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

Lecture - 28
Predictive Analytics (Contd.)

Hello everybody, this is Rudra Pradhan here, welcome you all to BMD lecture series. Today, we will continue predictive analytics 1 and that to discussion on multiple regression modeling, in the last couple of lectures, we have discussed this particular concept and that too to highlight, how to predict a particular variables as per the business requirement and then how to take a kind of management decisions and the discussion, which you had earlier is with respect to two variables and that too one is a dependent variable, which is a, which need to be predicted and then that is with respect a particular independent variable, how were in a kind of, business environment or in a real life scenario.

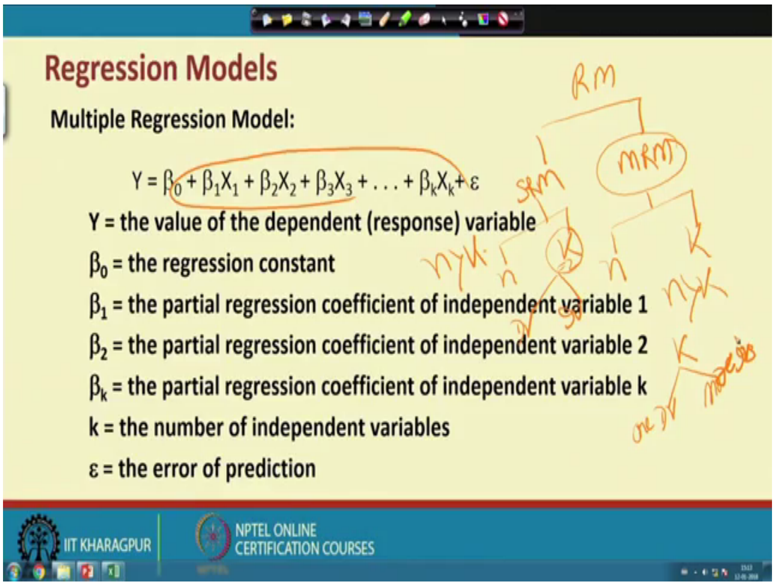
So, the prediction of a particular variable; reflected by various factors, which we have already discussing the last lectures, like the key performance of, kind of in a particular organization. Now, when you, when we are interested to predict a kind of variables, then we have to first identify, what are the factors through which we can do the kind of predictions.

So, there is a high chance that the prediction of that particular variable will be reflected by more number of variables and in a, of real life scenario or the kind of dynamics kind of environment. So, if you like to predict a particular variable, with respect to multiple independent variables, then the accuracy or the kind of optimality will be very perfect and that is why we have to very careful, how to address the kind of predictions, that too the predictive analytics, that too with the help of more number of independent variables.

So, the prediction of a particular variable with respect to one independent variable is very easy to handle or easy to address, but when will it, when we will be dealing this same problem means the prediction of that particular variable with respect to more number of variables, then there are lots of complexity you will find.

So, now, how will you address all these complexity and we will be dealing with all these problems. So, today we will discuss on that aspects and there is a high chance that the prediction of a particular variables with respect to more number of independent variables will give better kind of management decisions or means as per the particular business requirement.

(Refer Slide Time: 03:10)



Regression Models

Multiple Regression Model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k + \varepsilon$$

Y = the value of the dependent (response) variable

β_0 = the regression constant

β_1 = the partial regression coefficient of independent variable 1

β_2 = the partial regression coefficient of independent variable 2

β_k = the partial regression coefficient of independent variable k

k = the number of independent variables

ε = the error of prediction

IIT KHARAGPUR

NPTEL ONLINE CERTIFICATION COURSES

So, with this, a simple kind of background; so, will we start the kind of discussion and the literally framework of multiple regression model, is like this, and here Y is the dependent variables and then X 1 X 2 X 3 X 4 up to X k are independent variables, so; that means, technically. So, in the first instance, the regression modeling can be divided into two parts and the simple regression modeling, and then multiple regression modeling.

So, we have a construct called as a multiple variate regression modeling, which we are not discussing right now. So, the today's discussion is with respect to multiple regressions modeling in this case. So, here the kind of, there is an issue of n and there is an issue of k corresponding to the earlier discussion. This is also the game of n and k.

