

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

Lecture – 20
Linear Regression Modelling (Contd.)

Hello, everybody. This is Rudra Pradhan here. Welcome to Engineering Econometrics and today, we will continue with Linear Regression Modelling. In the last lectures, we have discussed something the reliability of the estimated regression line that is the bivariate regression modelling or bivariate converting modeling. And we specifically highlighted couple of test through which we can justify the fitness of the particular model. And this includes the first one specification test, second one the goodness of fit test then diagnostic test and the out of sample prediction test.

And, in the last lecture, we specifically connected to specification test and goodness of fit test where we have actually validated the parameters, significance of the parameters that is alpha head and beta head. And then we have validated the coefficient of determination r square and that too for goodness of fit measures. So, we have already discussed all these things and the procedure through which you can validate we have already highlighted. And to continue that lectures we can come to this particular discussion these are all we have discussed, and ultimately the standard structure after validation is like this.

(Refer Slide Time: 01:53)

Point Estimation for the Airline Cost Example

$$\hat{Y} = 1.57 + 0.0407X$$

For $X = 73$,

$$\hat{Y} = 1.57 + 0.0407(73)$$
$$= 4.5411 \text{ or } \$4,541.10$$

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, this is what the estimated equations. And ultimately, the issue of specification test and goodness of fit test we will not add something more in the estimated equation that is what we have here. The estimated equation is ultimately here and this equation will help you to predict the particular dependent variable with respect to independent variable. But, the specification test and the goodness fit test is the mandatory component to give the signal whether the model is absolutely for the predictions or not. Until and unless you check the specification test and the goodness of fit test, so, we cannot use this equation for the prediction and forecasting.

So, now, after the test that is both specification test and goodness of fit test we are now clear or we are now confident that this model is and three form all kinds of errors. Then, now we know means we are now in a position to use this model for forecasting. So, now what I will like to we justify here that this model is actually to predict Y with respect to X and we should know; what is the process of prediction and forecasting with this estimated a equation or estimated line. So, the ultimate the estimated equation is this one. So, now, this equation itself give a clear cut indication that we need to predict Y that is what Y hat is all about and that to with respect to X indications because this parameter is now known and specified.

So, this parameters value for this particular equation and for this problem for this sample will not change, it will be remain constant. Only change will happen in the X in X side

and then by default the change happening in the Y head so; that means, if we change any X then Y head will be also change in there is a no change in the data itself; that means, it is all about the change with respect to future data. So, we have a 12 data points, now if you add any one more data point let us say 13 by putting a X in any value. So, the average or the lower one or higher one corresponding to the twelfth unit or eleventh unit and then if we will check; how is the value of Y that is Y predicted.

For instance, let us put you know X equal to 73 and then we have actually Y head that is equal to 1.57 into beta coefficient into 73 after simplification. So, the figure is coming 4 point 5411. So, that means when the passengers you know numbers we will change to 73 the airline cost will be actually moved to 4541 dollars. So, this is how the predicted structures. So that means it is now a good way to derive the predicted value corresponding to the change of X indication.

So, regression line or you know regression modelling is a very beautiful component to work out all these you know predicted value for the dependent variable with respect to change of any independent variables. So, let us see what is more about this particular process.

(Refer Slide Time: 05:46)

Confidence Interval to Estimate μ_y : Airline Cost Example

$$\hat{y} \pm t_{\frac{\alpha}{2}, n-2} S_e \sqrt{\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{SS_{xx}}}$$





where : X_0 = a particular value of X

$$SS_{xx} = \sum X^2 - \frac{(\sum X)^2}{n}$$

For $X_0 = 73$ and a 95% confidence level ,

$$4.5411 \pm (2.228)(0.1773) \sqrt{\frac{1}{12} + \frac{(73 - 77.5)^2}{73,764 - \frac{(930)^2}{12}}}$$

$$= 4.5411 \pm 1220$$

$$4.4191 \leq E(y_0) \leq 4.6631$$





And, in this context in this context so, this is what the kind of more some something more about the regression outputs. Ultimately, we have already discussed a concept called as you know confidence interval. So, whatever we are discussing with respect to

