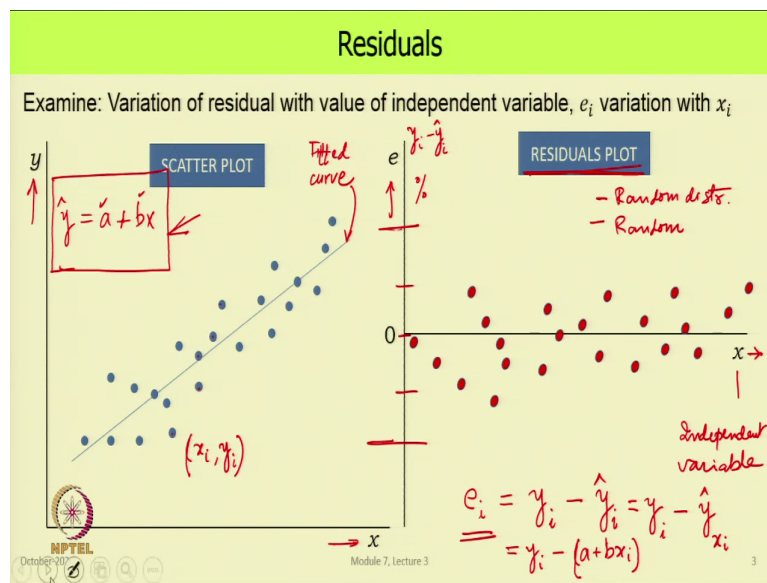


Introduction to Uncertainty Analysis and Experimentation
Prof. Sunil R. Kale
Department of Mechanical Engineering
Indian Institute of Technology, Delhi

Module - 07
Data Analysis
Lecture - 26
Correlation Related topics

Welcome to the course Introduction to Uncertainty Analysis and Experimentation. Today we will continue with the discussion on Data Analysis and look at Correlations and a few Related Topics.

(Refer Slide Time: 00:32)



We will start with by looking at residuals. We came across residuals in our discussion on regression, where we defined the residual e at a point i as value of y at that point minus the

predicted value of y_i for that same x_i . So, you can put it that way or make it more clear, we could even write this as $y_i - \hat{y}_i$; which means a predicted value from the sample correlation at x_i .

And we saw one use of the residual in developing the methodology for getting the regression constants. Now, we look at a different aspect which is by looking at the plot of residuals. So, this picture on the left side, we have all these which are our data points; x axis is x , y we have plotted on the y axis.

So, each one of these represents a combination x_i, y_i . And to that we did a regression and our fitted curve is over here. Now, when we look at the regression, we do not see the points from which it came; we just have before us a relation, which is y is equal to or \hat{y} is equal to $a + bx$, where a and b are known to us.

So, when we just look at this and we have no information of how that was the data from which this came, then there are situations where we could lead be in trouble. So, what we do is, we make the residuals plot. So, that is shown here on the right side. On the x axis, we are still plotting the independent variable; but on the y axis now, we are plotting the residual e which came from this relation. So, that is what we have done.

So, this is $y_i - \bar{y}_i$, \bar{y}_i we could also write this as $a + bx_i$, and a and b are have already been computed from the data that we have. So, now when we look at this what we should see is that, these residues are fairly randomly distributed above 0. So, this is 0; so there are as many above 0 as below 0 and there is no discernible trend or pattern that we see in this.

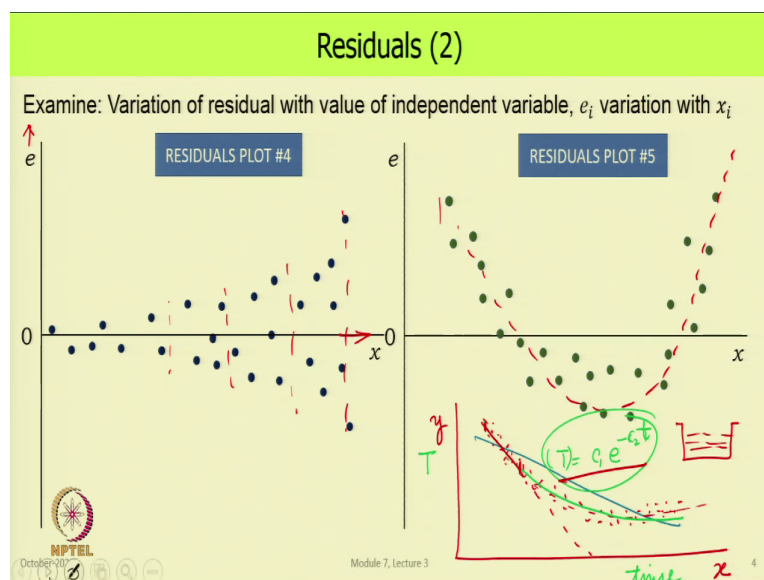
So, across x and at any x , there is randomness. If we, if this is how our residuals look; then it tells us that the data to which we had fitted a linear regression, that was justified and the fact it also tells us that there is the magnitude of the residuals, that can tell us how good our fit is.

