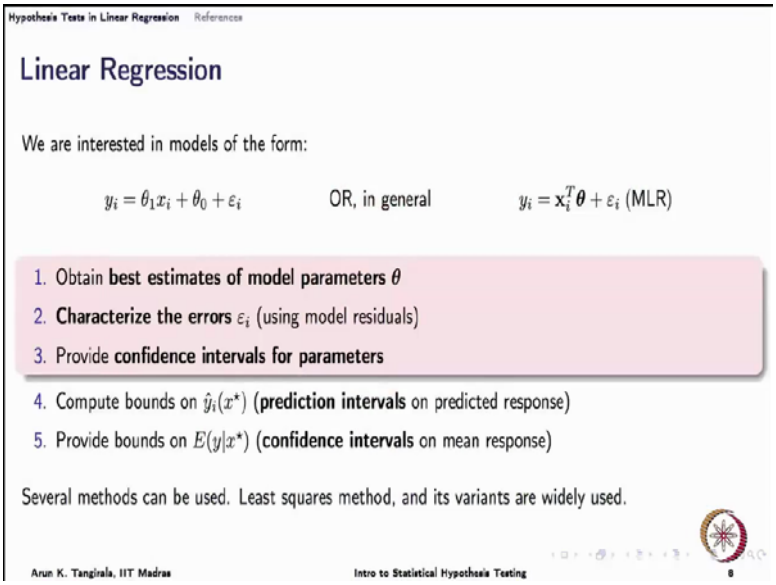
Introduction to Statistical Hypothesis Testing
Prof. Arun K. Tangirala
Department of Chemical Engineering
Indian Institute of Technology, Madras

Lecture - 16
Statistic for linear regression

So, let us move on to linear regression which is perhaps, perhaps one of the most widely encountered data analysis exercises, where you are trying to fit a model between 2 or more variables; typically falls into the predictive analytics (Refer Time: 00:30).

(Refer Slide Time: 00:23)



The simplest of the linear regression problems is where you regress a variable y on to a variable x. Now, in this slide we have used a notation y subscript i; that means, we say that the i th observation of a variable y is a linear function of the i th observation of another variable x. Strictly speaking, this is not a linear model; that is, theta 1 x i plus theta naught is not a linear function; it is an affine function. But with some abuse of terminology, we somehow tend to call this as linear; if this where to be strictly linear, theta naught should have been 0. Nevertheless, we will live with that kind of a terminology, and still call it linear regression. And here, linearity is with respect to the regressor x.

Now, this x is also known by several other names; very common alternative name that you find is explanatory variable, because it is a variable that explains what is happening in y. How do you choose this y and x? Again, generally, you look at the physics of the process and you say, x is probably causing y. If I take, for example, the relative humidity of the atmosphere, let us say, ambiance around us, then, we know from physics, relative humidity is a function of the temperature. So, y could be relative humidity and x could be temperature; or, if I believe that a gas is following an ideal gas law, and then say that, temperature is linear function of the pressure at fixed volume, then, I could set y to be the temperature and x to be the pressure, or even vice versa, depending on which you are going to change in your experiment independently, and which of the other one you are going to record.

So, there is an implicit assumption in the linear regression problem that, either x is the physical cause, or that, x is that variable that we are changing independently in an experiment, we have the freedom to do so, and y is being recorded, and therefore, it has some error, and that error is captured in epsilon i, of course, in the i th observation, but, there is so much more in this epsilon i, not just experimental error. It is possible, it is very likely in reality that, y is not purely a linear function of x, and such modeling errors also go and sit in epsilon. And, apart from modeling errors and sensor noise, there could be effects of unmeasured causes, unmeasured disturbances, that also get dragged on into epsilon. So, this epsilon is kind of a lumped error that we are accounting for in this model. But, we will not worry so much about that. I am just trying to give you a feel of where these models arise, and what the interpretation of this model is. There is so much more one can talk about linear regression; but the focus of this course is not on linear regression; it is rather on the hypothesis test that one would get to see, encounter, in linear regression.

In general, there may be more than one regressor, or more than one explanatory variable, or causal variable, in which case, x is a vector, and therefore, theta becomes a parameter. And, one can also absorb the theta naught into the vector of parameters theta, by thinking of another factor which is always held at a value of unity. That is, imagine that I have another regressor which is constantly held at a value of unity for all observations. Then, we can think of that also as a regressor, include that in x. And, that is what we have done

on the right hand side here, when we write y i equals x i transpose theta plus epsilon i. You could think of this as a re-representation of the model on the left, or in general, a model for what is known as multiple linear regression, when you have multiple regressors contributing to y. So, this is the linear regression problem for you. Of course, the non-linear regression counterpart would involve non-linear functions of x, and which we do not discuss at all in this course.

Now, the problem of linear regression, that is, from data, it consists of many different sub-problems, and obviously, the first one that we want to do is fitting. What do you mean by fitting? We want to choose this thetas; these are the parameters to be chosen freely, and so, we may, we have to make a decision; and these decision variables have to be chosen in an optimal manner. So, one of the first problems in linear regression, always linear regression arises in the context of data driven modeling. So, we want, we want to obtain best estimates of the model parameters theta.

