

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

Lecture – 18
Linear Regression Modelling (Contd.)

Hello, everybody. This is Rudra Pradhan here. Welcome, to Engineering Econometrics. Today, we will continue with regression modelling and that too continue with the process of the bivariate estimation and in the last lecture, we have discussed the problems relating to airline cost determinations. The objective of this problem is to predict the cost airline cost subject to airline passengers. That means there are there are couple of factors or variables which can affect the airline cost, but number of passengers is the one of the most important variable that can affect the airline cost drastically. So, that is how in the last lectures we have discussed this problem.

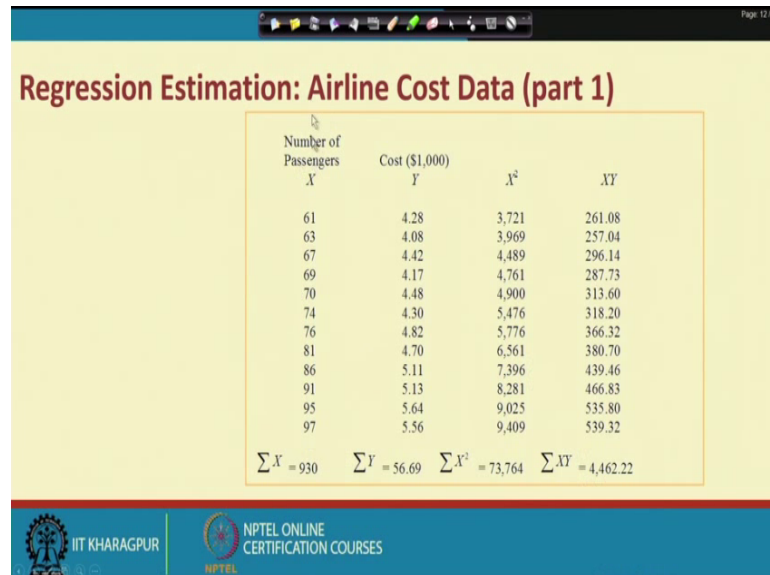
And, further we have used simple regression modelling that is the bivariate a econometric modeling, where Y is the dependent variables and that to the airline cost and X is the independent variable that to the airline passengers. And since we are dealing with linear regression modelling we have used a linear regression modelling that to Y as a function of X and to the simple equation is equal to alpha plus beta X where Y is treated as a airline cost and X is treated as a airline passengers and having a particular sample we have already discussed in details.

And, in the last lectures I have given you clarity about the mathematical procedures and that too the application of the technique called as OLS that is the ordinary least squares and the process of OLS is to minimize the airline sum squares to get the values of the unknown parameters. And these unknown parameters will help you lot to predict the airline cost with respect to airline passengers.

We have already discussed all these details. And right now we will just see how software can help you to calculate these parameters. And that too bring the predicted line through which we can predict the airline cost subject to airline passengers and then check the kind of errors or the details before you go for the predictions and as per the requirement of this transportation sector.

So, let us move to this you know. So, this is what the kind of problems. So now, we have the problem like this.

(Refer Slide Time: 03:15)



Regression Estimation: Airline Cost Data (part 1)

Number of Passengers X	Cost (\$1,000) Y	X^2	XY
61	4.28	3,721	261.08
63	4.08	3,969	257.04
67	4.42	4,489	296.14
69	4.17	4,761	287.73
70	4.48	4,900	313.60
74	4.30	5,476	318.20
76	4.82	5,776	366.32
81	4.70	6,561	380.70
86	5.11	7,396	439.46
91	5.13	8,281	466.83
95	5.64	9,025	535.80
97	5.56	9,409	539.32

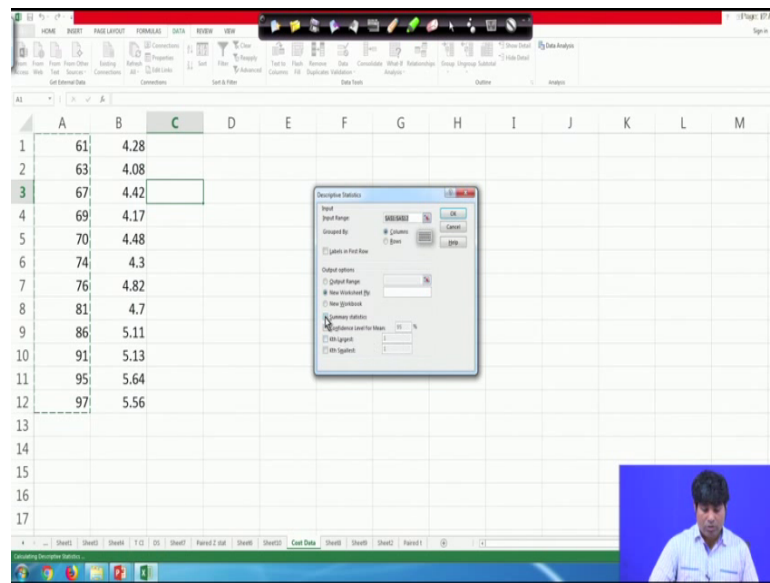
$\sum X = 930$ $\sum Y = 56.69$ $\sum X^2 = 73,764$ $\sum XY = 4,462.22$

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, this is the number of passengers and this is the cost and we declare this is a dependent variable and we declare this is the independent variable. And as a result we have already a discussed the manual procedure through which you can we can process and get the values of these parameters unknown parameters that too alpha cap and beta cap with the help of airline passengers data and airline cost data. And, without any softwares you have to follow these particular mathematical procedure to head this figures.