And it also tells us that, since there is no variation in the distribution of residuals with x ; we are fairly confident that what regression we got, which is this one is justified. So, this would

be an ideal case. Some of these residuals would also be non-dimensionalized and made as a percentage, and all good experimental data would always be accompanied by the residuals plot.

So, a very carefully done, very thorough experiment will have residuals in a smaller range; others may have the residuals falling in a larger range. So, it tells us both the goodness of the data and the quality of a data compared with others data. But this need not always be the case. And we will now see different things that are possible in the residual plot that we could see and what their implication is.

(Refer Slide Time: 06:17)



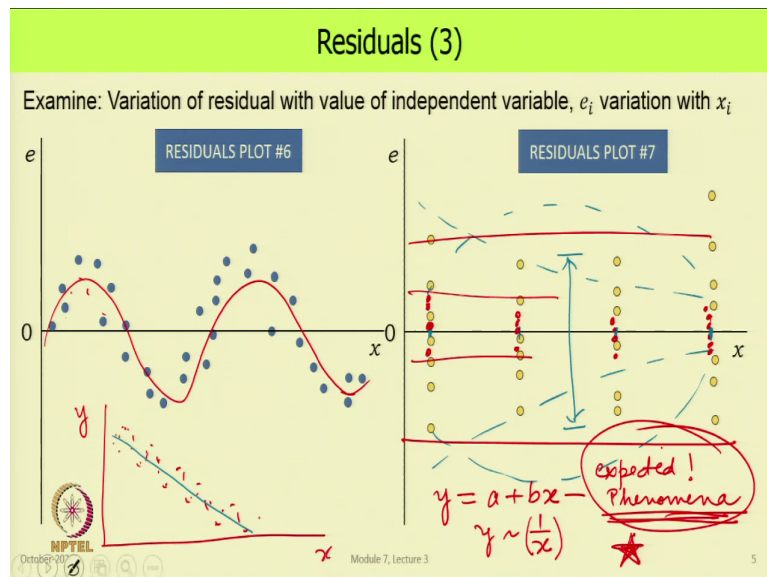
So, here the first plot here and here what we see, this is x in this direction and this side is e . And what do we have? We see a trend that, as x goes on, in general the residuals spread is

increasing; which means that the linear regression that we have fitted to this data is either not a very good assumption or the data itself needs to be relooked at to see why this is happening.

So, this is one example. The other possibility is shown here, where what we see is that the residuals we; here they are in this range then there we can see that it has got some sort of a trend coming up over here. This trend tells you that this is not good and this could happen if our data that we were originally working with was spread out like this. So, it was not, it look like that; but to this we fitted a regression, a linear regression which was like this.

So, this is x and this is y, our data would have looked like this; when we fitted this, our residuals on this side they are positive. So, that is what they do here; then they go negative, come back to 0 and then go on. So, what it tells us is that, a linear fit to this data is not a good idea and you got to do something to capture this variability.

(Refer Slide Time: 08:42)



So, this is another example. Here is yet another example, where the residuals seem to follow a trend something like periodic trend like this. So, this tells you that the raw data itself had some trend and fit in a linear regression to it was not a good idea. So, it could have been that the data was like that; this is y, this is x and to this when we fitted a linear regression, it gave residual plot like this.

Just by looking at the line or looking at the data, we may not have been able to discriminate this; but when we look at the residual plots, we could immediately say that look there is a problem. We need, we should not have put a linear regression, something else was the better fit.

And on this plot, there is something that you would see in most experiments, where we have few values of the independent variable and at each point we do the experiment several times.

So, we end up getting residuals like this. So, this spread of the residuals tells us that, if there is no trend in this we are ok; but if the residuals were say decreasing like this, then we would have to worry about it or if it was bulging in the middle, then also we would have to worry about it.

