

**Engineering Econometrics**  
**Prof. Rudra P. Pradhan**  
**Vinod Gupta School of Management**  
**Indian Institute of Technology, Kharagpur**

**Lecture – 16**  
**Linear Regression Modelling**

Hello, everybody. This is Rudra Pradhan here. Welcome to Engineering Econometrics and today, we will start with a Regression Modelling and in this unit, we have five different lectures and we start with simple regression modelling, the kind of you know requirements, the kind of you know need to solve the engineering problems and some of the basic kind of you know checks. We are suppose to do before we analyze any kind of you know engineering problems by using the regression modelling.

(Refer Slide Time: 00:57)

**Learning Objectives**

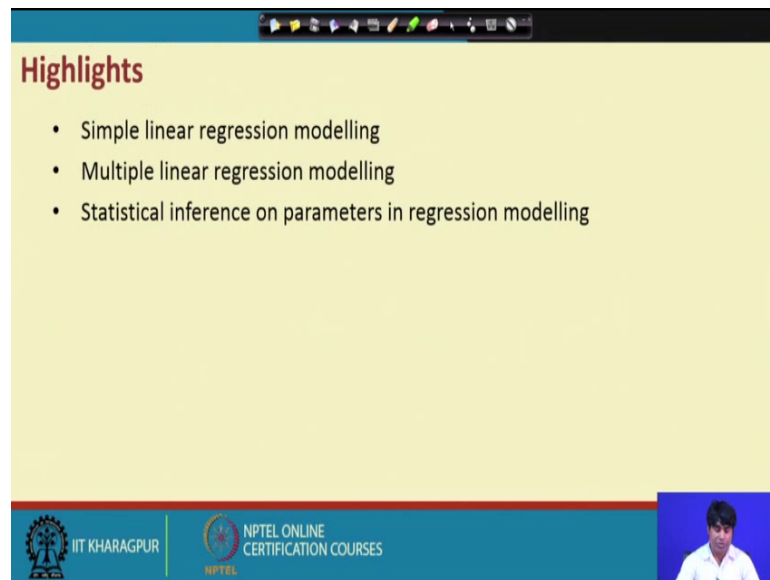
- Knowing the simple regression modelling from a sample of data, and interpret the slope and intercept of the equation.
- Knowing the importance of residual in regression modelling.
- Compute a standard error of the estimate and interpret its meaning.
- Compute a coefficient of determination and interpret it.
- Estimate the values of *DV* using the regression model.
- Knowing the structure of multiple regression modelling from a sample of data, and interpret the slopes and intercept of the equation.
- Knowing the structure of multivariate regression modelling from a sample of data, and interpret the slopes and intercept of the equation.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, first of all what are the things we are supposed to learn in this particular you know unit. So, knowing the simple regression modelling from a sample of data, interpret the coefficients like you know slopes and the intercept. It is basically a structure of mathematical modelling and the idea is just to connect the data and theory with this mathematical modelling that becomes a statistical modelling and here the basic objective of this statistical modelling or econometric modelling is to derive the parameters value on the basis of the mathematical formulation of a model based on particular engineering theory.

And, second knowing the importance of the residual in regression setup that to connecting the game between dependent variable and independent variable. Then we like to know standard error of estimate and how to interpret this particular you know item. And, compute a coefficient of determinations that is one basic or you know key indicator in the regression modelling to interpret or to analyze the engineering problem and estimate the values of the dependent variable using the regression equation or regression modelling. Ultimately the idea behind this particular modelling is to give comment about the dependent variable based on the information available on independent variable. And, knowing the structure of multiple regression modelling and then knowing the structure of multivariate regression modelling.

(Refer Slide Time: 02:51)



**Highlights**

- Simple linear regression modelling
- Multiple linear regression modelling
- Statistical inference on parameters in regression modelling

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Ultimately if you if you know go through various you know literature or the kind of you know cases and the application. So, we have two things to highlight here simple regression modelling, multiple regression modelling, then multivariate regression modelling and finally, statistical inference on parameters connecting to regression modelling in a bivariate setup, multiple setup and multivariate setup.