Now, in all of this, one has to remember that, the truth is probably far more complicated than what we are postulating here. The model that we have written is a postulate of how we believe y is evolving as a function of x, and the reality may be quite different; then, why do we work with these models because, they are going to help us in making predictions. So, that is a prime purpose for which we are doing all of this; however, that epsilon keeps reminding us that, despite the best of our efforts, we will not be able to make an accurate prediction, because, there is a randomness in y. We will assume for now, that, x is not random, and that the entire randomness is only in y.

So, the bottom line, or the summary is, despite obtaining best estimates of the model parameters, and then using that model for predictions, we are going to make mistakes in the prediction. There is going to be always a left-over term, which we call as a residual, and from these residuals, we try to understand the nature of epsilon, right; epsilon is an unobserved variable; as I said, it is a lumped error. We use the so called, the model residuals, to characterize the errors, epsilon. What we mean by characterization is; what is the variability in it? What is the kind of probability distribution that characterizes the randomness in epsilon, and so on? But typically, those are the 2 main characteristics of interest, the variance, we will assume that, epsilon is of 0 mean; if it is not, then theta

naught would take care of it; so, not to worry. And secondly, that, we assume that, epsilon has a certain PDF or maybe, we try to learn from the residuals. Typically, you will see later on that, we assume epsilon to be following a Gaussian distribution, but that is not required for obtaining the best estimates.

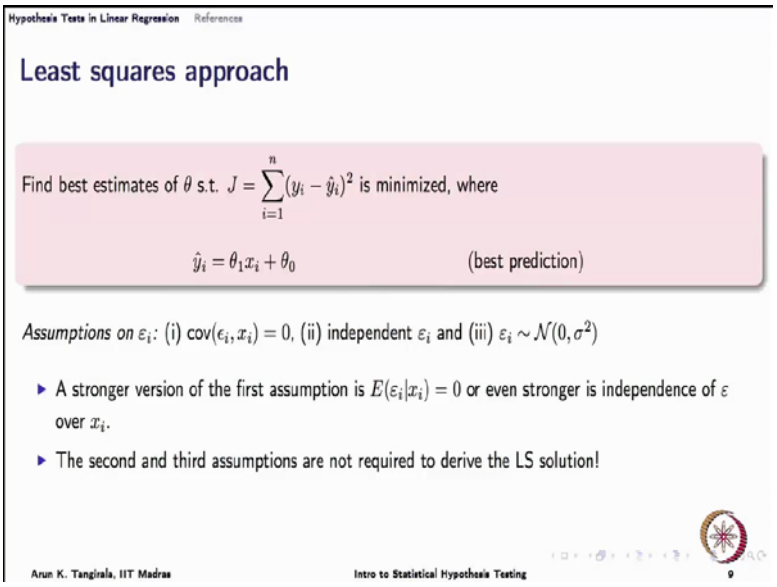
Now, we will come to what we mean by best shortly. So, let us move on to the third sub-problem of interest, which is in providing confidence intervals for the parameters. Now here, what we mean by confidence intervals for parameters is, assume that, now the truth also evolves in this way. Let us say, for the sake of discussion that, the truth is also linear, and what we have with us are the estimates of those true parameters, theta 1 and theta naught; and now, we want to provide confidence intervals for those true values from the obtained estimates. This may be contradictory to what I said earlier, that the truth may be more complicated; typically, more complicated; yes, but very often, when we want to test how good a regression method is, how a data fitting method, or a parameter estimation method is, we assume that, let us say that, we assume that, the truth is also very close to, structurally close to what we have postulated, at least under those conditions, under those idealistic conditions, where structurally the model and truth are similar, we should be able to say something about the truth; otherwise, if the truth is going to be different from what I postulate, then the item number 3 really does not make any sense at all. And, in those cases, anyway we will not even talk of confidence intervals for parameter.

So, the item number 3 applies to those cases, where we believe that the truth structurally has the same relationship, that is, under true conditions that, structurally y and x have the same linear relationship. And then, of course, there are the other two problems of constructing prediction intervals; that is, once I fit a model, I make a prediction, using the model, of course. And then, also, I would like to make a prediction for the average of y; one is a prediction of y, which is called a point prediction at any given observation, given x i; suppose, I tell you what is the temperature, you have to tell me what the relative humidity is; that is called prediction. Then, suppose I give you relative humidity, I want you to tell me what is the average sorry, I give you the temperature, what is the average relative humidity; that is called the mean; you are predicting, not the point; at that particular point you say, across all possibilities of x, what would be the, not possibilities of x, across all possibilities of epsilon, what would be the average of y; you

would average that, and that is called the mean response. One is called the point response, and the other is called the mean response and we may want to also provide confidence intervals on the mean response. Remember, the mean is no longer a random variable. Therefore, we talk of confidence intervals; whereas, the response at a certain x i is going to be a random variable, because we are not averaging it over the space of epsilons. Therefore, we talk of prediction intervals. However, in this module, we are not going to talk of prediction intervals, or confidence intervals, because, the objective of this module is to go over, or go in detail, on some of the popular hypothesis tests that one encounters in linear regression, without which the linear regression exercise is incomplete.

