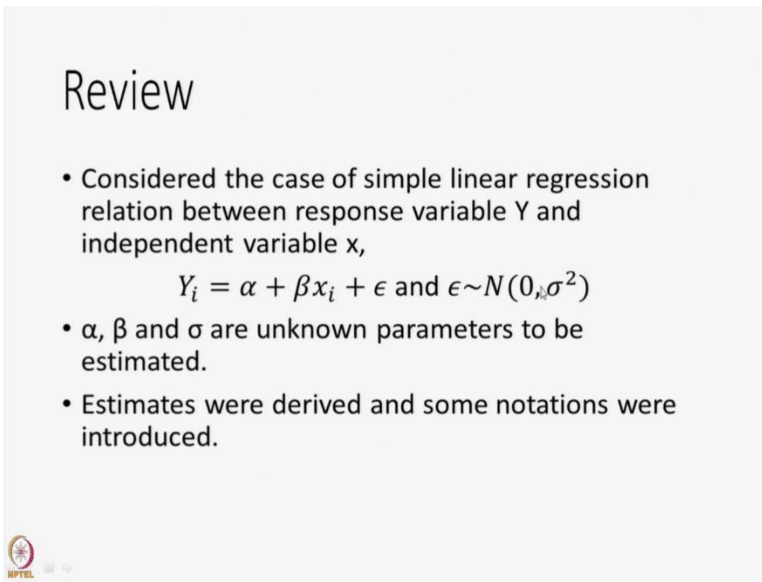
Dealing with Materials Data: Collection, Analysis and Interpretation
Professor Hina A. Gokhale
Department of Metallurgical Engineering and Materials Science
Indian Institute of Technology, Bombay
Lecture 82
Regression Analysis - 2

Hello and welcome to the course on Dealing with Materials Data. In the previous session, we have started working with Regression Analysis and we are going to go further on it. First let us review what we did in the past.

(Refer Slide Time: 00:34)

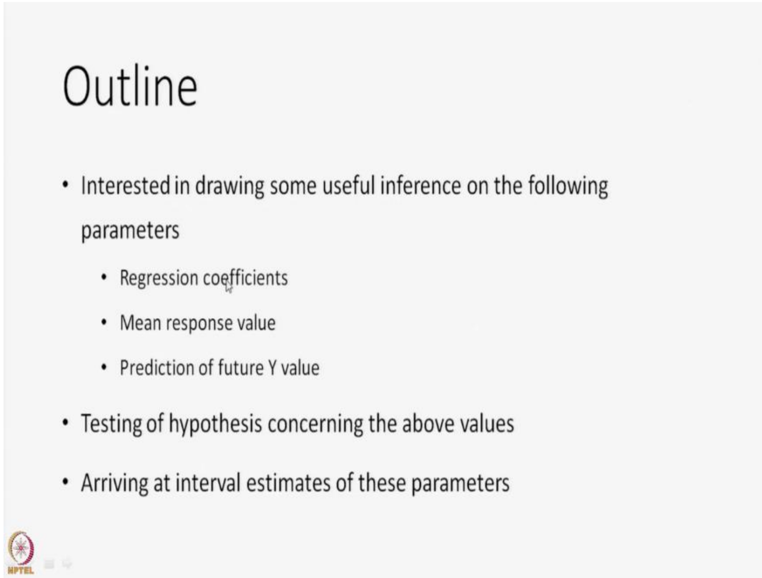


We considered the case of simple linear regression relationship between a response variable Y and then independent variable X. We denoted it as,

$$Y_i = \alpha + \beta x_i + \epsilon \text{ and } \epsilon \sim N(0,\ \sigma^2)$$

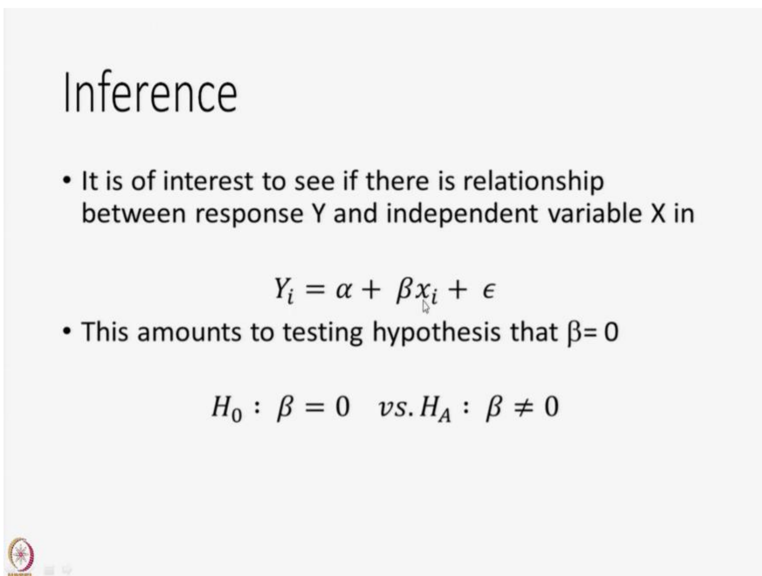
It is a very natural assumption to make with respect to random errors. There is a slight change in notation from the previous class, please note we previous session we said that Yi is equal to beta 0 plus beta 1 xi, I am for simplicity changing it to alpha plus beta xi. So, in this case as you can see alpha beta and sigma are unknown parameters that need to be estimated and we had derived those estimates and we had to use some notations which was freshly introduced.