(Refer Slide Time: 03:25)

**Regression Modelling Basics**

- Regression analysis is the process of constructing a mathematical model or function that can be used to predict or determine one variable by another variable.
- It is a game between dependent variable and independent variable(s); and to predict DV by the help of IV(s).

Handwritten notes on the slide:

- LM (Linear Model)
- Non-linear
- RM (Regression Model)
- LM (Linear Model)
- Non-linear
- BV:  $Y = f(X)$
- ML:  $Y = f(X_1, X_2, \dots)$
- MLM:  $(Y_1, Y_2, Y_3, \dots)$
- $= f(X_1, X_2, X_3, \dots)$

Footer: IIT KHARAGPUR, NPTEL ONLINE CERTIFICATION COURSES

So, that means, technically. So, regression modelling can be of you know three types can be of three types that is actually the kind of you know structure between dependent variable and independent variables. Let me highlight the basic you know structure.

So, here first of all what is regression analysis or regression modelling it is basically a process of constructing a mathematical model or a function that can be used to predict or to determine one variable by another variable. So, it is basically a game between dependent variable and independent variables and to predict dependent variable by the help of independent variables. So, so far as a classification is concerned so, regression modelling first of all can be divided into two parts linear regression modelling and non non-linear regression modelling. And, in this unit specifically we will highlight linear regression modelling and after this particular unit we will touch upon the non non-linear regression modeling.

In the linear regression modelling so, one of the basic assumptions is the connection between dependent variable and independent variable in the form of linear setup and in the case of non-linear regression modelling, the connection or the establishment of dependent variable and independent variable in the form of non-linear equation. But, here in the case of you know linear regression modelling so, we have three different structural together.

So, that means, technically within the linear regression modelling we have three different setups whatever we have discussed you know in the last slides. So, that is the simple regression modelling, simple linear regression modelling that to simple linear regression modelling, sometimes it is called as a bivariate econometric modelling and then it is called as a multiple econometric modelling and then multivariate econometric modelling, ok.

So, that means, there are three different you know setups ah. The first setup is the bivariate structure, the second setup is the multiple structure then the third setup is the multivariate structure. In the bivariate structure the minimum requirement is the two variables and that to one is the classified as a dependent variable, another is the classified as you know independent variable. In the case, for instance, it is a in the bivariate setup so, we can write simply like you know  $Y$  equal to function of  $X$ , that is the only one variable. And, in the case of multi multiple regression modelling so, we have  $Y$  equal to function of  $Y$  equal to function of  $X_1, X_2$  and so on.

And, in the case of you know multivariate; multivariate regression modelling this is multiple regression modelling multivariate regression modelling. Here it is the game between  $Y_1, Y_2, Y_3$  as a function of  $X_1, X_2, X_3$  and so on. So, that means, we have more number of dependent variables and more number of independent variables. In the case of multiple regression modelling so, we have one dependent with the many independent variables and in the case of you know bivariate econometric modelling or simple econometric modelling or simple regression modelling we have one dependent variable and one independent variable that is why we write you know  $Y$  equal to  $f$  of  $X$ . So, it is one to one game then one to many game then many to many game.

So, these are the basic classification of you know econometric modelling that too engineering econometrics in a linear setup. So, in linear setups we have a three different engineering econometrics structures. So, bivariate format, where we have one dependent variable with the one independent variable; multiple format or multiple you know econometrics modelling, so, here we have one dependent with the many independent variables and then in the third setup we have many dependent variable with the many independent variables. So, which particular technique we have to choose that depends upon a particular you know engineering problem.

So, looking into the particular you know engineering problems we may we may use bivariate econometric modelling or we may use multiple econometric modelling or we may use multivariate regression modelling. It is the process of you know simple to complex. The bivariate econometric modelling or bivariate regression modelling which otherwise called as you know simple regression modelling is very simple and very easy to handle and very easy to interpret. Then the complexity will start when we are connecting many independent variables with dependent variable and again the complexity will more interesting and more significant when we have a set of more dependent variable with the more independent variables.