So now, with the help of softwares we can easily get this a parameters below and the kind of regression line to predict the airline cost subject to airline passengers. So, let us see how is this particular the structure.

(Refer Slide Time: 04:11)



So, as usual the problem we have discussed. So, this is the airline passengers and this is what the airline cost in a 1000 dollars and we need to regress Y upon X where Y is the airline cost and X is the airline passengers and this is what declared as a X and this is declared as a Y. And here linear regression modelling will be Y equal to alpha plus beta X and our job is to find out Y cap that is the predicted Y or regression line which is equal to alpha cap plus beta cap X. And, to know the alpha cap value and beta cap value we use OLS technique and then the processes is to minimize error sum squares and the way you will process you can get the values of this parameters.

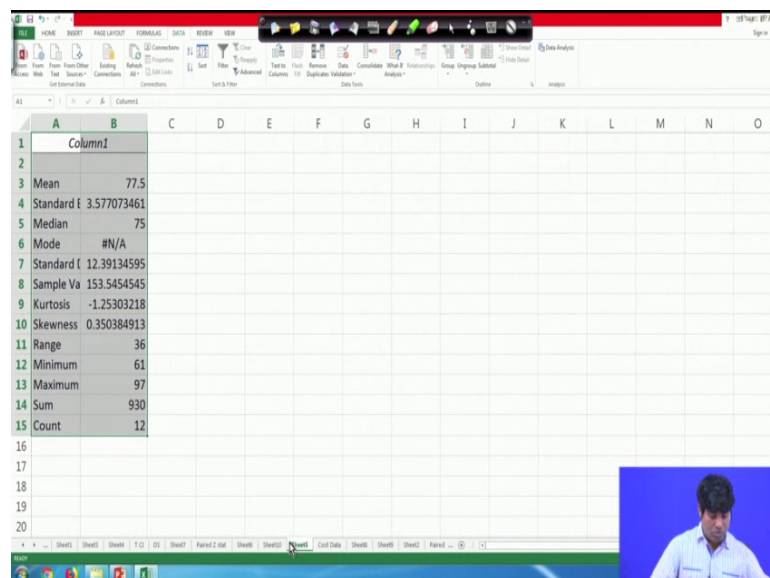
And, mathematical we have already checked and we have also tested and obtained the values with respect to this data. So now, same data we can used through a software and try to head this figures. So that we can get to know; how is this particular process because this is a small data set and that too with two variables it is easy to you know process manually. But, when your data size is substantially very high let us say 2000 3000 like this and when we are dealing with multiple regression modelling or multi way regression modelling that times you cannot easily you know handle when the problem manually and soft the requirement of software is a very significant during that time. So, let us try to acquainted with the kind of in a software how to deal with this problems and how to get this a requirements.

So, we can use it standard you know econometric software, but in the mean time since it is the beginning. So, I will use excel because excel spreadsheet has a data analysis package and the entire engineering econometrics more or less covered and bounded with regression modelling and excel data analysis package is having a regression technique component we can directly use excel and get these a values of these parameters. And later on the same problem we can actually handle through different softwares, and then check the robustness of the softwares and the kind of results.

So, here in the excel spreadsheet after entering the data you just click the data here. Then the moment you will click the data here, so, you will find there in the extreme right there is a data analysis pack. So, you just click then you will get the box like this and the box will give you the there are lots of things you will find here and starting with correlation, co-variants, descriptive statistics. So that means, technically we can check whatever we have discussed here descriptive statistics. Let us see one by one. So, this is descriptive statistics.

So now, for that you can give an indication about both the variables and then put. So, it will give you the, ok. So, this is what the descriptive statistics, summary statistics.

(Refer Slide Time: 07:59)

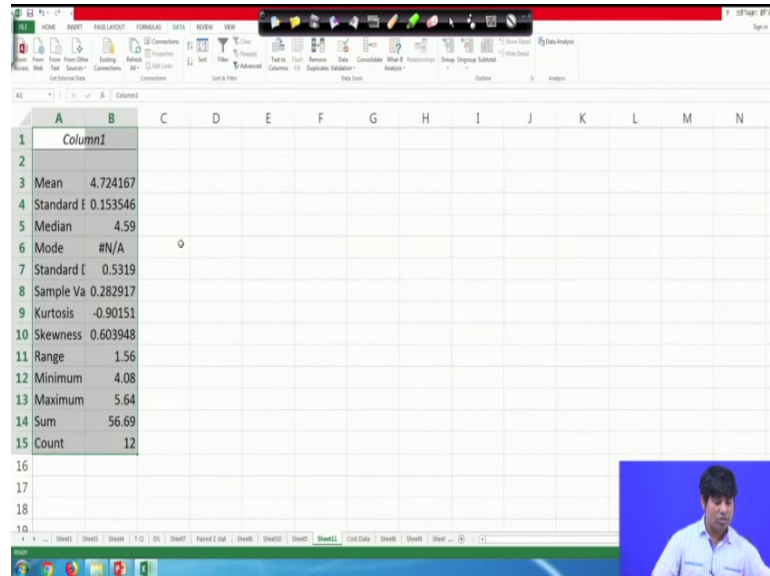


	Column1
Mean	77.5
Standard Error	3.577073461
Median	75
Mode	#N/A
Standard Deviation	12.39134595
Sample Variance	153.5454545
Kurtosis	-1.25303218
Skewness	0.350384913
Range	36
Minimum	61
Maximum	97
Sum	930
Count	12

So, this is with respect to X variable and these are all summary statistics these are all summary statistics and for X variables and again you can go to the kind of cost data. Again, then again you go to the data analysis package that to descriptive statistics, and

