

Business Analytics for Management Decision
Prof. Rudra P Pradhan
Vinod Gupta School of Management
Indian Institute of Technology, Kharagpur

Lecture – 27
Predictive Analytics (Contd.)

Hello everybody. This is Rudra Pradhan here and welcome to B M D lecture series. We are in the process of discussing predictive analytics 1. And in the last lecture, we have discussed about the regression analysis and that too we have highlighted.



There are three different ways; we can address this technique and the kind of you know predictive problems or predictive analytics and in this three are typically bivariate, trivariate and multivariate and in the last lecture, we have discussed about the bivariate structure and where we have actually two variables and that two; one dependent variable and another independent variable and the idea, is to predict the dependent variable subject to independent variable and we have already discussed, you know the structure through which you do the predictions. A prediction of Y subject to the X availability or you know X information. And in the last lectures we have discussed a problems that to airline problems, relating to the game between cost and number of passengers with the hypothesis that you know number of passengers you know and the cost are positively related to each other.

Now, regression will help you to give you the exact kind of you know quantification through which the particular passenger level will give you the particular you know cost package. So, we have already discussed and now will be actually verify, how these are all coming into the reality. So, in the last lecture what we have already discussed.

(Refer Slide Time: 01:50)

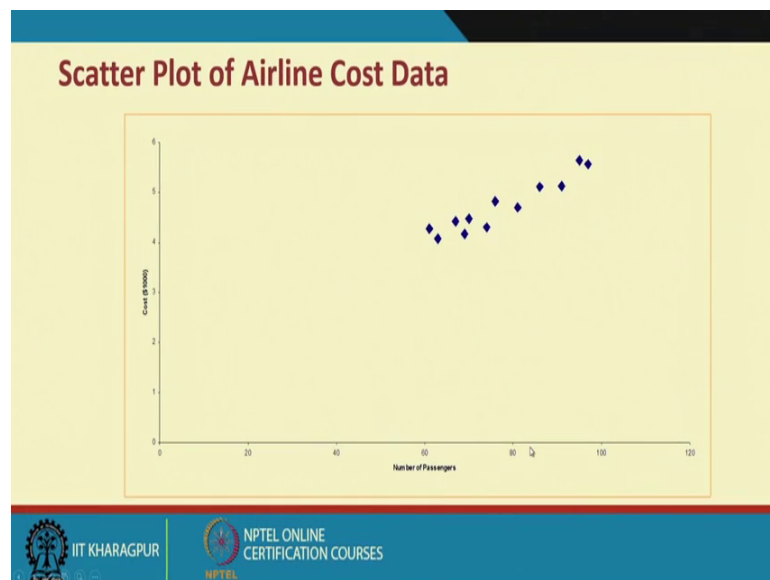
Example: Airline Cost Data

Number of Passengers X	Cost (\$1,000) Y
61	4.280
63	4.080
67	4.420
69	4.170
70	4.480
74	4.300
76	4.820
81	4.700
86	5.110
91	5.130
95	5.640
97	5.560

 IIT KHARAGPUR  NPTEL ONLINE
CERTIFICATION COURSES

This is what the problems which we have actually going to address and. In fact, we have analyzed this problem and going for a plotting.

(Refer Slide Time: 01:56)




(Refer Slide Time: 01:58)


Equation of the Simple Regression Line

$$\hat{Y} = b_0 + b_1 X$$

where : b_0 = the sample intercept
 b_1 = the sample slope
 \hat{Y} = the predicted value of Y




IIT KHARAGPUR




NPTEL ONLINE
CERTIFICATION COURSES

(Refer Slide Time: 02:00)

Ordinary Least Squares (OLS) Analysis

$$b_1 = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sum (X - \bar{X})^2} = \frac{\sum XY - n\bar{X}\bar{Y}}{\sum X^2 - n\bar{X}^2} = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sum X^2 - \frac{\sum X^2}{n}}$$
$$b_0 = \bar{Y} - b_1 \bar{X} = \frac{\sum Y}{n} - b_1 \frac{\sum X}{n}$$


IIT KHARAGPUR





NPTEL ONLINE
CERTIFICATION COURSES

(Refer Slide Time: 02:01)

Solving for b_1 and b_0 : Airline Cost Example (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

And then we have derived the output and then finally, this is the analysis.

(Refer Slide Time: 02:03)

Solving for b_1 and b_0 : Airline Cost Example (Part 2)



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

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

$$b_1 = \frac{SS_{XY}}{SS_{XX}} = \frac{68.745}{1689} = .0407$$

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

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

 IIT KHARAGPUR
  NPTEL ONLINE
CERTIFICATION COURSES

And this is how the kind of know cross check.

(Refer Slide Time: 02:07)

Airline Cost: Excel RA 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

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

Coefficients

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

 IIT KHARAGPUR
  NPTEL ONLINE CERTIFICATION COURSES



And finally, we have actually stopped here with having you know excel regression analysis output.

So, now my idea here just to show how actually these are all coming, you see on starting with only you know two variable information, see here. So, this is actually the original problem.

(Refer Slide Time: 02:28)

Example: Airline Cost Data

Number of Passengers X	Cost (\$1,000) Y
61	4.280
63	4.080
67	4.420
69	4.170
70	4.480
74	4.300
76	4.820
81	4.700
86	5.110
91	5.130
95	5.640
97	5.560

 IIT KHARAGPUR
  NPTEL ONLINE CERTIFICATION COURSES

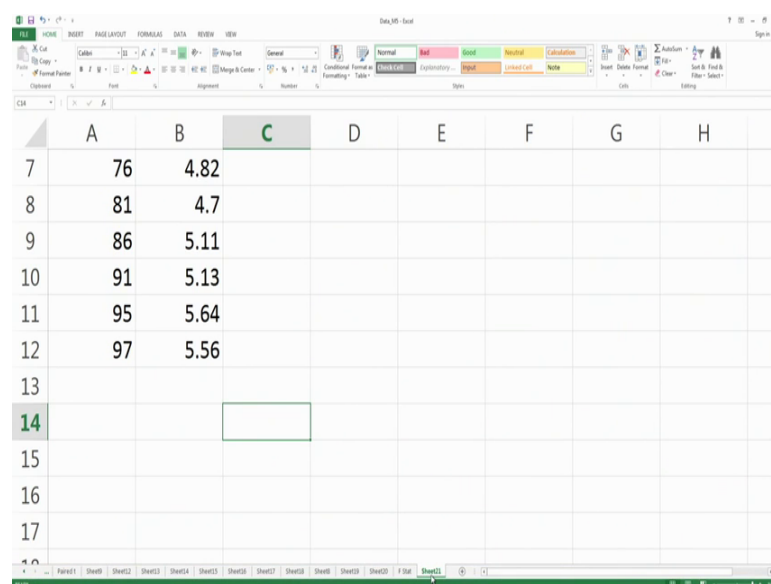
And you have only you know information about you know sample size n equal to 12 year for X and 12 year for Y . So, the first n condition is satisfied and second order, a

second and condition is also satisfied that you know n greater than $2k$. So, having these two information. So, we have actually plenty of numeric regression output here so. In fact, you know this is actually simple way, we have a calculated for a , b_0 and b_1 , but our the, you know process. We have actually plenty of output and actually these are all output to not actually manually calculated. So, it is obtained through this software and that too excel spreadsheet.