But if they are all pretty much in the same range; range means this and this (Refer Time: 10:37) randomly distributed, then we are we can be confident that what we have and what we have done the regression on is pretty good. So, this regression versus a data set, which would have looked like this say, one point here, one point there, and there and here also these are the points and here these would be the points and here these are the points; they all both these distributions of residuals, tell us that we are in good shape.

But the second one which I have drawn just now is much tighter than the first one, which tells you that this second data that we have drawn here is of a superior quality than the one that we had in the first place. So, all these plots of residuals bring us to a very important conclusion, which is that when we decide to do a linear regression; we should be pretty sure that this is what we should expect.

For example, in the previous plot, in this one, this was the behavior and this could be an experiment very simple one like we have a body which has been heated or something that has hot water and it is sitting in air and cooling. And we say that, y is our temperature and x is time.

And we know the curve is going to be like this, a Newton's law of cooling tells us. So, to fit a linear curve to this would fundamentally be wrong; we should have said that, I would expect a trend like that. So, it could be a decaying trend, which would be some constant c_1 times e to the power minus c_2 times t , T or T temperature difference.

Then the proper correlation would have been to linearize this thing and then make a fit; then we would have got this line and our residuals would have looked fine. So, this is the first thing that, what is expected is dictated by the phenomena that we are studying. If we expect an exponential decay, then if you had done this exponential fit for a very short data set, you collected data only for a very short time and you got these points; then even if you do a linear fit, it may look actually good.

But if you look at the phenomena that we are studying; then you say no, this is not right. I definitely cannot go beyond this, because it will completely give an incorrect interpretation of what is actually happening. So, this is a very important thing that, we should always make sure that the chosen correlation, the chosen regression follows at least qualitatively a phenomena that we expect to see.

If something goes as y goes as $1/x$, then fitting a linear fit to this is not going to be right; we would have had to convert this to a new variable, then do a fit and then look at it again. So, be very very careful in selecting where to put a linear regression and where if possible put a transform form. Or maybe linear regression is not the right one and it may go, have to go for higher order linear regression may be; second order, third order or something else.

So, this is something one should always keep in mind in selecting the type of function on which you will use for making other regression. So, this is something that we are seeing now

having studied how to make the regression that, deciding what the equation should be is a very fundamental important issue.

(Refer Slide Time: 15:24)

Correlation


Determine the distribution of two related variables, and the degree of association between them.

Two variables, X and Y, obtained independent of one another all other variables same

> Data for a variable from experiment (X), and from a numerical simulation (Y)

Assume: X and Y have bivariate normal distribution

Set of data pairs : X and Y :: (x₁, y₁), (x₂, y₂), (x_i, y_i) (x_n, y_n)

 MPTEL
October-2020

Module 7, Lecture 3

6

Now, we move on to correlation. It is quite distinct from what is the regression. What it tells that, it determines the distribution of two related variables and the degree of association between them. So, we are not doing a linear fit parse, linear fir or any other type of a fit; but you are seeing I have two sets of numbers and what is the degree of association between them.

Now, these two numbers in our context when we are looking at experiments would be some measured parameter or a result and we are looking at strength of association between them. But in general correlations have very large applications in many other fields and the people will try to correlate anything with everything possible.

So, what we have is that, we have two variables X and Y, obtained independent of one another by keeping all other variables the same. For example, you could have data from one variable from an experiment; we measured something and then we solved the governing equations, either explicitly or through a numerical scheme and got a value of the same variable, now from the numerical simulation.

And then we ask, can I be sure that the two are how well they are connected? So, this is the basic assumption we make in this that, X and Y have bivariate normal distribution and we our data set that we work with is X and Y which is $x_1, y_1; x_2, y_2$ like that for n pairs of data.

(Refer Slide Time: 17:45)

Correlation (2)

Degree of association is correlation, symbol ρ

- Dimensionless (unless!) ✓
- Can be positive or negative ✓
- Values between -1 and +1 ✓ $-1 \leq \rho \leq 1$
- High absolute value \Rightarrow High degree of association and vice versa
- If $|\rho| = 1 \Rightarrow$ Perfect correlation ✓
- If $|\rho| = 0 \Rightarrow$ Variables are independent


X, Y

}

X_i - expl $\rightarrow X$

X_i - Num. calc. $\rightarrow Y$

$|\rho| \approx 1$ $|\rho| \approx 0$
0.001



October-2020

Module 7, Lecture 3

7

And now we say what does the correlation tell us? The first look at the degree of association of the two parameters is correlation and its symbol for the population correlation is rho. First

of all generally, rho is dimensionless as long as X and Y are the same parameter, but coming from two different sources.