Now, the question therefore, the question therefore is, how do I obtain best estimates of model parameters? That is item number 1 for us. There are several methods in the literature, but one of the most popular methods that we use, that you are probably also familiar with is, Least squares method. What does a least squares methods rely on? It relies on the principle of minimizing the squared distance between the vector of observations and the vector of predictions.

(Refer Slide Time: 12:19)



.

So, if I make a prediction, let us say, I give you x i; what would be your prediction? Call that as y hat of i, that is, if I give you x i, and if I give you model parameters theta 1 and theta naught, what would be the prediction? The prediction would be theta 1 x i plus theta naught, assuming that, epsilon has 0 mean, and that, you cannot predict epsilon using x; that is very important. So, that is a first assumption we are making, that, there is nothing in x that you can use for explaining epsilon; that is what we mean by co-variance between x i and epsilon i being 0. In fact, a stronger version of this assumption is that, the conditional expectation of epsilon, given x is 0, or that epsilon is even independent of x. What we are assuming is that epsilon and x are uncorrelated; there is no linear effect of x on epsilon. But, a stronger assumption would be there is no non-linear effect at all; there is absolutely no effect of x on epsilon. Under what conditions this would be true? We say this conditions are true under so called, open loop conditions; no feedback; that means, feedback between what, between y and x; x produces y; y in turn should not produce, cause x; y in turn should not cause x. If that is the case, then, the first assumption breaks down. So, we will exclude those situations.

Now, again recapping what we just said, when there is no correlation between x i and epsilon i, given x i and model parameters, y hat would be theta 1 x i plus theta naught; this is what, this is how I would predict. Now, this is for the i th observation; I have n observations from a sample of size n, and there is no point in just trying to fulfill 2 observations or 3 observations. I would like to actually fulfill all the n observations. Now, we know from the model that, whatever I try and do, I will never be able to fulfill even a single observation, so to speak; even if I do, the max that I can do is, I can fulfill 2 observations; that is, I can, I cannot fulfill every observation, but I can fulfill 2 observations. Why? Exactly; why do I say that? Because, there are 2 unknowns; theta 1 and theta naught, I can randomly pick any 2 observations and force fit this theta 1 and theta naught on to those 2 observations, randomly picked observations. But then, the predictions on the remaining n minus 2 observations will take a beating. In any case, the predictions on the observations will take a beating; in this case, it will exactly fit 2 observations, but the remaining n minus 2 predictions will not be good enough.

In the least squares method, none of the predictions is going to be accurate, but collectively, the predictions are going to be such that, the sum square errors is going to

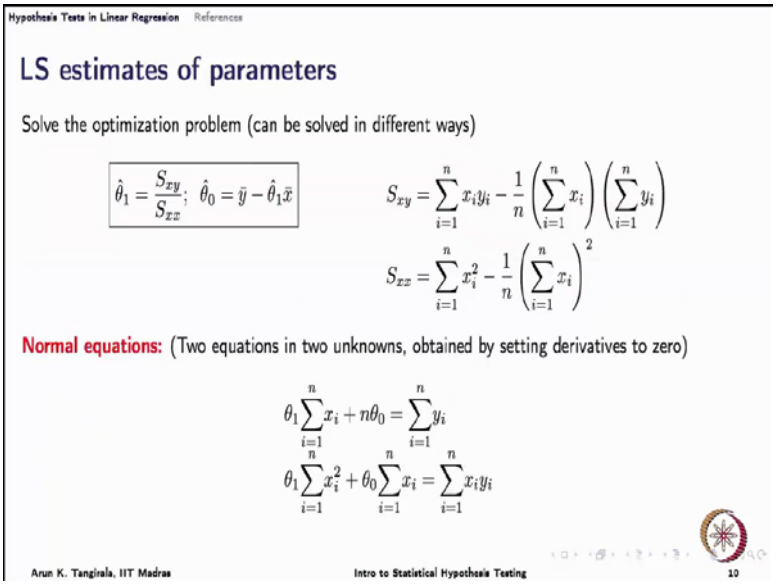
be a minimum, and that is the basic principle of least squares approach. And therefore, mathematically, what we say is, find theta such that sigma y i minus y hat whole square, y i is your observation, y i hat is your prediction and we are going to now collectively minimize this. So, we want the vector y. What is vector i? It consists of the n observations and the vector of predictions, consisting of predictions of those n observations. We want them to be as close as possible in the, in a Euclidean sense; that is, the distance, we want to be as low as possible in a Euclidean sense.

This is a standard optimization problem. It is a convex optimization, or a quadratic objective function, and, it is fairly easy to derive the solution. Now, very often you will see in text, the assumptions number 2 and 3 that is epsilons are independent. What do we mean by epsilon i is independent? That is, epsilons; epsilon i would correspond to error in the i th observation and when I look at the n observations, I have n errors. What we mean by independent epsilon i is, none of the errors, that is, an error in the i th observation is not going to influence the error in any other observation, in any way, non-linear sense also. But, that is not required to derive the solution, least square solution. And, the third assumption that you would see is that, epsilon i falls out of a Gaussian distribution of 0 mean and variance, sigma square. Or you can say, sigma square i also; if you are looking at epsilons falling out of a different distribution, that is same distribution, but different variance in the, from observation to observation, that is also admissible.