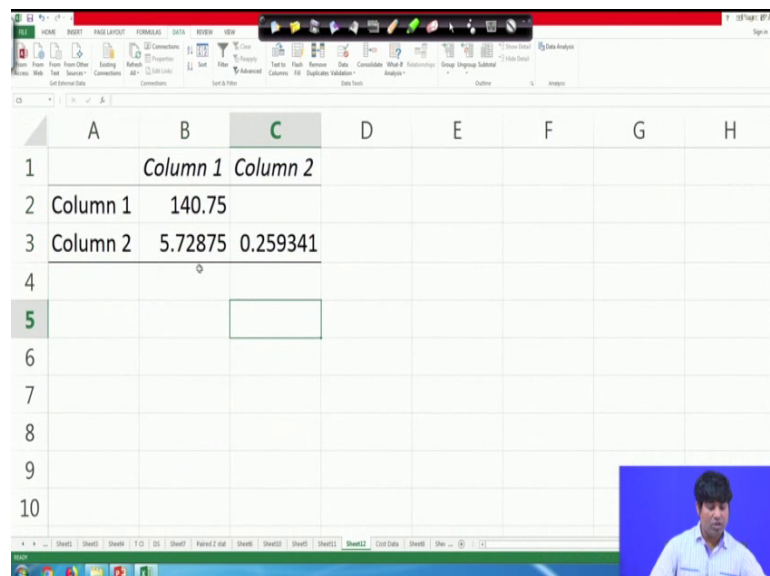
change the variable options, you can give the indication about the second variables and again put and you will get the, ok. So, this is what the range you know put, ok.

(Refer Slide Time: 08:43)



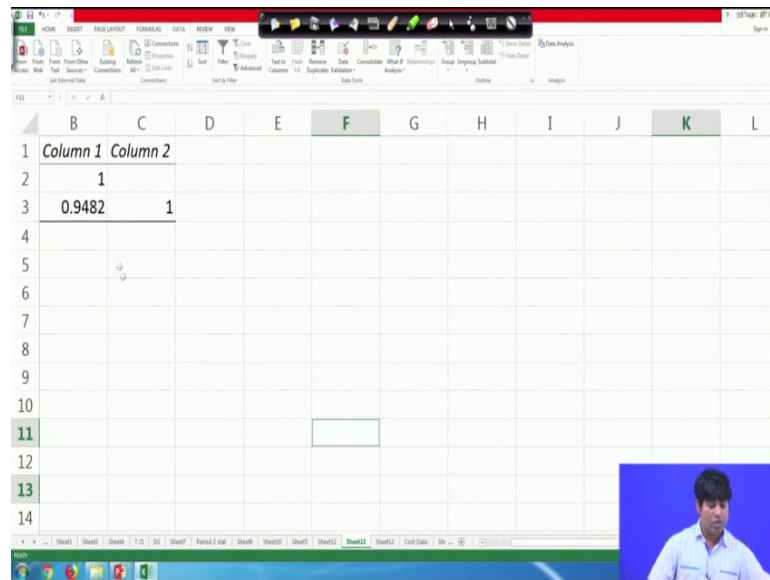
So, this is how the second variables range, ok. And, again go to the cost data and data analysis package. So, this is how the descriptive statistics figure and again you can check the covariance, just give the indications. And in that case both the variables simultaneously can be indicated and put ok, you will get the covariance page at here, this is the covariance matrix.

(Refer Slide Time: 09:07)



And, for here for from this covariance it is clear that you know Y and X are positively related and again we can test through correlations. So, the correlation can normalize the particular structures. So, again go to the cost data and then you can indicate the variable range and you will get this is what the correlation matrix.

(Refer Slide Time: 09:38)



And, here you will find the correlation between X and 1 correlation between X and Y is actually coming 0.95; that means, they are highly correlated. So, that means, the identification of airline passengers to airline cost determination is very perfect theoretical supported and statistical also supported. So now, our requirement is not to check the only relationship our check is to predict the airline cost with respect to airline passenger. So, that means, technically before we go for the prediction of airline cost with airline passengers we like to check the details about the kind of descriptive statistic that is the Y X variable that is the airline passengers.

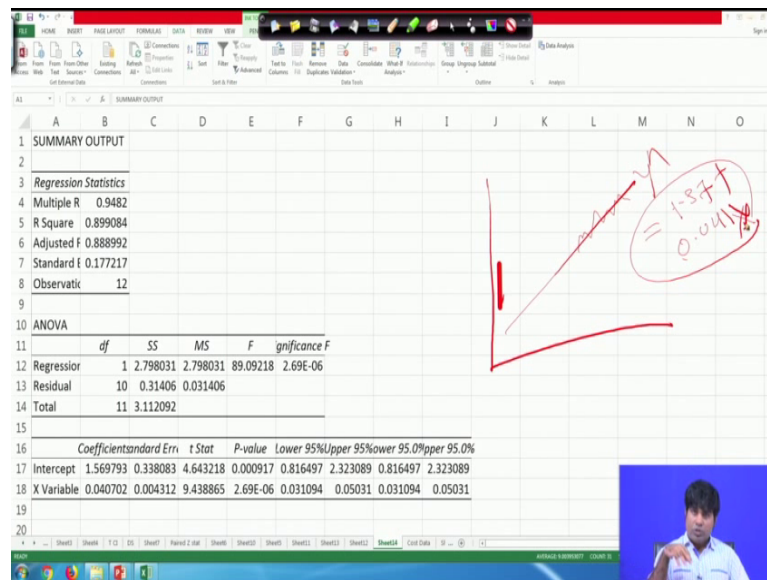
So, this is the main figure standard deviations median the spreadness, skewness, range, minimum, maximum sum count. So, to count is the number of size similarly we have the information about the second variable. So, these are all the variables against mean, median, modes then range maximum minimum count. So, 12 here at the sample size is also 12. So, that means, this will give you some kind of background since the range is actually 36. So that means, it is clear that the num items are you know informations are

not same. So, there is a difference and that will also be verified through the standard deviation which is not exactly equal to 0.