(Refer Slide Time: 01:52)



In the present class session, we would like to look into some useful inference on the parameters first the regression coefficients. Then the mean response value to the, of the regression and prediction of the future Y value. This we will do by test of hypothesis, concerning above values and we will arrive also simultaneously at interval estimates of this parameters at some confidence level.

(Refer Slide Time: 02:29)



First we start with inference. You see we assumed that Yi and xi has a linear relationship. I think it is important to test whether the, any relationship ever exists and if you want to test such a hypothesis, it amounts to testing hypothesis that the parameter beta is equal to 0.

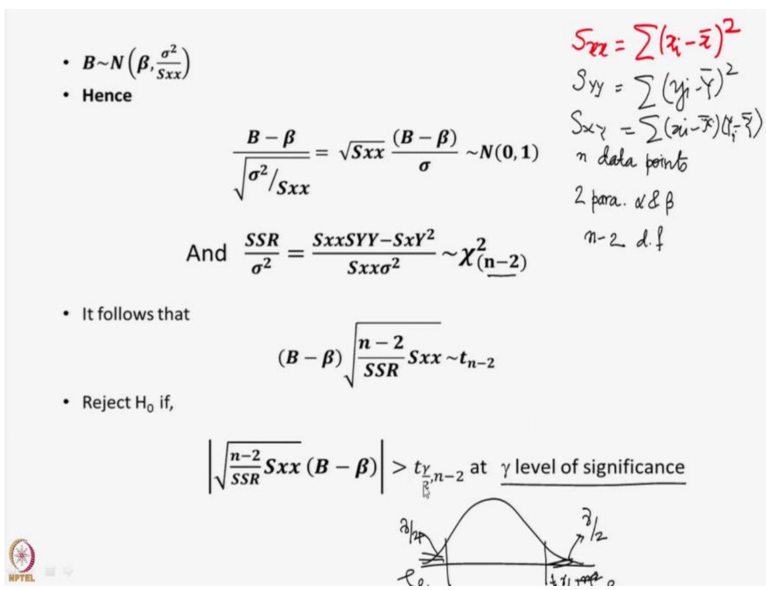
$$Y_i = \alpha + \beta x_i + \epsilon$$

This amounts to testing hypothesis that β= 0

$$H_0 : \beta = 0 \quad vs.\ H_A : \beta \neq 0$$

If because if the beta is 0, Yi is simply alpha plus a random error. So, it becomes a completely or random process without having any systemic change in it. So, we want to test hypothesis as 0 as beta is equal to 0 verses alternate hypothesis of beta as not equal to 0

(Refer Slide Time: 03:21)



In the previous session, we saw that beta estimated by B is distributed as normal distribution with a variance of, with a mean value of beta and variance of sigma square by Sxx. It is a good idea to recall. It is a good idea to recall that Sxx is defined as

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n}(x_i^2) - n\bar{x}^2$$

Hence, we can easily say that

$$\frac{\boldsymbol{B-\beta}}{\sqrt{\boldsymbol{\sigma^2}/\boldsymbol{Sxx}}} = \sqrt{\boldsymbol{Sxx}}\,\frac{(\boldsymbol{B-\beta})}{\boldsymbol{\sigma}} \sim \boldsymbol{N}(\mathbf{0},\ \mathbf{1})$$

Now sigma square is also unknown and that can be estimated as sums of squares of the residuals divided by sigma square which

$$\frac{\boldsymbol{SSR}}{\boldsymbol{\sigma^2}} = \frac{\boldsymbol{SxxSYY - SxY^2}}{\boldsymbol{Sxx\sigma^2}} \sim \boldsymbol{\chi}^2_{(\mathbf{n-2})}$$

Please recall last time we also introduced why it is n minus 2. Why it is n minus 2? Because we have n data points. We have estimated two parameters alpha and beta and therefore remaining degrees of freedom becomes n minus 2 and therefore it is n minus 2. Please recall just as Sxx, we can defy SxxYY as summation of Yi minus Y bar whole square and Sxy will also be defined as summation xi minus x bar times Yi minus Y bar.