So, here in the simple regression modeling, k is equal to actually 2 that is the number of variables and the n is the sample size and the requirement is n, should be greater than to k and out of this total number of variables, one will be dependent variables and the other one is the independent variables and in the case of multiple regression modeling.

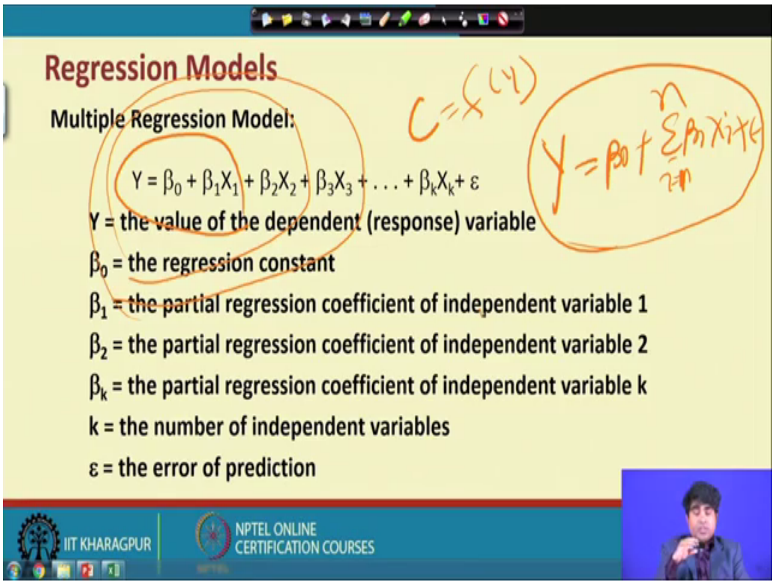
So; obviously, the same rule will be applied, n substantially greater than to k; however, in the case of k. So, the cluster will be; so, one dependent variables, and then more number of more numbers of more independent variables. So, right this is have the typical structures which you like to follow. So, now, so, how many independent variables, so, will be there in particular system the exclusively depends upon the kind of business environment or the kind of business problems.

So, will you have actually also the kind of structures were we can actually include one after another variables and then also exclude one after another variable depending upon the particular requirement or the kind of the kind of management a requirement. So, in the case of we know, in this particular case. So, we have to see how best we can actually address the problems and then will do the predictions with respect means prediction of Y with respect to the several independent variables. So, now; that means, technically, how many variables will be there in systems, the number of independent variables that to predict the variable Y?

So, that exclusively depends upon the kind of theoretical background about the business problem and from the theory, you to just identify the independent number of the independent variable, which can actually reflect the dependent variable. So, it should not be actually the kind of arbitrary choice. So, it is exclusively depends upon the theoretical background behind this prediction and then we will try to incorporate all these independent variables and when we are incorporating all these independent variables, so that times, so we have to see actually, so, how best we have to fix all these variables subject to the validity and reliability is concerned right.

So, let us see how is the kind of structures and in the first instance; so, the regression framework will be written like this.

(Refer Slide Time: 06:47)



Regression Models

Multiple Regression Model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k + \varepsilon$$

$C = f(Y)$

$Y = \beta_0 + \sum_{i=1}^{n} \beta_i X_i + \varepsilon$

Y = the value of the dependent (response) variable

β_0 = the regression constant

β_1 = the partial regression coefficient of independent variable 1

β_2 = the partial regression coefficient of independent variable 2

β_k = the partial regression coefficient of independent variable k

k = the number of independent variables

ε = the error of prediction

And if you put in a kind of, in a summation form, then this will be the regression modeling, will be Y equals to simply beta 0 plus summation beta i beta i Xi i equal to 1, to 1 to n 1 to n and then the error term. So, this is how the general form general format of multiple regression modeling. So, now, when you put i equal to 1 and, then the model will be restricted to this much, when you put i equal to 2, then the model will be extent to this much, when we will you put i equal to 3, then this model will extent to this much like this.

So, this is a kind of continues process and yes, mathematically it is very easy to include one after the other variables, but we are not dealing with mathematical problems. We are dealing with the business problem and we will include the variables, which are actually relevant for this kind of, for the prediction of that particular variable and corresponding to earlier discussions. So, the beta 0, beta 1, beta 2, beta 3, are the kind of parameters and these are all called as a partial regression, coefficient of independent variables from starting from 1 to k and beta 0 is the regression constant that is actually interceptor and then there is a kind of error terms. So, obviously in fact, obviously, like to know.