And, in the econometrics engineering econometrics we have a structure called endogeneity issue and a in the kind of you know multivariate regression modelling we will find such kind of you know endogeneity issue to address a particular you know engineering problem. So, where we will find you know bivariate kind of you know linkage or bidirectional kind of you know linkage in fact, in the bivariate setup typically simple regression modelling and then multiple regression modelling. The requirement is the univariate kind of you know linkage and that is it is called as a unidirectional linkage, where once you fix the dependent variable and that particular variable cannot be further you know treated as a independent variable.

But, in the case of multivariate regression modelling so, this a the particular dependent variables can be considered as a you know you know predicted variables and the sometimes it can be having role of you know predictor rules. So, that is why it is more complex and you know more complicated. So, in order to understand the particular you know setups how to you know develop this model and how to connect to a particular in a engineering problem and what kind of you know results you can obtained and how to interpret the results for analyzing these engineering problem so, these are the things first you first we like to know and then we move with you know bivariate to you know simple to multiple and then multiple to multivariate. That is the procedures we have to follow to do the kind of you know things as per the particular you know requirement.

So, basically we will be dealing with the simple regression modelling, multiple regression modelling and multivariate regression modelling. In the simple regression modelling it is one to one game, multiple regression modelling it is one to many game and multivariate regression modelling it is it is the game of many dependent variables to

many independent variables. So, this is the second level classification of regression modelling in the basket of engineering econometrics techniques ah. Then and then we will analyze or we will connect with different kind of you know engineering problems and to see how this these problems can be analyzed and what kind of you know inference we will obtain to comment the engineering problems and so far as a solution is concerned or the kind of you know future forecasting is concerned.

So, here there are two parts ah. Basically the so far as the objective of this particular technique is concerned so, one part is the estimation that is the predictions and the other part is the forecasting. So, with the given in you know setups we first try to find out the predictor line and then we will do the forecasting for the future requirement. So, it will give you some kind of you know future path on the basis of the estimated models or estimated regression line. So, how is this particular steps and what are the ways we can you know develop such kind of you know scenario, so, let us see or we will analyze a you know as per the particular you know requirement.

(Refer Slide Time: 12:56)

**Simple Regression Modelling**

- Bivariate (two variables) linear regression -- the most elementary regression model
  - dependent variable, *the variable to be predicted*, usually called  $Y$ .
  - independent variable, *the predictor or explanatory variable*, usually called  $X$ .

The slide includes a hand-drawn path diagram in red ink. It shows a box labeled 'Y' with an arrow pointing to a box labeled 'X'. Below 'Y' is the handwritten label 'DV' (Dependent Variable), and below 'X' is the handwritten label 'IV' (Independent Variable). To the right of the boxes is the handwritten equation  $Y = f(X)$ . The diagram illustrates the relationship between the dependent variable Y and the independent variable X, where Y is a function of X.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, in the mean times, so, what I like to say that you know we can start with a simple regression modelling. So, what I have already mentioned it is the game between two variables, one is dependent variable and one is the independent variable. So, that means, usually the structure is like this the structure is actually ok. So, the structure is like this. So, this is actually the particular structure here. So, here actually the path diagrams will

be like this Y is a dependent variable and then this is actually X is a independent variable. So, this is the dependent variable and this is the independent variable. Otherwise if you if you do not write dependent variable identity or independent variable identity, the simple you know integration can work out that is the kind of you know X to Y. So, that means, in other we can connect like this you know X to Y. So, in functional forms we can write like this Y of Y equal to f of X, so; that means, how X can influence the Y.

This is the basic understanding about the simple regression modelling, where Y is declared as a dependent variable and X is declared as you know independent variables and our job is to predict the Y on the basis of you know X availability that is what called as a predictor or explanatory variables in a particular you know engineering process. So, likewise so, we have to actually move on to this particular technique and to analyze the requirement.