So, now I will first show you how these results are you know coming, then I will you know analyze these results and interprets as per the program requirement and then and then we will come to a position to go for some kind of you know management decision as per the problem requirement. Here, the management decision requirement is what should be the cost structures with a particular, you know passengers package right. So, if the number of passengers would be like this, let us say 10, 10,000. So, what should be the cost package subject to the passed information and the use of regression analytics. So, we like to know the exact quantifications.

So, before that we know, we should understand these outputs and then these output will help you to predict, the kind of you know accuracy. So, let us go to the excel spreadsheet and then I will connect you, how these are all coming.

(Refer Slide Time: 04:01)



	A	B	C	D	E	F	G	H
7	76	4.82						
8	81	4.7						
9	86	5.11						
10	91	5.13						
11	95	5.64						
12	97	5.56						
13								
14								
15								
16								
17								

In fact, we have already discussed the excel spreadsheet. You know this is actually the problem.

(Refer Slide Time: 04:08)

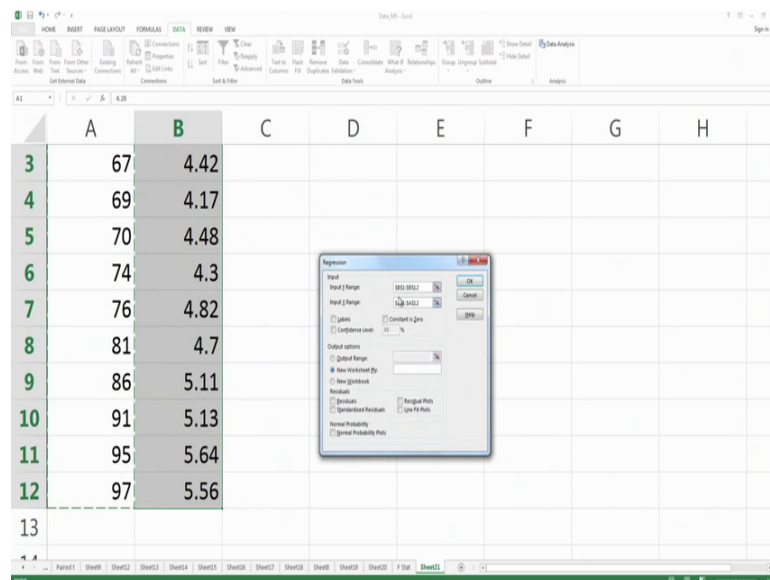
	A	B	C	D	E	F	G	H
1	61	4.28						
2	63	4.08						
3	67	4.42						
4	69	4.17						
5	70	4.48						
6	74	4.3						
7	76	4.82						
8	81	4.7						
9	86	5.11						
10	91	5.13						
11	95	5.64						

So, that is the, this is the X series. Now, you know number of passengers and that too we have actually A 12 informations and then this is the cost of this, you know airline and that too have actually 12 information. So, same sample size and there is no additional requirement and two variable sample size is actually n equal to 12 so; that means, technically 12 greater than to 2. So, it is satisfied and then you go to the data analysis package. So, just click the data analysis. So, you will find plenty of actually tools are there and. In fact, we have solved lots of problems, by this you know spreadsheet.

So, now our requirement to know the predictive structure and that to the working structure of the regression analysis and then you look for the regression component. So, here is the regression component and then after highlighting the regression component, you just put and then you will find the a menu box, in the menu box the requirement is, you have to give two different ranges for X ranges and Y ranges; that means, the first sample to the last sample. So, in this case it is A 1 to A 12 and then against the input range will be for independent variables. So, B 1 to B 12. In fact, we have discussed this particular structure, in the case of you know covariance and correlations, we have already highlighted this particular. You know analysis, but here the requirement is regression. So, where we are doing some kind of you know predictive structures, it is not just, you know kind of, you know association.

So, now you know putting these inputs and the kind of you know structuring. So, we will have this kind of you know regression output. So, let us see how it is actually coming. So, you can first indicate the sample points, for this is variance. So, this is variance and then this is actually X range. So, this is what the X range after putting this one.

(Refer Slide Time: 06:11)



So, this is B 1 to B 12 A 1 to A 12. So, that you have to cross check. If you, if I will put here B 12 and I will put here A 11 then and then this package will not run and; that means, software will show you the error.

So, now what will you do here, after putting the kind of an indication you just put.

(Refer Slide Time: 06:30)

The screenshot shows an Excel spreadsheet with the following data:

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%	Lower 95.0%	Upper 95.0%
Intercept	1.569793	0.338083	4.643218	0.000917	0.816497	2.323089	0.816497	2.323089
X Variable 1	0.040702	0.004312	9.438865	2.69E-06	0.031094	0.05031	0.031094	0.05031

So, you will find here plenty of results you will find plenty of results here. So, these are all you know regression results. So, whatever results we are finding from this you know spreadsheet, the same results are already you see here. So, I hope you can see this result. So, this is summary output and this is the anova output and this is how the regression, you know outputs right. So, these are all regression output. So, now, going to this, you know our, you know these structures, the same thing I have actually copied here and this is how the A results fall together in this results. So, what we have here so. So, now, now this is the result. So, what will you do with this results, we will check one by one and then will come to this kind of in a prediction and the kind of you know management decisions.

