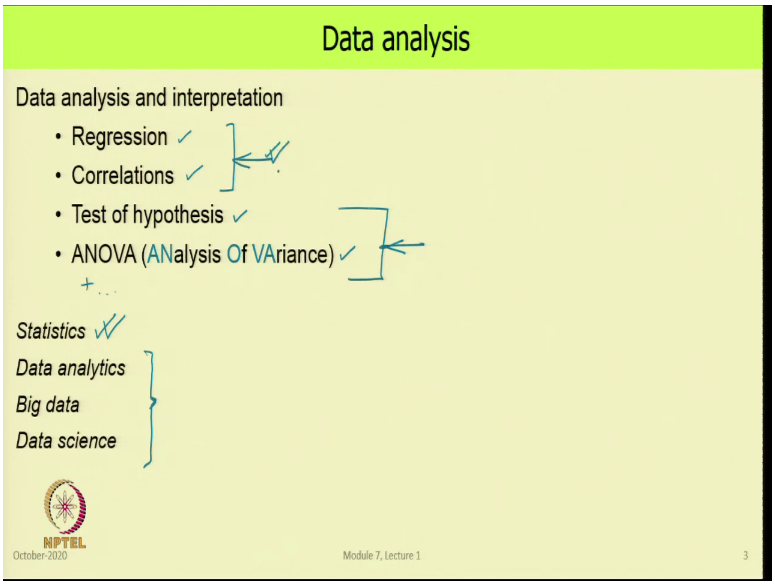
# Introduction to Uncertainty Analysis and Experimentation

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

## Module – 07
## Data Analysis
## Lecture – 24
## Regression introduction

Welcome to the course Introduction to Uncertainty Analysis and Experimentation. In this module we look at Data Analysis and we begin with an introduction to regression analysis.

(Refer Slide Time: 00:33)



Data analysis

Data analysis and interpretation

- Regression ✓
- Correlations ✓
- Test of hypothesis ✓
- ANOVA (ANalysis Of VAriance) ✓

+ . . .

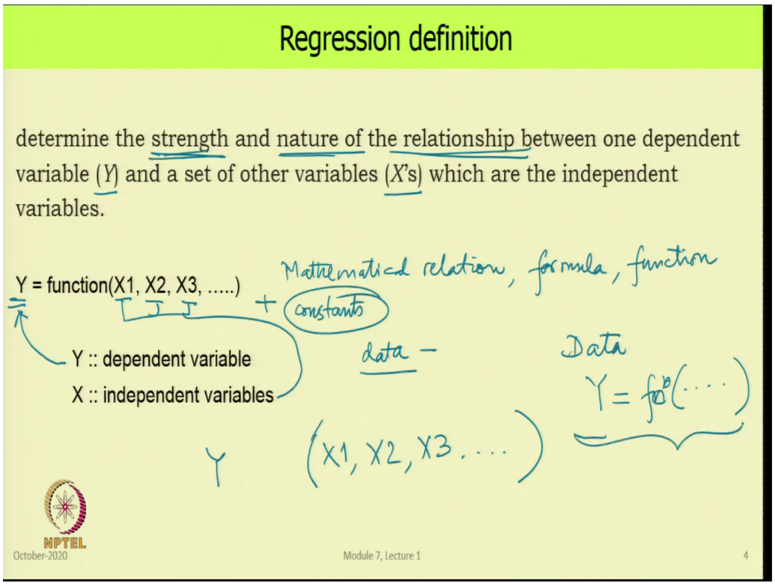
*Statistics* ✓✓
*Data analytics*
*Big data*
*Data science*

October-2020 Module 7, Lecture 1 3

In general data analysis encompasses many techniques and methods like regression, then correlations, tests of hypothesis, analysis of variance and many more topics here. All these largely rely on knowledge required of statistics. And these days we have even seen these develop as different branches by themselves which is data analytics, big data and called data science.

So, they use a variety of techniques to arrive at to make sense out of data. In this course we will limit ourselves to the first two only. These require more in-depth knowledge of statistics and can be taken up in our later course. But for many of the basic things one does in science and engineering these two we encounter quite frequently.

(Refer Slide Time: 01:42)



## Regression definition

determine the strength and nature of the relationship between one dependent variable ($Y$) and a set of other variables ($X$'s) which are the independent variables.

Y = function(X1, X2, X3, .....) + constants

Y :: dependent variable

X :: independent variables

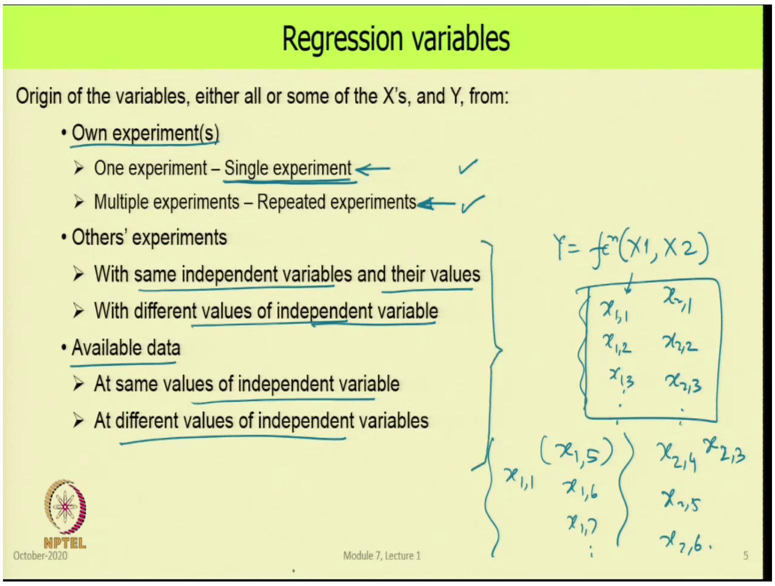
So, we begin with regression and start with the definition of regression and here we have written a basic definition of what regression is and it says, the regression determines the strength and nature of the relationship between one independent variable denoted by Y and one or more which is the set of other variables which are the X's which are the independent variables.

So, it says that we make a relation a mathematical relation or a formula or you can call it a function usually a continuous function, where Y which is the dependent variable is a function of various independent variables denoted here as X1, X2 and X3. So, this function can take many forms, but as long as we have a function like this what we are saying is that when we generate this type of a function from data.

In this case it could be a combination of data from experiment and or other sources also, then we make a function with these things and we generate some constants and that becomes our complete regression function. So, that is what we are meaning by first a making a relationship between these variables and then so that this is the nature of the relationship that what mathematical shape it takes.

And then we also have to talk about the strength of this relationship, that how strongly are these two sets connected to one another which is to say Y, how strong is Y connected to X1, X2, X3, and so on. So, this is what the idea of regression is and what we are looking for in this set of lectures is that I have data from my experiment or from others experiments or some other published data and I want to generate an expression Y as a function of these variables. So, this is our objective.