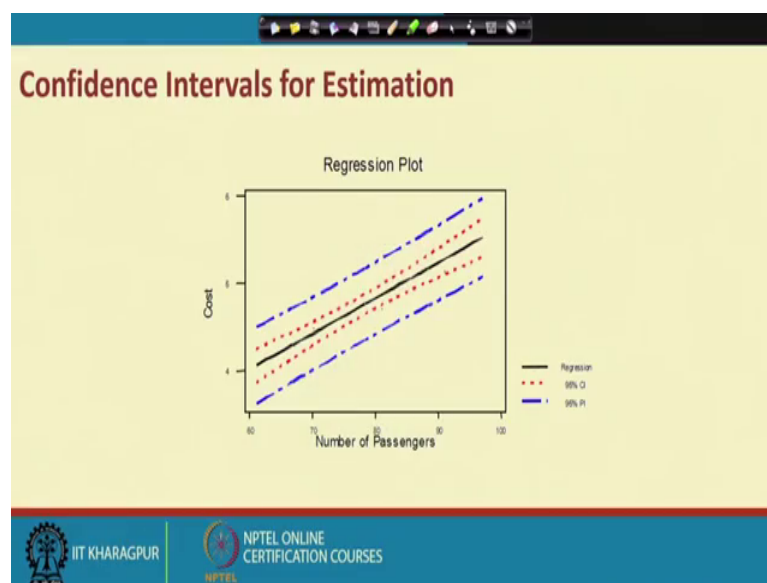
parameters and these are all called as you know average values. So, when we are considering about the average value. So, then we will sometimes you know talk about the confidence interval. So, the confidence interval ultimately depends upon you know plus minus you know something with respect to t statistic and that too with respect to degree of freedom and the probability level of significance.

So, this is actually this is what the kind of mean figure and mean figure and ultimately the confidence interval will be this, this is a t value and this is what actually the kind of standard error of the estimates. So now, with the help of t value and standard error of the estimates we can get the confidence interval and that too so, with average. So, we can have a some kind of compromise with left side and sometimes compromise with right side.

So, that means, corresponding to 4.5411. So, the lower limit will be 4.4191 and the upper limit will be 4.6631. So, that is the confidence interval through which actually the prediction can be generated we cannot sometimes you know very stick to the particular line estimated lines, but somehow you can give some kind of confidence interval, because ultimately the it is very important so far as you know application is concerned, because some variation maybe allowed and that variation should be in a kind of limit or what we called as you know confidence intervals.

So, in order to clarify this, let us you know see here you know graphical visualization and that is what actually here.

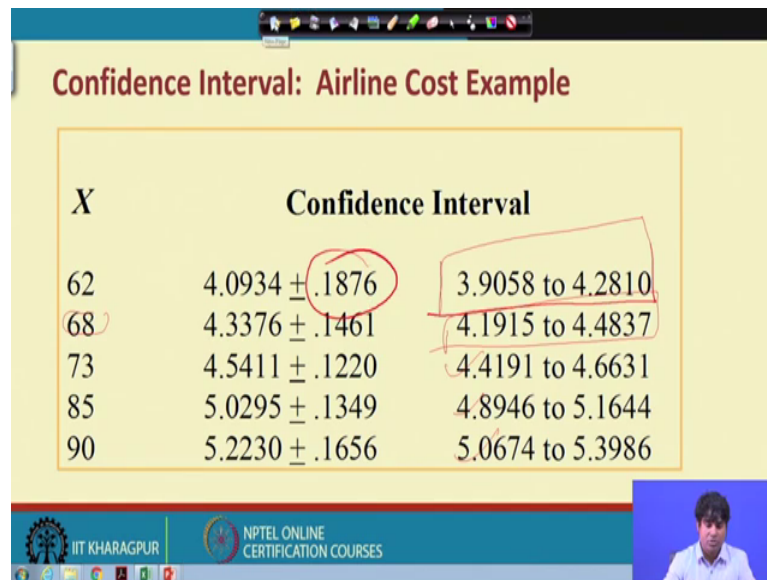
(Refer Slide Time: 07:55)



So, see here so, this is what the predicted lines the black one, and then we have the kind of confidence intervals. So, the confidence interval may again vary with respect to the probability level which you like to fix. So, because the standard three probability levels we which we usually if you know apply for testing the parameters and the model 1 percent, 5 percent and 10 percent. So, if you put 1 percent then the confidence interval will be somehow at a particular positions. Then, when we change the probability level to 5 percent, then the confidence interval will again change compare to one percent and then if you apply 10 percent then again it will change.

So, ultimately whether it is 1 percent 5, percent and 10 percent so, there will be a confidence interval corresponding to the average line. So, ultimately the question is what interval we have to we know apply for this particular problems. So, it depends upon you know where you are coming means where the parameters are coming statistically significant. If it is at 1 percent then you can fix the confidence interval at 1 percent, if the parameters are statistically significant at 5 percent then you can fix at that particular confidence again if it is at 10 percent then you can fix there itself at the 10 percent. So, likewise so, this will be give you the clarity.

