Introduction to Uncertainty Analysis and Experimentation

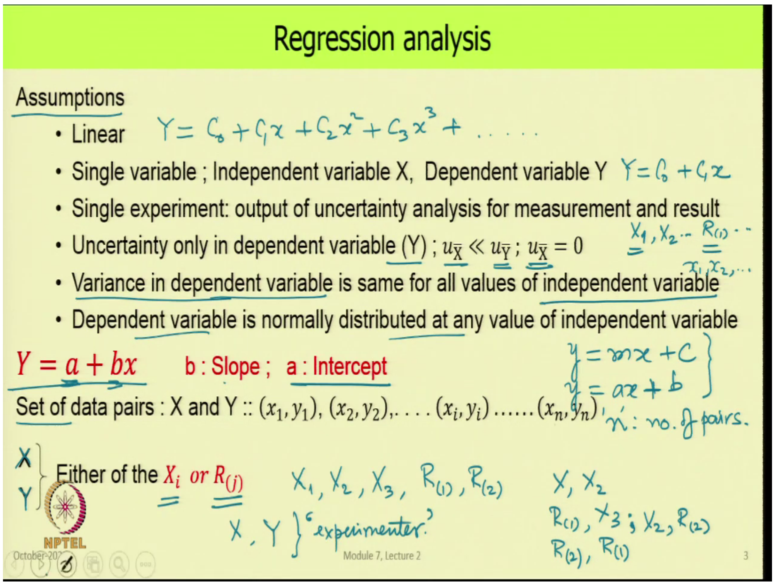
Prof. Sunil R. Kale
Department of Mechanical Engineering
Indian Institute of Technology, Delhi

Module - 07
Data Analysis
Lecture - 02

Regression Analysis - Linear, single variable

Welcome to the course Introduction to Uncertainty Analysis and Experimentation. In this module on Data Analysis, we will look at Regression development which is particularly linear regression in a single variable. So, in the previous lecture we had developed some of the arguments from which we will develop the regression and here those assumptions I have been listed.

(Refer Slide Time: 00:44)



Regression analysis

Assumptions

- Linear $Y = C_0 + C_1 x + C_2 x^2 + C_3 x^3 + \ldots\ldots$
- Single variable ; Independent variable X, Dependent variable Y $Y = C_0 + C_1 x$
- Single experiment: output of uncertainty analysis for measurement and result
- Uncertainty only in dependent variable (Y) ; $u_{\bar{X}} \ll u_{\bar{Y}}$; $u_{\bar{X}} = 0$ $X_1, X_2 \ldots R_{(1)} \ldots$ $x_1, x_2, \ldots$
- Variance in dependent variable is same for all values of independent variable
- Dependent variable is normally distributed at any value of independent variable

$Y = a + bx$ b : Slope ; a : Intercept

$\left.\begin{array}{l} y = mx + C \\ y = ax + b \end{array}\right\}$

Set of data pairs : X and Y :: $(x_1, y_1), (x_2, y_2), \ldots . (x_i, y_i) \ldots\ldots (x_n, y_n)$, 'n': no. of pairs.

$\left.\begin{array}{l} X \\ Y \end{array}\right\}$ Either of the X_i or $R_{(j)}$

$X_1, X_2, X_3, R_{(1)}, R_{(2)}$

$\left.X, Y\right\}$ 'experimenter'

X, X_2

$R_{(1)}, X_3 ; X_2, R_{(2)}$

$R_{(2)}, R_{(1)}$

October-202

Module 7, Lecture 2

3

The first thing we said we will develop only linear regression and then which means that our result the function Y on for which you make a regression this is like C 0 plus C 1 x plus C 2 x square plus C 3 x cube and there could be more terms. We further qualified by saying that we will look at only a single variable where the independent variable is X and the dependent variable is Y.

So, this expression above will then get reduced and will become just C 0 plus C 1 x. Then we said that we will look at the single experiment that is output of uncertainty

analysis of the measurement and result when we do one experiment we collected all the data, we got the values of X i which in this case is only X 1 and X maybe X 1, X 2 like that and results R 1, R 2 like that. And, we picked up one of these and one of any one of these any pair, any combination of two from these.

Then we said that uncertainty is there only in the dependent variable Y. In other words, uncertainty in the independent variable is very much less than uncertainty in the dependent variable. And for all practical purposes we could say that u X bar which is uncertainty in the dependent variable or variance in the dependent in the independent variable this is 0.

We also then said that irrespective of the value of the independent variable; that means, whether we took x 1, or x 2 whatever values we took the variance in the dependent variable y will be the same at every point. So, that is the next assumption that is here that the dependent variable the variance in the dependent variable is same at all values of the independent variable.

And, then we also said that variance in the dependent variable is normally distributed at any value of the independent variable. So, now, what we have before us is that we have a regression of the form Y is equal to a plus bx; b is called the slope, a is called the intercept. You will find in some books this is written as y is equal to mx plus c, somewhere it is also written as y is equal to ax plus b, some other symbols are used [noise.].

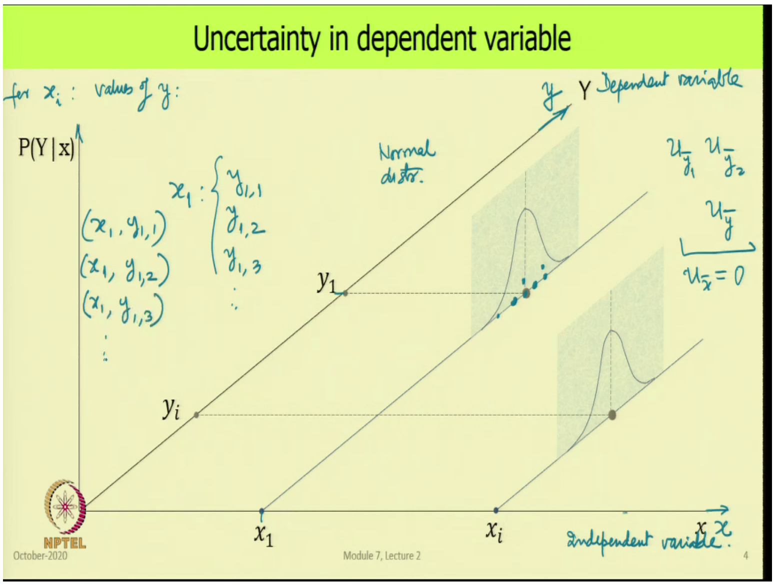
We will use this particular form for developing our analysis. The only restriction here is a is our intercept and b is the slope. The input data points that we have is a set of data pairs in X and Y, which we will denote as x 1, y 1, x 2 y 2; that means, value of y at x 1, value of y at x 2, value of y at x i, value of y at x n. So, there are n number of pairs.

And, when we map this X and Y that we have got here now into what we got from our experiment. We said that either of these could be either of the X i's or either of the R j's. So, in an experiment we could have a three parameters say X 1, X 2, X 3 and using that we generated two results R 1, R 2. And what we are saying is any of these could be our independent variable X and any of these remaining could be our dependent variable Y.