(Refer Slide Time: 04:51)



## Regression variables

Origin of the variables, either all or some of the X's, and Y, from:

- Own experiment(s)
  - One experiment – Single experiment ✓
  - Multiple experiments – Repeated experiments ✓
- Others' experiments
  - With same independent variables and their values
  - With different values of independent variable
- Available data
  - At same values of independent variable
  - At different values of independent variables

October-2020

Module 7, Lecture 1

5

$Y = f^n(X1, X2)$
$x_{1,1}$ $x_{2,1}$
$x_{1,2}$ $x_{2,2}$
$x_{1,3}$ $x_{2,3}$
$x_{1,1}$ $(x_{1,5})$ $x_{1,6}$ $x_{1,7}$
$x_{2,4}$ $x_{2,3}$ $x_{2,5}$ $x_{2,6}$
NPTEL

Now, what about where these variables are from? In general, the dependent and independent variables could be from one or more sources. And here are the various possibilities. All of them could be from our own experiments and within this we could have done just one experiment which we have so far called singles experiment and we saw in the uncertainty analysis of a result what was the techniques we used for a single experiment.

We also saw a technique where we said there multiple experiment repeated experiments. So, our own experiment could have also generated data from this, so, that is also ok. Now there could be situations where we have maybe some of our own experiments and some data from other's experiments. So, if others have done an experiment with the same independent variables and the same values.

For instance, so we have say Y is a function of X1 and X2 and we collected data at x 1, 1, x 1, 2 and x 1, 3 and x 2, 1, x 2, 2 and x 2, 3. So, these are the independent values that these variables were taken used in the experiment. Then if the other experiment done with a similar setup or maybe slightly different setup, also repeated the experiment at these very data then we can also use that data for generating a regression.

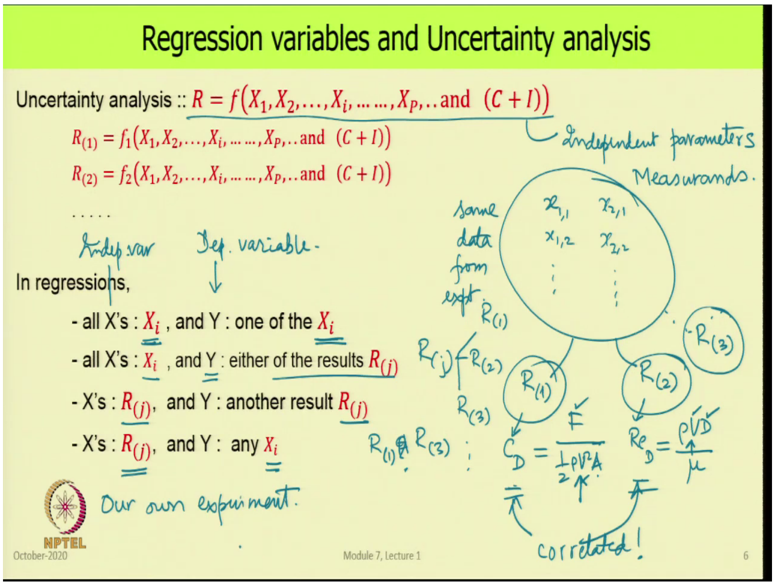
Or it could be a case that we took data at certain values of these, but somebody else using the same apparatus or maybe we ourselves using the same apparatus took data at some other values of the independent variable. So, instead of this we took it at say x 1 comma 5, x 1 comma 6, x 1 comma 7 and so on and this at x 2 comma 4, x 2 comma 5, x 2 comma 6.

Where these are different from any of these or it could be that they are all different or they could be that some of them are the same. So, you could have this data set and x 1, 1. And here which are completely different from here or either you could have a set x 2, 3 and the rest of these.

Any of these combinations are possible. Their implications and regression are different. So, that is what we have written here that the available data could be at the same values of the independent variable as we have seen just now here or at different values of the independent variables which is what we saw in this second set over here.

So, all these possibilities are there. And in general, we can use the regression for this and in the current course we will limit ourselves to either of these, this or this. And we will say that I have generated all the data either from a single experiment or I have generated data from multiple experiments and now I want to develop a regression. These require advanced techniques in analysis and so they are left for an advance course.

(Refer Slide Time: 08:36)



## Regression variables and Uncertainty analysis

Uncertainty analysis :: $R = f(X_1, X_2, \ldots, X_i, \ldots\ldots, X_p, ..\text{and } (C+I))$ — Independent parameters Measurands.

$R_{(1)} = f_1(X_1, X_2, \ldots, X_i, \ldots\ldots, X_p, ..\text{and } (C+I))$

$R_{(2)} = f_2(X_1, X_2, \ldots, X_i, \ldots\ldots, X_p, ..\text{and } (C+I))$

. . . . .

Indep var Dep. variable.

In regressions,

- all X's : $X_i$ , and Y : one of the $X_i$
- all X's : $X_i$ , and Y : either of the results $R_{(j)}$
- X's : $R_{(j)}$, and Y : another result $R_{(j)}$
- X's : $R_{(j)}$, and Y : any $X_i$

Our own expriment.

October-2020 Module 7, Lecture 1 6

So, now let us connect that we have done regression we are doing regression now and we have already done uncertainty analysis, what is the connection between these two? So, from an experiment we have seen that this is the basic nature of our formula. We said that R is expressed as a function of X 1, X 2, X i so on X p and some constants and independent numbers.

Now, what we meant? X 1, X 2 and all this X p were independent parameters which we were also say were independent measurands. So, they were different parameters that were measured during the experiment using some set of instruments. So, we got all the data. Our data from the experiment is for variable 1 x 1, 1, x 1, 2 and so on, for x 2 variable we got data x 2, 1, x 2, 2 and so on and like that for every measurand.

This was our complete raw data set that came out of the experiment. And then we put it in a result relation. We looked at the analysis for a general case where there is a result relation called R and we did not put any restrictions on what shape it took or why it was there. Now when you look at this we say that in done experiment after all this data was collected I could do a calculation for one result we will case this R 1 based on certain combination of these parameters.

Then I will use some of the other parameters and I will calculate another result which I will call R 2. And like that we could calculate many other result relations. So, the point is that from the same set of data that we got from the experiment we could generate different results and an example was there that if we make measurements you know wind turbine on a body experiencing aerodynamic drag we got one result as the drag coefficient.

Which we defined as F upon half rho V square into area and the other variable we could make was R e D which was rho V D upon mu. In the experiment we measured velocity, we measured area which also the diameter and we got this and we measure the force. From the same data set we took rho, V and D and got drag coefficient. We took only V and D and got Reynolds number.

So, we could think of drag coefficient as result formula number 1 and R e D as result formula number 2. The issue that comes up here and why we are not going to be looking at these type of situations is that when we do the uncertainty analysis uncertainty in V is coming here, uncertainty in V is coming here. So, the same uncertainty is coming in this and also in this result. That means that the uncertainties in these two parameters are correlated.

And in this course, we are not looking at correlated uncertainties. So, now we come back and say we are what am I going to plot make a regression of something against something? And so, I need to define what is the independent variable and what is the dependent variable. And we have many possibilities over here and we will go through some of these.

The first possibility listed here is that, all the independent variables are our measured parameters X i's and we make when we pick up this thing and make a plot against one of these. So, an example could be that we measure displacement of a spring. So, the weights that we put on the spring would be our independent variable X i and the dependent variable that we got was the deflection that we got.