(Refer Slide Time: 09:24)



X	Confidence Interval	Confidence Interval
62	$4.0934 \pm .1876$	3.9058 to 4.2810
68	$4.3376 \pm .1461$	4.1915 to 4.4837
73	$4.5411 \pm .1220$	4.4191 to 4.6631
85	$5.0295 \pm .1349$	4.8946 to 5.1644
90	$5.2230 \pm .1656$	5.0674 to 5.3986

And, for that this is what actually these are X values. So, now ultimately the confidence interval depends upon you know X values and the t statistics and the probability level of significance and the degree of freedom. So, this part ultimately depends upon depends upon the kind of X values, standard error and that too t values degree of freedom and that too n and k.

So, if you simplify for when X equal to 62 then the cost factor will be ranging from 3.91 to 4.28 when X passenger size will be 68 then your cost factors airline cost factor will vary from 4.19 to 4.48. Likewise when X will be changed 68 to 73, then 85 and 90 so, the variation will be also happening like that. So, these are all you know different kind of a you know confidence interval just to predict or to make the accuracy of the models and to analyze the problems as per the particular requirement.

So, now this is equally actually highlighted in the context of this problems graphically and that gives the clear cut indication that the you know the regression output or regression line in the that is the predicted lines will be lying in a kind of confidence interval. So, they small variation of this particular prediction will not get affected. And, ultimately in order to again stimulant the particular process we can go through the out of sample prediction test and that too very easy for a time series data where what we can do.

So, having the actual data whether it is a 12 points or you know 120 points, so, you try to use 80 percent of the data for estimations and keep you know 20 percent data reserved for the validations. So, ultimately we are doing the validation with the estimated you know models that too the used data of the 80 percent. Then once you get the you know a estimated model and validate the estimated model then just that line will be extended towards the right that is towards the you know futures. And, ultimately in the future structures 20 percent data already reserved that is the actual data point.

Now, this is predicted data and predicted kind of line and the actual availability is there. So, you just draw the line extend that lines and then compare the estimated line or the kind of predicted line with the actual points and then check the kind of difference and the kind of error. So, if that particular error and that particular difference is very very minimum and close to 0.

So, that means, this line can be considered as the best line that is the structure what we called as you know out of sample prediction test and it is very easy for you know means easy to use in the context of time series data or time series modelling, but it is not so easy in the case of cross sectional modeling. The reason is that in the case of cross sectional modelling the reserve of 20 percent data or 10 percent data or the 25 percent data for validations is always question mark. So, which particular data points should be reserved.

For instance let us say 12 data points, so, it is a cross sectional data of different industry and when there are twelve different industry data are there and which you know industry data that is out of twelve 80 percent of data will be used for estimation and remaining 20 percent of the data out of this twelve industry will be kept for validation. So, this is always actually question mark why not the kind of reverse? If you take that first you know first you know eighty percent of the data will be estimation and a next the kind of 20 percent of data for validation, but in the cross sectional data. So, the ordering of the variable is you know so much fixed actually.

So, for instance company 1, company 2, company 3, company 4 so, there is no hard and fast rule that you have to arrange like that. So, the company 3 can go fast company 4 can go fast like that. So, as a result cross sectional data. So, the ordering is not so consistent. So, any company can come fast any company can late. Ultimately, the data structure will not change, estimation process will not change, but what will be question mark is the

data which you keep for you know reserve and that too for you know validation of the estimated model.

So, that is how. So, it is always suggestive to apply out of sample prediction test in the context of time series data and it should not be you know use you know significantly in the context of cross sectional modeling, because of these issues as. But, in the case of time series modelling or time series data so, this is one of the mandatory test and this test in fact, gives better indications so, as far as a model validation is concerned. Because the accuracy of the estimated model will be more you know perfect in by the help of out of sample prediction test.

Because, it is comparing the estimated with the actual availability, but when you use the entire 100 percent of data for estimation then you extend the lines and that too generate the predicted points or predicted line and where the actual data will not be there. It will be you know actual data will be coming later stage, but for the time being it is not readily available to compare and you know conclude. So, that is how you know you can apply you know out of sample prediction test for the time series data and against for you know large sample size. Having 12 12 data points and it is not suggestive to keep you know 20 percent data reserved for the comparisons and only 80 percent of the data to be used for you know estimation.