However, the assumption number 2 and 3 are not required to derive the least square solution. Many a times, in many texts this can be confusing. We need assumptions 2 and 3 later on, when we talk of sampling distributions or distributions of theta 1 hat and theta naught hat. For deriving the solution, the first assumption is sufficient, because, that tells us that the best prediction is what we have written there on the slide, and then, we can proceed to the best estimates of theta 1 and theta naught, using standard optimization techniques. What is a standard optimization technique? You differentiate the objective function with respect to the decision variables, take partial derivatives and set the derivatives to 0, because at the extreme, minimum or maximum, the slope of the objective function is going to be 0, in the parameters space here. So, it is a 2-dimensional parameter space. Once we do that, we do end up with what are known as normal

equations.

(Refer Slide Time: 18:41)



So, let me draw your attention to the equations at the bottom and then we will go to the solution at the top. These are the 2 equations that one encounters when we set the partial derivatives of j with respect to theta 1 and theta naught respectively to 0. So, you would have 2 equations in 2 unknowns, and that will then get me unique estimates of theta 1 and theta naught. That is the beautiful thing about a linear regression problem with a least squares approach. Whenever I use a linear model, so to speak, linear predictor and I am minimizing the sum square errors, I am guaranteed a unique solution for theta 1 and theta naught, which is not necessarily the case in a non-linear regression problem, where I cannot. In general, derive an analytical solution and I have to, therefore, turn to numerical optimizers.

Here, the solution is arrived at in an analytical way, and the solution is unique. And, that solution is given at the top here, in the box; theta one 1 hat is S xy by S xx. What are this numerators and denominator? S xy is this summation. Remember, when we were discussing correlation, I did mention that, you would see summations similar to those that you see in correlation linear regression as well. You see some kind of expressions like this. I will bet you may not see x bar and y bar straight away, but you can guess here,

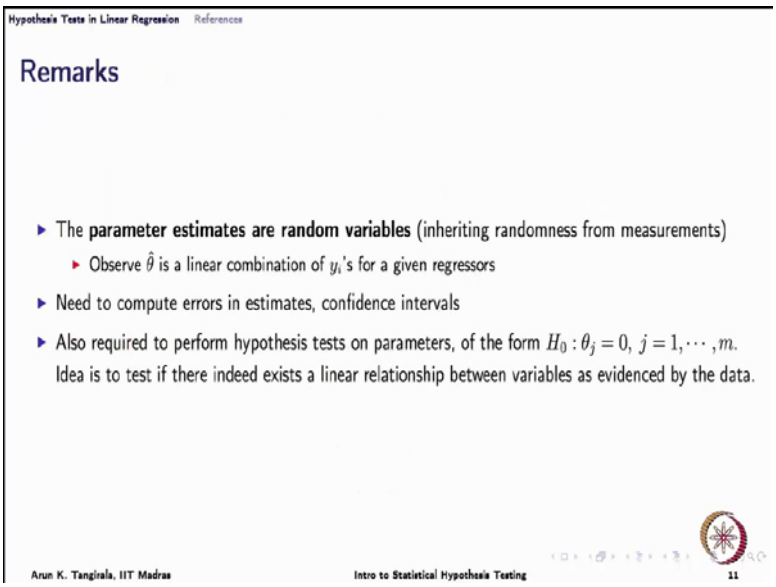
sigma x i by n is x bar; or you can say, sigma x i is n times x bar and sigma y i is n times y bar. So, you can write S xy as the sum of the products, minus n times the x bar, n times x bar times y bar; you can write that way. And, you can see that, S xx is nothing, but S xy, but with y replaced with x; that is all it is. This, the relation here, sorry, the expression for theta i hat, looks somewhat similar to the correlation coefficient; particularly, the numerator; not the denominator. In the denominator of correlation coefficient, we have products of square root of the sigma square hat, x times sigma square hat y. That is, we would see square roots of sum of S as square root of product of S xx times S xy. We do not have that here, that you should note. But, the numerator is the same. Therefore, we should expect, if there is 0 correlation between y and x, that means there is no correlation between y and x, we should expect theta 1 hat to be 0, or theta 1 truth to be 0, sorry. So, the true theta 1 would be 0, if there is no correlation because, not because the expression for theta 1 hat and the correlation coefficient are identical, but because the numerators are more or less identical. If the numerator in the correlation coefficient is 0 or you can, that is, there it is an estimate it would never be 0; but let us say, ideally speaking if that is 0, then theta 1 hat would also be 0.