So, why there is a intercept and why there is a kind of error involvement in the particular process, because the first instance is, you are predicting Y, then without the kind of independent variables, the Y can also be there in the system. For instance, let us take an example of consumption predictions, with respect to incomes.

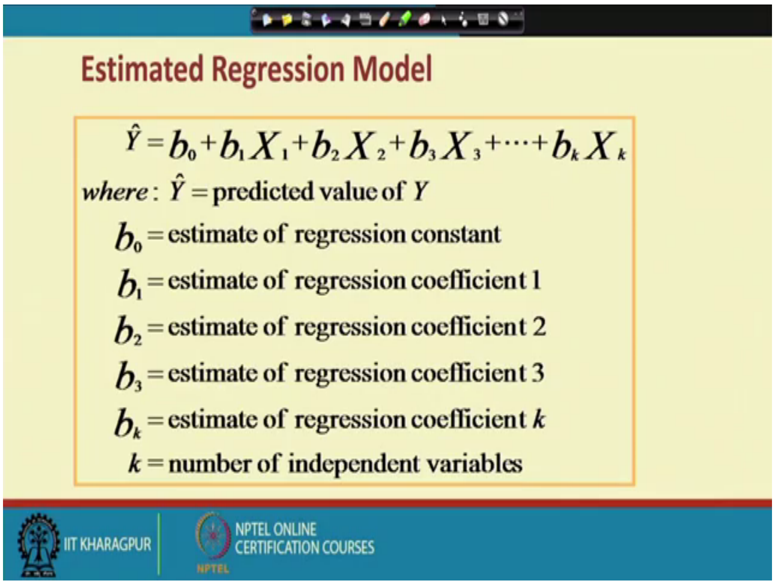
So, now, consumption typically depends upon the income of a particular persons, but you will find in real life scenario, lots of people are there, they had no income, but by the way they must have actually consumptions. So, in that case, so, when there is no income, still there is a consumptions and that consumption will take care by the intercept so; that means, a. So, the consumption of a particular individual will not depend upon his or her income, but it will be depend on somebody's income or something like that right.

So, that is how; so, the intercept will be there in the system and in the meantime. So, error term in Y error in the system of regression modeling, because of many reasons and one of the typical, a typical reason is that, we are not in a position to capture all the independent variables to predict the dependent variables. Since, we are not in a position to incorporate all the independent variables. So, the error term can take care that particular machine variables and sometimes when you are means, you are dealing with the data sometimes; there may be error in data entry or something like an choice of the particular functionality and choice of a particular model.

So, like that there are many different reasons, through which you actually, we have to incorporate the error term in the systems and in fact, we have already discussed this particular aspect and typically, the regression framework is dependent variables, independent variables and the parameters. Out of all these parameters, one is the independent; with the all independent variables that is what intercept and the other variables are reflected by the coefficients like beta 1 beta 2 beta 3 and then the error term. So, this is the framework of multiple regression modeling and now, how will it deal with this models to predict the business environment or to do the kind of management problems, kind of investigation. So, we will get to know in details, so, now, the kind of structure will be like this.

So, corresponding to the previous model; so, this what the estimated model, so; that means, we have a structure of population model and then this is sample specification and here, so, with a particular samples or the kind of structure. So, we have to estimate the models and the way, we will estimate the model, the error terms will be removed in the process so; that means, technically we will first build the model like the previous ones and then here actually, there will an error terms and then these parameters are unknown.

(Refer Slide Time: 11:20)