So, that is a one on one. So, like this we could have more combinations or we could have all our measurands X i and Y which is our dependent variable could be any of the results that we have generated. So, this is what is called as R j. It could be any of these results that we have which we have just seen here and called them as R 1, R 2, R 3 and so on.

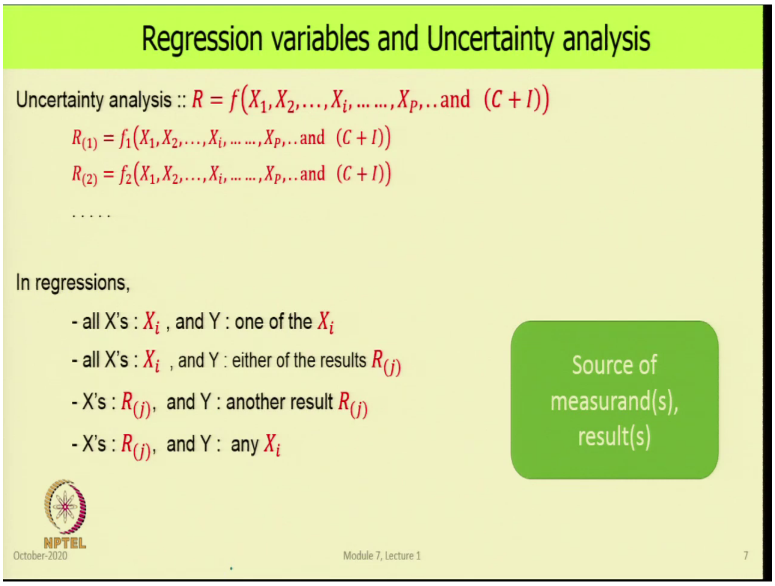
That means we are making a regression in this case between one measured parameter and another measured parameter or in this case against the measured parameter and a result from the same experiment. The third possibility is that we could make a regression between any of the result relations and another result relation.

So, we have R 1, R 2, R 3 like that. So, in this case we could make a regression between R 1 and R 3. So, that would be a mathematical function that we are looking for and we get one more thing I put here is that we would make a regression between any of our result and any of our measured parameters.

So, any of these are ok and possible and we often do various things like this in the type of work that we come across. And we have excluded in particular data from other's experiments. So, we are all saying that all these thing we are doing from our own experiment.

Later on we will see how we can modify this and say compare same independent variable between our experiment and somebody else's experiment, but results from our experiment and result from the other experiment.

(Refer Slide Time: 15:37)



## Regression variables and Uncertainty analysis

Uncertainty analysis :: $R = f(X_1, X_2, \ldots, X_i, \ldots\ldots, X_p, ..\text{and}\ \ (C+I))$

$R_{(1)} = f_1(X_1, X_2, \ldots, X_i, \ldots\ldots, X_p, ..\text{and}\ \ (C+I))$

$R_{(2)} = f_2(X_1, X_2, \ldots, X_i, \ldots\ldots, X_p, ..\text{and}\ \ (C+I))$

. . . . .

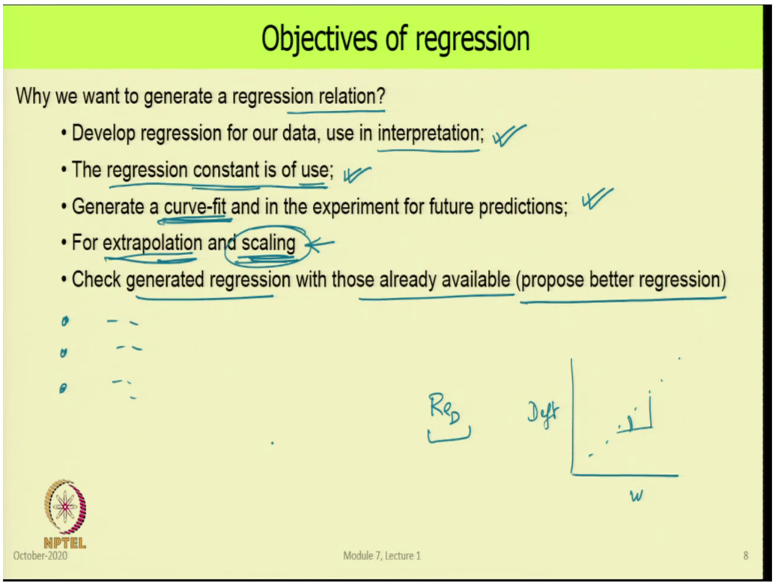
In regressions,

- all X's : $X_i$ , and Y : one of the $X_i$
- all X's : $X_i$ , and Y : either of the results $R_{(j)}$
- X's : $R_{(j)}$, and Y : another result $R_{(j)}$
- X's : $R_{(j)}$, and Y : any $X_i$

Source of measurand(s), result(s)

So, next is how do we get the various uncertainties on this.

(Refer Slide Time: 15:41)



Objectives of regression

Why we want to generate a regression relation?

- Develop regression for our data, use in interpretation;
- The regression constant is of use;
- Generate a curve-fit and in the experiment for future predictions;
- For extrapolation and scaling
- Check generated regression with those already available (propose better regression)

$Re_D$
Deff
w
NPTEL
October-2020
Module 7, Lecture 1
8

Next, we see what are the objectives of regression. So, why do we want to generate a regression relation? And here there can be many options or possibilities and reasons for doing this and I have listed some of them here we can add more. First, we want to develop a regression for our data and use it in our own interpretation.

The regression tells us how is what the functional relationship between one variable and another variable and that gives us maybe insight into the type of experiment we are doing and the type of result we are looking for. So, that is one possible reason for doing it. The second reason is the constants that come out of a regression analysis they are of use to us. For example, if it was a spring deflection and weights that were put here and deflection of a spring then you would get some points like this.

And we are interested in the slope that is what we call the spring constant. So, that is an example of where we say that the regression constant is of use to us. A third reason could be that we make a regression we call it a curve-fit. So, between data point that we have we made a curve and this said that then we say I will use this in my experiment for future predictions.

So, the example of this is that we could calibrate a thermocouple and then say that is my experiment I will measure the e m f and use that curve-fit to calculate the temperature. So, that is what we call curve-fit. Then we could like to do something on scaling and extrapolation. If a parameter has got dimensions, we generally cannot scale it, but if you make a non-dimensional number out of it like so, many non-dimensional numbers we come out in various engineering disciplines.

So, like Reynolds number then we know that if we did an experiment with air at a particular Reynolds number then what we saw there is likely to be the same case with water at the same Reynolds number. So, that is helps us in scaling and even also that if we took a correlation for a range maybe we would be able to predict something bigger than that would be extrapolation.