So, from this it is easy if you, if we replace, if we replace this sigma square, this sigma square in this formula by the estimated value of sigma square as SSR or rather the sigma square by SSR divided by this particular value it turns out that this will be distributed as t distribution with n minus 2 degrees of freedom. Because this sigma square estimate and the beta estimate of that is B.

These two are in the SSR and B are two independently distributed SSR is distributed as a Chi square, B is distributed as a normal and therefore the ratio is going to be distributed as a t distribution with n minus 2 with a appropriate multiplication constants, which are given here and therefore we can say that, we can reject the null hypothesis of beta is equal to B is equal to beta by stating that absolute value of this statistic is greater than t at n minus 2 degrees of freedom with gamma level of significance.

$$(\boldsymbol{B-\beta})\sqrt{\frac{\boldsymbol{n-2}}{\boldsymbol{SSR}}\boldsymbol{Sxx}} \sim \boldsymbol{t_{n-2}}$$

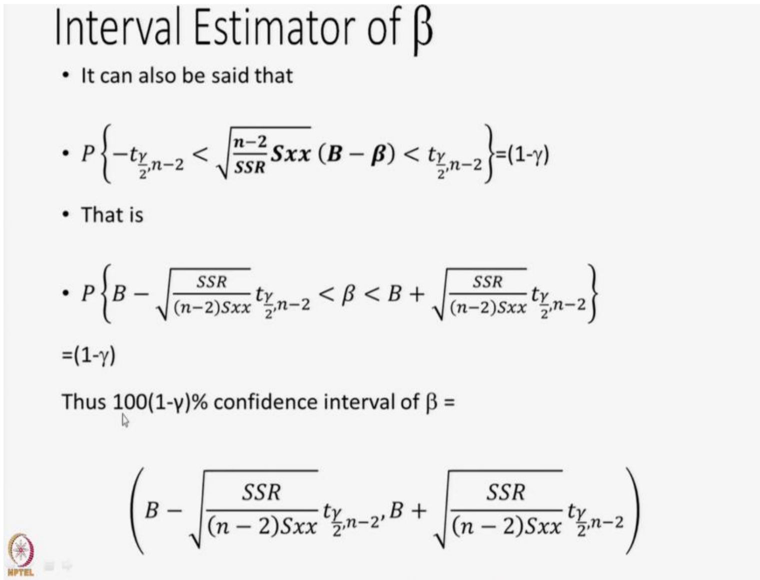
- Reject H_0 if,

$$\left|\sqrt{\frac{\boldsymbol{n-2}}{\boldsymbol{SSR}}\boldsymbol{Sxx}}\,(\boldsymbol{B-\beta})\right| > t_{\frac{\gamma}{2},\,n-2} \text{ at } \gamma \text{ level of significance}$$

Please recall that you have to take gamma by 2. Because when you take alpha by 2, your t distribution is also a symmetric distribution and this is your rejection area that is, this is your critical area see and this area total you would like it to be gamma. So, this is gamma by 2 and this is gamma by 2, gamma is your level of significance and therefore it is T this value, which is defined as T gamma by 2, N minus 2 is the value you are taking and therefore it works out that this is the rejection region or critical region.

So you reject 0, if square root n by 2 by SSR multiplied Sxx multiplied B minus beta, absolute value is greater than the cut off value at t at gamma by 2, n minus 2 degrees of freedom. If it is large you are going to reject the hypothesis at gamma level of significance.

(Refer Slide Time: 08:53)



This also gives a rise to the interval estimator of beta. Because this is the acceptance region of the hypothesis. This is the acceptance that is a critical region that is a rejection region of hypothesis. This is an acceptance region of hypothesis and this probability is 1 minus alpha and therefore 1 minus gamma sorry and therefore the 100 times 1 minus gamma percent confidence. So, if gamma is say 0.01, then 99 percent confidence interval for estimator of beta can be given in this manner.