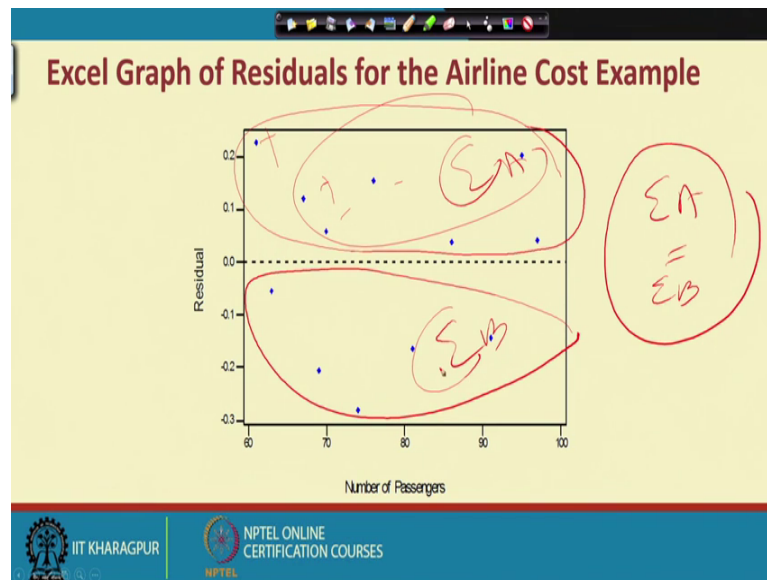
So, now, first things first check is actually check the coefficients and this is coming, actually positive and; that means, the theoretical, the expectation which you have in this problem is the cost and passengers are positively linked and the coefficient is showing that you know there is a positive link and if the theoretical expectation is not fulfilled through this. You know data visualization or data estimation then; that means, we have to think twice before you go for some kind of you know predictions. So, you know some kind of, you know management decision. So, this is going, you know in favor of our theoretical understanding. So, that is why we can move and then we will be analyze as per the problem requirement and here.

(Refer Slide Time: 08:05)

Residual Analysis: Airline Cost Example			
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
$\sum (Y - \hat{Y}) = -.001$			

So, the same problem. So, this is the X information and this is the Y information, which I have already, Y 2 which I have already. This is the X information, this is Y information and this is Y predictor and this is what the error component Y minus Y that is the kind of you know different and we expect that you know the sum of this error term should be equal to 0 or you know means almost or converge to 0 and here, in this case, it is coming minus 0.001. It is mostly, because of you know rounding up, you know error and in reality, it will be you know approximately equal to 0. So, means if it is actually coming 0, then by default this line will be called as you know line up the best pit or you know best predicted lines. So, will be means, see this is a first hand kind of you know observations and then we will be go for you know some kind of, you know other diagnostic checks through which you can justify that, you know this particular equa estimated equation or this particular line is the perfect line for the kind of you know pretty kind of a requirement.

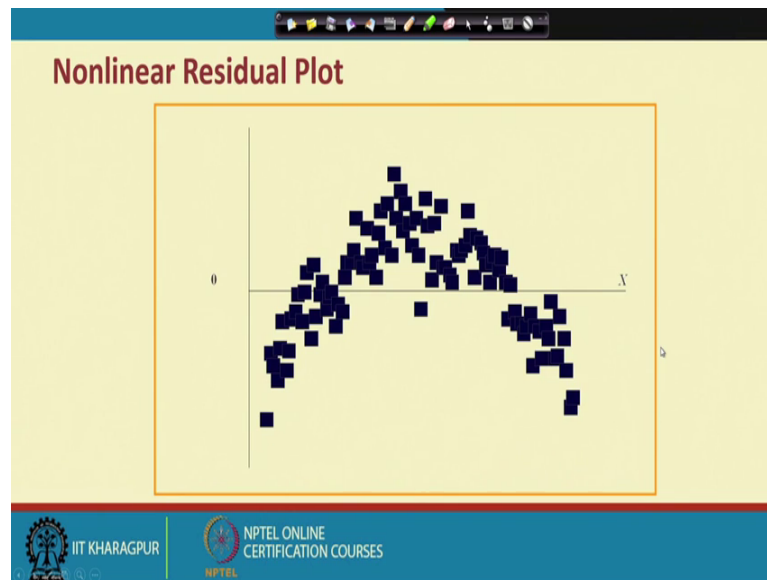
(Refer Slide Time: 09:18)



So after knowing this particular structure, so, let us move to; you know further processing. So, what will you do here? So, first will plot the residuals. So, what I have already mentioned that, you know once you plot, you know get the error component then 54 of you know fear, some of the planar mostly 50 percent of the points will be positive side and 50 percent of the points will be in the negative side. So, you take all these sums. So, you add up all these sums then you know take this sum here, and take this sum here, and this, this, this is a part A and this is part B. So, some A should be equal to some B. So, if that is the case, then you are in the right track and then we will proceed for the kind of you know predictions, but the fact is that you know sometimes A, particular in this case. It is coming minus 0.001.

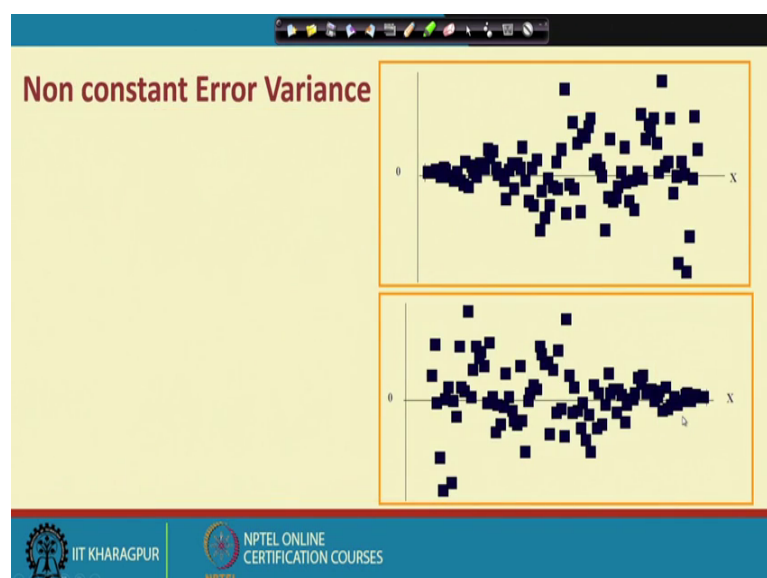
So, this is mostly, because of you know round off error, graphically you know just you check the kind of, you know structures here. So, this for a 7 points and then this is the 5 points. So, there is the points may be up and down, but actually some by default will be you know matching each other. So, that will be the kind of you know signal, which must be, which we must have, before you go for some kind of you know predictions, by this you know technique and knowing up, knowing all these things, then again you move to the, you know process, for instance.