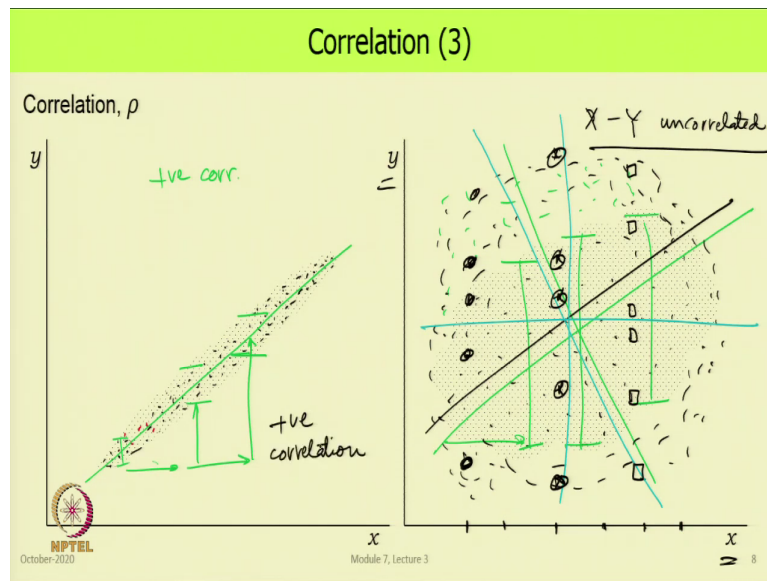
So, it could be a parameter X_i one coming from experiment and this we call X; other is same parameter X_i , but coming from a numerical calculation, this we denote as Y. So, in this case rho will then be dimensionless. But if you are correlating X_i with some other parameter which has some other dimension or with some result, which has some other dimensions, then rho will not be dimensionless.

But generally one does not do that, one looks at this or it could be from our experiment and X_i other fetched data could be from somebody else's experiment. So, that qualifies as X and Y. So, this is one thing; the second thing we are saying is that, this can be positive or negative and we will see in a minute what that means.

Rho will have values always between minus 1 and plus 1. So, the definition is such that, it cannot be greater than 1 in any case. If we have a high absolute value; then we say, we have a high degree of association. So, this basically means that, mod rho is getting close to 1; 0.9, 0.95, 0.99 like that.

But if rho is very small; that means mod rho is of the order of maybe 0.1 or 0.001 something like that, then we say that the degree of association is very weak. If mod rho is 1, we have a perfect correlation; if mod rho is 0, that means the variables are independent of one another, that means there is no correlation between what happens to one and what happens to the other. The two are completely independent and different numbers.

(Refer Slide Time: 20:44)



So, these are the characteristics of rho and here are some pictures that tell us what things look like. Here are a bunch of data points, which you are saying are spread out and this is a general pattern that they have. Of course in this case, we seem to have lots and lots of data points, may not always be the case. And by looking at it, we can say that I see something happening this way; that in general when x increases y also increases.

Of course, it does exactly go that way, in that the increase could be within a range. Say in this case this much, in this case this much, at the lower end it could be smaller. So, this would qualify as a positive correlation. Now, let us look at another case and in many experiment this actually happens, all these dots are the data points.

And we ask what is the correlation of x with y? And one could argue, but not very convincingly that, while I think they lie on a line like this; one could equally will argue that

they could lie on the line like this. And in any case, the spread is very large; that if I go that much in x , we get a very big variation in y .

And that variation in y is remaining almost constant; so that means there is a relatively weak correlation in this one. And an extreme case would be that, all these points just make, just fill up the whole chart. So, in addition to all this point that have been put up here, these points are also there.