(Refer Slide Time: 14:43)

**Regression Modelling: Basic Requirements for Fitting (Part 1)**

- At least two variables in the system.
- Declaration of DVs and IVs.
- Must have theory and/or logical thought of integrating IV with DV.
- Declaration of sample size.
- Data visualization.
- Reporting descriptive statistics.
- Reporting correlation statistics.
- Check the multicollinearity issue.

Handwritten notes on the slide:

- $n/k$  and  $n > k$  with arrows pointing to "Declaration of sample size."
- A scatter plot with a regression line and a circle around a data point.
- Equation:  $Y = f(X)$
- Equation:  $= \alpha + \beta X$

Footer: IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

And, a in the process of regression modelling whether it is a bivariate structure or multiple structure or multivariate structure we have some kind of you know basic requirements or what is called as you know necessary requirement and that too we must have before we enter to the game and once you enter to the game then we have also some of the basic requirements we are supposed to check and you know clear for you know for further advancements.

So, what will I do here is, we like to know what are the minimum requirements before the game and what are the minimum requirements after the games. Then the process will complete and you know as per the particular you know problems requirement. So, first of all the first and requirement is at least two variables in the system otherwise this technique cannot be applied. So, at least two variables in the system if that is the case it is called as a simple one or bivariate one and if it is more than two variables then either it will be multiple one or multivariate one.

So, depending upon the theory and the kind of you know structure so, we can connect either with you know multiple regression modelling or multivariate regression modelling. Then the second one is the second requirement to start the process is the declaration of you know dependent variable and independent variable, because here the input the need is that you know in the case of you know simple and multiple that is one basket and multivariate is another basket. In the first basket where we have a two different structure the simple and multiple, but the requirement or the structure is the unidirectional where every time an independent variables will influence dependent variables, but dependent variable simultaneously will not influence the independent variable.

But, in the case of you know second basket where we have a multivariate. So, here all the variables may cause each other and that is what is called as you know a simultaneous equation modelling or it is called as you know structural equation modelling and that is how this process is very complex and for that the declaration of dependent variable and independent variables are you know must. If you if in a particular problem if your declaration is one dependent variable with one independent variable, then you can use simple regression modelling if your declaration is the one dependent variable with multiple independent variables then the requirement is multiple regression modelling and if your theory will give you the indication about more dependent variables and more independent variables then the requirement is the structural equation modelling or simultaneous equation modelling in other words called as a multivariate regression modelling.

And, the third requirement is a you must have theory if not then you must have logic to integrate dependent variable to independent variables without logical support or without theoretical support we should not start with the this kind of you know predictions or this



kind of you know modelling ah. Otherwise the particular process will be called as you know spurious regressions or sometimes called as a nonsense regressions. So, before we start establishing the relationship mathematically or functionally so, we must have a theory behind it. So, that means, if you say Y and X. So, there is a there is you know logical reasons how X can influence Y or how Y can influence X when in the case of you know multivariate structure.

So, as a result you should you know you should be you should be very clear and means you must be very focus and then apply as per the particular you know requirement. For instance, we like to predict productivity of a particular industry and that too with the help of you know infrastructural availability. So, here the assumption is that more is the infrastructure both in terms of quality and quantity, more is the kind of you know productivity of a particular industry. So, that means, technically infrastructure in terms of quantity and quality can have positive impact on productivity.

So, that is the basic theory and yes, theoretically we convince that you know having more infrastructure or better infrastructure both in terms of quantity and quality, so, the productivity of a particular industry will be up. And, if the infrastructure availability both in terms of quantity and quality are lacking or at low level then the productivity of a particular industry will be low. And, we have lots of literature in our or we can have a lots of you know theoretical support to justify this particular you know fact. Since we have the understanding and we have the literature to say that you know infrastructure can be the cause and the effect will be on productivity so, we can you know regress infrastructure with you know productivity of an industry.

Then, the issue will come whether the productivity will influence either means only through infrastructure or any other component. So, then the complexity will start and you have to we know find out the kind of you know you know means it is a kind of you know search process to justify with that you reach a restricted model to bivariate state of where you can integrate the productivity with infrastructure only or we need to predict the productivity with respect infrastructure and sum of the other factors as for the particular you know production industry and sometimes there is a possibility that you know the productivity of a industry also influence infrastructure.