This is for a variable 1 and this is for variable 2 and again in the variable 3 you know in the in the third case we have calculate the correlation matrix where we have found there is a significant relationship and that relationship is it a positive. In fact, this is what the first end output that can be statistically again checked and to know the details and in fact you know before we go for airlines cost predictions. So, we have the clarity both theoretically with the descriptive statistics with covariance matrix with correlation matrix. So, with the data visualization again now we will check you know how is the kind of modelling procedure through which you can we can actually predict the airline cost with respect to airline passengers.

And, for that the standard technique will be regressions, again you just go to the data analysis and then you choose the regression. So, this is what the regression and now, in the regressions we will find there are two boxes here. So, input Y range and input X range. So, Y range is the dependent variable and X is the independent variable. So, now, in this case this is declared as a X and this is declared as a Y. So, as a results you will give an indication about the Y series and X series. So, since it is the dependent variable. So, we can give the indication like this and then this is for the X indications, where give the X indication here, and then what will you do we will just put, and you will find the regression output.

(Refer Slide Time: 13:10)

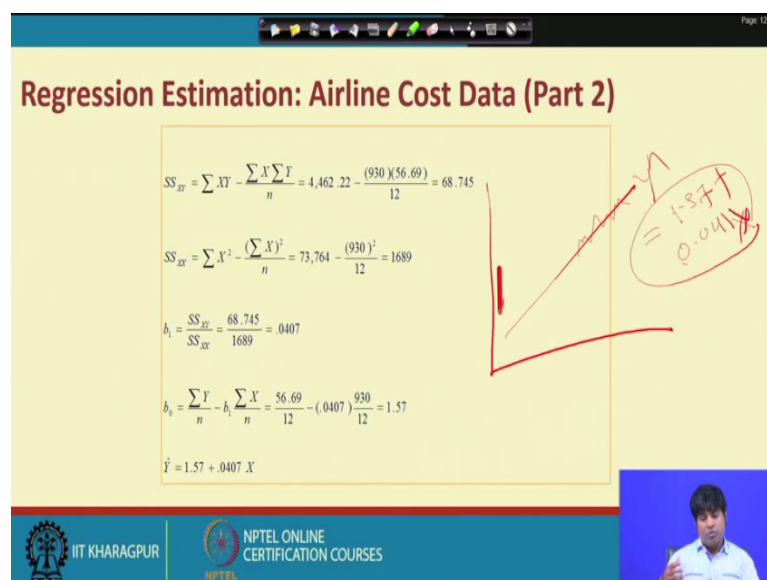


Handwritten notes on the Excel screenshot: A red arrow points from the 'Multiple R' value (0.9482) to a circled calculation: $1 - 0.9482^2 = 0.0417$. Another red arrow points from the 'X Variable' p-value (0.05031) to the same circled calculation.

SUMMARY OUTPUT					
Regression Statistics					
Multiple R	0.9482				
R Square	0.899084				
Adjusted R Square	0.888992				
Standard Error	0.177217				
Observations	12				
ANOVA					
	df	SS	MS	F	Significance F
Regression	1	2.798031	2.798031	89.09218	2.69E-06
Residual	10	0.31406	0.031406		
Total	11	3.112092			
Coefficients					
	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	1.569793	0.338083	4.643218	0.000917	0.816497
X Variable	0.040702	0.004312	9.438865	2.69E-06	0.05031

In this particular in this what called as a summary of regression output and that too with the help of you know and data analysis package through excel softwares and till now whatever we have discussed that is with respect to this version only, ok; so this version only. So, that is the how to get the alpha coefficient that is 1.57 and X variable 0.04. So now, if you will go to the kind of slides we will find the results more or less, ok.

(Refer Slide Time: 13:57)



Handwritten notes on the slide: A red arrow points from the b_1 calculation (0.0407) to a circled calculation: $1 - 0.9482^2 = 0.0417$. Another red arrow points from the b_0 calculation (1.57) to the same circled calculation.

Regression Estimation: Airline Cost Data (Part 2)

$$SS_{xy} = \sum XY - \frac{\sum X \sum Y}{n} = 4,462.22 - \frac{(930)(56.69)}{12} = 68.745$$

$$SS_{xx} = \sum X^2 - \frac{(\sum X)^2}{n} = 73,764 - \frac{(930)^2}{12} = 1689$$

$$b_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{68.745}{1689} = .0407$$

$$b_0 = \frac{\sum Y}{n} - b_1 \frac{\sum X}{n} = \frac{56.69}{12} - (.0407) \frac{930}{12} = 1.57$$

$$\hat{Y} = 1.57 + .0407 X$$

So, this is where we get beta coefficient 0.0407 and beta coefficients is 1.57 this is alpha and this is beta. And, same thing we obtained here interested to that is alpha coefficient is

1.57 and beta coefficient is 0.0407. So, that means, whatever I you know answers we have obtained through the process manual process it is actually same with the help of also softwares. So, that means, software now to have these values through software is very you know very effective because within less time we will get these results and then came you know then analyze as per the particular requirement.

So, now, in addition to these values of you know alpha coefficient beta coefficient there are couple of other items are available in the you know excel spreadsheet that to the data analysis package, but these are all not unnecessary component. These are all you know useful for the kind of regression modelling that too predict the cost factor airline cost with the airline passengers, but the interpretation of these items are different, over the time we will be connect and you know discuss.