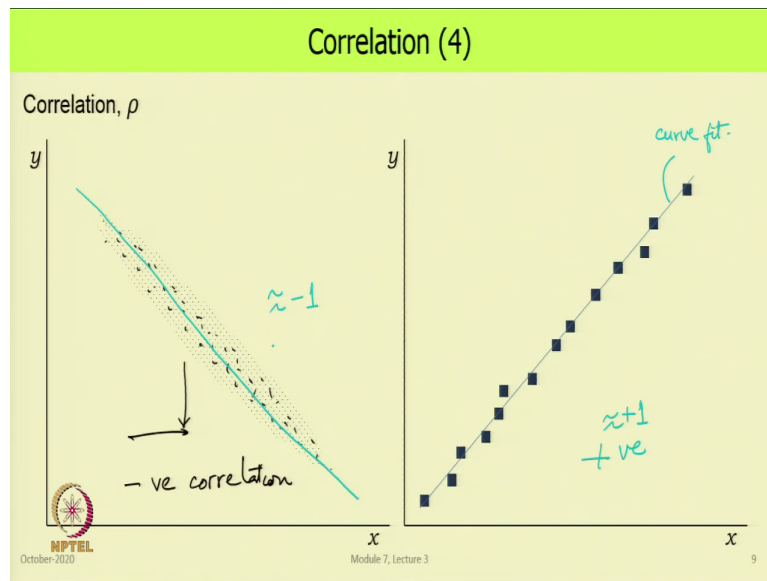
We would looking something like a big blob of data point spread all over the place. And now we say, one what is the correlation between the two and if you were to plot calculate that, it will be very very low. And then if someone says, you know I have got this data, but I want to do a linear regression on this; then we can see here that this line could be as justified as this line, as this line or even this line.

So, we have absolutely no way to ascertain whether any regression on such a paired data even makes any sense. So, for all practical purposes, we will treat this data as X and Y are completely uncorrelated. In practical work this could happen, where while X would be selected at different intervals.

The experiment is such or the phenomena is such that, there is a huge variability in the data from the same experiment when it is done repeated number of times. So, if it is an this is the type of data you could get out here, it could be this data. In this case, it could be here; this is completely a natural outcome of the experiment.

The phenomena is such that, trying to correlate this with this for that phenomena actually tells you that, you know there is no point doing it; the two are not related to one another, they do not influence each other. So, look for something else. So, this is what some correlations look like.

(Refer Slide Time: 25:59)



Here is an example, where you have points over here. And if we do a, they are well correlated that is one thing we see and what it tells us is that, as x increases, y is decreasing. So, this is an example of a negative correlation. The earlier picture that we had this one; this was a positive correlation. And again if we wanted to fit a regression that is how it would go?

In many real word things, this is the type of data we get or maybe some of these points are even spread out more; we draw a line and this is our regression or the curve fit. And we can even see whether these points are worthy of being considered to be influencing each other and we can calculate the correlation between them.

We will see a pretty good correlation and a positive 1. So, here it will not only be very close to 1, but it will be plus 1; in a if the same distribution were like this, it could be close to minus 1.

(Refer Slide Time: 27:42)


Correlation (5)

Sample correlation coefficient (Estimator of ρ):

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

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

From slope of least squares line:

$$r = b \sqrt{\frac{\sum (x_i - \bar{x})^2}{\sum (y_i - \bar{y})^2}}$$


Module 7, Lecture 3 10

Now, we come to the formula for calculating the correlation coefficient. The correlation coefficient it is r and this is an estimator of rho. Rho is the correlation that comes from the population, r is coming from the sample. And the formula are given here that, r is summation x i minus x bar y i minus y bar whole square root x i minus x bar over y i minus y bar.

So, we can x bar is nothing, but the mean of the x values, y bar is mean of the y values and using this, we can calculate the correlation coefficient r. One can see a similarity with least squares line, we have just drawn that that is r is b times these terms square root x i minus x

bar square over y_i minus \bar{y} square. So, you could even do this thing that, look at the slope of the line and then this term and that tells you what is the correlation coefficient there.

(Refer Slide Time: 29:08)

Topics in statistics

- ✓ Normal distribution
- ✓ student's t-distribution : Small sample size

- ✓ Chi-square distribution
- ✓ F distribution
- ✓ Binomial distribution
- ✓ Rectangular distribution
- ✓ Triangular distribution

Test of hypothesis →

Before modification } $-15.0 \pm 0.3 \text{ km/L}$
after modification } $-15.2 \pm 0.4 \text{ km/L}$
at what C.L.?

MPTEL
October 2011

Module 7, Lecture 3

11

So, that was for correlation coefficient. Now, coming to some other topics related to data analysis, we will not go into any detail of this; but it is worth mentioning how statistics is very useful and very relevant to this type of work. We have been talking of normal distribution, student's t-distribution this we have done.