Estimated Regression Model

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + \cdots + b_k X_k$$

where: $\hat{Y}$ = predicted value of Y

b_0 = estimate of regression constant

b_1 = estimate of regression coefficient 1

b_2 = estimate of regression coefficient 2

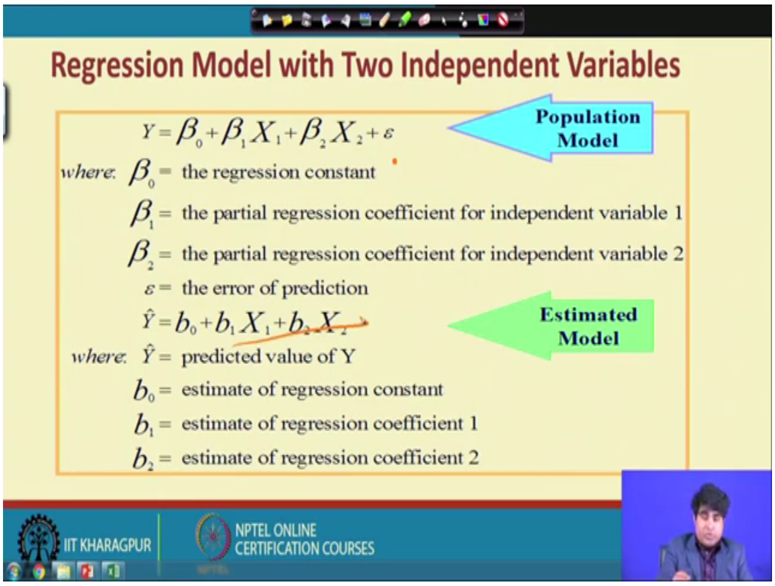
b_3 = estimate of regression coefficient 3

b_k = estimate of regression coefficient k

k = number of independent variables

Now, with the sample data or the sample problem; so, we have to estimate the particular models and then by default error, term will be removed and finally, the coefficient will be, they have to predict the kind of dependent variables.

(Refer Slide Time: 11:35)



Regression Model with Two Independent Variables

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

where: β_0 = the regression constant

β_1 = the partial regression coefficient for independent variable 1

β_2 = the partial regression coefficient for independent variable 2

ε = the error of prediction

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2$$

where: $\hat{Y}$ = predicted value of Y

b_0 = estimate of regression constant

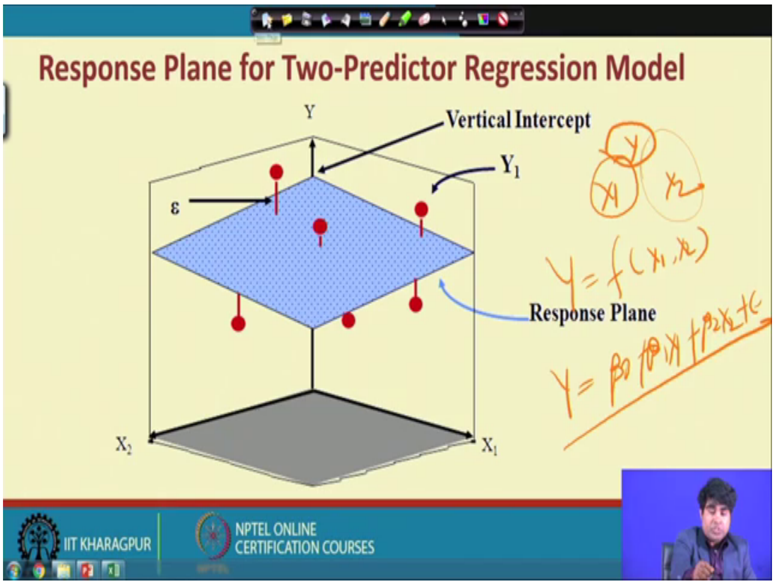
b_1 = estimate of regression coefficient 1

b_2 = estimate of regression coefficient 2

So, that is the usual structure and which we have already discussed in the last lecture and for the simplicity point of view. So, let us start with two independent variables, because we have discussed the kind of regression modeling, with respect to one independent variable, in the previous lecture.

And, here we will just extend by putting one more independent variable, in the system, and we will see how the particular structure, then by default is, we can actually extend the particular process of the variables, then slowly you can actually drop, if it is actually not coming important for this prediction. So, we get to know in details in later stage, but in the meantime; so, this is how the particular structure and in this structure; so, this is the kind of population model and this is what a sample specific model, and here error term will be in not there, in the systems, because it will take care with this data or the sample problem. So, now, corresponding to these, so, will see how is the particular structure?