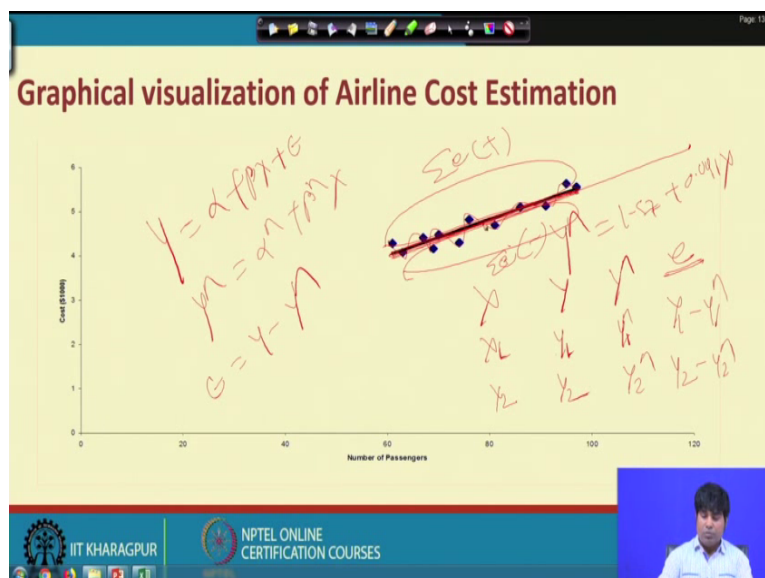
But, in the meantimes we get to know how to get the alpha coefficient that is the intercept and how to get the beta coefficient that is (Refer Time: 15:27) coefficient that is you know beta and we are acquainted with the system manually and we are acquainted with the particular process again through softwares. So, both cases the results are matching and it gives the kind of structure to predict the airline cost with respect to airline passengers. So, that means, from this result we can have the predicted lines where Y equal to 1 plus 57 plus 0.041 X . So, that is what the predicted line.

So, that means, technically. So, here predicted structure will be like this is the predicted line. And of course, the data point are looking like actual data point will be like this and this is Y cap and that too it is nothing, but intercept is 1.57 plus this is 0.0, 0.041 X . So, this is what the predicted line. So now, you put X value any X value you can get the Y predicted. So, that is that means, technically that is what is our objectives. So, if the passenger will increase or decrease what should be the airline cost? So on that basis we can we can you know apply some kind of strategy that may be helpful for the airline industry or as per the kind of problem requirement, but we you should know equality process through which you can obtain these figures, right.

So that means, technically we are now acquainted with the system how to obtain the values of these parameters through the process of less and that too with the help of you know manual process and again through the help of you know softwares. So now, after

getting this predicted lines. So, next job is you just check; how is the kind of behavior. So, that it till now we have discussed this much.

(Refer Slide Time: 17:45)



And, now this is what actually the predicted lines where we have the predicted lines that too 1.57 alpha coefficient and 0.041 X that is what the black line indicates the estimated line and these are all actual lines. If we will join our actual data points it will like you know zig zag, but the predicted line will be straight lines that is what it is called as a line of the base fit. So, without putting the X value you just extend this lines you can get the filtered values of you know Y corresponding to the X indication.

So now, from this estimated values alpha and beta we can obtain error terms because initially we start with the process that Y equal to alpha plus beta X plus error term epsilon and here Y hat equal to alpha hat plus beta hat X and here error terms epsilons will be Y minus Y hat. So, that means; this is the actual and this is the predicted; So, that means, for let us say X is a variable here and Y is the dependent variable and Y predicted that is the predicted dependent variables. So now, for every X let say X 1, so, we have Y 1 we have Y Y 1 hat equal to X 2 we have Y 2 and Y 2 hat. So, that means, for every X we have actual Y and predicted Y.

So now, the difference will give you the error terms e or you know epsilons that is Y minus Y Y 1 minus Y 1 hat then Y 2 minus Y 2 hat and so on, then you can create a series. So, this is how we can obtain the error terms. So, initially we start with the Y and

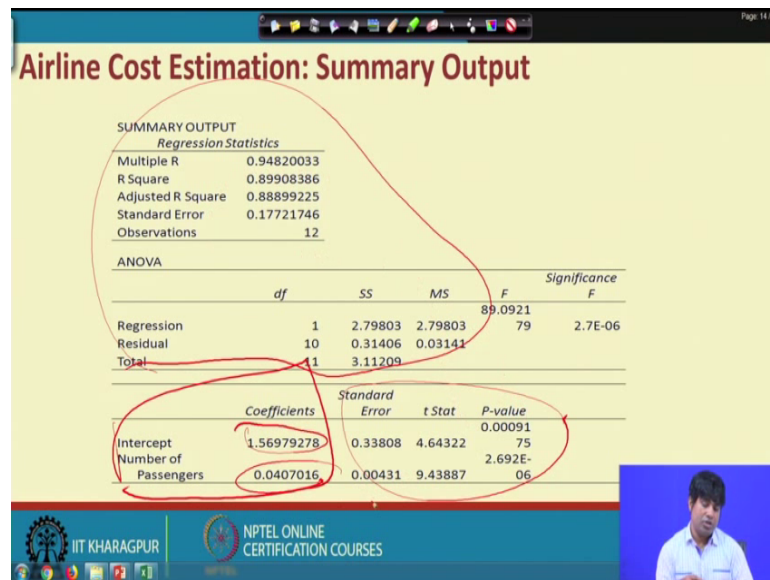
X and over the process you can have the excel sheet Y X Y cap and error terms. So now, as per the requirement of the regression modelling what we have discussed earlier there is a certain checks before the process and there are certain checks after the process. And before the check we have already discussed in details that is the uniform sampling then the minimum sample size the kind of theoretical understanding. So, these are all we have already you know discussed.

