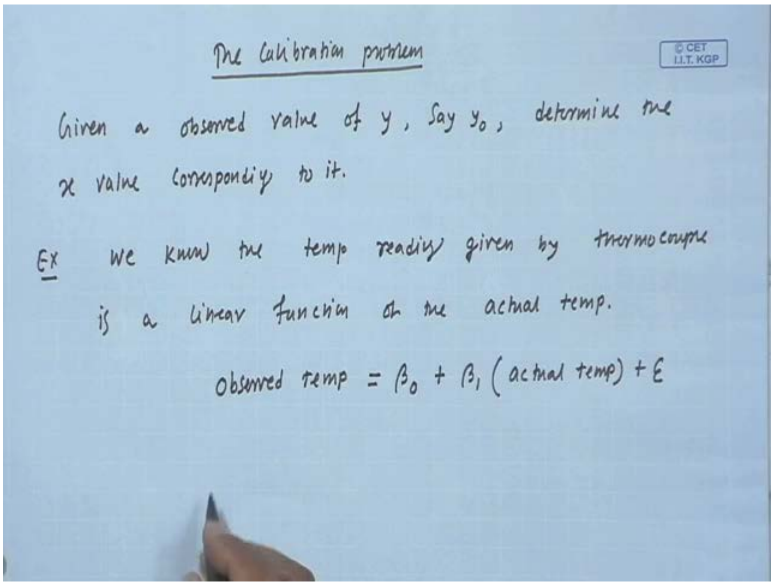
(Refer Slide Time: 46:25)



Finally, the technique says that if variability in the measurement error that is sigma a square is small relative to the variability of the x value, then it suggest that the measurement error can be ignored and ordinary least square method can be applied. So, here we have learned how to fit a model in the presence of measurement error in the regressor variable. And next what we will be doing is, we will be talking about the calibration problem which is also called the inverse problem.

Here, usually you know given a value of regressor variable x we estimate the response variable y. Here, the problem is just opposite, you are given value of y you have to estimate the corresponding regressor value that is called the calibration problem. So, here is the calibration problem.

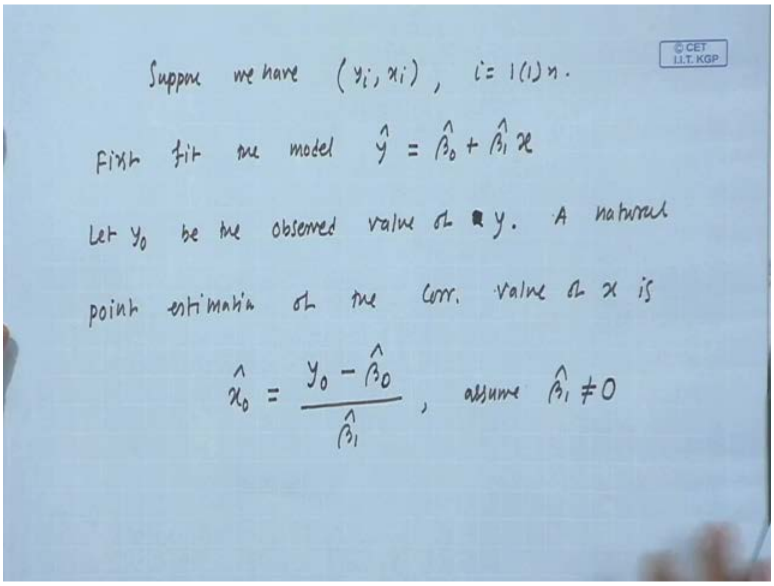
(Refer Slide Time: 48:37)



The statement of this problem is that, given a observed value of y say y naught, you have to determine the x value corresponding to it. Let me give an example now why we need this calibration problem. Example is we know that the temperature reading given by a thermocouple is a linear function of the actual temperature. So, what I mean by this is that the observed temperature, this observed temperature given by this thermocouple is a linear function of the actual temperature.

This observed temperature is equal to beta naught plus beta 1 actual temperature plus epsilon. In such situation you know the observed temperature and you want to know the actual temperature, so given a value of y say sum of the observed temperature y naught you can determine the corresponding x value that is the actual temperature. This is what the purpose of this calibration problem is and how do we solve this problem is that.

(Refer Slide Time: 51:37)



Suppose, we have some y i x i values for i equal to 1 to n. So, what you have to do is that, you just first fit the model y hat equal to beta naught hat plus beta 1 hat x. And let y naught be the observed value of x sorry observed value of y. And what you want to know is that, you want to know the value of x for which y equal to y naught, from this straight line fit. So, a natural point estimation of the corresponding value of x is say call it x naught hat is equal to, just put y naught here and then what is the x naught hat corresponding is equal to y naught minus beta 1 sorry beta naught hat by beta 1 hat, assuming that beta 1 hat is not equal to 0.

This is a very simple problem like the calibration problem is also called the inverse problem. So, here given a value of x sorry given a value of y say y naught you have to find the corresponding x. Once you have a fitted model there is no problem finding a point estimation for the corresponding x values.

So, what we have learned in this topic is that the usual situations in almost all cases what happened is that y is regressor variable, which is a random variable and x is regressor variable and it is usually a deterministic variable or controlled variable. And, here we have learned about two variations of this situation like when x and y both are random variable then how to fit the model. And also we have learned when there is a measurement error in regressor variable how to deal with that situation also, we need to stop now. Thank you.