Instead of 12 data points if you have 120 data points then first 100 data points can be used for model building or model estimation have the estimated model and last 20 data points should be kept reserved for you know testing. Again, this is the first instance how you can do this kind of structuring and ultimately the best structure will be means in suppose you like to check the robustness of the results and modelling. So, you can use you know 100 data points and 20 for you know testing reserve 20 for testing and then check the kind of structuring and finally, again you can use entire the 100 percent data 120 data points and get the estimated line and compare the previously estimated line that too with respect to 100 data points.

And, then the estimated line with respect to 120 data points and then you know just you know move these two lines in a kind of plane. So, you will check whether they are you know converging each other or not. So, if that is the case then increase or decrease of sample size will not affect the particular process so far as the estimation and the kind of

forecasting is concerned. Sometimes we use actually different a kind of robustness check of course, we are going through specification test, goodness of fit test, diagnosis test, out of sample prediction test, but after going through all these particular test structures. So, still we may have a doubt about the validation of the model.

And, to again minimize this particular doubt or you know issues and the kind of error. So, you can every time go through robustness check and the robustness check is not a kind of single component. There are multiple ways you can check the robustness and one of the way to check the robustness is the applying you know different sampling you know sample size in the estimation process. For instance original sample is having 120 and you can start the modelling with you know 70 data points, 80 data points, 90 data points, 100 data points 120 data points and check every time how is the estimated line.

So, estimated line at 70 percent 70 percent data, estimated line 80 percent data estimated line 90 percent data and estimated line you know entire data that is 100 percent data and then check whether the particular lines are exactly parallel or not. So, there will become closer actually of course, an increase of sample size always having better accuracy of the model. But, ultimately after a particular point of time increasing sample size will not actually change the results too much, but decreasing the sample size may also affect the particular process.

So, it is not mandatory that every time increase sample size will increase the you know level of the model accuracy. So, model accuracy if you say model accuracy then highest is at the 1 percent level, after that an increase of sample size may not you know affect the particular process too much, because ultimately we have reach at the highest level and, but it will be very you know problematic or kind of top of issues when we have actually multiple regression modelling and multivariate regression modeling. In that context there are couple of things simultaneously you have to work out when you target a particular variable then the other variables you have to compromise.

So, that means, you will be find in a particular situation where some of the variables will be statistically significant and some of the variables are not statistically significant. So, in that context we will use you know increased sample size or sometimes we are searching for the optimal sample size where the model will be perfectly fit and where the all the variables will be statistically significant. But, when you are a bivariate you know kind of

environment, so, increasing sample size is not a big advantage, because for bivariate models at a minimum level of for optimum level of sample size it will give you statistical significance at the highest level. After that increase in sample size will not add any extra value to the particular process.

Of course, the estimated coefficient will increased, so, that that will give you the indication about the percent contribution of independent variable to dependent variable. Ultimately, significance of the particular models or the kind of line will not you know drastically effect or it will not affect at all. Once you reach 1 percent levels then an increase of sample size will not add any extra value to this validation. Ultimately the coefficient value will be changed because that is the only way to interpret the kind of problem so far as you know prediction and you know forecasting is concerned.

(Refer Slide Time: 21:38)

Pearson Product-Moment Correlation Coefficient

$$r = \frac{SSXY}{\sqrt{(SSX)(SSY)}}$$

$$= \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2 \sum (Y - \bar{Y})^2}}$$

$$= \frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sqrt{\left[\sum X^2 - \frac{(\sum X)^2}{n} \right] \left[\sum Y^2 - \frac{(\sum Y)^2}{n} \right]}}$$

$$-1 \leq r \leq 1$$

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

And, for that the kind of structuring is like this. One more important is here the kind of comparison between correlation and regressions, because correlation like you know covariance and correlation are very closely related and they depend each other similarly correlation and regression are also close related to each other and that too for a bivariate framework you know at least.

And, in the regression setup for a bivariate kind of environment one of the most important indicator is the r square capital r square that is the coefficient of determination. And, in the context of correlation coefficient r is the indicator a fitness indicator to

measure the association. So, here r^2 will measure the association capital R^2 and here small r^2 will indicate the measures of association. So, what is more important here to bring this concept for a bivariate setup so, small r^2 always equal to capital R^2 .

So, as a result; so, far as a measuring the fitness of the model is concerned or you know relationship is concerned then having r^2 small r^2 and capital R^2 will be bring the same kind of conclusion. And, what is more important the correlation coefficient you know r^2 small r^2 will not exactly same of capital R^2 in the case of multiple regressions and multivariate regression. So, that will be different.