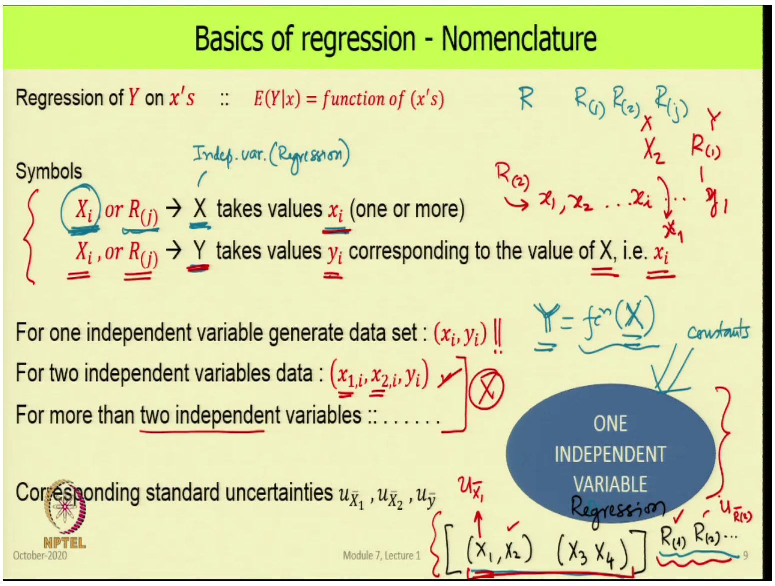
And yet one more objective could be that we generate a regression from our experiment and we want to see how good does it compare with some regression that is already available and maybe we can even prove that we have done the much better experiment and the existing relation which is already there has been improved and we propose a better regression.

We see this things happening all the time in various disciplines and every time a better regression comes this like we have a slight improvement in whatever object the engineering we were doing. And like that you can add many more reasons why we would like to do a regression.

In this course, we will look at this and this as our primary objective and we will also see why curve-fit comes in, we will not go too much into the detailed of scaling for which you would like to do non dimensional analysis or similitude use similitude theory that we will not do in this course.

The last one requires comparing one regression with another regression, it requires more advance techniques that we will not go into at this point.

(Refer Slide Time: 19:41)



## Basics of regression - Nomenclature

Regression of $Y$ on $x's$ :: $E(Y|x) = function\ of\ (x's)$

Symbols

$X_i$ or $R_{(j)}$ → X takes values $x_i$ (one or more)

$X_i$, or $R_{(j)}$ → Y takes values $y_i$ corresponding to the value of X, i.e. $x_i$

For one independent variable generate data set : $(x_i, y_i)$ !!

For two independent variables data : $(x_{1,i}, x_{2,i}, y_i)$

For more than two independent variables :: . . . . . .

Corresponding standard uncertainties $u_{\bar{X}_1}$, $u_{\bar{X}_2}$, $u_{\bar{y}}$

ONE INDEPENDENT VARIABLE

October-2020

Module 7, Lecture 1

9

Now, before we go into the mathematics of it, we need to do something about a book keeping. What is happening now is that, we will use a bunch of symbols many of which we have already used in our earlier analysis. So, there could be a good deal of confusion that we could generate by saying what does this symbol mean.

And say in the case of regression analysis this is what it means, but in uncertainty analysis it meant something else or it could be a symbol that we use in regression whereas, in statistics we have already used that symbol for something else. So, we spend some time to do this clarity on book keeping and be clear that in this analysis which is the regression analysis symbol that we use here it has the either the same meaning or slightly different meaning if and if so what it is?

So, here we are. From the experiment we got symbols X i and R j although we did not use R j that much we have now introduced it saying that result was R and if we generated multiple result from the same experiment we just called it R 1, R 2 like that and we have called this R j. And we have two variables there in regression the independent variable X and the dependent variable Y.

And the main thing about what we will study here is that we will look at the simplest possible case where we have only one independent variable. So, there will be many instances where that could be two or more independent variables those we are leaving for a later discussion not now. So, with one independent variable we only have Y as a function of X. That means, everything else that is there in this function they are constants.

So, if you want to look at our result formula if we were looking at the result formula could have more than one functions the way we have looked at it earlier, but now we pick up one here and one here. And as we have seen in the earlier slide what is it that X could be X and what is it that we could consider as Y and here is what we are seeing.

That X could take any of the values of x i, but what we are saying it X is one of the independent variables X i or a dependent variable of the experiment, which is like we set some values say the heater power and measure the temperature. So, in that context the temperature became a dependent variable, but now once we have got the data the temperature is also part of the X i and this is what it could be here.

So, what we are looking at is that what we considered as the independent variable for regression is slightly different from what we talked of as independent variable in an experiment. In the experiment we said that I have X 1, X 2 as my independent variables which meant that by doing the experiment, I could set the value of these independently.

And then the experiment will do what it has to do and I have some more measurements that I take X 3 and X 4 parameters they are indicated to me by the experiment and those are dependent parameters of the experiment. However, all the data that we got from the experiment was all of this and using some combinations of this we calculated our results R 1, R 2 and so on.

In the contest of regression any of these could be our independent variable and that is what is written here. We are saying now that any of X i's which were here or any of the results that we have computed which is here this could be our X, X is our independent variable now we are talking of regression. And this way we will denote that it takes some values which we denote as x subscript i.

So, x i now has a very different meaning. It is the value of the independent variable that we have chosen. It does not mean that these are the values of the variable x i or say if this is x 1 does not mean that it is the value of variable X 1 which was this variable here. This means that I am taking a value and assigning a value to either of these X i's or to any of the R j's which I have selected. So, this could be say R 2 and the first value of R 2 that we use that will be small x 1 for our analysis.

The second one would be x 2 and so on and that is where we get x i and it goes on there. So, x i is denoting one value of the independent parameter whatever it may be. So, this is our independent variable x. Now for the dependent variable that is Y and a few minutes back we saw that Y could be any of these, Y could be any of the parameters of the experiment independent parameters or the dependent parameters or it could be any of the results that we calculated from the experiment.

So, that is what it is said here that any of the X i's or any of the R j's we will call that whatever we take we call that as Y. And Y takes the values of y i; that means, whatever we took there whatever values we had for those cases they are y 1, y 2 like that. Further with the restriction that this y corresponds to the same value of X, x i.

So, our first number could be a result R 1 and we pick up and we want to make a regression with X 2 then the first value of R 1 which will be x small x 1 must correspond to that same data set from which this both calculation was done. So, it will be X first value of X 2. So, that is what it is. X 2 and R 1 is what we want to make a regression.

The first value of R 1 this is Y, say this is X, the first value of this we will call it y 1 and it must correspond to the first value of X 2 which we call x 1. So, this is an important distinction we are making of what we mean by independent variable integration, what we mean by the parameters and the result as being a set from which we pick out an independent variable and a dependent variable.