So, then the complexity will be more in complexity very high and it will be more interesting. In that case we will be move into the structural equation modelling or you know simultaneous equation modelling, but in the first instance you start with the simple one by integrating independent variable to dependent variables and that to the kind of you know requirement of you know or the ability of both theory and logic. Sometimes inbuilt theory is there you need to just verify the theory and sometimes it is exclusively new kind of you know thought and for that you have to develop some kind of you know logic before you start doing the modelling because modelling is a kind of you know mathematical process so, but econometrics is not only mathematical process it needs to justify the facts and you know figures. So, in addition to mathematical you know concept we have to integrate with the statistical way of you know thinking and then theoretical way of you know thinking, then we can merge and as a result we will have a engineering econometrics modes of you know thinking.

So, then in the next requirement is the declaration of you know sample size. So, in this case in this case there are two things sample size with respect to variables and then the kind of you know the kind of you know the type of you know on data. So, maybe time series cross external pool or panel. So, we have you know you know we must have a very clarity here on the basis of you know  $n$  and  $k$   $n$  is the sample size and  $k$  is the number of variable in the systems of course, in the bivariate setup or simple set up there are two variables, but in the case of you know multiple and multivariate setup you have a you know number of variables, in that case  $k$  will be greater than to 1.

So, that means, so far as sampling sample size is concerned. So, we have a you know optimum balance or you know good balance between  $n$  and  $k$ . So, the regression requirement is that  $n$  should be substantially greater than  $2k$ . So, that means, the sample points will be substantially higher than to number of variables in a particular you know setup. For instance, if it is two variables the sample size should be more than 30 or more more than you know 50. But, if  $k$  becomes you know 20 so, your sample size should be actually 200 or you know 2000 like this. So, it depends upon you know the kind of you know problem to problem, but every times you know the essential requirement is that  $n$  should be substantially greater than  $2k$ , even  $n$  cannot be equal to  $k$  ah. So, you must be very careful.

So, the difference between  $n$  and  $k$  that is the sample size and the number of variables should be substantially very high and positive. So, higher the difference higher is the accuracy of the model and accuracy of the fit and accuracy of the regression modelling lower the difference lower is the accuracy and it will take you to the low fit and the fit may not be statistically significant or not justified. So, that is why one way to justified you are in a relationship or modelling is to have a more sample size that too the big difference between  $n$  and  $k$  at a positive a no way. So, that means, technically  $n$  should be substantially higher than  $2k$  and the difference between  $n$  and  $k$  that is the that is what is called as  $n$  minus  $k$  is called as a degree of freedom.

So, other way to represent the kind of you know scenario is a higher the degree of freedom higher is the model of model accuracy lower the degree of freedom lower the model accuracy. So, we have to bring high model accuracy by having more degree of freedom and that too the difference should be substantially high between  $n$  and  $k$ . And so, that is one of the strong and basic requirement in any kind of you know regression modelling whether it is a simple structure or multiple structure and multivariate structure.

In fact, in the case of multivariate structure  $k$  can be further divided into the kind of you know dependent variable and a independent variables while in the other case in the simple and multiple so, every time the kind of you know dependent variable is one, but in the case of you know multivariate. So, the dependent variable and independent variables maybe you know different. So, so, we must be very careful how to fix the sample size and the kind of you know degree of freedom ah.

And, next requirement is the data visualizations. So, that means, means we start with the simple ones where we have a two variables  $Y$  and  $X$  and the process of econometrics is that not to build the model you know mathematical model. We have to transfer into a statistical form of the models and to validate this model. So, we must have a you know data behind all these you know variables.

So, so, so we start with the simple one let us say  $Y$  equal to function of  $X$  that is that is a is of it is a kind of you know functional form ah, but ultimately we start with a simple way of you know you know thought that you know since it is a linear regression modelling we may start with let us say  $Y$  equal to  $\alpha$  plus  $\beta X$ , but in reality even we have a problems we may know whether it will come to the linear modelling or it will

come to the non-linear modelling. So, so what is actually required or how to clarify the particular you know issue is you know you have to go for you know data visualization.