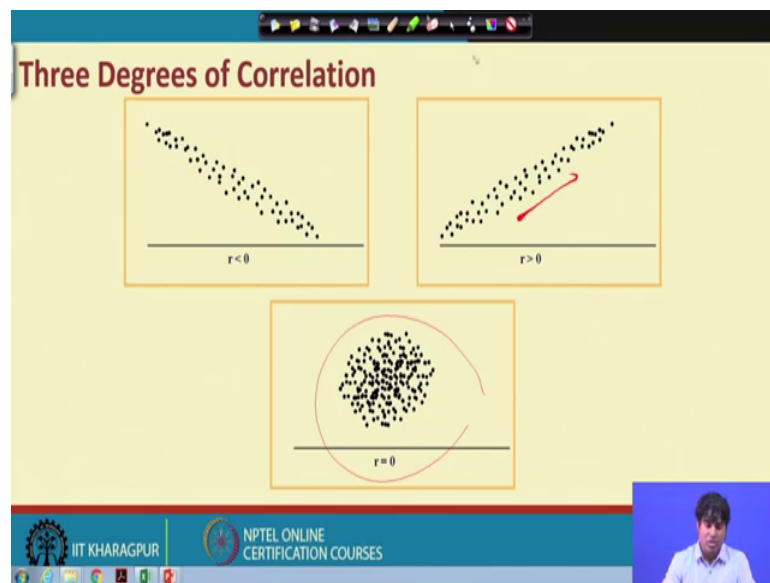
And, another important is correlation coefficient r^2 will also lie between 0 to 1 and in fact, only correlation coefficient will be in between minus 2 1. So, when we are squaring the correlation coefficient that becomes you know small r^2 it will be in the range of 0 to one and in the context of regression modeling so, the indicator is a r^2 coefficient of determination which is also between 0 to 1.

So, the indication is that when small r^2 close to 1, then they are perfectly correlated and when small r^2 is equal to 0 then there no relationship or you know association. Then when r^2 is coming actually close to 1 then the interpretation is that the independent variable is you know hundred percent affecting the dependent variable. That is the percentage variation of X and Y that is the kind of interpretation you have to bring here. So, when it is a actually one then percentage variation of X influence Y a 100 percent. So that means, you know if you increase one you know one level of input, then obviously output will be also increased 1 percent level.

So, if that is how the kind of signal you will find from the r^2 then that to the game between Y and X. So, when r^2 is close to 1, then the then the accuracy of the increase or decrease depends upon the X to Y exclusively depends upon r^2 strategy. When r^2 will be low and some kind of lower accuracy will find to predict X and y, but ultimately both are you know good indicators to know the relationship between two variables in one case we are drawing the association kind of structure in another case the causality kind of correlation kind of things where the influence of X and Y is derived.

And, but ultimately there you know same at a particular context. Ultimately for bivariate kind of environment so, there is no big difference between Y and X. So, far as a association is concerned because while you know regressing Y and X ultimately it is a kind of kind of association structure. Ultimately more important is the cause and effect kind of environment in the regression case and while it is the correlation case it is kind of simply degree of association.

(Refer Slide Time: 26:23)



And, so, the degree of association will be like that it may be negative it may be positive it maybe linear sometimes it maybe non-linear. But till now we are discussing something related to linear correlation, linear regression and, if the points will be coming like this. So, this will be negative the kind of structure and this is positive kind of structure and here this is there is no association at all. So, it is a kind of scattered and will not give any kind of indication or association and that can be also predicted through you know kind of Y indications, ok. So, this is what the kind of structuring about the correlation and regression that too in a bivariate setup.

So, ultimately we have discussed the regression modelling in the bivariate context that too with respect to two variables, where one is dependent variable another is independent variable and we have we get to know how to estimate the process, how to obtain the estimated line, how to validate the estimated line through specification test, goodness of fit test and somehow the out of sample prediction test. But, the diagnostic a part of this

particular estimation procedure is little bit complicated which we can discuss in the later stage. And, because it includes so many components and various requirements; you know requirements. Until and unless these requirements you cannot proceed for the diagnostic test whether it is bivariate structure or multiple structure or multivariate structure.

So, till now whatever we have discussed that is with respect to specification test, goodness of fit test and out of sample prediction test. So, the diagnostic test part will covered in the latest stage. So, first we will go through multiple regressions and after knowing the multiple regression structure some of the structure of the multiple regression is very important while doing the diagnostic check. And after the multiple regression modelling we can come to the particular test that is the diagnostic test to know how is the validation of the estimated model through this through this particular test.

So with this, we will stop here.

Thank you very much. Have a nice day.