But, after the process of you know estimations we like to check certain things before you go for the predictions and the kind of forecasting. The first check is the error component. So, ultimately the requirement is to check whether the mean error is equal to 0; that means, some of the error term should be equal to 0. So, from this from this figures we can easily understand that you know some of the data points are you know above and some of the data points are you know below. So, now any the above data points will give you the positive error sum that is summation of you know error term positive side and these are will give you now error terms negative.

So, these two should be matching, so that these total error will be equal to 0. So, if that is the case then this particular line can be declared as the line of the best fit, otherwise it may be the estimated line, but it cannot be considered as the best fitted line. So, technically the process of OLS is that you know it will be the line of best fit, but there may be some error because you know lots of you know other issues like you know different functional form the variable inclusion, exclusion, sample size, so many things are there. So, that the particular predicted line may not be the best you know because of this you know regions.

But, ultimately whether the particular predicted lines will be called as you know line of the best fit. So, we have to check certain things then finally, we can give the signal that yes, as per the obtained line from the values will declared as the line of the best fit and then you can go ahead with the prediction as per the particular engineering requirement. So, obviously, the process is like this. So, you have the predicted line here and now what will you do. So, you just check the kind of a regression output.

(Refer Slide Time: 22:25)



Airline Cost Estimation: Summary Output

SUMMARY OUTPUT

Regression Statistics

Multiple R	0.94820033
R Square	0.89908386
Adjusted R Square	0.88899225
Standard Error	0.17721746
Observations	12

ANOVA

	df	SS	MS	F	Significance F
Regression	1	2.79803	2.79803	89.0921	2.7E-06
Residual	10	0.31406	0.03141	79	
Total	11	3.11209			

Coefficients

	Coefficients	Standard Error	t Stat	P-value
Intercept	1.56979278	0.33808	4.64322	0.00091
Number of Passengers	0.0407016	0.00431	9.43887	2.692E-06

This is what the regression output which just you know initially tested through excel software and what we have here. So, till now we have discussed up to this much this is the beta this is alpha coefficient and this is the beta coefficient and in addition to this alpha coefficient beta coefficient we have actually this much of extra you know outputs and these are all extra outputs. Till now, we have not discussed anything about all these items, in the later stage we will discuss and all are you know highly required to validate and to justify the line of the best fit; and that too the line which can predict the airline costs with the availability of airline passengers. That means, we like to predict Y subject to the X availability.

So, now, what are these items we it means the other items that is the estimated regression requirements. So, we will discuss in detail in the later stage.

(Refer Slide Time: 23:39)

Residual Analysis: Airline Cost Data

Number of Passengers X	Cost (\$1,000) Y	Predicted Value \hat{Y}	Residual $Y - \hat{Y}$
61	4.28	4.053	.227
63	4.08	4.134	-.054
67	4.42	4.297	.123
69	4.17	4.378	-.208
70	4.48	4.419	.061
74	4.30	4.582	-.282
76	4.82	4.663	.157
81	4.70	4.867	-.167
86	5.11	5.070	.040
91	5.13	5.274	-.144
95	5.64	5.436	.204
97	5.56	5.518	.042

$\sum (Y - \hat{Y}) = -.001$

Footer: IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

In the mean times we just like to know the something about the line of best fit that to check the error procedures. What we have discussed in the in the last you know lectures that you know we have X information, we have Y information by using OLS process and the kind of fixing the linear regression modelling. So, we can obtain the Y predicted which is nothing, but you know $\alpha \text{ hat} + \beta \text{ hat } X$ and $\alpha \text{ hat} = Y \text{ bar} - \beta \text{ hat } X \text{ bar}$ and $\beta \text{ hat} = \text{covariance of } XY / \text{sigma square } X \text{ variance of } X$ and then we can we can actually generate this predicted line.

Now, with the help of you know Y actual and Y predicted Y we can get the error component that is the difference between Y actual and Y predicted Y actual Y predicted with different sample points. So, as a result if you check the residuals you will find some are positive and some are negative. For instance, this is positive this is negative, this is positive this is negative, this is positive this is negative, again this is positive this is negative, this positive this positive.

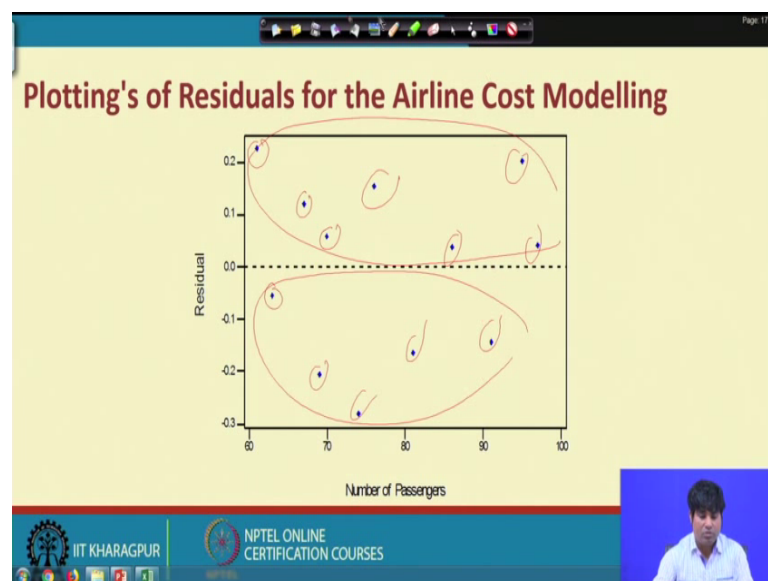
So, that means, more or less some is up some is down. So, that means, it is very much you know consistent as per our requirement because the line a line on the best fit should be in the middle of this points. As a result one will be positive one will give you positive error another will give you negative error and if you know take this sum this sum should be close to 0. In fact, this is the error sum is error sum and it is coming minus 0.001. That means, it is close to 0 and the it should be exactly equal to 0, but due to this round up you

know rules so, we are obtaining you know some of the error term equal to minus 0.001 which is very very close to 0.