(Refer Slide Time: 10:34)



So, after you know getting the error term first observation, you have to check whether you know this sum is coming equal to 0 or not then and then next things we have to observe that you know how these errors are, you know coming into the pictures, if the errors are not actually scattered here and there, then there will be, they really put a kind of in a functional form and if that is the case, then this will be having actually kind of, you know bad signal. So, there may be you know related to each other and their relations may be linear one may be non-linear one, but it is not actually good signal for the perfect prediction or the kind of, you know best fitness is concerned.

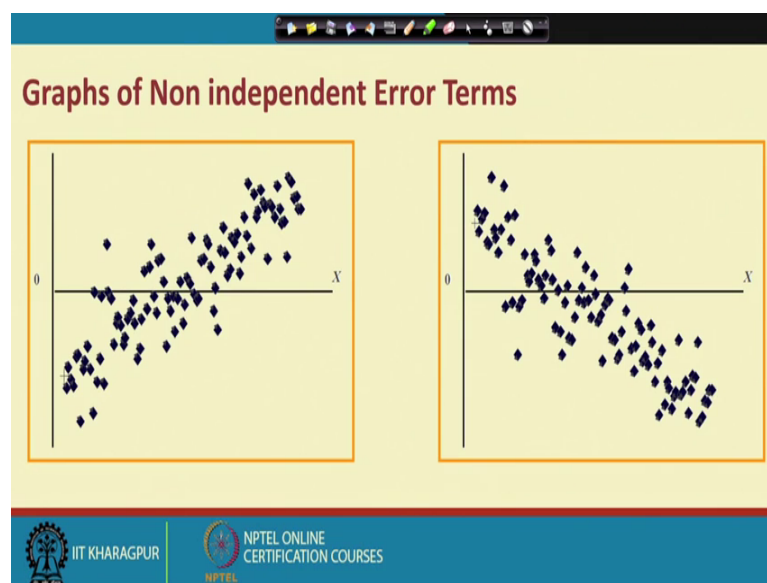
(Refer Slide Time: 11:26)



So, this is one of the, you know a set of the possible plotting of the residuals, but actual plotting depends upon the kind of, you know data process and similarly the plot you know corresponding to this plotting. So, you like to check the kind of you know error variance and sometimes the error variance, our different situations you know if you would check, then it may be actually constant or there is some kind of you know variance, you know variations. So, if they are you know constant then; that means, it is, if you are in a right track and if there is a kind of you know variance, kind of, you know difference then it will be having actually some kind of negative impact, because we have actually in the process of this particular technique for predictive kind of, you know requirement.

So, we have a component called as a homoscedasticity and there is a component called as a heteroscedasticity. So, the literary meaning in the business analytics contest homoscedasticity means homogeneous variance and heteroscedasticity means heterogeneous variance and. So, far as variance technique is concerned, that too for any kind of you know predictive kind of an analysis, see error component that is the error variance should follows actually homoscedasticity, if that is the case, then you know it will be give you the kind of you know best requirement, because we are using the kind of you know well structure.

(Refer Slide Time: 12:44)

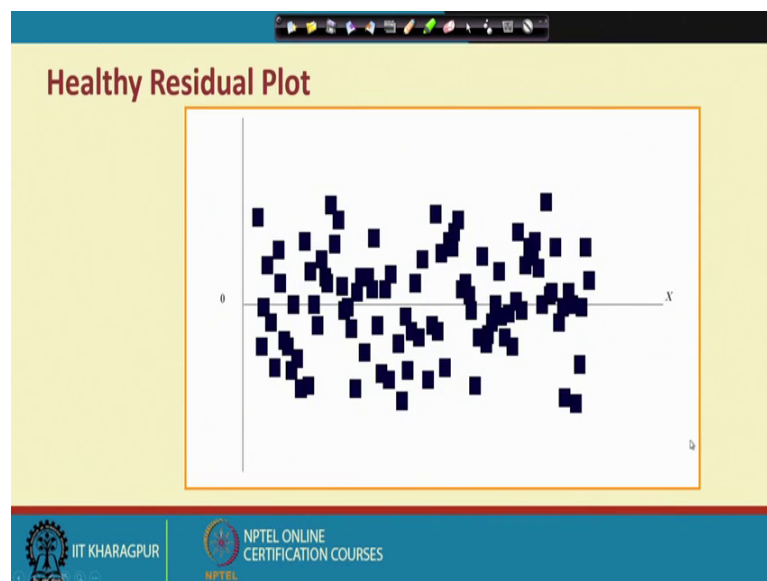


So in the case of you know, if they are not actually A, they have a means, they have no relationship, then there may be called as you know A and they are independent to each other so; that means, technically. So, we have A 2, you know means typically three issues. We have to actually highlight here first, structure is some of the error, term should be equal to 0 and error terms are. So, not actually correlate each other and then and the error variance should be homogeneous in nature right. So, that is actually homoscedasticity.

So, now these three things should be actually there, when you are actually using some kind of you know, because we are receiving this error term on the basis of you know various mechanisms, but the various requirement like that you know, your error sum should be equal to 0 and there should not be any correlation among the error terms and error variance should follow the kind of you know homoscedasticity that is actually normally distributed with the mean 0 and unit variance like, which we have already discussed the concept, you know normal distribution and that too in the case of you know inferential analytics.

So, now, knowing all these things let us see, these are all various looks, you know visual looks about the error term and this is what we call as, you know healthy look.

(Refer Slide Time: 13:57)



About the residual force plots, because you know ultimately we start the game with the Y information and X information, then with the help of you know regression analysis, we

are getting the Y head and then we are getting the error component and then with the error components, by default will give you enough exposures about your. You know predictions, actually our predicted line will be \hat{Y} , here that is \hat{Y} k f, but the accuracy or the kind of you know. So, far as you know fitness, you know you know fitness or perfect prediction is concerned. So, it exclusively depends upon the indication from the error term only. So, if the error indication is very good as for the problem requirement then; obviously, the model estimations model, reliability can model prediction will be very perfect right. So, this will be the kind of you know structure before you go for some kind of you know requirement ok.