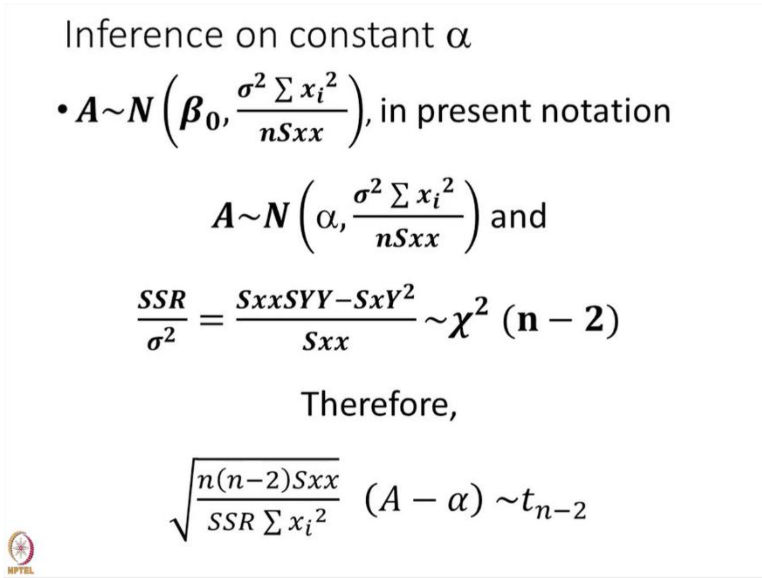
$$P\left\{-t_{\frac{\gamma}{2},n-2} < \sqrt{\frac{\boldsymbol{n-2}}{\boldsymbol{SSR}}\boldsymbol{Sxx}}\,(\boldsymbol{B}-\boldsymbol{\beta}) < t_{\frac{\gamma}{2},n-2}\right\}=(1-\gamma)$$

$$P\left\{B-\sqrt{\frac{SSR}{(n-2)Sxx}}\,t_{\frac{\gamma}{2},n-2} < \beta < B+\sqrt{\frac{SSR}{(n-2)Sxx}}\,t_{\frac{\gamma}{2},n-2}\right\}=(1-\gamma)$$

Thus 100(1-γ)% confidence interval of β =

$$\left(B-\sqrt{\frac{SSR}{(n-2)Sxx}}\,t_{\frac{\gamma}{2},n-2},\ B+\sqrt{\frac{SSR}{(n-2)Sxx}}\,t_{\frac{\gamma}{2},n-2}\right)$$

(Refer Slide Time: 09:36)



Let us move on in the same fashion with respect to constant alpha. Now when we want to test whether alpha is equal to 0 or not. What we are really testing is, if there is a linear relationship then the Y insert intercept of the line is 0 or not. It means that does the line pass through the origin or not and therefore here or gain, we have A distributed as

in the previous notation as beta 0 with sigma square summation xi square divided by nSxx in the new notation. Because we have replaced beta 0 by alpha it becomes A is distributed normal with mean value of alpha and variance of sigma square with this multiplicand

$$\boldsymbol{A \sim N\left(\beta_0,\ \frac{\sigma^2 \sum x_i^2}{nSxx}\right)}, \text{ in present notation}$$

$$\boldsymbol{A \sim N\left(\alpha,\ \frac{\sigma^2 \sum x_i^2}{nSxx}\right)} \text{ and}$$

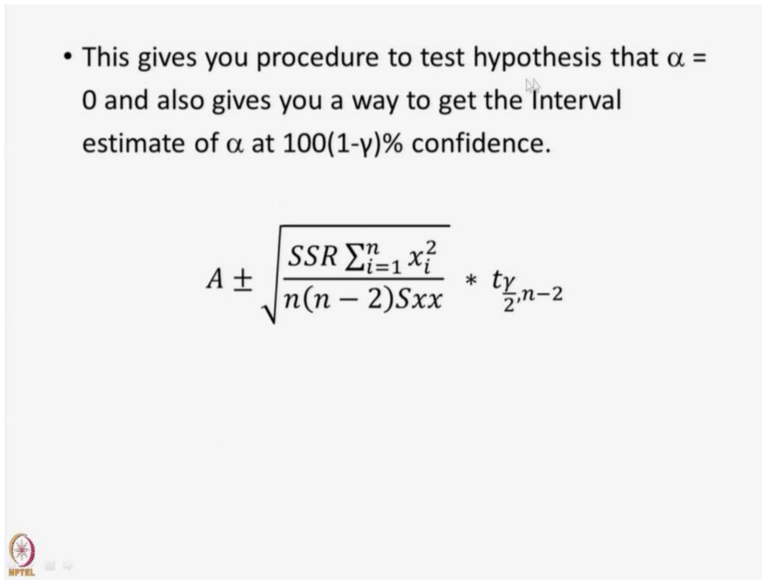
$$\boldsymbol{\frac{SSR}{\sigma^2} = \frac{SxxSYY - SxY^2}{Sxx} \sim \chi^2(\mathbf{n-2})}$$