You can also show that, theoretically, the optimal estimate of theta 1, what we mean by theoretically is not from observations, but given the population if I give you all the observations, then you can show that the optimal estimate of theta 1 is related to the correlation between y and x, and you can show that, when the correlation goes to 0, theta 1, optimal estimate of theta 1 will also go to 0. Estimates never go to 0. So, that was only an ideal discussion earlier. So, bottom line is, if truly there is no correlation between y and x, the true value of theta 1 would also be 0, which means, performing a hypothesis test on theta 1 equals 0 would amount to saying that there is no, or performing a hypothesis test of the type rho equals 0; that is why we went through the correlation exercise first, prior to fitting a model. If the data passes the significance test for correlation, that means passes meaning if the null hypothesis that correlation is 0 is rejected, then we believe that, we can fit a linear model; then, there is a case for fitting a linear model.

The optimal estimate for theta naught is kind of intuitive. It is based on the means of y and x, and the optimal estimate of theta 1. You should verify indeed that, this is the

solution that you get by solving the so called normal equations that we discussed earlier. And, these are expressions that we have been probably seeing from high school days. These are all nothing complicated. We are going to stick only to the single regressor case. For the multiple linear regression problems we do not perceive that here, but in the general regression course, or a course in estimation theory perhaps, we would; very good.

(Refer Slide Time: 24:09)

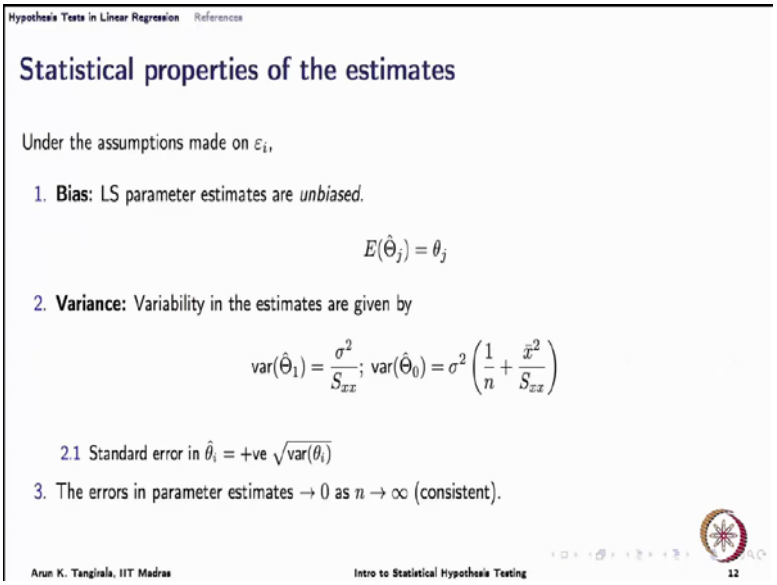


So, before we proceed to discussing the hypothesis test, we should remember that, the parameter estimates are random variables, because, they are being derived from data. One hat and theta naught had to be random. Well, the randomness in y; even if you assume x to be deterministic, which we normally assume in a classical linear regression problem, y has randomness in it. Therefore, both the parameter estimates inherit that randomness.

And therefore, both the theta hats are random variables in their own right. They have a mean; they have a variance, and so on. And, we need to look at 2 things; whether the average of theta 1 hat gives me the so called true theta, and what is the expression for the variance of theta hat. Why do we need both these expressions? One, we want to see if the least squares estimator is unbiased, right; an unbiased estimator is one which produces a

truth on an average. And two, we need the variance for many reasons. One, we want to know is there any control over the variability in my theta hat. What we mean by variability in the theta hat is, one data record will give me one estimate of theta hat; another data record for the same experiment, under the same conditions, we know because of randomness, will produce a different value of theta hat and so on. We want that variability in the estimates to be as low as possible. Is there any control over it? Or, simply the process dictates the variability in theta hat? More importantly, for hypothesis testing of the type theta equals 0 that is, significance test for the parameters remember, we need to know the variance of the parameter. We know this already from our hypothesis test on mean remember, when we conducted a one sample test for the mean with variance unknown, we needed to estimate the variance. So, here also the parameter of interest is theta, and we want to ask if the average theta that is the truth is 0 and to be able to conduct such a hypothesis test we would need to know the variance of theta hat.

(Refer Slide Time: 26:32)



So, let us look at the bias, without going through any rigorous proof, I am stating straightaway, under the assumptions that we have made here the Gaussianity assumption on epsilon is not required. All that is required is the first assumption that we have made, epsilon and x are uncorrelated, and that epsilon is of 0 mean; that is enough to guarantee unbiasedness of the least squares parameter estimates. So, expected value of theta hats

are the truths themselves. Here, we are assuming the truth also is structurally the same as our model; that is, the underlying relationship between the y and x is truly linear otherwise, all of these discussions are meaningless. Then, we look at the variance. Once again, without any proofs I am giving the expression for the variance of thetas. So, the variance of theta 1 hat is given by sigma square over S xx and variance of theta naught hat has a slightly more complicated expression to it, but it is also a function of sigma square, and the expression is exactly sigma square times 1 over n plus x bar square over S xx.