(Refer Slide Time: 15:00)

Standard Error of the Estimate

Sum of Squares Error

$$SSE = \sum (Y - \hat{Y})^2$$

$$= \sum Y^2 - b_0 \sum Y - b_1 \sum XY$$

Standard Error of the Estimate

$$S_e = \sqrt{\frac{SSE}{n-2}}$$

Handwritten notes on the right side of the slide:

- $H_0: b_0 = 0$
- $H_1: b_0 \neq 0$
- $H_0: b_1 = 0$
- $H_1: b_1 \neq 0$
- $F \text{ test} = ?$
- $t \text{ test} = ?$

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

When we have actually in the process. So, we have actually on the basis of the theoretical look hypothesis building then the a kind of you know functional form then the estimation will get some kind of an estimated output, but before you use this estimated out for for any kind of in a prediction. So, that need to be tested and like we have already discussed you know inferential analytics having actually sample parameters. So, we need to actually test through various test statistics like you know we have we have gone through after this testing through jet statistic we have gone through the t statistic high square statistic and f statistics.

So, now these are also the values which you are receiving from, you know beta 0 b 0 or b 1 etcetera, these are all called as a sample statistics right. And these sample statistic need

to be tested against before you go for any kind of, you know predictions or any kind of, you know generalizations like you know creating confidence interval for population parameter or any kind of you know comments, you are putting as for the management requirement or any kind of no management decision you need to take. So, that need to be tested a first right. So, earlier we have already gone through the process of jet statistic, t statistic, I square statistic, f statistic, one sample case, multiple sample case, and the same things. Now, you know highly required, here is to validate the models, because here I array, we are actually you know in a simple structures, you are, you know just calculating the test statistic and you know testing it with the help of you know fixing alpha degree of freedom knowing, the critical value, calculated value, and comparing taking the decision. It is also same process, but it is in minor, different kind of you know setups or you know structure all together and like the previous, you know examples in the inferential analytics once you get the sample statistics. So, that need to be tested through z or t. So, every times. So, you need actually standard errors. So, here accordingly, we have also standard error. So, these are all actually standard error of you know error terms and standard error of the estimates right.

So, basically. So, our model is nothing, but actually called as a \hat{Y} equal to b_0 plus one X right. So, b_0 . So, this is the sample estimation and this is the sum. So, that need to be tested. So, usually if you go through, you know excel output or you know regression output. You will have actually two different, a three different sets, all together model summary anova and the model parameter estimate. You know estimated parameters information. So, technically we are supposed to check the validation of this model and that too we have to check the first, you know reliability and validity of these parameters and then we have to check the goodness fit overall fitness of this particular you know linkage and that will be established through anova the model summary.

So, now the first hand requirement, is these parameters need to be checked and there should be statistically significant and for that actually this is the, you know sample statistic that need to be connected with, you know testing procedure and for that the best test statistic, which you can follow up is called as a t statistic so; that means, technically. So, t of β_0 need to be significant and t be 0 here t of β_1 should be statistically significant so; that means,. So, you know assuming that you know. So, the null hypothesis will be setting like that you know b_0 equal to 0. The corresponding

alternative hypothesis will be b_0 equal to b_0 not equal to 0 similarly for b_1 . So, null hypothesis will be b_1 equal to 0 and alternative hypothesis b_1 not equal to 0. So, this is how the per kind of you know testing procedure and for that.

So, b_0 by the operation of you know standard errors and depending upon the alpha, where alpha signal and that is the predictive level, which you will like to fix and then the kind of, you know degree of freedom, which depends upon the sample size and the number of variables involvement. So, you can get the critical value and connect with the kind of, you know calculated a value and then finally, you can take the decision, whether to reject the to null hypothesis and accept the alternative hypothesis.

So, for that we need to have it, this kind of inner structure, and let us see, how is this particular, you know this is actually. So, S S is nothing, but the ((Refer Time: 19:29)) sum of this pair of this errors and standard error of the estimate is nothing, but actually square root of you know S S E divided by n minus 2, actually in the generalization process. It is n minus k , because it is 2 bivariate models. So, by default k here equal to 2 here.

(Refer Slide Time: 19:48)

Determining SSE for the Airline Cost Example

Number of Passengers X	Cost (\$1,000) Y	Residual $Y - \hat{Y}$	$(Y - \hat{Y})^2$
61	4.28	.227	.05153
63	4.08	-.054	.00292
67	4.42	.123	.01513
69	4.17	-.208	.04326
70	4.48	.061	.00372
74	4.30	-.282	.07952
76	4.82	.157	.02465
81	4.70	-.167	.02789
86	5.11	.040	.00160
91	5.13	-.144	.02074
95	5.64	.204	.04162
97	5.56	.042	.00176

$\sum (Y - \hat{Y}) = .001$ $\sum (Y - \hat{Y})^2 = .31434$

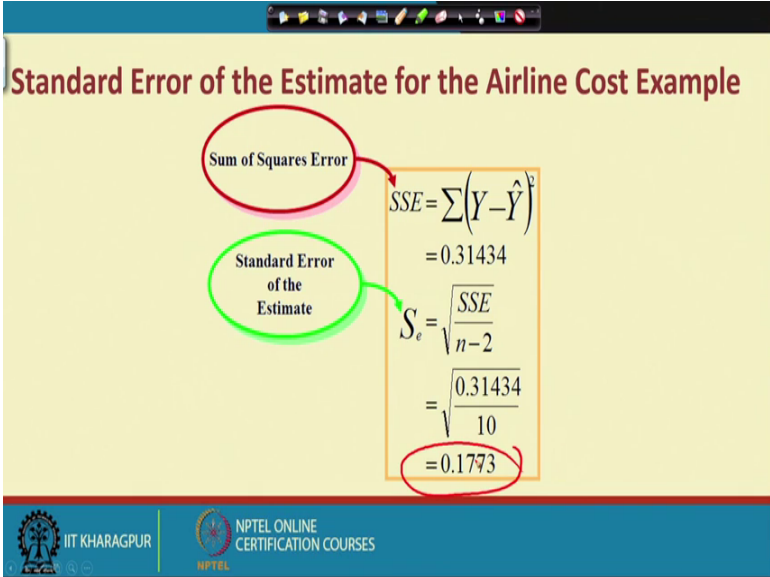
Sum of squares of error = SSE = .31434

So, this is what the some of you know, we like to get this. So, you know sum of squares of errors. So, that is actually is. So, this particular part actually. So, once you will be find the residuals. So, squaring the kind of you know residuals and then take this on. So, this is actually a you know kind of actually, we called as you know residual sum of this

sphere so; that means, technically, like in the anova, we have already discussed you know. So, some square totals, then you know some square within the group, some square between the groups.