This selection depends on the experimenter you need knowledge of the experiment to decide which is the independent variable and which is the dependent variable. So, X could be X 1 Y could be X 2 one possibility that you are looking at correlating and measurement with another measurement.

One could be R 1 and other could be a measurement say X 3. So, we have making a correlation of one result with a measurement or it could be that you are making a correlation between two results. And, in this case we could also have the opposite say X 2 and R 2. So, any of these combinations are possible. It is at the experimenter you have to make this decision.

(Refer Slide Time: 06:19)



Uncertainty in dependent variable

October-2020 Module 7, Lecture 2 4

So, let us illustrate what those assumptions meant and this is something that we looked at in the earlier lecture let us quickly revise that. What we are saying is that we have a 3-dimensional plot on this axis here we have x on this axis here, we are plotting y and, this is in the x-y plane and normal to that, here we are plotting probability of y for a given x.

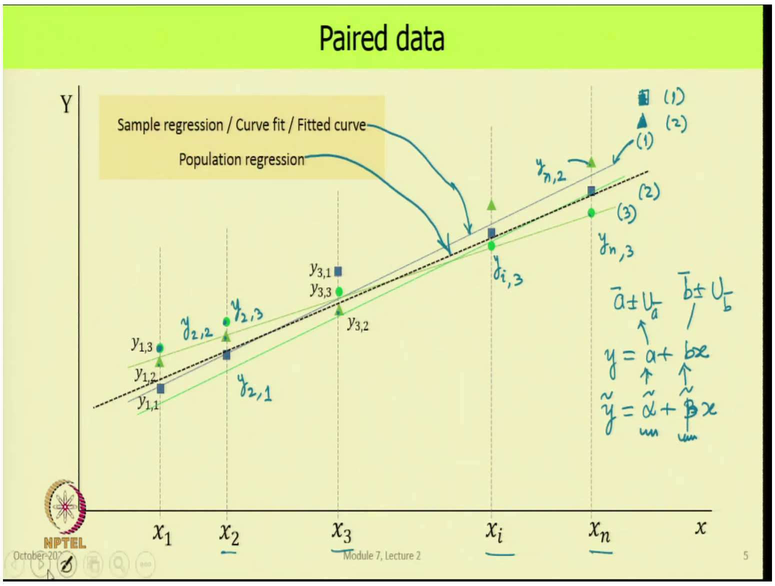
The idea being that values of y for any given x, x i values of y can vary from experiment to experiment. So, what we have here is a plot of that variation in y. So, at this x 1, this is y 1, this is y 1 this is x 1 and this is our data point. This is the mean value. When we did the experiment we did not get this x 1, but we got once we got this value, once we got this value, once we got this value.

Another time we got this value, sometimes we repeated that value, the another time we got this value and next time we got this value. So, like that if we keep plotting it in the limiting case when we have a population this will end up becoming like a curve and this will be a normal distribution. This is the assumption we wrote few minutes back. So, all the y i's will follow this.

So, each one of these points this was all y 1, but something came from the first experiment, something came from the second experiment, something came from the third experiment and so on. So, this is what these points are. But, they are all at the same value of x, x 1 and we will have some variance in y. This is from our symbols that we have used so far standard uncertainty in this will be u y bar or more strictly if we are saying it is same at any value of x.

So, at x i it will be this curve. So, we utilize u i y 1 bar or u y 2 bar they are all same and we can denote this as u of y bar and this is under the assumption that u of x bar is 0. So, that sets up our nature of the problem that we have variability in the dependent variable y, this was our dependent variable and there is little or no uncertainty in x which is our independent variable.

(Refer Slide Time: 09:41)



Paired data

Sample regression / Curve fit / Fitted curve

Population regression

So, now let us see what happens. What we are saying here is that in the first experiment you would get this x 1 and this y then next time you do an experiment you would get the same x 1, but maybe this y. So, our data pairs will be x 1, y 1, 1 from the first experiment ; same value if we do it x 1, y 1 from the 2nd round; x 1, y 1, 3 from the 3rd round and so on.

And, this next plot we will start putting it them together and we say that x 1 we got y 1 from the first experiment, y 1 from the second experiment, y 1 from the third experiment and this we will do for all different values of the independent parameter which here are indicated as x 2, x 3, x i, x n.

So, this square represents the first experiment. So, if we have to make a legend we can say that this is a square and this is experiment number 1 for which we have these data points, y this is y 2 for the first experiment, y 3 for the first experiment, y 4, y i, y m. And, with all these this experimenter who got only this pair can make a regression and get this line here which is regression line number 1.

The first experiment gave us this regression. We will see in a minute what equation it is that we put for this. So, now, we move to the argument that let us look at the second experimenter this person got the triangular data points and we call this experiment number 2.

So, this was y 1 from the second experiment, this was y 2 from the second experiment, this is y 3 from the second experiment and like that this is y n from the second experiment this point. So, using these points only the only the triangles we can make yet another regression not knowing that a regression on the first experiment was already there.

And completely independent of that and so, we get; get one more of the a correlations or the regressions and that we call as number 2. And, then we make say that there is a third experiment which has been done and so, at x 1 the value is this circle here y 1, 3; here there is y 2, 3; this circle is y 3, 3 this circle is y i, 3 and this circle is y n, 3.

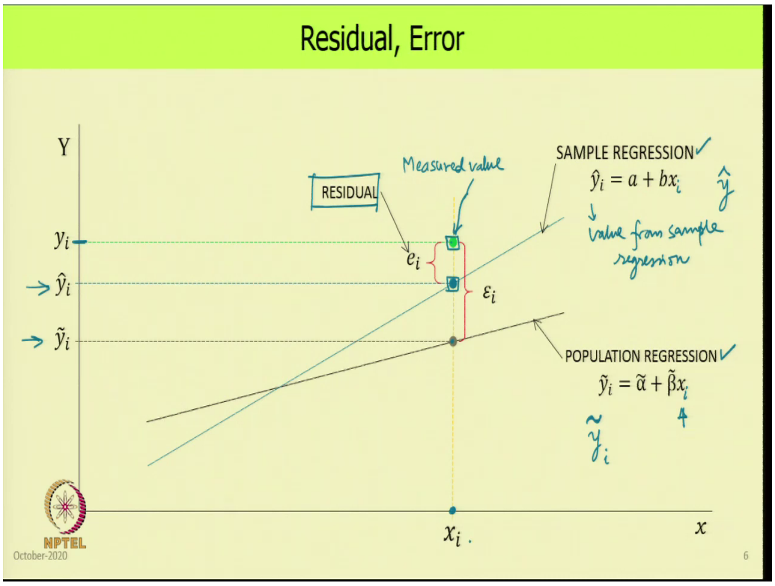
And, using all of these the third person can also make a regression and say this is our third regression. So, we would have three regressions from three different sets of data and each one of these; these are called the sample regression or the fitted curve. When we say we do a curve fit, this is what we mean. This is somewhere referenced as fitted curve. It is rigorous name is sample regression.