But then there are many other distribution that we can use for different types of purposes; the chi-square distribution, the F distribution, binomial distribution, rectangular and triangular distributions, we have come across a little bit in uncertainty propagation. And then there is a major set of things for which we require this knowledge, which is test of hypothesis.

And that is what many science and engineering experiments are all about. For example, if we make a modification to a process and we want to quantify its improvement; we (Refer Time: 30:14) what was the parameter that was there before the modification and what was there its value after modification.

And now we want to make a conclusion. So, we set up a test of hypothesis saying that, this is more than this. So, this must be the case and then we do a calculation and find, whether that the null hypothesis is true or not. So, if in one case you had say fuel economy for a car as 15.0 plus minus 0.3 kilometers per liter and in another case we had 15.2 plus minus 0 point say 4 kilometers per liter.

Can we say that this is better than this? And if so, at what confidence level? So, this is something we cannot straight away answer from what we have learnt so far; it needs an understanding of the test of hypothesis. So, there is lot more beyond what we have looked at in this course and in advance topics of uncertainty analysis, many of these will be required.

(Refer Slide Time: 31:42)



I will now introduce something quite different, which is verification and validation of simulations. So, quite often it happens that when we do an experiment, we take the data and then we do a calculation, analytical calculation of solving the governing equations or we do a numerical calculation.

And then we say that for the same conditions, for the same parameter, what is my experimentally measured value and what is my value from the computation? How do I connect these two? And the question would be, can I say that the numerical model is faithful in predicting what the experiment is telling us?

So, there are two aspects to this; one is called verification, the other is validation. And developments in this field have been very recent and I have put up here three of these and

there are about four more of these; standards put out by the American Society of Mechanical Engineers, this is called V and V verification and validation 20.

And this is standard for verification and validation in computational fluid dynamics and heat transfer. Then there is V and V 10, which is standard for verification and validation of computational solid mechanics. Then there is V and V 40, which say assessing credibility of computational modeling through verification and validation application to medical devices.


With these days when we are looking at things like ventilator and what not, this is exactly what we have to be able to implement. So, these are not very broad rigorous detailed things of what to do and how to do; but they give a broad procedure on what it means and what you should be actually looking at. Its relevance to experimental work is there at one point; but it is good to see what these two mean and where that experimental work comes in.

(Refer Slide Time: 34:01)

Verification and Validation of simulations (2)

Verification *in a simulation.*
 Whether equations are solved correctly ?
 Equations, BCs, ICs, Difference technique, ++
 $\frac{dy}{dx} \approx \frac{\Delta y}{\Delta x}$
 $\pm 7\%$

Validation
 Process of determining the degree to which the model is an accurate representation of the real world.
 Compare with experimental data, self-consistency, ++
 - correlations
 - require experiment.
 $() + () + () = 0$
 $(\dots) + ()$

 NPTEL
 October-2020

Module 7, Lecture 3

13

And for that we will look at the definitions of these two terms. So, verification means that in a simulation, whether we have solved the equations correctly? And this includes setting up the equation itself. For example, we take a fundamental equation which is quite complex and then we make some assumptions and drop of some terms, make it simple and solve it.

So, we modified the equation. So, the goodness of the equation that we have used, that is one first parameter. The next thing that we have to decide is what are the boundary conditions and what are the initial conditions? So, boundary conditions could have multiple methods; you could have boundary at a value of the at the boundary or the gradient at the boundary or a mix of those two.

And that two we will end up in a real world making some approximation about that boundary. So, that is this one. Initial condition is there in the case when there are transient phenomena and something is changing with time; we come across the initial conditions and we say how well am I setting my initial condition to start the calculation.

Then comes a very big thing, which is what is the difference techniques that we have used? So, we represent an analytical equation which was in the form of differentials by difference and from there we are saying, no I can calculate dy by dx in so many ways; in our analysis for sensitivity coefficient, we said this is Δy by Δx and some of the example that I showed, we just took a central difference scheme.

But there are so many other schemes possible. So, how do we make the differential into a difference that affects the answer that we get? And then there are many more things that come in. So, in verification, we have to establish how good is this? It is not a question of whether it was exactly correct or wrong; but how good is it?

Where are the pitfalls? Is it that something is completely wrong, that is what verification tells us. So, there is very little by way of experimental work that comes into at this stage. But validation is quite different; validation is the process of determining the degree to which the model is an accurate representation of the real world.