So, this particular structure also follow here. So, typically it you know it is called as a t S S equal to E S S plus R S S so; that means, technically some of a squares total and explained sum of squares and residual sum of here. So, the residual sum of square is nothing, but actually this part that is sum square errors right. So, residual means it is the difference between the actual and the predicted and this is the explained sum square so; that means, the Y predicted results and. So, so this is how the t S S, t S S means this part actually. So, from this if you sparing with you, know the kind of mean and then taking this sum, you will get the t S S component. So, the t S S component, then the residual component will give you the explained item so; that means, out of these three, any two can be calculated and then can be used further for checking the validity or the fitness of this particular, you know model outcomes right. So, with this we can actually move furthers and get to know how is this happening.

(Refer Slide Time: 21:25)



Standard Error of the Estimate for the Airline Cost Example

Sum of Squares Error

$$SSE = \sum (Y - \hat{Y})^2$$

$$= 0.31434$$

Standard Error of the Estimate

$$S_e = \sqrt{\frac{SSE}{n-2}}$$

$$= \sqrt{\frac{0.31434}{10}}$$

$$= 0.1773$$

The slide includes logos for IIT Kharagpur and NPTEL Online Certification Courses at the bottom.

So; that means, your first requirement for testing procedure is a standard error of the estimates. So, that depends upon actually the sum square sum of square errors. So, thats nothing, but actually sum of Y minus Y hat squares and then followed by degree of freedom. So, finally,. So, the component is coming actually, this is what we called as you

know standard errors and after knowing the standard errors. So, the simple you know the t value of this particular parameter is nothing, but you know b 0 by standard error and b 1 by standard errors right so. In fact, actually b 0 and b 1 standards errors. you know slightly different, but the particular structure is like these to know and to get check the kind of you know reliability.

So, now moving forward again so.

(Refer Slide Time: 22:14)

Coefficient of Determination R^2

$$SS_{TY} = \sum (Y - \bar{Y})^2 = \sum Y^2 - \frac{(\sum Y)^2}{n}$$

SS_{TY} = explained variation + unexplained variation

$SS_{TY} = SSR + SSE$

$$1 = \frac{SSR}{SS_{TY}} + \frac{SSE}{SS_{TY}}$$

$$R^2 = \frac{SSR}{SS_{TY}}$$

$$= 1 - \frac{SSE}{SS_{TY}}$$

$$= 1 - \frac{SSE}{\sum Y^2 - \frac{(\sum Y)^2}{n}}$$

$0 \leq R^2 \leq 1$

Handwritten notes on the slide include: $TSS = ESS + RSS$, $\frac{ESS}{TSS} + \frac{RSS}{TSS} = 1$, and $R^2 = \frac{ESS}{TSS}$. A circled '4' is also present.

So, this is what the particular in a structure, which was actually highlighting and. In fact, what I have already mentioned that you know. So, t S S equal to E S S plus R S S so; that means, technically. So, so t S S is nothing, but actually here in this particular structure. So, it is represented by S S Y Y so; that means, actually it is depends upon you know Y value generally sum of squares Y series then sum square s s r that is called as actually explained variations, that is explained sum of squares, then the particular structure is called as the unexplained variations, that is actually R S S that is residual part right. So, unexplained variations.

So, now . So, now, if S S S S Y can be actually, you know divided a both the sides so; that means, technically. So, E S S by t S S plus R S S by t S S is equal to you know ones and. In fact, R square is nothing, but R square is nothing, but you know explained sum square by total sum square. So, total sum square. So, that is what actually the values of you know R square, the R square component that is capital R square, which is

represented as a coefficient of determination. So, the typical interpretation is, you know it is a percentage variation of you know Y sub you know percentage variation of you know explained item subject to total items right. So, this is the way through which you can actually adjust the kind of inner validation or goodness fit of the model. Usually R square value, you know positive, and the range of the R square will be 0 to 1 there. Like this we have already highlighted here and the value of R square close to 0 means, there is a low fit and the value of R square close to 1 means, it is a high fit, and if exactly equal to 0, then there is no association at all. So, there is no linkage between Y and X , if it is actually equal to one, then they are perfectly fit and they have actually very strong and you know perfectly ((Refer Time: 24:28)) and each others means having 0 is very rare and having 1 is also very rare, but most of the instances, it will be in between 0 to 1, but we try to have you know high R square, because R square will give you better validations and you know high predictions, but that is the necessary condition, the sufficient condition depends upon the significance of the parameters, that the significance of the variables and having our square and most of the variables are not significant, then this is a problematic case.

So, if R square will be high, then most of the variables must be a significant, if R square is low, then most of the variables there is a high means, if in most of the instance may be not statistically significant, but if we, most of the variables are not statistically significant and then by default R square will be low, but if R square is higher, then most of the variables should be statistically significant, that is the usual structure or you know a strategy of you know the checking, the reliability part and if other way around then; that means, there is a problem in the particular, you know process or in either in the estimation or you know fixing of the problem identification of variables with respect to data. So, many things will be there, but by the way these are the process, through which you have to you know check and you know validate the model and again moving forwards.

So, we recheck, you know how is this particular inner structure of this model.


(Refer Slide Time: 25:54)

Coefficient of Determination for the Airline Cost Example

$$SSE = 0.31434$$
$$SS_{YY} = \sum Y^2 - \frac{(\sum Y)^2}{n} = 270.9251 - \frac{(56.69)^2}{12} = 3.11209$$
$$R^2 = 1 - \frac{SSE}{SS_{YY}}$$
$$= 1 - \frac{.31434}{3.11209}$$
$$= .899$$

89.9% of the variability of the cost of flying a Boeing 737 is accounted for by the number of passengers.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES



And in this case, we have already calculated a sum of square error terms. So, now, you know having information about SS_{YY} and our know kind of you know $s s e$. So, you can able to calculate the kind of you know r square and in this case the r square of a loop prints typically in this particular problem the r square value is coming a zero point eight nine nine so; that means, ninety percent and so, this is the ratio between $E S S$ by $t s s$ explained sum of square by total sum of squares so; that means, technically ninety percent of the variability of the cost of flying means is accounted for the number of passengers of course, the cost may be affected by other variables, but in this problem. So, ninety percent of the cost mostly you know derived by the number of passengers and that is the that is what the first and kind of you know observation sorry you know kind of you know hint through which you have to do the kind of you know predictions and the kind of management decisions.