Therefore,

$$\sqrt{\frac{n(n-2)Sxx}{SSR \sum x_i^2}}\ \ (A - \alpha) \sim t_{n-2}.$$

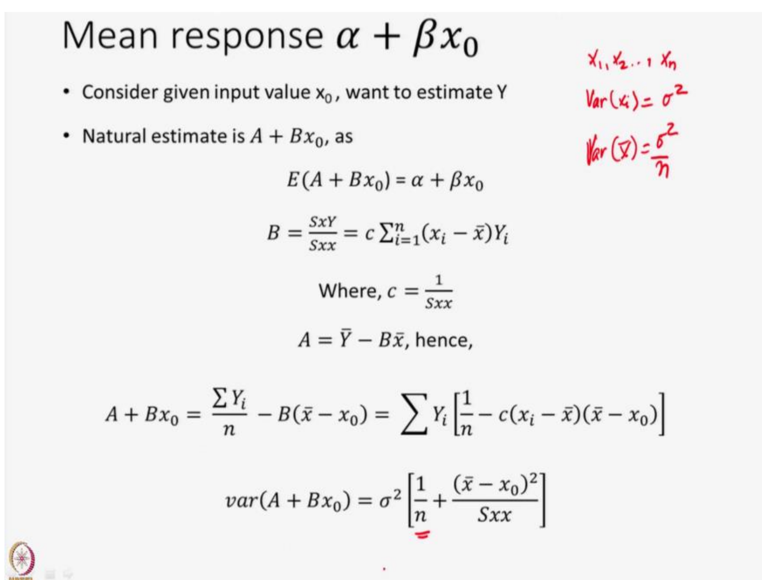
$$A \pm \sqrt{\frac{SSR \sum_{i=1}^{n} x_i^2}{n(n-2)Sxx}} \ * \ t_{\frac{\gamma}{2}, n-2}$$

(Refer Slide Time: 11:05)



And therefore once again we can set up the procedure to test the hypothesis that alpha is equal to 0 and it also gives you, gives us a way to get an interval estimate of alpha at 100 times 1 minus gamma percent confidence in this manner.

(Refer Slide Time: 11:26)



Now we come to the next interest. See a number of times our interest lies in considering the output or the response. When the input value is x0, some given value x0. This can be said that suppose you are conducting an experiment with some variation in temperature. So, temperature becomes your independent variable and you are checking out some response variable Y. Sometimes you are interested in finding out what happens at a given particular temperature. Now, this can looked into it in two ways.

The first way is what we are trying to look into here and this is called the mean response. This is called the, this is called the mean response. Now this mean response we call it. Because the natural estimate of alpha plus beta x0

$$E(A + Bx_0) = \alpha + \beta x_0$$

$$B = \frac{SxY}{Sxx} = c\sum_{i=1}^{n}(x_i - \bar{x})Y_i$$

Where, $c = \frac{1}{Sxx}$

$A = \bar{Y} - B\bar{x}$, hence,

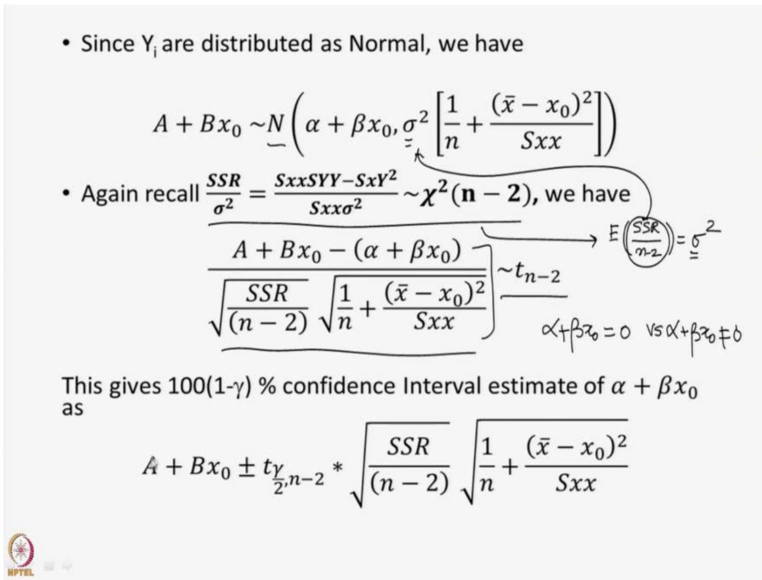
$$A + Bx_0 = \frac{\sum Y_i}{n} - B(\bar{x} - x_0) = \sum Y_i \left[\frac{1}{n} - c(x_i - \bar{x})(\bar{x} - x_0)\right]$$

$$var(A + Bx_0) = \sigma^2 \left[\frac{1}{n} + \frac{(\bar{x} - x_0)^2}{Sxx}\right]$$

At the same temperature when you are conducting an experiment several times, the average value of that experimental result is going to be summation Yi and then it turns out that this summation is going to be the expected value that is the expected value of that Yi is going to be A plus or rather alpha plus beta x0.