So, data visualization will give you the clue that you know whether the particular functional relationship will be linear one or non-linear ones if it is linear ones then whether it will come you know in a downward shape or you know straight line shape or something like you know kind of you know 45 degree angles whatever may be the way. So, it will give you some kind of you know you know a basic inference about the variables and the kinds of you know functionality. Then, of course, through regression modelling we can actually validate the particular you know thought or you know bring the strength of strength to this particular you know thought.

So, one of the basic requirement again in the process of regression modelling is the data visualization. So, having good visualization in fact, I have already you know mentioned in the second unit, data visualization is not a one step process. There are various ways you can you know visualize the data a simple line diagram to bar diagram something like that and you know like to know how is the particular you know structure. Understanding the data, the structure of the data, the nature of the data, the sample size, the sampling structure all these are you know basics for you know handling the particular you know issue, without which it is very difficult to address the things as per the particular you know requirement.

So, data visualization is one of the strongest requirements for you know regression modelling, because it will take you to the path of either simple one or multiple one or multivariate one. And, that too whether you are in the process of the linear modelling or in the process of you know non-linear regression modelling and the next requirement is reporting the descriptive statistics against each variable. So, without knowing the descriptive statistics you are not sure whether the particular data corresponding to that particular variable is very consistent and very stable.

Ah Sometimes the data may have outliers problem. So, outlier is a concept where you know most of means a it is a data point which is a exclusively or highly distance from other data point. For instance, so, in a kind of you know plotting most of the data points are here in this particular boundary, but one data point will be here. So, now, if you are you know clubbing this all these data points in analysis definitely the problem will be or

the kind of you know so, it will be biased because it will give you lots of volatility or you know variation. So, that means, the dispersions of this particular you know set dataset will be very high. And, if you remove this data points and just keep you know take this much of data point in the variation will be low and this will strengthen your you know modelling process.

But, the question is you know there are two things to identify the outlier and second thing whether to remove the outlier or you know whether you are supposed to do something to you know materialize the kind of you know a requirement. So, it depends upon the data structure. Ah, If it is the the case of you know cross sectional data then it is very easy to drop that data point and then handle the problems with you know other data points, but in the case of you know time series data you cannot just drop this data point because it will it will bring you the inconsistency in the process of you know time series data.

And, in the same times whether it is a cross sectional kind of you know setup or time series setup if it is the presence of you know outlier. So, one solution is it to remove the outlier and handle the problem that will give you some kind of you know unbiased results. Otherwise, it will give you to some biased result and in other instance you can remove the outlier and then this then you solve the problem and there is a high chance the results will be completely unbiased. Provided there may not be any other obstacles and these obstacles we will discuss in later stage.

And, in the third case so, there will be a you know issue of some kind of you know outlier, but with the you know manipulation of data or normalization of data the degree of you know influence that too the degree of influence of this particular in outlier may be minimized. For instance, the first in data and you can analyze and then block transport data or you know first difference transport data if you analyze then you will find big difference. In the low transfer data the violability will be very low and in that case the impact of outlier will be minimizing there.

Or in the case of original data the in fact, of you know outlier will be very high and in this case the particular you know model fit will not be very good and it will not give you some kind of you know unbiased results, but in the context of you know normalized data where the impact of outlier is the minimize, so, you may get you know some kind of you

know perfect result and that result you know maybe completely unbiased and that that depends upon you know you know means how to understand how to you know you know bring that particular you know setup so, the typical requirement is to check the descriptive statistics.

And, so, the descriptive statistics that too mean median mode maximum minimum then the standard deviation variance skewness all these things you know basic statistics you have replot it will give you some kind of you know indication. Sometimes there is high chance that you know in a particular variable all the data points are you know same. In that context so, the standard deviation equal to 0 and if all the data points are same whether it is a with whether it is with respect to dependent variable or independent variable. So, you are not in a position to you know do the regression modelling because it going against the systems.