So, that means, technically the estimated lines and that too the values of these parameters α and β hats are perfect in one sense, but we cannot say absolutely perfect until unless you go through other checks. In the first check that is with respect to error check error term check so, that is the sum of the error sum should be exactly equal to 0 which is actually happening here. So, the declaration is that the values of the estimated parameters which you obtain through OLS mechanisms are perfectly to predict the airline cost with respect to airline passengers.

Now, now with each processing; so, what will we do next is to check the error behavior of course the error sum is actually coming 0, but we like to check the behavior of the error terms with respect to the X and Y involvement.

(Refer Slide Time: 26:53)

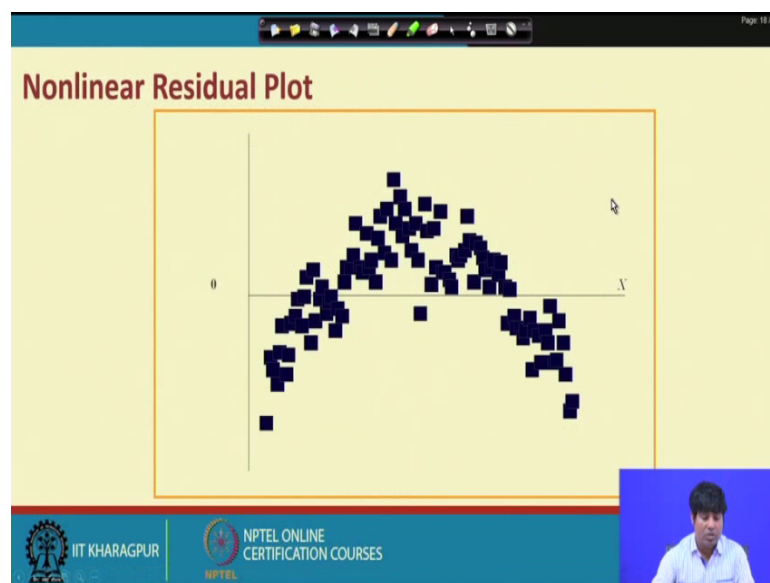


So, in the next slides we will check the kind of plotting this is what the plotting of you know residuals and a number of passengers and you know with respect to the a residuals. So, because it is the number of passengers which can you know decide the airline cost. So, when you find the error points are you know error points error you know drastically different, ok. So, some error point is here, some point error point is here, some error points are here, it is a kind of heterogeneous you know kind of spread. So, it is not

showing any kind of direct functionality and this will give you some kind of indication that you know the model is a somehow, ok.

The one way to justify is that you know in more or less same number points in the upside and some are some of the numbers are you know in the downside. So, we have this sum that side and we have this sum this side and these both these both should actually matching each other, that is how the indication here. That means, looking this figures it seems that you know it is somehow to go for the predictions.

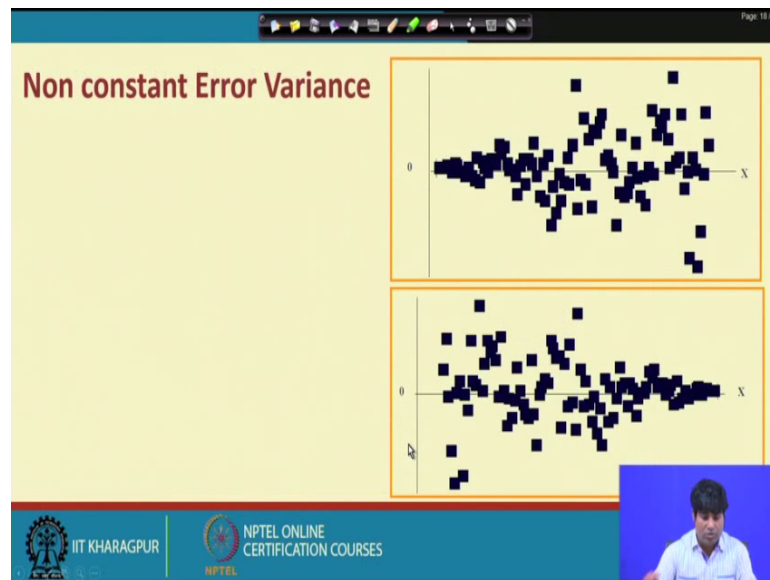
(Refer Slide Time: 28:16)



And, then in the next slides you can actually check the means this is what actually the kind of behavior and if you plot these you know points it looks like you know non-linear residual plotting sometimes it is coming the residual plotting and this is another way to justify the a non-linearity. So that means, technically like we have checked the form data visualization to know the nature of Y and X .

So, next step after the estimation is it to check the error sum and the plotting of the error term whether the error terms are equally linear one or it behaves like you know non-linear. If the model is a linear then error should be should be a linear if it is non-linear then the modelling process will be you know different and that too it may affect the process of you know reliability check.