So, the S the estimator of that is going to be Yi multiplied by the constant as it comes here. Looking at this and then the variance of this value will be exactly sigma square multiplied by this quantity, this you have to simply calculate it, from that 1 over n comes because of it is a y bar. So, therefore the quantity 1 over n comes here. Please remember, let us I think it is a time to recall that if X1, X2, Xn are given and you know that variance of xi is sigma squared. Then the variance of X bar is given as sigma square by n and therefore here this 1 over n term comes.

(Refer Slide Time: 15:04)



Since Y are distributed as normal,

$$A + Bx_0 \sim N\left(\alpha + \beta x_0, \sigma^2\left[\frac{1}{n} + \frac{(\bar{x} - x_0)^2}{Sxx}\right]\right)$$

- Again recall $\frac{SSR}{\sigma^2} = \frac{SxxSYY - SxY^2}{Sxx\sigma^2} \sim \chi^2(\mathbf{n-2})$, we have

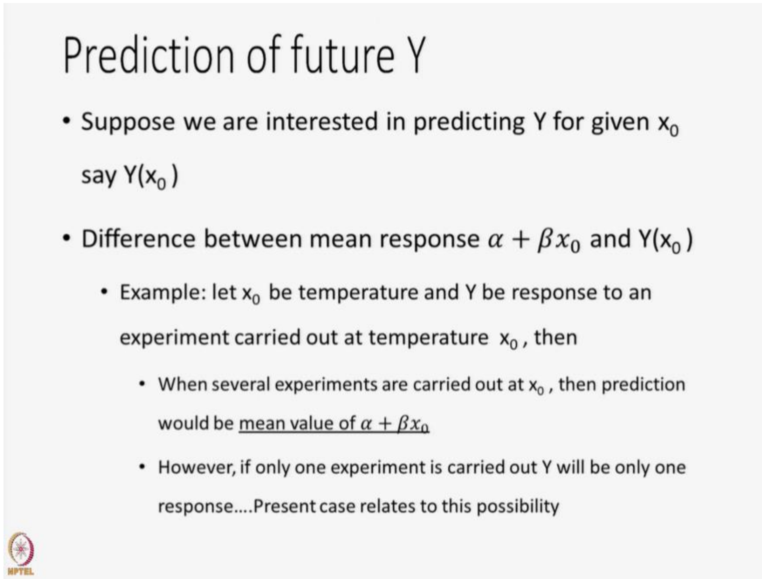
$$\frac{A + Bx_0 - (\alpha + \beta x_0)}{\sqrt{\frac{SSR}{(n-2)}}\sqrt{\frac{1}{n} + \frac{(\bar{x} - x_0)^2}{Sxx}}} \sim t_{n-2}$$

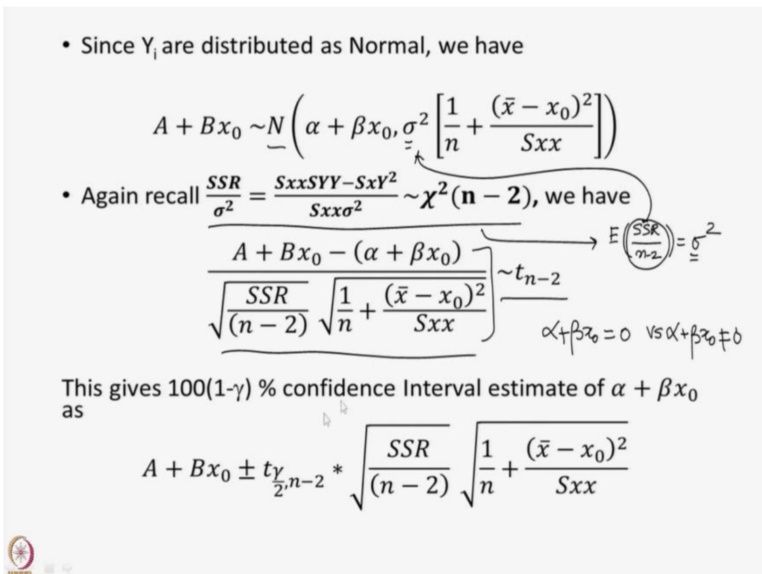
This gives 100(1-γ) % confidence Interval estimate of $\alpha + \beta x_0$ as

$$A + Bx_0 \pm t_{\frac{\gamma}{2}, n-2} * \sqrt{\frac{SSR}{(n-2)}}\sqrt{\frac{1}{n} + \frac{(\bar{x} - x_0)^2}{Sxx}}$$

And to test the hypothesis now we have a test statistic and the same test statistics will give us a way to have the interval estimator at 100 times 1 minus gamma percentage confidence of alpha plus beta x0 that is the mean response value at x0 as this.