And, what we want is that if I can consider all this type of data together, then I would get yet another regression which is this thick dotted black line and this is our population regression. We will never know the population regression like we have argued in the earlier cases of uncertainty analysis.

So, our objective then becomes that from our sample regression or from our curve fit we get some regression constants which in our expression will become y is equal to a plus bx. And, the population regression we could say y tilde is equal to alpha tilde plus beta tilde x and our a's and b's are point estimators of alpha tilde and beta tilde.

And, to know where these lies we will finally, have to express these as an interval estimate some a bar plus minus capital U a bar and in this case b bar plus minus capital U b bar that is our objective. Now, what is a how do you calculate a bar b bar and how do you calculate their expanded uncertainties and those are estimators of this and this.

(Refer Slide Time: 15:35)



Residual, Error

SAMPLE REGRESSION ✓

$\hat{y}_i = a + bx_i$

POPULATION REGRESSION ✓

$\tilde{y}_i = \tilde{\alpha} + \tilde{\beta}x_i$

To do that, we first make one important definition. And, what we are plotted here is just one case, one point on the x-axis it is the independent variable x. And here we have taken one point x i and from the experiment we got this point which is y i. So, we can say that this one, this is our point that came from the measured value the experiment. This is what we know.

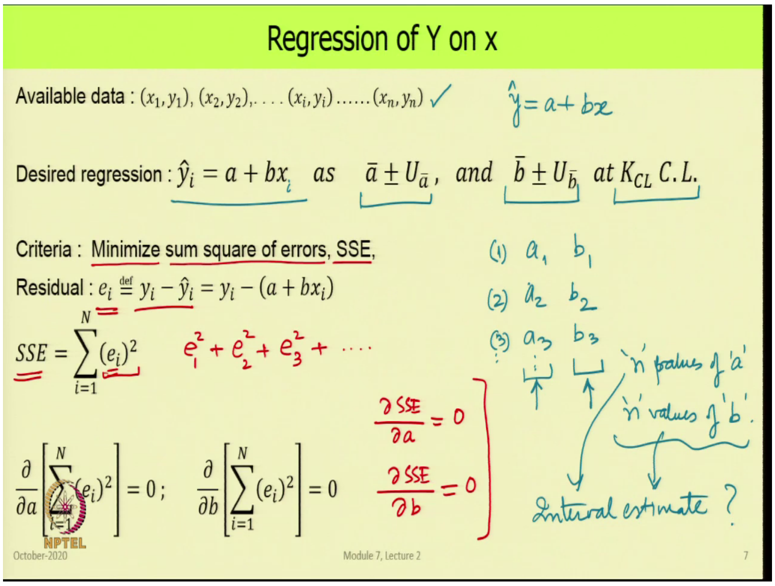
When we develop our sample regression we get this line here, which is the sample regression line. And we denote that as y hat i is equal to a plus bx we can put i there which means the hat denotes that this is a projected value from sample regression and y hat has been i-th denote that this is the i-th reading.

For x i the corresponding value from the correlation is y i which is this point and this is here as y i hat. That the point we have got from a sample regression. And, so, there is a difference between these two this is what we call e i we call it the error, but the correct technical name there is that this is our residuals. And this plays a very important role in many things that we do, many thing that we interpret.

The next thing that we have on this curve is yet another line which is this line, which is our population regression which means that if we had infinite number of y data points for every x i and using that we develop our correlation. Then we would get this expression y i tilde is equal to alpha tilde plus beta tilde x i. That means, we are putting in a value of x i and the number that comes out we are denoting it by y i tilde the tilde denotes that this is from the population regression.

So, hat denotes data coming from the sample regression the tilde denotes data coming from the population regression and this is our point. So, the difference between this value and this value this entire thing this is what we called as epsilon i and e i is an estimator of epsilon i. So, this is an important definition and we will write down the expression for residual in a minute.

(Refer Slide Time: 19:09)



Regression of Y on x

Available data : $(x_1, y_1), (x_2, y_2), \ldots (x_i, y_i) \ldots\ldots (x_n, y_n)$ ✓ $\hat{y} = a + bx$

Desired regression : $\underline{\hat{y}_i = a + bx_i}$ *as* $\underbrace{\bar{a} \pm U_{\bar{a}}}$, *and* $\underbrace{\bar{b} \pm U_{\bar{b}}}$ *at* $\underbrace{K_{CL}\ C.L.}$

Criteria : Minimize sum square of errors, SSE,

Residual : $e_i \overset{\text{def}}{=} y_i - \hat{y}_i = y_i - (a + bx_i)$

$$SSE = \sum_{i=1}^{N} (e_i)^2 \qquad e_1^2 + e_2^2 + e_3^2 + \cdots$$

$$\frac{\partial}{\partial a}\left[\sum_{i=1}^{N}(e_i)^2\right] = 0; \quad \frac{\partial}{\partial b}\left[\sum_{i=1}^{N}(e_i)^2\right] = 0 \qquad \left.\begin{aligned}\frac{\partial SSE}{\partial a} &= 0 \\ \frac{\partial SSE}{\partial b} &= 0\end{aligned}\right]$$

(1) a_1 b_1

(2) a_2 b_2

(3) a_3 b_3

'n' values of 'a'

'n' values of 'b'.

Interval estimate ?

October-2020 Module 7, Lecture 2 7

So, how do we develop regression of Y on x? We have all these data points x 1, y 1, x 2, y 2, x i, y i, x i, x n, y n this is what we have. The desired regression is y hat is equal to a plus b i or this or the better way to write is that y hat is equal to a plus bx and we want the answer as a plus minus U a bar b plus minus U b bar at certain confidence level.

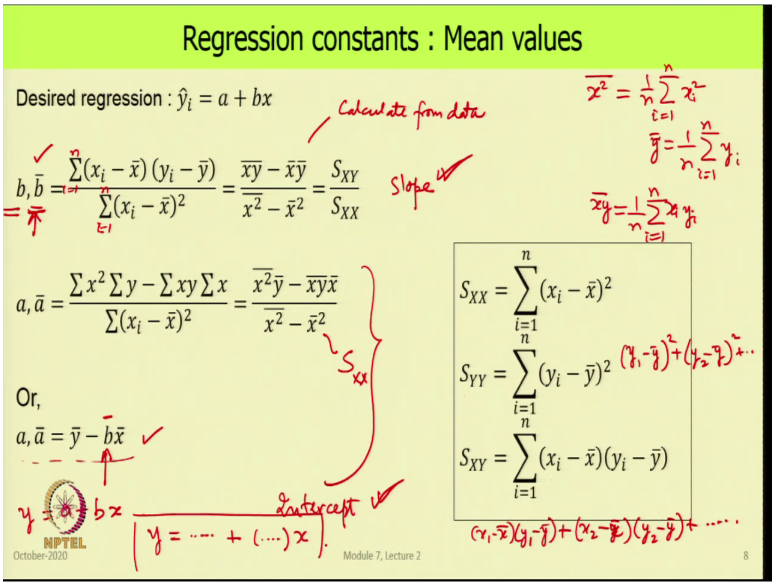
And, the strategy we do is that a and b will take different values depending on which regression you are looking at. So, from first data set we will get a 1 b 1; from the second data set we will get a 2 b 2 like that. So, all these a's they will themselves form a sample and all the b's will also form yet another sample. And we look at the treatment of these two variables and work out what these constant values could be.