So, now again moving forwards. So, we will see you know,

(Refer Slide Time: 27:12)

Hypothesis Tests for the Slope of the Regression Model

$H_0: \beta_1 = 0$ $H_a: \beta_1 \neq 0$	$t = \frac{b_1 - \beta_1}{S_b}$ <p>where: $S_b = \frac{S_e}{\sqrt{SS_{xx}}}$ $S_e = \sqrt{\frac{SSE}{n-2}}$ $SS_{xx} = \sum X^2 - \frac{(\sum X)^2}{n}$ β_1 = the hypothesized slope $df = n - 2$</p>
$H_0: \beta_1 \leq 0$ $H_a: \beta_1 > 0$	
$H_0: \beta_1 \geq 0$ $H_a: \beta_1 < 0$	

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

what are the wage; you have to again further address and although, we have already highlighted R square. So, I. In fact, I have only highlighted the, you know the significance of the parameters which we have actually gone through the process so; that means, the technically once you get the estimated values of the b_0 and b_1 that needs to be tested and the process of testing is t statistic, which we have already highlighted so; that means, technically, in the process of this you know regressions. So, we have actually regression coefficients that and that is the regression outputs for these. You know parameters and then the anova which is nothing, but you know explained sum of square, residual sums of squares, and total sum of square, and the corresponding degree of freedoms and then finally, the R square, the coefficient of determinations.

So, now like the parameters below b_0 and b_1 , that need to be tested and that is through t statistic like which we have already highlighted here and I have already explained and R square is also reported and R square also need to be tested and the test of R square followed, you know used to followed by f statistics and the parameter testings usually follow the t statistics. So, this is what the procedure, through which you have to check the significance of the parameters and then by default you go for the kind of, you know testings right.

(Refer Slide Time: 28:36)

Hypothesis Test: Airline Cost Example (Part 1)

$H_0: \beta_1 = 0$	$df = n - 2 = 10 - 2 = 10$
$H_1: \beta_1 \neq 0$	$\alpha = .05$

$t_{.025, 10} = 2.228$

If $|t| > 2.228$, reject H_0

If $-2.228 \leq t \leq 2.228$, do not reject H_0

Handwritten notes: $n=12$, $df=n-k=12-2=10$

Logos: IIT KHARAGPUR, NPTEL ONLINE CERTIFICATION COURSES

So, now this is how you have to fix the kind of hypothesis position, which I have already highlighted and then the same procedure. You have to calculate the t statistic by you know the estimated value divided by standard error, that will give you the calculated t statistic, then depending upon the alpha and degree of freedom and then you will get the critical value. So, now, as usual the inferential analytics, which you have discussed, if the calculator test statistic will overtake, the critical value which is actually calculated here, from the table means obtained from the tables with respect to two tailed tests and where alpha equal to 5 percent and the degree of freedom, actually a sample observation, is you know 12 and since it is a two variable model. So, your degree of freedom is nothing, but color the n minus k, that is actually 12 minus 2 and that to 10. So, that is how we have derive here 10 and. So, this is actually critical value. And now, the calculated value is coming why; you know coming when the basis of the sample information or the kind of an estimation, then we have to check.

And I having this kind of you know value. So, this is how the, this is the kind of you know requirement and accordingly the decision can be taken into consideration whether to reject or to accept the null hypothesis. So, you know. So, one way it is a actually critical value, then we will check the calculated value and you know see the kind of you know decisions right.

(Refer Slide Time: 30:10)

Hypothesis Test: Airline Cost Example (Part 2)

$$t = \frac{.0407 - 0}{.1773 \sqrt{\frac{73,764 - \frac{(930)^2}{12}}{12}}} = 9.43$$

Since $t = 9.43 > 2.228$, reject H_0

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, this is how the t value and the t calculated is coming a nine point four three and a this side. So, this this is actually so; that means, technically. So, this is actually the signal about the null hypothesis and this is the estimated coefficient and this is what the standard error and then the t value is coming nine point four three which is substantially higher than the critical value as a result you are rejecting the null hypothesis; that means, the you know the coefficient is actually statistically significant and it it gave the kind of inner strength for the kind of you know predictions and that is what the the kind of a not test you know the kind of you know requirement and then we will be will be moving towards you know again these problems.

So, the testing of the overall fitness of the model which I have already highlighted that you know the r square needs to be tested against you know you have actually a the f statistics which is actually the ratio between residual sum of square by explained sum of squares and then that will be adjusted with two degree of freedom and then we we have actually you know f critical and then the decision would be on the basis of you know the critical value and calculated a value like the previous you know t statistics. So, having the kind of you know sample informations and the critical value. So, we can actually move to the decisions and lets see how is the decision kind. So, this is what actually.

(Refer Slide Time: 31:44)

Testing the Overall Model (Part 2)

ANOVA					
	df	SS	MS	F	Significance F
Regression	1	2.79803	2.79803	89.092179	2.7E-06
Residual	10	0.31406	0.03141		
Total	11	3.11209			

$$F = \frac{\frac{SS_{reg}}{df_{reg}}}{\frac{SS_{err}}{df_{err}}} = \frac{MS_{reg}}{MS_{err}}$$

$$F = \frac{\frac{2.7980}{1}}{\frac{0.3141}{10}} = \frac{2.7980}{0.03141} = 89.09$$

$F = 89.09 > 4.96$ reject H_0

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

The anova table all together and. So, this is, and this is what the kind of you know sum of square that is a typically explained sum of square this is a residual sum of squared and this is total sum of square. So, the past check will be this plus this will be this like you know anova which you have already discussed and subject to this with you know degree of freedom will get mean sum of square and the ratio between the two will give you the f value right. So, so technically.

So, it is nothing, but actually 2798 and divided by 0.3; 0.314 and subject to the degree of freedom 1 and 10. So, you will get actually f f calculated is a 89.1 and the f critical which you have already fixed you know four point nine six calculated is hour taking the f critical. So, we are actually rejecting; that means, our fitness indication is that you know it is this highly and statistically significant at alpha equal to 5 percent right and accordingly you can actually you know go for the kind of inner predictions. So, in the case of you know parameters it is coming statistically significant and the overall fitness it is also coming statistically significant.

So, that means, but the cases giving you know green signal to justify that you know the model or the estimated outputs are you know fair enough to go for the kind of you know predictions and. In fact, this is not the only things, only checks we are supposed to do and then it will get a complete green signal, for the kind of enough prediction, on the management decision will have a thorough check with other diagnostics and then it will

give you some kind of you know more and more accurate kind of you know picture, or you know structure, through which you can go for further prediction and the kind of you know analysis.

(Refer Slide Time: 33:42)

The slide is titled "Point Estimation for the Airline Cost Example". It displays the linear regression equation $\hat{Y} = 1.57 + 0.0407X$. Below this, it states "For $X = 73$ ", followed by the calculation $\hat{Y} = 1.57 + 0.0407(73)$ and the result $= 4.5411$ or $\$4,541.10$. The slide includes logos for IIT Kharagpur and NPTEL Online Certification Courses at the bottom, and a small video inset of a presenter in the bottom right corner.