(Refer Slide Time: 17:35)



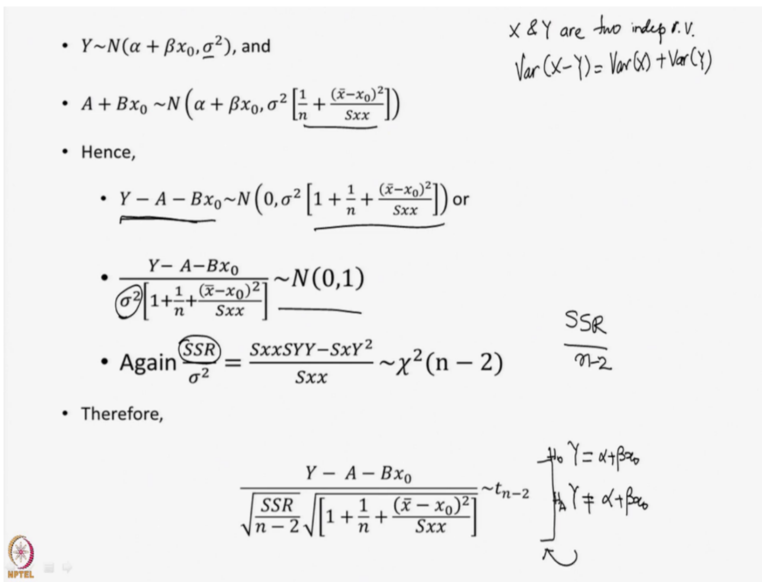


Now we move to the prediction of future Y. See here there is some confusion can exist, so I would like to say spend some time in clarifying this. Suppose we are interested in predicting Y for a given x0. It means that a good example would be a weather prediction. Suppose I want to predict weather tomorrow morning. Then in that case. I have to know the weather conditions which are independent variable x0 for tomorrow morning and I should also know that in general when such conditions prevailed in past what is the average response?

So in the previous session, what we have estimated is an average response, mean response when the value of the weather condition is like x0 and now what we are trying to do is we are trying to predict what is exactly going to happen tomorrow. So, it is going to depend on both. So, this is the difference between the mean response $\alpha + \beta x_0$ and the exact response $Y(x_0)$. Here I have given a example, once again of the temperature where Y be a response and x0 be a temperature. When the experiment is carried out, then several experiments are carried out at x0.

We would like to estimate mean value of alpha plus beta x0. However, when one experiment is carried out. So, like one weather forecast is to be carried out, in that case we have to work out with the mean response for not the mean response. But the one response of Yi and the case we are talking about is this case.

(Refer Slide Time: 19:43)



So, let us start Y is distributed as normal

$$Y \sim N(\alpha + \beta x_0,\ \sigma^2), \text{ and}$$

$$A + Bx_0 \sim N\left(\alpha + \beta x_0, \sigma^2\left[\frac{1}{n} + \frac{(\bar{x} - x_0)^2}{Sxx}\right]\right)$$

Therefore, if you take the difference of the two, the difference is distributed since the mean values are same. The difference of a normal random variable is also normal and the difference is going to be the mean difference of the two means which is 0 and the difference of the variable has the variance as summation of two random variables. Please recall, please recall that if X and Y are two independent random variables. Then variance of X minus Y is variance of X plus variance of Y.

So, the same principle is being used here and you get this quantity. It is this, plus one of this and therefore you find that, if you divide this quantity with its variance, it is distributed as standard normal distribution variate, which is normal with mean 0 and variance 1.

$$Y - A - Bx_0 \sim N\left(0, \sigma^2\left[1 + \frac{1}{n} + \frac{(\bar{x} - x_0)^2}{Sxx}\right]\right) \text{ or}$$

$$\frac{Y - \ A - Bx_0}{\sigma^2\left[1 + \frac{1}{n} + \frac{(\bar{x} - x_0)^2}{Sxx}\right]} \sim N(0,1)$$