And the real world is established by the experiment, experiment or the field data or any other measurements. So, what we do here? If we say we will take the output of a numerical model and compare it with experimental data. And that is where our correlations come in and there are many other forms of correlation for advance techniques.

We have not done that, that is for a higher level course; but that is what we would do. So, at least in correlations, we would be worried and to some extent maybe, we take data from an computation and then generate a regression. Then the issues of self-consistency; for example, in the real world, all systems must satisfy certain basic laws of physics, the law of motion, the first law of thermodynamics, Newton second law.

So, if you are doing a numerical simulation, then at somehow you have to establish that your results at any point are such that, yes we get self-consistency. In an like an experiment, even the numerical work; in theory if we get that something plus, something plus something is equal to 0 according to the physical law. In practice all of them will have variability, it will not be 0, it will have a finite value.

And then you have to go back to various techniques of statistics to prove that, statistically this and that is where it is expanded uncertainty comes in; that you prove that within this uncertainty, we say that this is zero. So, that is validation. Both these are now becoming more and more important and good work, high quality work establishes these very thoroughly.

And even if it is not exact; for example, if you do energy balance in this and you say that I got energy balance plus minus 7 percent, that may be the best that the numerical technique can give you. You may have be quite happy that 7 percent is a good number; but you should be able to establish this number from the numerical technique.

(Refer Slide Time: 39:26)

Summary

- Importance of residuals ✓ *Regression relation*
- Correlation ✓
- Various topics in statistics ✓
- Validation and Verification ✓

MPTEL
October 2020

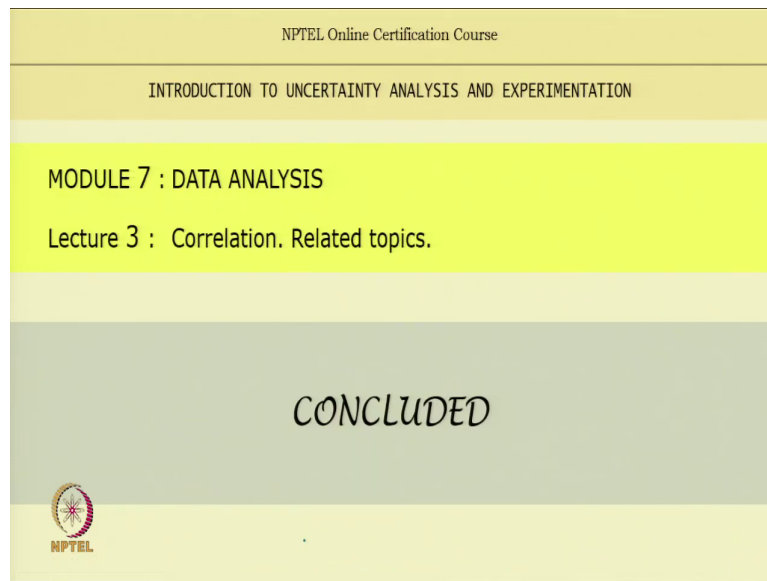
Module 7, Lecture 3

14

So, on that note, we will conclude this lecture. We started off by looking at the importance of residuals, and how to formulate the regression function. Some do's and do not we saw that, try to capture the physical behavior in the regression as much as possible; do not just blindly put in a linear regression.

Then we looked at correlations, looked at some basic aspects of that. Then we realize that there are many more topics in statistics which are required for many other things beyond just looking at uncertainty. And we concluded by saying, what is validation and verification in the context of engineering work and how this connects up with experimental work.

(Refer Slide Time: 40:25)




NPTEL Online Certification Course

INTRODUCTION TO UNCERTAINTY ANALYSIS AND EXPERIMENTATION

MODULE 7 : DATA ANALYSIS

Lecture 3 : Correlation. Related topics.

CONCLUDED



NPTEL

The image shows a presentation slide with a yellow and grey color scheme. At the top, it reads 'NPTEL Online Certification Course'. Below that, 'INTRODUCTION TO UNCERTAINTY ANALYSIS AND EXPERIMENTATION'. The main content area is yellow and contains 'MODULE 7 : DATA ANALYSIS' and 'Lecture 3 : Correlation. Related topics.'. A large grey section at the bottom contains the word 'CONCLUDED' in a stylized font. In the bottom left corner, there is a small circular logo with a star and the text 'NPTEL' below it.

On that note, we will conclude this lecture.

Thank you.