So, what we do is we, but with the residual that we just saw e i which is defined as y i minus y i hat. The earlier graph showed that the value that we have measured minus the value we get from our sample regression and this value from sample regression is nothing, but a plus bx i. So, residual is y i minus a plus bx i.

And, then our strategy is we minimize this sum square of errors, which is SSE. So, we are minimizing SSE Sum Square of Errors, which means we have taken each error e i squared it which is what we get here. This is for every point individually. So, this is e 1 square plus e 2 square plus e 3 square plus so on.

We square it and this is what we call SSE sum square of errors and our objective now becomes that we will minimize this SSE with respect to a and b. So, that means, that d SSE divided by da. This we want to set to 0 and d SSE divided by b del b this we want to set to 0. If we do that, we will get the values for a and b. We can do the simple mathematics, write the full expression, differentiate it and set to 0 and the answer that we get is like this.

(Refer Slide Time: 22:37)



Regression constants : Mean values

Desired regression : $\hat{y}_i = a + bx$

Calculate from data

$$b, \bar{b} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{S_{XY}}{S_{XX}}$$

Slope ✓

$$\overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$\overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i$$

$$a, \bar{a} = \frac{\sum x^2 \sum y - \sum xy \sum x}{\sum (x_i - \bar{x})^2} = \frac{\overline{x^2}\bar{y} - \overline{xy}\bar{x}}{\overline{x^2} - \bar{x}^2}$$

S_{xx}

Or,

$$a, \bar{a} = \bar{y} - \bar{b}\bar{x}$$ ✓

$y = a + bx$

$y = \cdots + (\cdots)x$

Intercept ✓

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

$$S_{YY} = \sum_{i=1}^{n}(y_i - \bar{y})^2 \quad (y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \cdots$$

$$S_{XY} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

$(x_1 - \bar{x})(y_1 - \bar{y}) + (x_2 - \bar{x})(y_2 - \bar{y}) + \cdots$

NPTEL

October-2020

Module 7, Lecture 2

8

That the value of b or rather we can say the mean value of b is this summation over i equal to 1 to n for the product of x i minus x bar y i minus y bar divided by i equal to 1 to n x i minus x bar square. And, in some notations that we develop we get a some short forms we define S XX; S as capital S is the sum is summation of x i minus x bar whole square. S YY S y i minus y bar whole square summed up over all the values.

So, for instance for this will be y 1 minus y bar square plus y 2 minus y bar square and so on. Where did we get the y bar? y bar was nothing, but 1 upon n summation of y i, i equal to 1 to n simple arithmetic mean of all the y values; x bar is the mean of all the x values.

So, we calculate this difference, square it. Similarly, we do it for every point data point and add them up that is called S YY. So, it has the dimensions of the square of the dimension of y. Similarly, this will have dimensions of square of the dimension of x and then we define S XY as summation of x i minus x bar into y i minus y bar.

So, what we do here is we take x 1 minus x bar into y 1 minus y bar add to that x 2 minus x bar times y 2 minus y bar and so on. This is our S XY. We do not need to do all these detailed calculation if we are working in a problem because these things are there in most spreadsheets, statistical packages, even in calculators. You just press input all the data points and just press one key and the answer comes out, but we should know what is it that they are doing.

So, in that definition the mean value of the slope so, this is slope b is S XY upon S XX and you can see here that this will have the units of the units of Y divided by the units of X. So, when we put it there this whole term will have units of y which is consistent. We can further expand this a little bit divide it by n define various means and we get xy bar minus x bar y bar upon x square bar minus x bar square.

So, like we defined here xy bar this is 1 upon n summation i equal to 1 to n product of x i, y i and x square bar is. So, x square bar is 1 upon n summation i equal to 1 to n x i square.

So, all these terms on the right side this we can calculate from data. We have all the values first we get x i y i pairs from that we can calculate all these summations or we can directly calculate these means and then we plug into this formula and get the value of the mean value of the slope.

Most textbooks and references you will find this written as just b, but in this terminology the symbols that we are using for uncertainty analysis this is the parameter whose value we want to estimate. So, we are saying that is the mean value plus minus some expanded uncertainty and in that nomenclature it will be called b bar. That will also help avoid confusion what are b and what is b bar then we want to calculate the intercept a.

For this we have two options: first we calculate b bar which is this one that we got here and then we use this formula which is a very simple straightforward one y bar minus bx bar which nothing came from y is equal to a plus bx to which I said a is y minus bx b we calculate this way we just put the values in here and we have the value of a bar.

The other option is to do a little more involved calculation where we do the same thing that we did with b. Simplify it and what you get is sigma x square bar into sigma y minus sigma xy into sigma x upon sigma x i minus x bar square and if you further simplify it this becomes x square bar y bar minus xy bar x bar upon x square bar minus x square. The denominator you can see is nothing, but S XX.

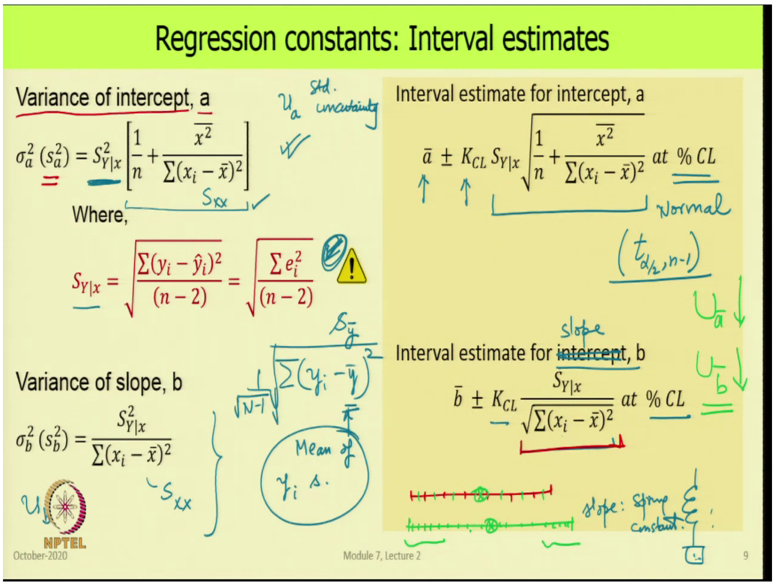
We have already calculate calculated all these parameters for the calculation of b. So, either we can just use those numbers in this formula and get a bar or we can do a this calculation and get the value of a bar. So, there we are we have got the value of the slope here and either of these methods we got the value of the intercept.

And, so, we can now write y is equal to sum number plus sum number times x or you can call this y hat if you want to do that. In the expression itself the hat will not be there when we calculate a parameter then we should qualify that this is coming from a sample regression. This is what our objective in this exercise was and we got it.