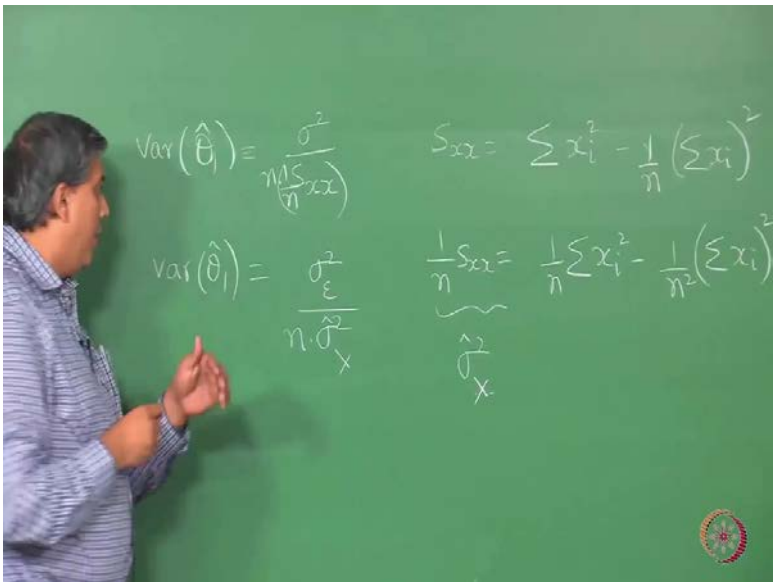
Earlier, we asked the question, is there any control on the variability; that means, can I drive the variability to 0; the data may be changing, but can I guarantee, can I somehow ensure that estimates do not vary? For finite samples, it is not possible; but at least for infinite samples that means, what I mean by samples is, observations, I am sorry; for infinite sample sizes at least in those cases is it possible? When I have a large sample size, can I have a very low variance of the parameter estimate? Because, the square root of this variance is a standard error, and I want the errors in the parameter estimates to be low.

Now, if you look at the factor, let us look at, for example, variance theta of theta 1 hat. The expression is sigma square over S xx. What is sigma square? It is a variance in epsilon which I have no control over. It depends on the process and the sensor. S xx is something that I may have a control over, if I am performing the experiment. What is S xx? It is this expression here, sigma x i square minus 1 over n sigma x i the whole square. So, if I want low variability in theta 1 hat, I want high values of S xx. All that I have to do is, therefore, choose high values of x i or choose high values of n, because, the more terms I have, the more would be the value of S xx; that is another way of looking at it. So, there are 2 things that I can perform in an experiment to ensure that I obtain parameter estimates of low error. This is a very fundamental result in design of experiments. It tells us that, if I want to fit a linear model, either choose a high amplitude for the independent variable, which sometimes may not be possible, because of physical limits.

Suppose, I am looking at an experiment where I want to determine the relationship

between temperature and pressure of a gas; I may not be able to, and let us say, pressure is the independent variable, and I can independently change it. There is a limit to which I can increase the pressure without, before I cause any hazard to the experiments, or to the neighborhood. But, there may not be a limit, any limit to the sample size, if one has sufficient time. So, there are 2 factors that one can vary, or the experimentalist can vary, and play around with it. In fact, if you write this carefully, this variance expression carefully, you can show that, essentially, there are 2 factors. One is the signal to noise ratio. What is a signal here? We say that, the signal here is x, you can say, and what do you mean by signal to noise ratio is, ratio of the variances of the signal, which here is x. S xx is not an estimate of the variance of x, if you think of x as random. But, it is 1 over n sigma, sorry, n times S xx can be written as 1 over n times, n times variance of x, and in, when you write it that way, you can show that, variance of theta 1 hat is a function of the sample size and the signal to noise ratio. What I meant is, take the denominator S xx, simply write it as n times 1 over n S xx; that 1 over n S xx is an estimate of the variability, that is, the level of fluctuations, or the power in x, you can say. And therefore, sigma square over 1 over n S xx would be called as the signal to noise ratio. Let me just write that on the board for you.

(Refer Slide Time: 31:49)



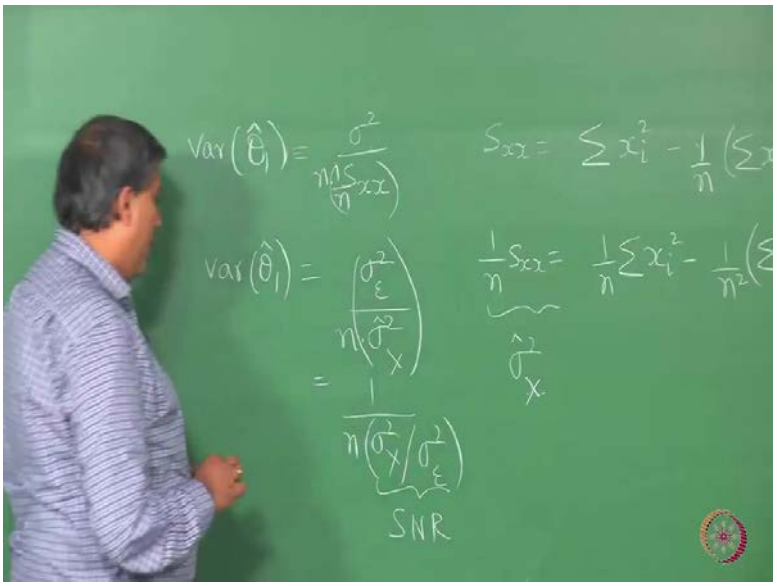
What we have is variance of theta 1 hat is sigma square over S xx. As I said, S xx is not

the measure of the power. Remember, we had S xx as sigma x i square minus n times, sorry, 1 over n times sigma x i to the whole square; this is what is S xx.