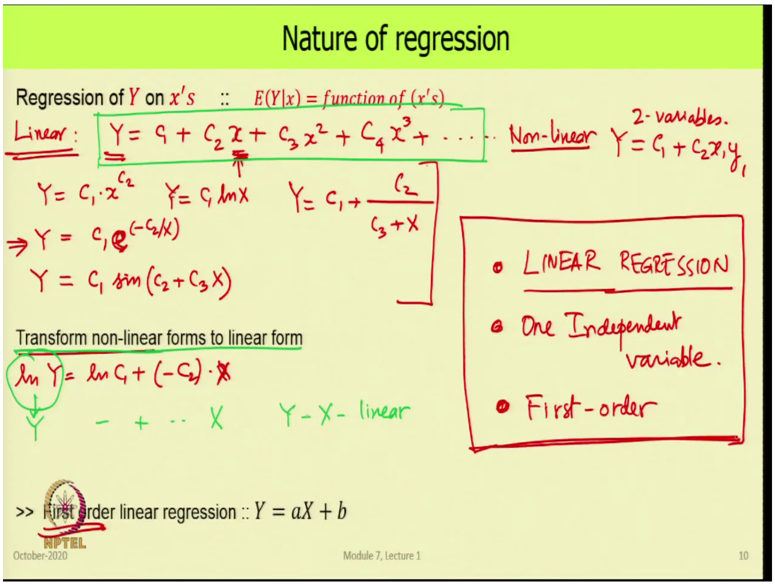
So, if there is only one independent variable we generated data set x i, y i, i depends how many readings we have taken. If there are two variables in the experiment and we want to make a regression on that it could be a variable x 1, x 2 which x 1 could be either of these which are listed here, x 2 could be any other one from this and y i could be yet another one from this set.

They all must correspond to the same experiment. So, that is an important part. And the same thing would happen if there are more than two independent variables. So, what we are saying is we are not going to be looking at this in this course. We will restrict ourselves to this one, one variable regression.

Now, we get one more data that we already done in the experiment and looked at in our uncertainty analysis. Input that we will require for doing the regression and that is that with each one of these we have an associated standard uncertainty u X 1 bar and similarly for each result your standard uncertainty say u R 2 bar.

All these values we are carrying forward from our previous analysis of uncertainty. And we will see how these become useful in our analysis. So, data that came out of this that is one and its uncertainty that is there that is the second one. This is our input data for performing a regression.

(Refer Slide Time: 30:13)



## Nature of regression

Regression of $Y$ on $x's$ :: $E(Y|x) = function\ of\ (x's)$

Linear : $\boxed{Y = C_1 + C_2 x + C_3 x^2 + C_4 x^3 + \dots}$ Non-linear $Y = C_1 + C_2 x_1 y_1$ 2-variables.

$Y = C_1 \cdot x^{C_2}$ $Y = C_1 \ln X$ $Y = C_1 + \frac{C_2}{C_3 + X}$

$\Rightarrow Y = C_1 e^{(-C_2/X)}$

$Y = C_1 \sin(C_2 + C_3 X)$

Transform non-linear forms to linear form

$\ln Y = \ln C_1 + (-C_2) \cdot X$

$Y$ – + ·· $X$ Y - X - linear

- LINEAR REGRESSION
- One Independent variable.
- First-order

>> First order linear regression :: $Y = aX + b$

October-2020 Module 7, Lecture 1 10

Now, let us look at what is the nature of regressions that we can have. In for a moment let us not just look at the linear regression, but look at type of functions we can get in general. So, the idea of we will first of course, restrict ourselves to a linear regression and linear means that Y is some constant plus C 2 times x plus C 3 times x square plus C 4 times x cube plus so on. This is a linear regression and linear regression of one on one, one dependent independent variable, one independent variable.

To give an example of a non-linear we could have Y is equal to C 1 plus C 2 x 1 y 1. So, that is what we are getting there but this is what, 2-variables. Then there could be other possibilities. We could have that Y is equal to C 1 times x to the power C 2. Now, this is not a linear form relation, it is also exponent here or it could be ln also.

That Y is equal to C 1 ln X or we could have Y is equal to C 1 e to the power minus C 2 by X, this is also non-linear or it could be a sinusoidal function sin C 2 plus C 3 X or even Y is equal to C 1 plus C 2 upon C 3 plus X. Now, quite clearly we have only restricted ourselves to looking at 1 independent variable, but we got many functions which are we will encounter quite frequently which are not linear.

So, what one does is to transform these into linear form and an example of that could be that say Y is equal to we take this case then we can take ln Y is equal to ln C 1 plus minus C 2 times X. So, what has happened now is that we called of relation between ln Y which we will call as our different Y, transformed Y as a function of something plus something plus X.

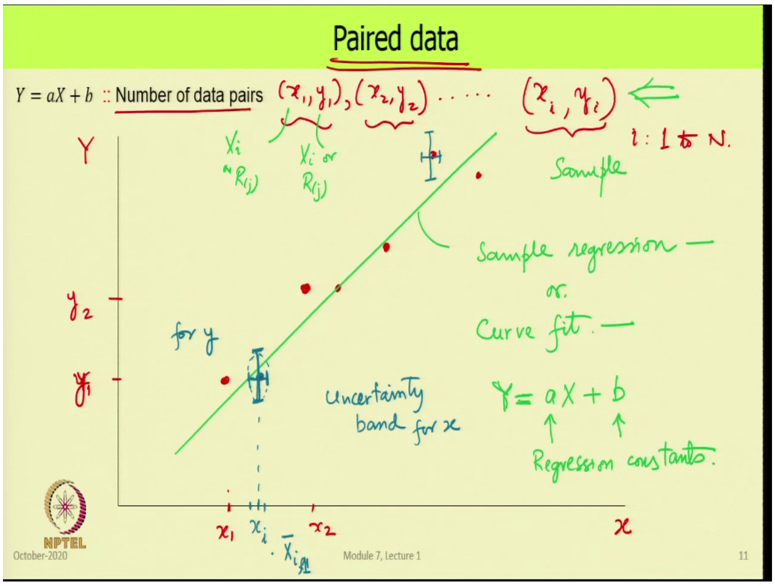
So, the relation between these X and Y, this is now linear and in this we can now apply all the theory that we are going to learn. So, in many situations one can do a transformation and bring a non-linear form to a linear form. So, that is a transformed variable and we get this thing.

So, this is one restriction that we will put in analysis that we do that we look at only linear regressions. And then one more thing linear regression would in general mean the expression that we wrote here and this is linear in 1-variable.

So, we put the second restriction now. It is a linear regression, one independent variable, and a third restriction that we will now put is that we will only look at a first order linear regression; that means, only this term is there these x square, x cube terms are not there. So, we look at first-order regressions. So, this course we restrict ourselves to regressions of this type.

There are many other forms which are beyond these there are techniques to do that, but we will not be discussing those in this course they are topics for an advanced course. So, let us see this.

(Refer Slide Time: 36:21)



## Paired data

$Y = aX + b$ :: Number of data pairs

October-2020

Module 7, Lecture 1

11

First what is the idea of having what we call paired data and what we plot here is x values of X here and Y and what we have is from the same experiment we picked out this data and we plotted them here. So, for X say we got value x 1 here and we got this data point here, we got this y 1.

So, we got these 2 numbers x 1 and y 1 from the experiment, the mean values and we have just plotted it on this chart. Then we take another pair x 2 and say this is y 2 and we make a, we see where they will fall and we say that this is my point corresponding to x 2, y 2. Like that we can put many more points. And they could lie whichever way they lie the experiment is telling us this is what it is.