So, this is half the story what we saw is that because all these a 1's came from different experiments; from the 3rd experiment we had a 3, b 3 and so on. We got n values n pairs or you can say n values of a, which are a 1, a 3, a i, a n and n values of b. And, all these values are likely going to be different.

So, we cannot say that we are going to pick up this particular value or that particular value our objective then becomes for each one of these what is the interval estimate and that is our next job.

(Refer Slide Time: 30:57)



Regression constants: Interval estimates

Variance of intercept, a

$$\sigma_a^2\,(s_a^2) = S_{Y|x}^2\left[\frac{1}{n} + \frac{\overline{x^2}}{\sum(x_i-\bar{x})^2}\right]$$

Where,

$$S_{Y|x} = \sqrt{\frac{\sum(y_i-\hat{y}_i)^2}{(n-2)}} = \sqrt{\frac{\sum e_i^2}{(n-2)}}$$

Variance of slope, b

$$\sigma_b^2\,(s_b^2) = \frac{S_{Y|x}^2}{\sum(x_i-\bar{x})^2}$$

Interval estimate for intercept, a

$$\bar{a} \pm K_{CL}\, S_{Y|x}\sqrt{\frac{1}{n} + \frac{\overline{x^2}}{\sum(x_i-\bar{x})^2}} \text{ at } \%\,CL$$

Interval estimate for ~~intercept~~, b

$$\bar{b} \pm K_{CL}\frac{S_{Y|x}}{\sqrt{\sum(x_i-\bar{x})^2}} \text{ at } \%\,CL$$

October-2020

Module 7, Lecture 2

9

I will skip all the detailed statistical theory that lies under these formula. You can read it from the book. For now, I will just present the final answer that we get and what we have is the variance of the intercept a, sigma a square or its indicator s a square is given by this formula.

So, here is a new term that is coming up. S Y given x square in bracket 1 upon n, n is the number of data pairs, we have no problem with that plus x square bar we calculated that we know how to do it over summation x i minus x bar square this is also known to us and again we recognize that this is nothing, but S XX.

So, using this S XX, S YY terms this makes it easy to write formula and later on when we do analysis of variants and many other things these sums come in very handy. The question is this came from the data, this all came from the data. So, everything in the square term square brackets this comes from the data, there is no problem with that.

What about this one? So, here there is a important definition coming up that we define S Y x as square root of y i minus y hat square upon n minus 2. Where y i minus y hat is summation of e i square divided by n minus 2; n minus 2 comes 2 comes. Because we got 2 less degrees of freedom, one was the mean and the other was the mean value of the other parameter.

So, this caution symbol tells us that this is not the same as what we have as S y or S y bar. This was standard deviation of the mean or the standard uncertainty, the square of this to the standard variants. But, in this definition we always calculated the value based on y i minus y bar square and then divided by n minus 1.

This was the mean of the y's. Here it is the residual where the difference is relative to the value that you get from the sample regression y i hat where this was y bar. We can see that if uncertainty in x is 0, these two values will be pretty much the same. In a minute we will come back to this.

Now, can we have a simpler relation because we cannot go about calculating y hat every time, but what we have here now, is that we have a formula by which we can estimate what we have called so far the standard uncertainty in a. So, we will try to go back to our definition that we use throughout our uncertainty analysis and that is u a.

Now, variance of the slope and we want to calculate what is now as u b and that is given by this relation here this is S Y x square upon summation x i's minus x bar square. Again, you can see the denominator is nothing but S XX. So, we got all the relations we wanted. We just need to get a little more workable formula for this one. So, we will see that in a minute, but to put things together now we got the mean value of the intercept from our first calculation earlier.

Now, we got the variance which is this much in a and this we multiplied it in our case by K CL which is related to a percentage confidence level. In many books you will find that K CL assumes that there is a normal distribution. And, this is what we have followed in all our uncertainty analysis saying that we will make life simpler for ourselves, we will assume it is a normal distribution instead of 1.96 for 95 percent confidence level we will take it as 2.

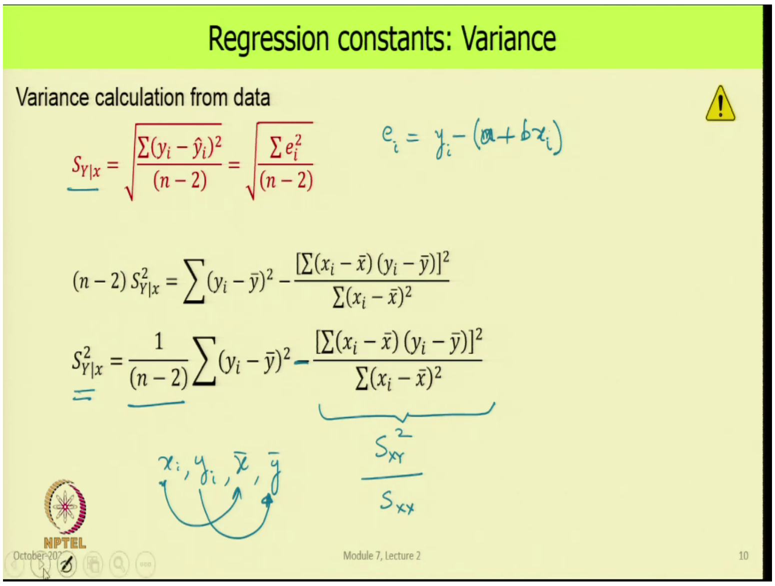
So, we are little more conservative on that. Strictly speaking this would have been like even in the other cases t alpha by 2 n minus 1. The value coming from the students t-distribution, but does not matter we will just round it off make it 2 and put our K CL over here like we have done throughout uncertainty analysis. We have earned a bit, but on the conservative side.

So, we are not under reporting the uncertainty which is something we should always avoid we are reporting it slightly larger than what it would be. And, then we can write the interval estimate for the intercept which is coming from this expression here b bar plus minus K CL square root of everything that we had over there S Y x upon summation x i minus x bar square at this is at a certain confidence level CL.

So, we got the expressions that we set out to get and in some experiments it could be that we were interested in the slope itself, we were not interested in the whole regression line. For example, if you have a spring and we had weights over there and measure the deflection, then in this case the slope that is of interest was is the spring constant.

So, that is an example where we are interested in the slope sorry, this is not intercept this should be slope. So, this is the interval estimate of the slope b, this is the interval estimate of the intercept a and that is everything that we were looking for. Only thing remaining is this symbol here caution that we do not have a workable formula for this y i hat has to be calculated separately can be simplified.