Now, moving forward again we will check the kind of enough validation. So, this is what the after knowing the kind of accuracy of the estimated parameters and the accuracy of the fitness, that is with respect to R square followed by f and again the parameters b_0 and b_1 s followed by t of b_0 t v of t of b_1 , then you know we are in a coming to a position, that you know the model is actually strong enough, you know to go for some kind of, you know predictions about this you know airline cost structure subject to the passengers right.

So, now what is the actually d prediction structure here? So, now,. So, the sample information gives the idea that you know this is what the predicted model. So, this is simply, actually in a quantitative model now and now, having any ex information; that means, once you know the exact number of you know passengers, then what should be the expected cost; that is how the management decision need to be, you know consider or need to be taken into consideration right.

So, now for any X. So, for instance, in this case, we are put in 73. So, now, if X equal to 73 then the estimated cost would be coming actually like this, you know 4,541 dollars right. So, this is how the kind of, you know this part is called as actually prediction part.

So, now,. So, this is actually kind of you know first having the information earlier that is how the learning can start and then it will be trained and then it will be you know predicted for the future. So, this is what the actually prediction is happening. So, again moving forwards.

(Refer Slide Time: 35:28)

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 ,

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

$$= 4.5411 \pm 1220$$

$$4.4191 \leq E(Y_{73}) \leq 4.6631$$

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, will see actually so; that means, technically if you move forwards, then you will ind the kind of you know structure so; that means, this is what the total predictions, sometimes what happens with the help of you know these sample estimates, you can the confidence intervals like we have already discussed in the case of an inferential analytics. So, now, here the sample information is actually Y estimated and then subject to the you know t values and the alpha and the degree of freedom. So, the Y estimate plus minus. So, this will give you the confidence interval of the two particular in our range.

So, now in this particular problem. So, this is the Y estimates corresponding to the you know X precisions that is 73 and then the plus minus of this you know adjustment will give you the confidence intervals. So, that is what the confidence interval about this one. So, this is the typical range, you know the kind of, you know prediction can vary and that is how the kind of you know structure, which you need a actually for any, the kind of you know; that means, what we can say that? This is what the management decision, you know you need to take or with the help of this particular you know structures or you know kind of, you know confidence intervals.

(Refer Slide Time: 36:50)

Confidence Interval to Estimate the Average Value of Y for some Values of X: Airline Cost Example		
X	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

So, now, we are in the previous case, we are putting X equal to 73 and we are getting the confidence interval from 4.0 to 24.66 in a round of structures against, if X will, each X will be changing. So, now, in this case if you know change X 62 68 73 85 90, this is how the sequence is there. So, every case having X putting X equal to 62 you will find you know Y estimate and then you will get a confidence interval similarly, putting X 68. So, you will find a Y estimate and gain again get the confidence interval. So, we will find, you know plus minus the kind of you know. So, lower limit and upper limit. So, like means you know. So, this will give you enough kind of you know exposures to take a decision or for you know; that means, you have actually lots of you know options to you, know to fix, the kind of you know structure as for the you know management requirement. So, this is how the management decision needs to be, you know taken into consideration. So, with the having of you know flexibility or the kind of you know structures. So, we are, if we are in a right track to, you know to understand the kind of in a situation and then we can take a decision as per the problem requirement or the business requirement.

So, with this the kind of you know structure. So, we can actually,

(Refer Slide Time: 38:26)

Prediction Interval to Estimate Y for a given value of X

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

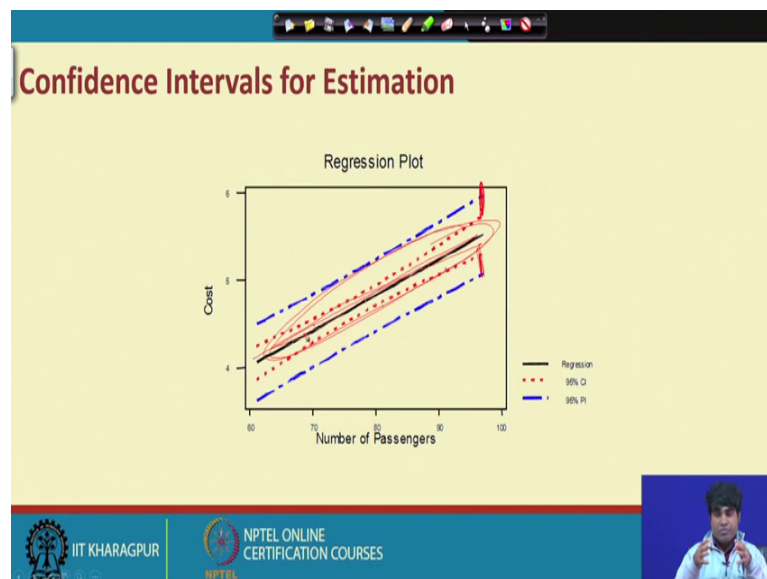
where: X_0 = a particular value of X

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

IIT KHARAGPUR NPTEL ONLINE CERTIFICATION COURSES

This is further kind of you know, in interval estimation predictions.

(Refer Slide Time: 38:29)



So, now, whatever confidence interval which you have already have. So, this is how graphically you can actually see and this is what the predictor line and suffer as you know confidently interval is concerned. So, this is the lower bound and this is what the upper bound so; that means,. So, your management decision need to be taken in this particular range and so; that means, technically one you know, you may be now in a position to know, that you know how regression can be very useful to give you a fair

kind of, you know scenario through which you can take. You know I cannot say the perfect you know, but it will give you know some kind of you know perfect accuracy to, you know to solve your management problem and the kind of decision which you can you know considered for the problem requirement or the kind of you know business requirement.

So, with the, with you know, with this you can actually move furthers and then you see a. So, with this actually we are actually, you know to get to know the kind of you know requirement so; that means, you know what we have already discussed that you know. So, with respect to the availability of two variables in this particular problem the, a cost and the kind of you know passengers. So, we like to know how we can actually start the particular you know process, then with the help you know very prediction, you know predictive operative analytic tools, that is you know regression analysis. So, you are in a position to know a lots of you know structure, through which you can take a fair judgment and then you can come into a position to take some kind of you know management decision, as far the business requirement. So, with this we will stop here, in the next class we will discuss details about the further complexity about this modeling.

Thank you very much; have a nice time.