Again

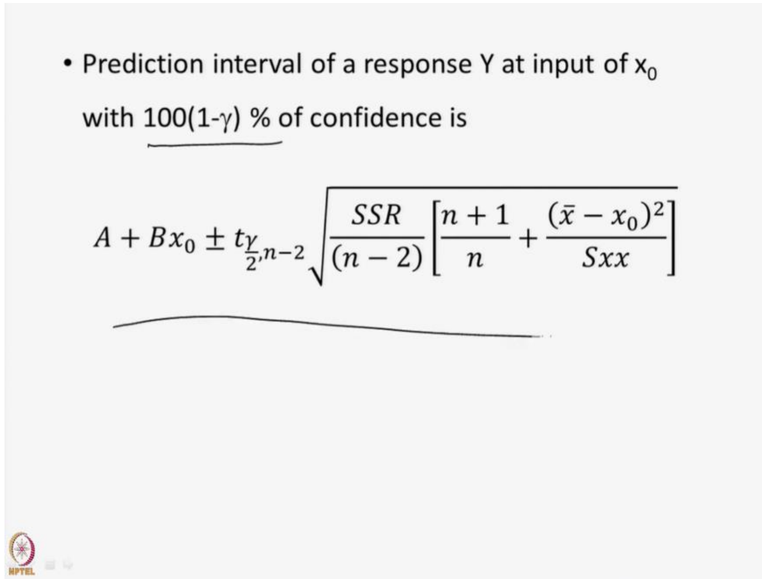
$$\frac{SSR}{\sigma^2} = \frac{SxxSYY - SxY^2}{Sxx} \sim \chi^2_{.}(n-2)$$

Again sigma square is not, note it has to be replaced by sums of squares of residuals divided by its degree of freedom n minus 2 and therefore this quantity, we find is distributed as t distribution with n minus 2 degrees of freedom. Because the quantity this and the quantity sums of squares of residuals are who independent random variables and therefore again.

We have a way to test the hypothesis that Y is equal to alpha plus beta x0 versus Y is not equal to alpha plus beta x0. These are the two hypotheses we can test using this t distribution with n minus 2 degrees of freedom.

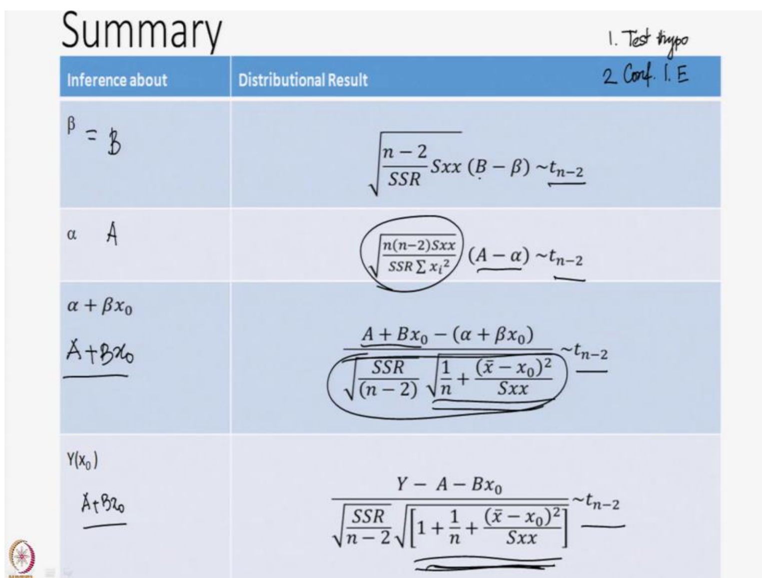
$$\frac{Y - \ A - Bx_0}{\sqrt{\frac{SSR}{n-2}}\sqrt{\left[1 + \frac{1}{n} + \frac{(\bar{x} - x_0)^2}{Sxx}\right]}} \sim t_{n-2}$$

(Refer Slide Time: 22:57)



And the same relationship or the distributional facility we can use to predict the interval response for Y at input level of x0 and with the 1 minus gamma times a 100 percent confidence and that can be given by this interval.

(Refer Slide Time: 23:23)



So in summary, let us go through it. We have done the inference about four quantities. These quantities are, these quantities are beta and alpha and then we also worked with the mean response which is alpha plus beta x0, mean response at a independent variable value x is equal to x0 and a prediction of Y at x0. These quantities can be written in a distributional way as beta will be distributed as B minus beta that is the estimated value of B.

Let us write it down, here we find that beta is estimated by B and that B has a distribution of B minus beta multiplied by this quantity as t distribution with n minus 2 degrees of freedom. Alpha is estimated by A and this A minus alpha multiplied by this quantity also has a t distribution with n minus 2 degrees of freedom.

The mean response of alpha plus beta x0 will be estimated by A plus B x0 that is here and that also along with this divider has the same t distribution with n minus 2 degrees of freedom and if you want to predict Y x0, it will also be predicted as A minus a plus b x0. Please remember this there is a sum difference between the two and this difference will be estimated by this. But the variance will be different. Note that the two variances are different, because we are estimating different quantity.

However, once the correct divider is used the distribution is again t with n minus 2 degrees of freedom. They give us two things, number one, way to test the hypothesis and it also gives you the way to have confidence interval estimator of each of these quantities. So, with this summary next we will move on to the next session of regression analysis.