(Refer Slide Time: 38:56)



Regression constants: Variance

Variance calculation from data

$$S_{Y|x} = \sqrt{\frac{\Sigma(y_i - \hat{y}_i)^2}{(n-2)}} = \sqrt{\frac{\Sigma e_i^2}{(n-2)}}$$

$$e_i = y_i - (a + bx_i)$$

$$(n-2)\,S^2_{Y|x} = \sum (y_i - \bar{y})^2 - \frac{[\Sigma(x_i - \bar{x})(y_i - \bar{y})]^2}{\Sigma(x_i - \bar{x})^2}$$

$$S^2_{Y|x} = \frac{1}{(n-2)} \sum (y_i - \bar{y})^2 - \underbrace{\frac{[\Sigma(x_i - \bar{x})(y_i - \bar{y})]^2}{\Sigma(x_i - \bar{x})^2}}_{\frac{S_{xy}^2}{S_{xx}}}$$

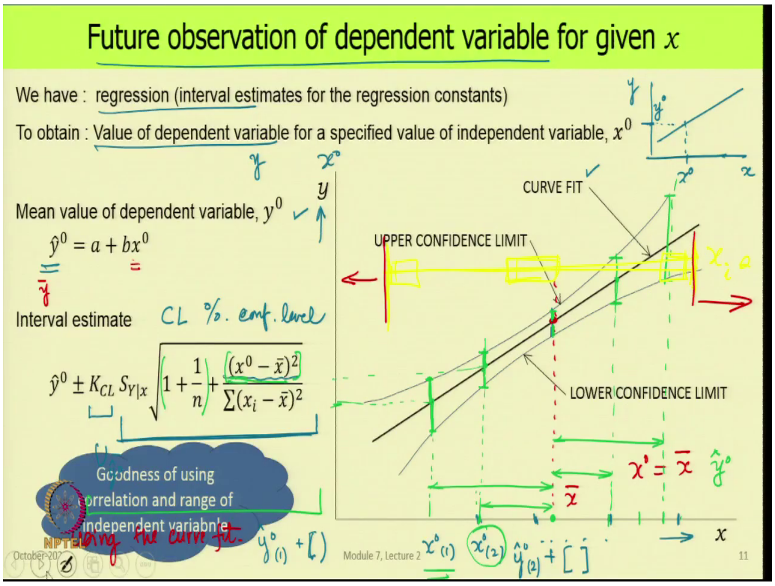
$x_i, y_i, \bar{x}, \bar{y}$

And, that is what we do over here. We pick up the definition of S Y x and from here we can simplify this put e i as what we have defined earlier as y i hat minus a plus bx i. No, sorry. no y i minus y hat which is what we had here. So, here y i minus a plus bx i and y i x i b i values are known to us; a and b has been calculated.

So, we can calculate this term, but we can further simplify it in terms of various summations. And this is the relations that we get that S Y x square is 1 over n minus 2 summation of y i minus y bar square minus summation x i minus x bar into y i minus y bar square upon x i minus x bar square. So, if you see the second term this is nothing but S XY square upon S XX.

So, this expression is in terms of all the numbers that we have with us, all the x i values, all y i values and also x bar and y bar which also came this came from here and this came from here. This we already calculated in our earlier calculation. So, we can just use it and calculate S square Y x. So, that gives us the complete methodology for getting the regression constants and their interval estimate.

(Refer Slide Time: 41:01)



Future observation of dependent variable for given x

We have : regression (interval estimates for the regression constants)

To obtain : Value of dependent variable for a specified value of independent variable, x^0

Mean value of dependent variable, y^0

$\hat{y}^0 = a + bx^0$

Interval estimate

$$\hat{y}^0 \pm K_{CL} S_{Y|x} \sqrt{1 + \frac{1}{n} + \frac{(x^0-\bar{x})^2}{\sum(x_i-\bar{x})^2}}$$

Goodness of using correlation and range of independent variable

Now, we will look at two things one we do in many practical cases we take a regression. And then we use it in the experiment itself by saying this is my independent variable value, what is the value of my dependent variable from the curve fit and we use that in our experiment.

So, this is what we call future observation of a dependent variable for given x. So, first we will see this; that means we are trying to predict what will be the value that will come out of this regression and it is interval estimate two thing we are doing together. In the next part, we will see that instead of asking for the future observation, we will say well, now y was distributed how can I predict it is mean value.

So, that, we will see in the next one. So, what we have? We have a regression interval estimates for the regression constant we got, and we obtain we want to obtain value of the dependent variable means y for a specified value of the independent variable x 0. That means, this is x, this is y we did the experiment and we produced a regression which was like this.

Of course, the regression constant had uncertainty. Then we picked up some value x 0 and using this value x 0 we went to a regression line and got this value y 0. So, what is this mean value and what is its uncertainty, that is what we will look at now. So, we have mean value of the dependent variable y 0 which is nothing, but we use the regression substitute x 0 value in value of x and we get the predicted value of y for that value which we denote by y hat because it is predicted value y 0 because it is corresponding to x 0.

So, this is the mean value of the dependent variable that we get, simple and straightforward. The interval estimate comes with a little involved expression one can go through all the statistics and rigorously derive this. I have skipped those steps and this is the final answer that we get that at CL percentage confidence level. The U in y hat 0 is this. K CL is our multiplication factor coming from confidence level.

This is an expression that we have got which is S Y of x which we have seen in the earlier slide how to calculate that square root of 1 plus 1 upon n plus x 0 minus x bar upon summation x i minus x bar square x bar and all the x i’s came from our data, x 0 is the input that we have given in this particular calculation.

Now, let us see the implication of this expression. Here we have plotted x on this axis and y over here. And our regression, sample regression which is our curve fit is this black line. So, that is our curve fit and now, we say if x 0 was here, here, here, there or there, what will be the uncertainty in x 0?

So, these are different values x 0 could take. We can say that this is x 0 1, we can say this is x 0 2 and like that and we are asking the question what is y 0 1 and it is uncertainty. In the second case what is y 0 from the second experiment, y hat these are predicted values and what is its uncertainty, right.

So, we will look back at this formula here and we find that there is x 0 minus x bar coming here. So, what we will do is we will say well I have all these x values from which I make this regression the mean of those x values I will denote somewhere there. So, this is x bar and that is what this line would look like at x bar the value of y predicted value of y will be that much.

So, if x 0 were same as the value of the mean that we had; the mean was say 15.2 and we got the correlation and then somebody say tell me the value at 15.2? Well, we will go back to the regression, put that 15.2 over here and get the value of y hat 0. This will match exactly with what we had got, but this need not be this may not be the same as the y bar value.

It will be the same as y bar value because this equation always passes to the mean point, but when you look at this expression this is x 0 minus x bar. So, if x 0 was equal to x bar, this first term here on the numerator this one this becomes 0. And, so, the uncertainty becomes square root of 1 plus 1 upon n.

And if you say that I have a very large sample of data pairs even this could be very small compared to 1, in which case this will also all of this will become 1 and we are left with just this little part here.

But, this part has become 0, but the rest of it remains and so, that we say is the uncertainty band that we have here and we have over there. So, we can say that this was the plus minus expanded uncertainty in the value of the predicted value of y 0. Now, let us say another point say this one x 0 2, this lies over here.

The mean value comes from the regression. So, that straightforward we substitute this value in the a plus bx relation and we get the value of the predicted y value which is this value. Then we say what is uncertainty here? Now, we go back to this expression and what we find that x 0 minus x bar is no longer 0 it is finite. And because it is squared it becomes a positive number, which adds up to this first two terms.