So, one of the requirement of regression modelling that too with respect to data there is a uniform data availability for both X and Y and within the X so, all the data points you know should be different there should not be similarity ah. If all the data are not you know different they should not actually same if all the data points are same so, that means, this is going against the regression modelling. As a result so, we we should not analyze and if you analyze the results will be completely a indeterminate and that too you know completely unbiased sorry completely biased. So, so, it will go against the regression modelling setup.

Of course, we have a concept called as you know means there is a concept called as a Gauss-Markov theorem and popularly known as BLUE theorem linear best linear unbiased estimator. And, most of the regression modelling output with the pass through this Gauss-Markov theorem that is how the validity check of the particular you know econometric modelling and where the particular model will be declared as a best if it is means the context of linear modelling. So, linear unbiased and must have minimum variance. So, the presence of outlier again and without checking the descriptive statistics so, the particular you know regression output may go against the BLUE theorem. If the particular output is going against the BLUE theorem then the particular output cannot be used for you know analyzing an engineering problems or the kind of you know any kind of prediction.

So, to normalize all these you know issues or you know bring the efficiency of this particular you know econometric modelling and that to the solution to the engineering problem. So, you are supposed to check the descriptive statistic against each variables and then you know then you like to manipulate or normalize as per the particular you know requirement. So, regression modelling will not have any kind of you know issue, whether you are using the actual data or you are using the normalized data or manipulated data.

So, ultimately we need to develop a model which can help you know some kind of you know solution to a particular engineering problem if that model is actually means the model must be very perfect. If the model is not perfect then you know then we should not you know think about you know perfect solution ah. So, the requirement of the perfect fit model depend upon the basic you know you know requirements of the basic checks. So, this will give you some kind of you know boost to the process of you know efficient model.

And, then we refer we like to refer the correlation statistics because regression is a a kind of you know modelling where we have a causality issue, but the process of causality you know we can start when there is a kind of you know relationship, that is how you know we connect with a theory and the kind of you know logic. So, the the reason behind you know establishing logic and connecting to the theory is that you know we like to justify that there is a there is a kind of you know relationship and a that relationship you know technically can be quantify before you go to the regression modelling with the help of an of you know covariance and correlation.

So, that is why before you start processing of regression modelling you are you are suppose to report the correlation statistics or covariance statistics, it will give you some kind of you know confidence that you know yes, there is a relationship and that too with the help of you know theory and logic and that too you know some kind of you know quantifications through covariance statistics correlation statistics. And, this is again one of the basic requirements of any kind of you know regression modelling that too whether it is a simple one, multiple one or multivariate one.

And, the last, but not the least final check is the multicollinearity issue. This is actually a typically econometrics terms or you know statistical terms ah. We have a separate lecture

for that, but in the meantime I let you know that multicollinearity is a problem where there is an existence of linear relationship among the independent variables. And, that too this particular component is always there or you know there is a chance of you know issue in the context of multiple and multivariate, but it is not the issue in the case of you know simple regression modelling or bivariate regression modelling.

So, the literary meaning of multicollinearity is the existence of linear relationship among the regressors that is the independent variable. Since we are declaring that you know  $X$  are independent variable against the dependent variable  $Y$  so, if it is a single  $X$ , then there is no chance. If there are multiple  $X$  that is the independent variables say  $X_1, X_2, X_3$  then if they are independent then the relationship between  $X_1 X_2, X_1 X_3$  or  $X_2 X_3$  should not be there.

So, that means, technically having correlation matrix you know that we have mentioned in the last step. So, it will give you the clear cut idea that you know whether there is a kind of you know relationship among the regressors and in reality it is very you know very rare chance that you know you will find a situation where all the independent variables are completely independent, but with the help of correlation matrix you can actually you can mark and you can justify. And, what is needed for this particular you know requirement for the regression modelling, you like to check that you know whether there is a correlation among the regressors, if not then this is a very perfect and that is the best and if it is then you are supposed to check whether they are statistically significant or not. If they are not statistically significant then you can go ahead with the particular you know regression process.