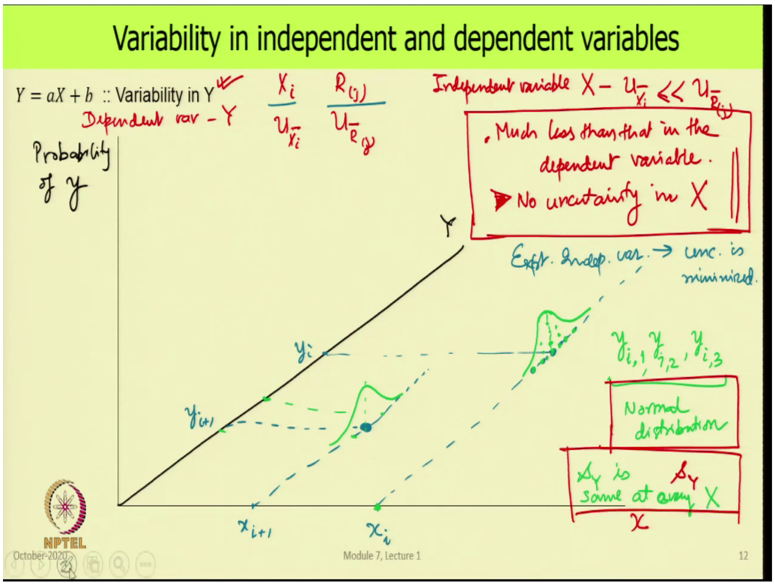
So, we got say that many points and this is our pairs of data that we work with and so for this we have this many number of data pairs which we call as x 1, y 1, x 2, y 2 and so on x i, y i, i goes from 1 to N. That is what is called paired data, pairs of data and now we say that my objective is in a linear regression what you may already come across that I want to fit a line to this.

So, this is the this is data which we call our sample and this curve that we make here this is what we call is a sample regression or this is also referred to as curve fit. So, this is the first thing that happens and the curve fit takes the form Y is equal to a X plus b, you may have come across this from your school days; a and b is what we call as the regression constants.

So, if we did the experiment once we got these data pairs we picked up this pairs I mean. So, what we have just argued that this could have been any of the X i or any of the R j's and this could be any other X i or any other R j and using these pairs we did our analysis and this is sample because we got a limited set of data that came from the experiment.

So, this is our idea of what a data pairs and how using one pair one set of paired data we get a sample regression or a curve fit. The method we have not looked at, but it will generate one sample regression or one curve fit for one sample of the form Y is equal to a X plus b, a and b are the regression constants.

(Refer Slide Time: 40:57)



Variability in independent and dependent variables

$Y = aX + b$ :: Variability in Y

Dependent var - Y
$\frac{X_i}{u_{\bar{X}_i}}$
$\frac{R_{(j)}}{u_{\bar{R}_{(j)}}}$
Independent variable X – $u_{\bar{X}_i} << u_{\bar{R}_{(j)}}$
. Much less than that in the dependent variable.
No uncertainty in X
Probability of y
Y
Expt. Indep. var. → unc. is minimized.
$y_i$
$y_{i+1}$
$y_{i,1}, y_{i,2}, y_{i,3}$
Normal distribution
$s_Y$ is same at every X
$s_Y$
$x$
$x_{i+1}$
$x_i$
NPTEL

October-2020

Module 7, Lecture 1

12

Now, we look at variability in Y. So, what we are looking saying is that from a data we knew that every one of those X i’s and R j’s, we learnt how in each of these measurands we could estimate the standard error in that X i and how using a result relation we could get the standard uncertainty in the result.

Now, what is the implication of this, on what we have we are going to see? So, we go back to this curve, what we are now saying is that this data point that I had let us take this one there was uncertainty in this value. So, there was uncertainty in the X value corresponding to this.

So, whatever this would have been say x i, which means that this value would have lied at 95 percent confidence level between this and this. We are plotting here the mean value that we got from the experiment. In some sense this is X i bar. At of the first one value of this which we said was comma 1.

So, what we denote here on this picture is that this value could have lied in this band and that is the way we will show this. And this is called the uncertainty band and at 95 percent confidence level it tells you where the data lies. Same is true with y, this will also have an uncertainty and this we will denote as like that. So, this is the uncertainty band for y. So, this is for x, this is uncertainty band for y and that is what came from our earlier analysis.

So, what it tells us is that we can expect that point in this two domain two variable domain to lie somewhere in this or maybe in a rectangle like that something like that. So, the question then is how good is this line that we have made. Knowing that there is variability of x and y in every point.

How do we handle this? So, what we do in our analysis is that we invoke the fact that the independent variable which we have called X which could be any of these X i's or R j's, this has got some uncertainty u, but this is very much less than the uncertainty of the other dependent variable which say is say R j.

So, we are saying that uncertainty in independent variable is much less than that in the dependent variable. Practically, we will get a finite value, but it could be that in one case the uncertainty is like 10 to the power minus 3 of that order whereas, this is of the order of 1.

Then we said that yes, the uncertainty in my independent variable is very much less than uncertainty in the result. And what it means is that, in the limiting case we could even say that there is no uncertainty or no error in the independent variable X.

So, this is one assumption we make one more. That uncertainty in X is negligible and in any case practically very much less than uncertainty in the dependent variable. So, that is how we would pick it and in many cases when we looked at the experiment design in the context of the experiment when we looked at independent and dependent variables. So, in the experiment we picked up the values of the independent variables in such a way that their uncertainty is minimized.

So, in general that would be a strategy for designing the experiment, then say that my independent variable I want to keep my uncertainty as small as possible in comparison to the other parameters. But in regression we could have said that any of those X i's or R j's that we pick as the independent variable we say its uncertainty is very small that is 1.

What that leaves us with that is that in the dependent variable there is finite uncertainty variability which we cannot ignore and we have to factor that into our regression analysis. So, here is what it means? So, here we have x and we will make a plot where we will plot Y on this axis.

So, you look at rather 3-dimensional plot and on this side is the probability of y. And what we do? We pick out one value of x say here x i and ask that I do the experiment many many times and I get the values of Y in each case and now I plot that. The mean value of Y lies over there.

So, in this plane XY plane this is our data point. But while x would have been the same value in every experiment from experiment to experiment Y value would come different. So, sometimes will be here sometimes here, sometimes there, sometimes there, maybe there, maybe there. So, what is we are saying is Y takes up different values. And there is the certain probability distribution to those values; that means, some values are coming up more frequently some are coming less frequently and like that.

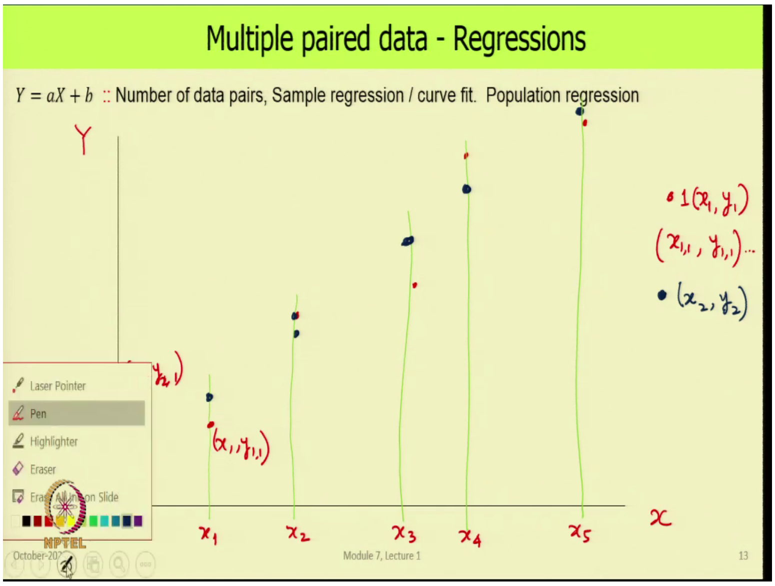
And if we have large enough number of values of Y we can do a curve-fit and for this analysis we will say that this follows a normal distribution. So, y values all of y i's y i, 1, y i 2, y i 3, i means corresponding to value of x i. All these values they follow a normal distribution. So, this is what we are getting. And the same thing if you have to look at any other value of x say we could have taken this value here and we could have got say this value of Y.