So, the uncertainty at this point is more than the uncertainty at the mean value. So, this is the larger uncertainty band over there compared to what we had over here. So, we would put up error limits like that. Then we take a third point and say I will come to x 0 1 bar. So, at this value we put it use the a plus bx relation and from there we get our corresponding y value which is over here. And, now we say that what is the uncertainty in this?

We will go back to this expression, now we put x 0 value that we have and we note that x 0 value in the first case now is greater than x 0 value in the second case all of which no, sorry. In this case it is. So, what we see here is that x 0 bar x 0 minus x bar which is this gap, this is more than the case in the second case.

So, we earlier had this as the gap x 0 minus x bar when it was the two point, when it is the first point the gap is this much it has increased. So, this numerator term over here has further increased which means that the overall interval estimate has also increased and so, that is what we see here this would be going from here to there. And now, what we see here is that the uncertainty as we are going away from x 0 on this side is increasing.

Now, we can take points on the higher side. So, if we are over there we get the mean value from the correlation and now, the difference is this much. This is then the numerator x 0 minus x bar and we get an uncertainty larger than what it was at x bar and which is what you see here.

And, if you go further away say over there and we go up and see what happens over there, this is now our term x 0 minus x bar it has become bigger. So, the uncertainty band has become bigger and so, we get the range something like that. These will all be symmetric.

So, here is what it tells us that when you use a linear regression for future value prediction your uncertainty at the value that you pick up is not the same everywhere closer to the midpoint of the regression. So, what it tells us x bar tells us that this is approximately somewhere in the middle range over which the correlation was made.

So, correlation was made in this range; that means, all the x i's were within this. And, when we use the regression for future value prediction at this approximately in the center region of this our uncertainty is going to be the smallest and as we go away from this either on the lower side. So, we take values over here or we take values over there, our uncertainty increases.

So, this is telling us something very important about how to use a regression to predict future values and one thing that will absolutely we should not do is that if the regression was made in these limits, we should be extremely cautious and maybe not do at all that use regression values for x values less than this or x values greater than this.

So, what it means is that if you calibrated something between these two values then use that regression for independent values between those values only not at higher values. For example, if you did calibration of a thermocouple from 0 degrees centigrade to 100 degree centigrade you should not use that regression to get the value at say 500 degree centigrade, that will be completely wrong.

Uncertainties of course, will be high maybe it would be completely wrong, maybe the linear assumption may not hold at all. All those things would come in. So, this is one big caution in using the regression or the curve fit. Never use it beyond the range over which it was generated in the first place, this is what we are saying.

All this also give you an idea that if I want to minimize the variability in a and b, what should I do and that you can look at these relations and say if I want to minimize the uncertainty in my slope what should I do. And, what one does? He said I will minimize this whole thing as much as possible how can I minimize x i minus x bar square? We take lots of data points, that is fine.

So, if this is a range over which you are doing calibration you could take equally space say 5 points, 7 point, 9 points. Somebody else comes and says no let us improve the uncertainty make it narrower, let us take twice as many points. This denominator term will increase further or our estimate of the interval of the U bar, U b bar this will go down which is good.

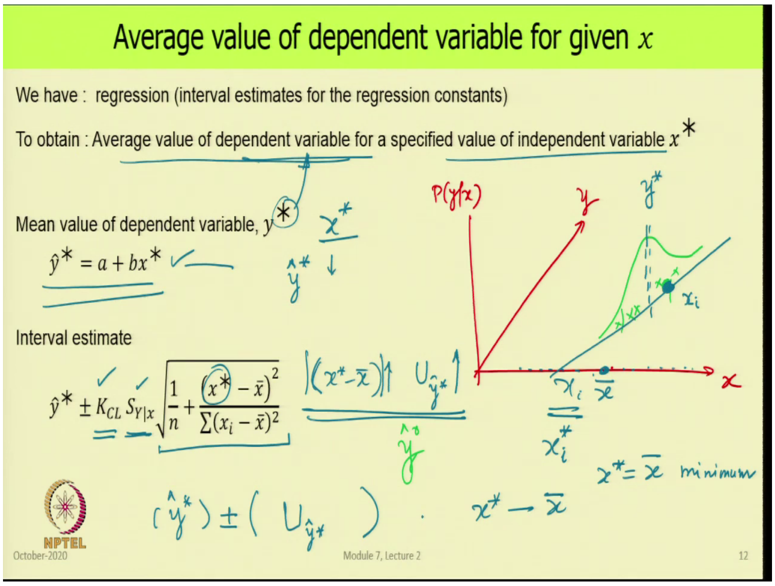
But, what it also tells us that well you need not take that many points at very small intervals. But do something else this is the range on which you want to take data points. Take lots of points over here, few points in the middle and then again take lots of points at the other end.

So, you are doing more measurements at the extremities of the range over which the regression has to be made. So, the x i's in general is larger in all the cases of course, x the mean value is the same. But, now that we have taken more values at the ends to make the regression your uncertainty in U b bar and also the uncertainty in a bar the U a bar both these will be going down.

So, this simple formula that we got is giving us a lot of insight on how to plan an experiment. So, if you look at the experimentation stages when we said that we have to make the test plan in the pre-uncertainty stage. Here is where some aspects of uncertainty are coming in and giving us very valuable information on how to do the experiment.

So, that was the point I was making that the goodness of using a relation, goodness of using a correlation that we have developed is best within the range and which the independent variables were chosen in the first place. So, this is what we were discussing this for.

(Refer Slide Time: 57:59)



Average value of dependent variable for given x

We have : regression (interval estimates for the regression constants)

To obtain : Average value of dependent variable for a specified value of independent variable x^*

Mean value of dependent variable, y^*

$\hat{y}^* = a + bx^*$

Interval estimate

$$\hat{y}^* \pm K_{CL}\, S_{Y|x}\sqrt{\frac{1}{n}+\frac{(x^*-\bar{x})^2}{\sum(x_i-\bar{x})^2}}$$

October-2020

Module 7, Lecture 2

12

x^*
$\hat{y}^*$
$|(x^*-\bar{x})|\uparrow \; U_{\hat{y}^*}\uparrow$
$\hat{y}_0$
$(\hat{y}^*) \pm (U_{\hat{y}^*})$
$x^* \rightarrow \bar{x}$
P(y/x)
y
y^*
x_i
$\bar{x}$
x
x_i^*
$x^* = \bar{x}$ minimum
NPTEL

And, now we come to the other aspect and which is like this that what we predicted earlier. And if you see that curve that we had made earlier, say this was our x, this was y and this was P y of x. And, what we had made a drawing there so, I will just try to quickly make it here. So, this is what we have x i and we said that over here we had a distribution of y values.

This was our mean value of y. Our first calculation we just saw calculating y hat 0, was told us that where is what is the likelihood of where the future value will lie. So, it could lie anywhere here. It could have been here, it could have been there, it could have been there or like that. And, what is the interval about which we can say it lies.

Now, what we are asking is that we made our measurement there, our population we did not have with it with us. So, this was not there. We at this point from which we made the this came out to be our x i that was used in making the regression. And now, we are asking where does this mean value of y lie? So, this is what we call y star.