Now, suppose I multiply and divide here with n, and, you know, 1 over n here multiply with n and divide by n. Then, 1 over n S xx is 1 over n sigma x i square minus 1 over n square, whole square; which of course, you can show is if you think of x as a random variable, we have been saying x is deterministic; but suppose, you think of x as a random, just for the sake of discussion, you can call this as an estimate. You can verify that, indeed this is the estimate of the variance of x using 1 over n as a factor. In that case, I can write this as sigma square. Let me remind you that, this is of epsilon and this is of x, and then here you have n. So, this is the variance of theta 1 hat.

Now, it is clear as to what factors really affect the variance in theta 1 hat. One is the sample size, and other is the signal to noise ratio.

(Refer Slide Time: 33:54)

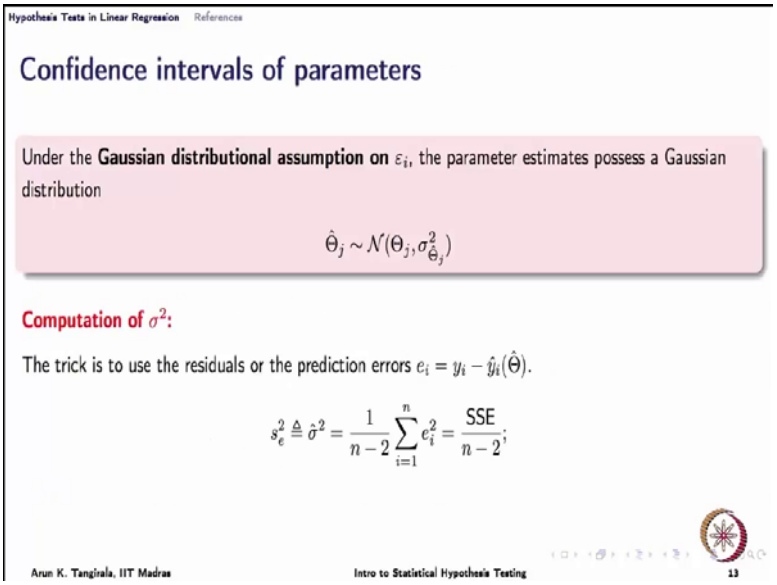


Or, the noise (Refer Slide Time: 33:50) were very low. We can write this further as, this is called the signal to noise ratio. Either I maintain a very high signal to noise ratio, the signal to noise ratio is a measure of the truth to the uncertainty, you can say; epsilon is a level of, sigma square epsilon tells me the level of uncertainty in y, or randomness in y,

and sigma square x tells me the power in x, which is the independent variable. As long as I maintain this very high or maybe you know, take to a very large value, I would get low values for the variance in theta 1 hat; consequently, low values for the errors in theta 1 hat. Alternatively, at a fixed SNR, if I increase the sample size to a very large value, then also, I get low errors in theta 1 hat. So, this throws some very valuable insights into the design of experiments. This is just for your information, Ok.

So, now, let us move on. We use these variance expressions for conducting our hypothesis test.

(Refer Slide Time: 35:04)



And, I am going to now move on to the hypothesis test, and constructing confidence intervals. We will conduct the hypothesis test using the confidence interval approach, but to be able to do that, we need the sampling distribution of theta hat. What we have derived are only the first and second order movements. I am straight away giving you the result for the distribution of theta hat. If you assume epsilon to be Gaussian distributed, then theta has to be, theta hat has to follow Gaussian distribution, because look at the expression for theta hat. Now, y is the only random variable here. So, for all our discussions, except for the one that we made just now, x is assumed to be deterministic. Even if it is not, for a fixed x, theta hat derives its randomness from y. So, if you look at

the expression for S xy, it is nothing, but a linear combination, or linear function of the observations. Therefore, if y has Gaussian distributed errors, a linear combination of Gaussian distributed errors is what is going to creep into theta hat, and we know that, a linear combination of Gaussian distributed errors is also going to be Gaussian.