(Refer Slide Time: 12:35)

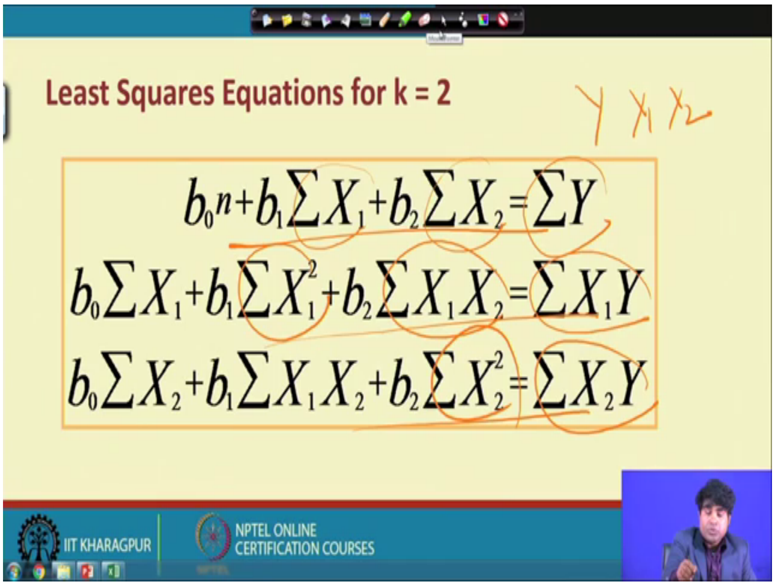


Response Plane for Two-Predictor Regression Model

So, typically the kind of requirement is like this. So, when you have actually dealing with the two independent variables, two independent variable typically, so; that means, if this is what the Y variables and this will be reflected by two independent variables, that is actually X 1 and then X 2.

So; that means, this is how the game all about since, we are dealing with two independent variables. So, there should not be any relationship between X 1 and X 2. So, if there is again relationship between X 1 and X 2 predicting Y. So, this will be again problems. So, which that problem, we will discuss in the latest stage, but in the meantime. So, you can now, fix the model like this Y, equals to function of X 1 X 2 and then the regression modeling will be Y equal to beta 0 plus beta 1, X 1 beta 2 X 2 and then plus error term. So, this is, have a simple mathematical representation of the prediction of Y subject to the availability of X 1 and X 2. So, now, we like to see how is the kind of estimation process?

(Refer Slide Time: 13:50)



Least Squares Equations for k = 2

$Y \quad X_1 \quad X_2$

$$b_0 n + b_1 \sum X_1 + b_2 \sum X_2 = \sum Y$$
$$b_0 \sum X_1 + b_1 \sum X_1^2 + b_2 \sum X_1 X_2 = \sum X_1 Y$$
$$b_0 \sum X_2 + b_1 \sum X_1 X_2 + b_2 \sum X_2^2 = \sum X_2 Y$$

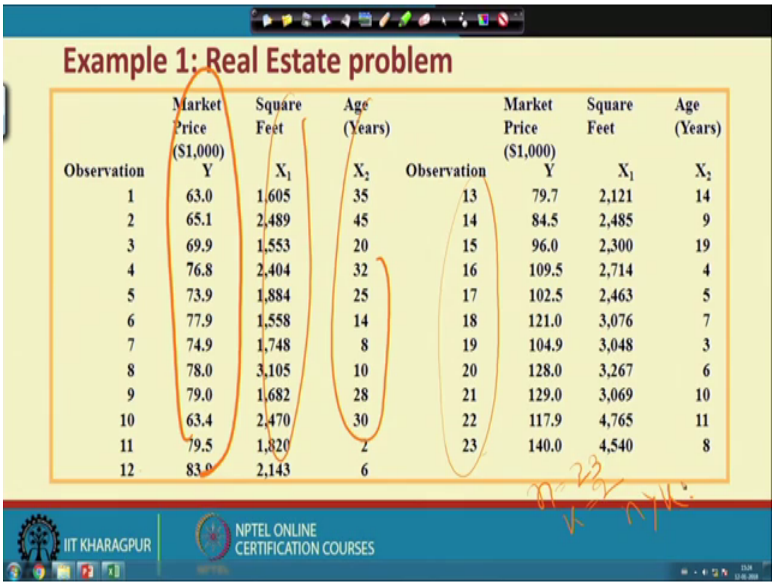
So, ultimately, the idea is a we have to apply technique and then we try to minimize the errors, sum squares, and the way we will minimize the errors sum squares, then the coefficient by default will be calculated, and in the last lectures, we have dealing with, in a two variables; that means, in that case it is least square equations, for k equals to 2; that means, two independent variables. So, now, in the earlier case, we have discussed with, one independent variables.