If they are statistically significant then you have to work out differently and the first step of the process is how to minimize this particular you know relationship or you know degree of you know association. And so, there are many solution tricks how to reduce this particular you know relationship means the quantification of this relationship and before you start the regression modelling. Until unless you minimize this particular you know relationship; that means, statistically significant to statistically not significant then you can go for regression modelling, if not then it is a big you know issue before you know process the data and to analyze the problems as per the particular you know engineering requirement. So, you must be very careful how you have to deal the situation.



(Refer Slide Time: 42:32)

**Regression Modelling: Basic Requirements for Fitting (Part 2)**

- Report the estimated parameters.
- Report error term (actual DV and predicted DV).
- Check the sum of error term.
- Check the variance of error term.
- Report the standard errors of estimates.
- Test the estimated model.
- Check the diagnostics of the model.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

And, then finally, regression modelling means what is the after process. In the in the process of modelling the idea is to you know develop you know mathematical model, then statistical models and when when we connect with data.

So, the modelling process will give you the parameters value and after getting the parameters value and that is how the regression modelling is all about. For instance, in the middle stage we have a regression output and before processing this output we have some requirement which we have already discussed and then in the process of regression modelling you will come with the some kind of you know output that is called as regression output or econometric output. And, then what kind of you know things you have to do you know; that means, what kind of you know basic checks we are suppose to do before you start processing the output for you know solving the engineering problems.

So, first of all you have to report the estimated parameters with the help of some kind of you know formulation or you know calculation, then report the error terms which is the difference between the dependent variable actual and the dependent variable predicted. And, third check is the you know requirement of you know sum of the error term if the model or the functional of you know goodness of fit or you know functionality of this particular you know process is accurate then the error term sum of the error terms should be equal to zero.

So, the issues here you know the error term is the difference between actual and predicted. So, there is there is a high chance that you know sometimes the difference will be positive and there is a the difference will be negative and since the actual plotting will be for instance, it will be like this the actual plot ok, the actual plotting will be the actual plotting will be like this and the predicted line will be like this. So, as a result we have some positive difference and we have some negative difference.

Now, the basic requirement of the regression modelling before we start the estimation we will be in the best if the line in the straight line or predicted line should be in between middle of these points. If that is the case then you will find 50 percent of the error terms are at the positive sides and 50 percent of the error terms are in the negative sides. As a result, if you know sum up then the sum will be equal to 0 if that is the case then your model will be considered as the perfect fit in the first instance. If not then you can actually rectify and try to structure restructure again, so that this particular structure can happen.

And, then check the variance of the error terms. So, the variance of the error terms should not be equal to zero. If that is the case then it is very difficult to again analyze the problems. So, there should be some kind of you know error variance, but that error variance should be at the minimum level and we our aim is to how to minimize this error variance or minimize the error terms.

Then, report the standard error of the estimates which is one of the basic requirements for inferential parts of the regression modelling because the first part of regression modelling is to have estimated output that is the process called as a first end output. Then before you use this output for solving the or analyzing the engineering problem you need to test it or; that means, technically we like to check the reliability part of the estimated model and for that standard errors will help you lot to validate the outcomes as per the particular you know requirement.

And, then we finally, test the entire model the estimated models because the estimated models will give you lots of you know outputs that that includes the parameters value, then the kind of you know goodness of fit that is the declared in the form of you know coefficient of determinations and some of the other statistic which we called as you know

analysis of variance error sum square, sum square totals the these are the items called as you know regression output which we will discuss in details in the next lecture.

But, ultimately after having all these outcomes, it is your duty to test and you know check whether it is on the line as per the requirement of regression modelling. And, then finally, you will go for you know checking the diagnostics of the model, where we have a multiple things we have to check and you know bring into the kind of you know consistency label before you a come before you analyze the problem and comment about the process and to you know go for the kind of you know prediction and forecasting.

And, with this we will stop here and we will be discuss in details about this particular processing in the next class.

Thank you very much. Have a nice day.