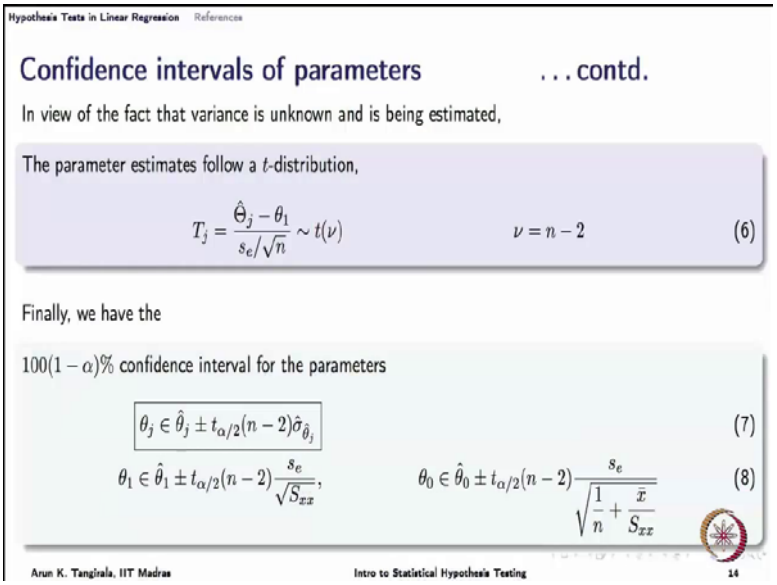
What about the case of non-Gaussian errors in y? In that case, the result that we had given for the sampling distribution of theta hat holds only for large n, by virtue of the central limit theorem. In any case, for the classical linear regression problem, theta hat is a linear function of the observations; that is an important point to observe. If epsilons fall out of a Gaussian distribution, then, regardless of the sample size, theta hat follows a Gaussian distribution. If epsilons fall out of a non-Gaussian distribution, then, only for large sample sizes, theta hat will tend to have a Gaussian distribution. We will assume that, epsilons are Gaussian, make the standard one; because, the goal is here to show you how hypothesis tests are conducted. Now, obviously to complete the discussion here we need an estimate of sigma square epsilon, right. And, early on, we said, one of the objectives that we have to take care of in linear regression is, to characterize the errors. We have fixed epsilons to be Gaussian, but we do not know the variance. Using the residual, that is the prediction errors, whatever leftovers that we have after we have made the best predictions, we can characterize epsilons; particularly, we can obtain an expression for the variance of epsilon.

And, once again, I am giving you straight away, a standard result that is available in linear regression literature. I will not try to prove that here; in a full estimation theory course I may prove it. So, you just take it for granted here, that, this estimator for the variance of epsilon which we call as S square epsilon, or sigma square hat, and which is equal to Sum Square Errors; Sum Square Errors, is your sigma y i minus y hat square on the training data, that is, the data that you use for modeling, divided by n minus 2. Now, why do we say n minus 2 here? We have actually used up 2 degrees of freedom in estimating theta 1 and theta naught.

Therefore, the leftovers, or the residuals that are available for the n observations indeed have only n minus 2 degrees of freedom; that means this actual sources of variability are only n minus 2. To begin with, we had n observations, n sources of variability. Now, we

have only n minus 2. And, once you have the sampling distribution, we known how to, you know how to construct the confidence intervals. It is straight forward.

(Refer Slide Time: 39:03)



We assume small sample size, Gaussian distributed errors, variance unknown. So, this is the classic case of your one sample t test. So, we construct a t statistic, and with, and say that, this t statistic theta j hat minus theta, sorry, theta hat i minus theta hat i, whatever j or i, by S epsilon over by root n follows t discussion with n minus 2 degrees of freedom, because this S epsilon follows a t, sorry, has n minus 2 degrees of freedom.

So, with the expressions for the variance of sigma square theta hat, in fact, sigma square epsilon that we discussed just now, and the distributional properties, we are now ready to march ahead, to write the confidence intervals for parameters. But, there is a small thing that one has to note, with respect to the sampling distribution of theta hat. When the variance is known, then theta hat follows a Gaussian distribution. But, we know from the one sample t test for mean that, when the variance is unknown and for small sample size, theta i hat would then follow a t distribution. If the sample size is large, again one can return to the Gaussian distribution. But, it is always better to work with the t distribution, and therefore, the statistic theta i hat minus theta i, strictly speaking, this theta is nothing, but your truth, the postulated value, divided by sigma hat theta i. What do we mean by

sigma hat theta i hat is, sigma, theta i hat is the standard deviation of the i th parameter and sigma hat is a estimate. How do you obtain the estimate? First, calculate your S square e, and then, take that and plug in to these expressions, wherever you see sigma square; that is how you calculate your sigma. And then, of course, take the positive square root; that is how you would calculate sigma hat theta i hat.

This follows a t distribution with n minus 2 degrees of freedom. This degree of freedom, as usual, with the t test before, or the t statistic before, comes from the denominator. Now, once we have the sampling distribution, we know from a previous lecture how to write the confidence interval. For the i th parameter, we have, because we are looking at a two-sided test all the time; we are asking if the slope is 0, and the intercept is 0; that is, theta 1 is 0 and theta naught is 0. Therefore, we are looking at a two-sided confidence region and the significance level being alpha. So, the 100 times 1 minus alpha percent confidence interval for the i th parameter is theta i hat plus or minus t alpha by 2, n minus 2 degrees of freedom times sigma hat theta i hat. And, I have given the specific expressions for theta 1 hat, I mean theta 1 and theta 2, the confidence intervals for not theta 2, sorry, theta naught. By substituting the expressions for sigma hat theta i hat from this slide that we had seen before for theta 1 hat, we use this expression, variance of theta 1 hat being sigma square over S xx. Of course, we replace the sigma square with its estimates and likewise, for theta naught.