So, this is the mean value of y i, y i was here, this is say y i plus 1 or whatever and then the same thing would happen that the values of Y would be varying and would take the shape of a normal distribution over there. So, some cases the Y value would be here, some cases the Y value will be there and there is a probability distribution because of which this function is coming like this.

So, this is what is happening that we have said that we will consider the case where variability in x is 0. So, we have fix this point and we get one line here otherwise this also would have been a big line and a Gaussian distribution around that and then this one.

So, we are make some more assumptions now that variability in the independent variable is very small compared with dependent variable and we will be considering it as 0 and that the variance in Y this is what we got which we will denote as s Y. We also assume that the distribution of Y values follow a normal distribution. So, the two more assumptions have come into the analysis.

(Refer Slide Time: 52:34)



Multiple paired data - Regressions

$Y = aX + b$ :: Number of data pairs, Sample regression / curve fit. Population regression

Now, let us see what happens because of this and again as before we plot x there and Y over here and we will now look at the case where we had one experiment and one set of paired data second experiment under set of paired data, but at the same values of the independent variable. So, let us try to plot these things.

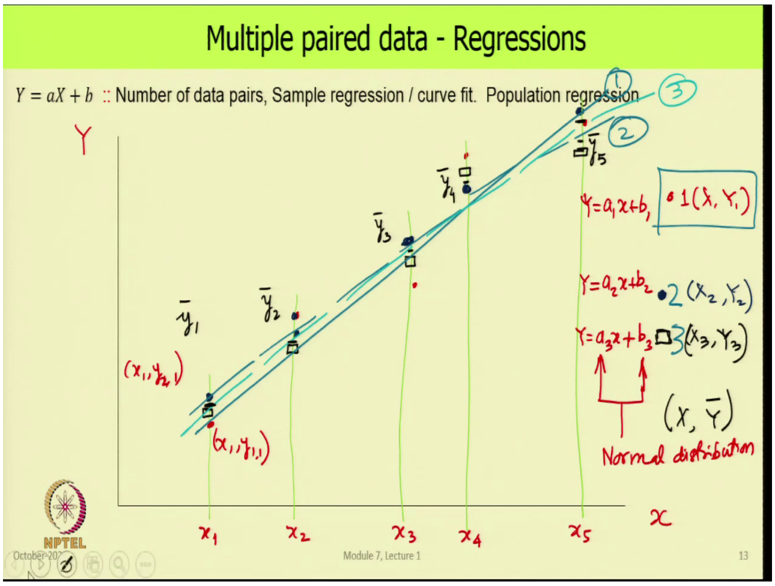
So, first the data points we got is say here, here, there and there. Say we can put one more there. So, this symbol this is data set number 1 which gives us x 1, y 1 pairs which for each point will actually end up being x 1 comma 1, y 1 comma 1. So, that is this point x 1 comma 2, y 1 comma 2 would be this point and like that. Then we say I did a experiment second time or somebody else did the experiment second time and we will plot their values.

So, this value came over there, over there, over there, here and say here. The point is that we are saying that all these have been made at the same value of x. So, x value did not change.

This was x 1, this was x 2, this was x 3, this is x 4, this is x 5. And the first data point that we just saw that would then be x 1, y 1, 1, the second point this would be x 1, y 1 comma 2, y 2 comma 1; this is from the second experiment.

So, this second data point set that we have which we have put here, here, here, here and here this is our second data set x 2, y 2 we can call this capital. So, that there is no confusion.

(Refer Slide Time: 55:32)



Multiple paired data - Regressions

$Y = aX + b$ :: Number of data pairs, Sample regression / curve fit. Population regression

October-20
Module 7, Lecture 1
13

So, this is X 1, Y 1 and the second one is the second set X 2, Y 2. Now, get a third set. We do the experiment yet once again and say now I have got a 3rd set and my data points are like this. So, this is my 1st data point, the 2nd at the same values of the independent variable, 3rd, 4th, and 5th. So, this becomes X 3, Y 3. Now, what happens?

We take the 1st person says I got my data over here and I will make my regression and so with these points some regression was developed and this is the regression number 1, which is this one. The 2nd person says these are my dataset, I will use only my dataset and I will make a regression out of it and so, this person makes a regression and his regression comes say something like that.

Third person says this was my data set, I will make my own regression and so this person makes a regression which goes like this. This is regression number 3. And so, as many sets of such data are there that mean regressions are coming out. Now, the question is what should we take as a representative regression for this data set.

At the outside we will not discard anybody's data as being bad or as we do not like the data anything like that we say that all the data are equally good. So, all the regressions are equally justifiable. So, this is what happens and what we have to do then is say that the I can do 2 things with this. What I will do is for each x 1, I got different values of y, I will take the average of that value y 1 bar and plot it somewhere.

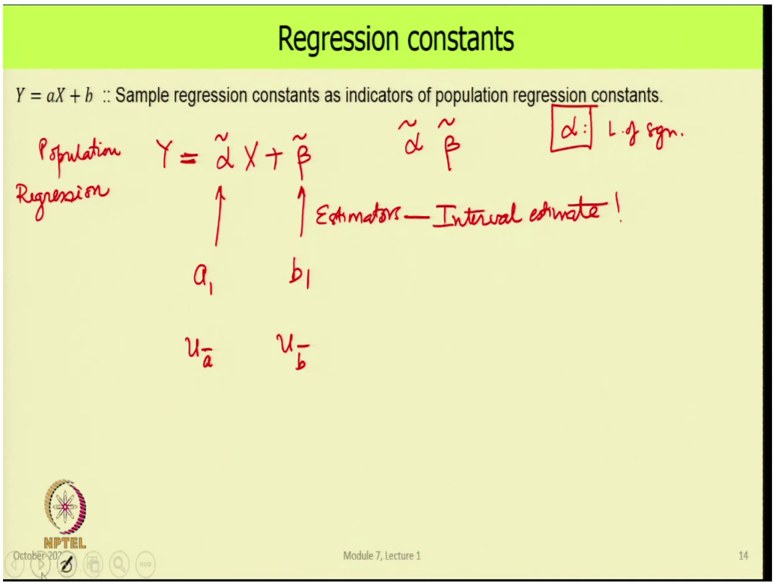
And says the number comes over there. I will do the same thing for the second one, I will get y 2 bar and plot it over there. Then calculate the mean of these 3 and plot it over there the mean of these 3 we plot it over there and so like that this is mean of y 5. I said now I have one set of consistent data, I will make a regression between which I will call X and the Y bars. In some situations one could do this provided there is not too much variability at this any of these points.