(Refer Slide Time: 29:18)



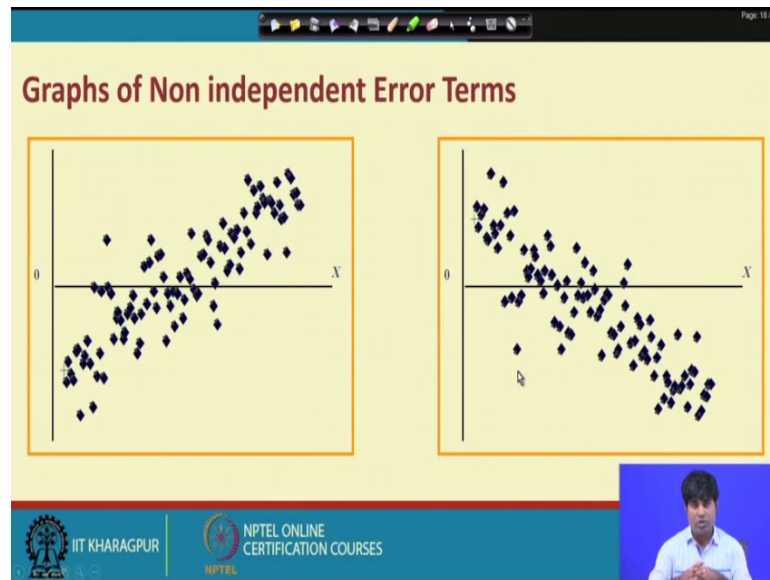
And, that too this is how the error behaviors. Of course, what we are supposed to do here we are checking the error sum which should be exact which should be come to 0. And then we check the plotting of the error terms which will give you the functional forms whether it is a linear spread or non-linear spread and, again we like to check the error variance, ok. So, which is nothing, but sigma square u and how to get this error variance we will discuss in details later.

But, the error variant should be actually homogenous, if error variance should be heterogeneous then it will give you some kind of econometric issues or some kind of statistical issues that is what is called as you know the component called as heteroscedasticity which will discuss in the later part, but one of the requirement of this particular OLS processing is that your error variance should be homogeneous in nature to justify the effect that the particular estimators or line of the best fit a line best line of the regression is the best one.

So, for that the error variance should be plotted and check whether it is a having homogeneous spread or heterogeneous spread. If it is a homogeneous spread is good for the OLS mechanism, if it is not homogeneous spread it will going against the OLS estimation. So now, if it is supportive then we can go ahead with the predictions, if it is not supportive to the OLS then we may change the particular route al and then we will go for the another you know another procedure to predict the particular requirement. So, we

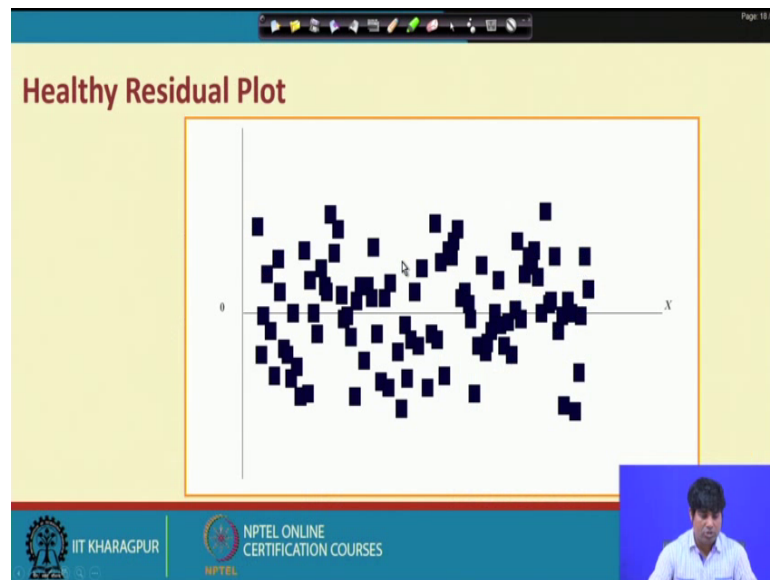
will discuss in details in the later stage, but it is the mandatory requirement that you know to check the nature of error component that too the sum the plot nature and the kind of error variance.

(Refer Slide Time: 31:30)



So, likewise this is another kind of plotting to check whether the error terms are actually a completely independent to each other and their error variants are homogeneous in nature. The error component, the individual kind of nature and their variants. So, the actual positions and the variance positions both are you know different error variants is something behavior of this particular plotting, but the positioning of this you know error term should be completely different with respect to Y and X , ok. So, this would be you know very much consistent as per the particular requirement.

(Refer Slide Time: 32:11)



This is actually what we can call as a healthy plotting. So, as per the OLS requirement if the plotting of the residuals will be like this gives the signal that you know they are independent and the error variance is somehow homogeneous in nature. If that is the case then again will go ahead with the predictions and the line will be declared as the best fit line or best fitted line. So, if the error sum is not coming 0 or the error variants is not homogeneous then by default the particular lines which we consider the line of then best fit may not be very handy to go for the predictions in the for these particular problem to predict the airline cost with airline passengers.

These are the process through which actually we can we can we can predict the airline cost with respect to airline passengers. That means, technically in this lectures we get to know how to estimate all these parameters manually and through softwares and get the estimated parameters value and the line of the best fit and obtain the predicted Y obtain the error terms and checking the behavior of the error terms. These are these are the minimum requirements you are supposed to do before you start the process of predictions and the forecastings.

Then, in the next class will continue here and that too we will check the reliability of this particular estimated model. Of course, we have slightly check the reliability with respect to error terms but, there are certain other items we are supposed to check before you start

the process of you know predictions and the kind of forecasting as per the particular engineering requirement.

With this, we will stop here.

Thank you very much. Have a nice day.