So, we take any value of x and say we will call that for this calculation it could just have been x i, but we will call it x i star and when we will just use the regression curve that we had earlier. So, this becomes y i hat star. So, this is the predicted value of the mean value of the variable. So, this is telling us average value of the dependent variable for a specified value of independent variable.

So, we are specifying x i star and said using this correlation I am predicting y i hat star. That is what this formula tells us. Now, the question is well we did not put this i there yeah. So, what we are saying is, I have a value of x which I will call x star and I want to know what is the interval estimate for the mean of this curve.

So, the mean value we calculate from our correlation which is y hat star, y hat because this is the predicted value star because it is we are differentiating this by calling it is star that it denotes the average value of the dependent variable. So, this value we got. That the mean will lie in an interval about y hat star plus minus something and that something turns out to be what is given in this relation here.

K CL is our confidence level, S Y x came from our earlier thing and now we have this value. And, this as x star minus x bar over summation x i minus x bar square plus 1 upon n. In the earlier expression we had 1 plus this, which means that the interval for the mean prediction is less than the interval for a future value production. And that is to be expected because the means will all lie in a smaller tighter range than the values themselves.

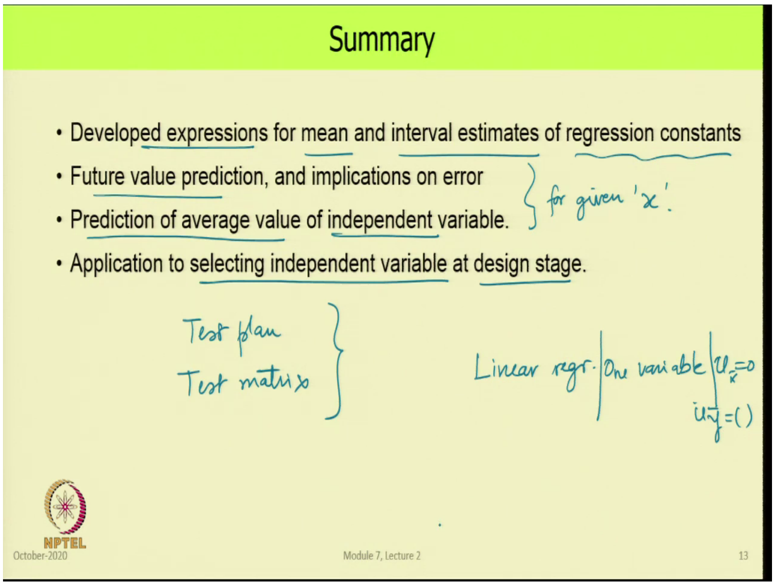
So, this is a we know x i x bar, we have been given the value of which we want to calculate the interval estimate that is x bar we put it in here. These are already known to us we can calculate the U of y star. And, again like the previous case here also if x star is very close to x bar, then we have the minimum uncertainty that you can expect the smallest interval at x bar equal x star is equal to x bar.

This first term will become 0 the second term remains, so, it is still finite. So, it is the minimum and as you move away from the mean value of x i so, this was the mean value x bar. As we take values on this side or we take values on this side as the number keeps increasing the interval estimate keeps increasing.

So, as x star minus x bar increases, so does U y star delta, this also increases. So, x minus x star we can even say absolute value of this because it is the square coming here the negative value does not matter. So, this is just like what we saw in the future value prediction in the earlier case.

So, this is another thing we would like to know from this. The third thing that we would have like to know, but we will not going to detail of it is that we have y equal to ax plus b, but now I am given the value of y what is the value of x. So, that is another set of calculations and formulae, we will not be going into it because most of the time we use this.

(Refer Slide Time: 65:05)



Summary

- Developed expressions for mean and interval estimates of regression constants
- Future value prediction, and implications on error
- Prediction of average value of independent variable.
- Application to selecting independent variable at design stage.

for given 'x'.

Test plan
Test matrix

Linear regr. | One variable | $u_{\bar{x}}=0$
$u_{\bar{y}}=()$

October-2020 Module 7, Lecture 2 13

So, that completes our discussion on developing regressions. We have looked at the statistical basis and developed a set of ideas as to how and why we developed regressions in a particular way. And, having done that, we developed expressions for mean and interval estimates of the regression constants.

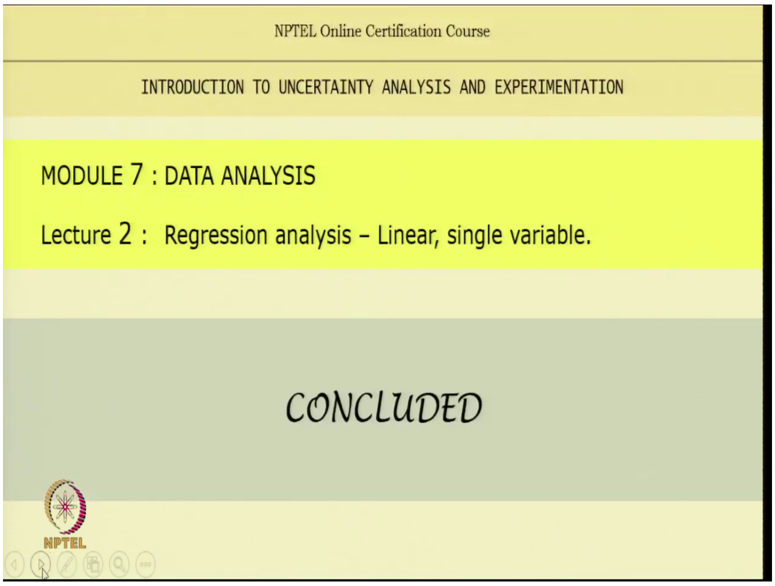
We are restricting our discussion to linear regressions in one variable and with u x bar equal to 0 but, u y bar is finite. So, we got all the relations for that then we looked at future value prediction that using a regression how can I predict a value for a value of the independent variable what will be the corresponding value of the dependent variable and what will be its uncertainty.

And, we saw that using that regression in its central portion of x values gives us the minimum uncertainty using it at the far ends it gives larger uncertainty. Then, we looked at the expression for calculating the average value of y or the independent variable for a given x; x which is the independent variable.

And, we also looked at the formula and saw its implication on designing, how the experiment should be done, where the data should be taken. And particularly, how do we select independent variables at the design stage and at the stage where we are developing the test plan and the test matrix.

So, in this experimentation part, in that thing we can make use of this knowledge and say first where should the data points be taken and how many data points should be taken. What we have seen here gives us a good idea of how to make a decision on that.

(Refer Slide Time: 67:24)



NPTEL Online Certification Course

INTRODUCTION TO UNCERTAINTY ANALYSIS AND EXPERIMENTATION

MODULE 7 : DATA ANALYSIS

Lecture 2 : Regression analysis – Linear, single variable.

CONCLUDED

With that we come to the end of this lecture, which was on developing the complete treatment and developing of relations for linear single variable regression analysis.

Thank you.