But still a word of caution in actually doing this type of a regression, strictly not recommended. So, the second thing to do is that we say that I have all these regressions and what I will do is I will write the equation for each one of them separately. So, here is what happens, the 1st person gets Y is equal to a 1 x plus b 1. So, their regression constants are a and b, the 2nd person gets Y is equal to a 2 x plus b 2, the 3rd person gets Y equal to a 3 x plus b 3 and as many such regressions were there.

What this now tells us is that all these a 1, a 2, a 3 they are like numbers coming from a statistic they are not the same. So, this forms a new variable and b 1, b 2, b 3 similarly come from another distribution and we presume that both these have a normal distribution. This will be the case if all the y's had normal distribution with similar parameters.

So, what we said that in the earlier thing that each one of these are same in addition to saying that they have normal distribution, we can also add here that their standard deviation is same at every X. So, this is yet another condition we are putting. These two will come from normal distributions and so they would have their own population and from there we could get something else.

(Refer Slide Time: 61:38)



## Regression constants

$Y = aX + b$ :: Sample regression constants as indicators of population regression constants.

Population Regression $Y = \tilde{\alpha} X + \tilde{\beta}$ $\tilde{\alpha}$ $\tilde{\beta}$ $\boxed{\alpha:}$ L. of sgn.

$\uparrow$ $\uparrow$ Estimators — Interval estimate !

$a_1$ $b_1$

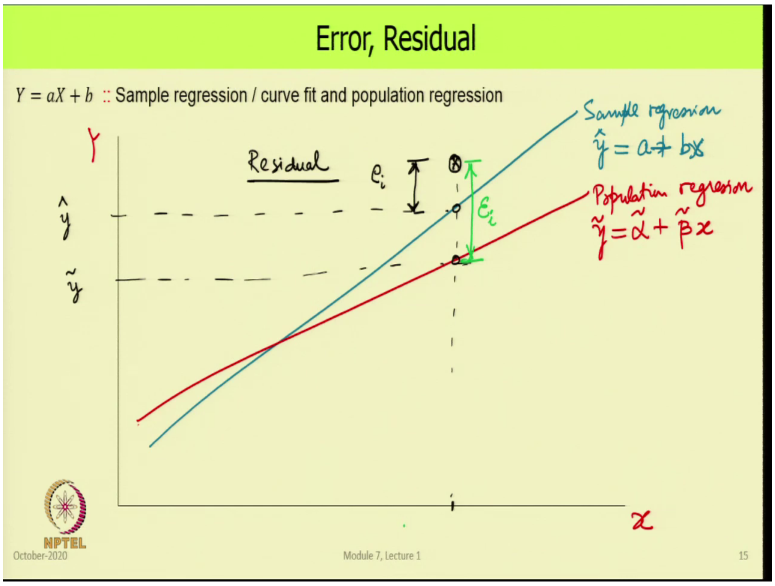
$u_{\bar{a}}$ $u_{\bar{b}}$

So, let us see what happens in that. So, we have all these regression constants coming in and we have if we had a population; that means infinite number of experiments from which the Y values came then we can make a population regression of their nature that Y is equal to alpha X plus beta tilde.

So, this is the population regression and the symbols that I have used their constants alpha and beta, but I put this tilde on top of that for a particular reason that the symbol alpha is used in statistics for the level of significance. So, this will lead to lot of confusion if you were to look at it in that way and to distinguish it from the level of significance we have made it alpha tilde and beta tilde.

So, our experiments whatever a's we get a 1, b 1 they are estimators of alpha tilde and beta tilde and we have to give an interval estimate. This is essential now. So, this is one important thing that we have come up with and this means that we have to calculate u in a bar and u in b bar.

(Refer Slide Time: 63:54)



## Error, Residual

$Y = aX + b$ :: Sample regression / curve fit and population regression

October-2020

Module 7, Lecture 1

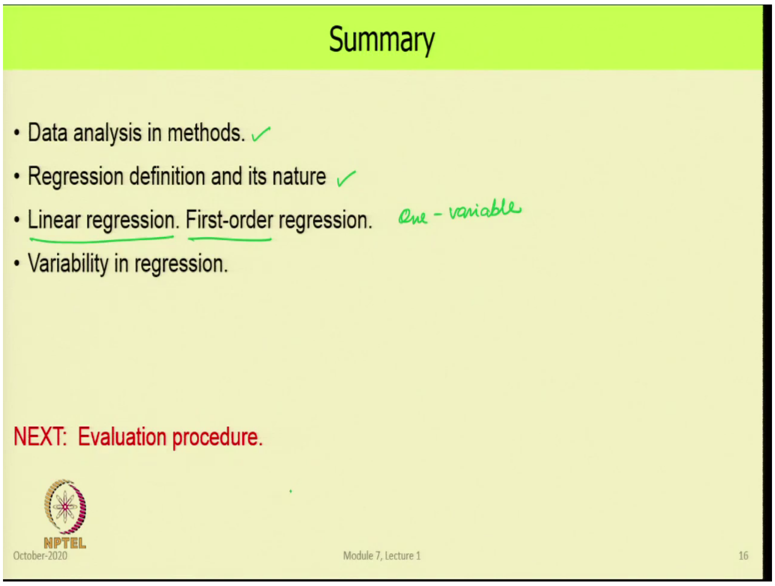
15

And finally, we will define one more term and then we will stop, which is error and the residual. So, we are plotting here Y, this side is x and from the discussion that we just had we can now make 2 lines. One is this one which we will call the sample regression, which we will denote as y hat as the sample regression predict value, this is a x a plus b x and then we draw a second line which is the population regression which we will call as y tilde is equal to alpha tilde plus beta tilde times x.

And we if you take any value of x say here then we get 2 points, one from the population regression and this is what we call the corresponding y tilde and this we call it as y hat. But our data point could have been somewhere there. So, this is the value from the sample regression, this is the value from the population regression and this is our data point. So, now we define two important an important parameter.

That this difference is called the error e i and this we call as the residual. So, this is an important parameter that we will see later on why it is it gives as lot of information about what is type of regression we made. And we define one more the difference between this and the population regression value this we denote as epsilon i and e i is an estimator of epsilon i and they both follow statistical distribution and that gives us lot of valuable information in what we want to do. So, what we have seen here this expression.

(Refer Slide Time: 67:03)



Summary

- Data analysis in methods. ✓
- Regression definition and its nature ✓
- Linear regression. First-order regression. One - variable
- Variability in regression.

NEXT: Evaluation procedure.

October-2020 Module 7, Lecture 1 16

So, in developing the regression constants, so, with that we come to the end of this lecture. We have seen what are the various methods of data analysis, what is regression and its nature, how we can convert a non-linear case into a linear regression and we will focus on linear regressions, first order regressions and in one-variable.

And finally, we said how we will treat variability in the regression analysis by saying that the independent variable has little or no uncertainty, the dependent variable has uncertainty and we will use this information to further develop our ideas and how to calculate the regression constants, how to calculate a future predicted value, how to calculate a future predicted mean.

So, all that will come in the next lecture that is the evaluation procedure. Now, on that note we conclude this introductory lecture on regression.

Thank you.