So, that is why we dealt with the two different structural equations. Now, since there are three parameters and three variables Y X 1 and X 2.

So, you must a have actually three structural equation. So, these are the equation through which actually, you have to address the kind of problem and then after simplifications, you will get the values of these parameters. So, typically, so, you have actually Y X 1 and X 2. So, now, so, you need actually, summation X 1 summation X 2, then summation Y, so; that means, with respect to data. So, you can find out the summation, then you need summation X 1 squares and summation X 1 X 2 and summation X 1 Y then finally, you needs summation X 2 square and summation X 2 Y.

So, after go to the excel spreadsheets and manually, you can calculate all these things. After getting all this value, you can just put it here, and then solve these equations and then by default, you will get the coefficient of b 0 b 1 b 2, these are all the parameters, through which we need to actually do the predictions that is with respect to Y. So, corresponding to this equations; so, we can connect with a problem and then get to know how this actually the process is happening.

(Refer Slide Time: 15:46)



Example 1: Real Estate problem

Observation	Market Price ($1,000) Y	Square Feet X_1	Age (Years) X_2	Observation	Market Price ($1,000) Y	Square Feet X_1	Age (Years) X_2
1	63.0	1,605	35	13	79.7	2,121	14
2	65.1	2,489	45	14	84.5	2,485	9
3	69.9	1,553	20	15	96.0	2,300	19
4	76.8	2,404	32	16	109.5	2,714	4
5	73.9	1,884	25	17	102.5	2,463	5
6	77.9	1,558	14	18	121.0	3,076	7
7	74.9	1,748	8	19	104.9	3,048	3
8	78.0	3,105	10	20	128.0	3,267	6
9	79.0	1,682	28	21	129.0	3,069	10
10	63.4	2,470	30	22	117.9	4,765	11
11	79.5	1,820	2	23	140.0	4,540	8
12	83.0	2,143	6				

$n=23$ $k=2$ $n>k$

IIT KHARAGPUR

NPTEL ONLINE CERTIFICATION COURSES

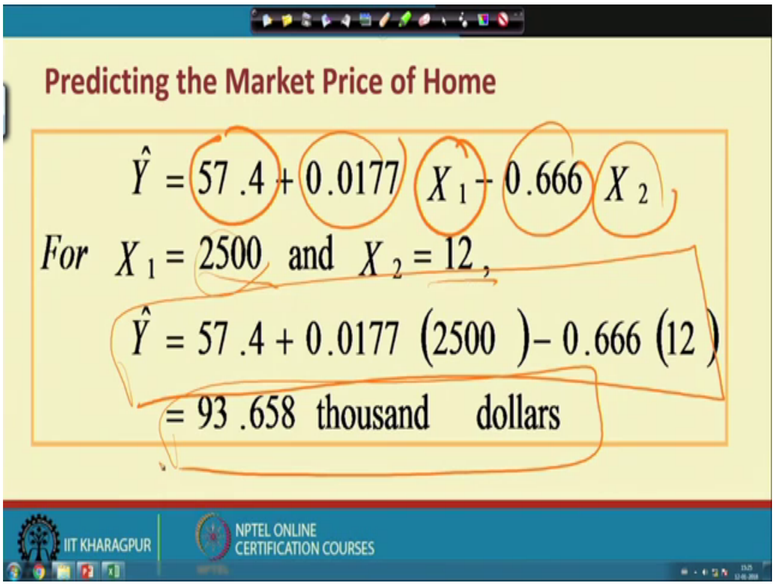
So, here let us take example. So, this is a real estate problem and in the real estate problems. So, every times our job is to predict the housing price.

So, that is the market price of a house and it is reflected by two independent variable that is the size of the house and that is measured with respect to square feet and then age number of years depending upon, the kind of building constructions.

So, this can be the another variable, through which you can do the predictions so; that means, technically. So, we have actually, we have actually dependent variable here and then two independent variable X 1 and X 2 and 12 and 23. So, total 23 sample observations and since so; that means, n equals to 23 here and k equal to k equal to 2 that is two independent variables; that means, some total variables here is a 3 and as a result. So, n is greater than k. So, accordingly, we can actually do the processing; that means, technically, this is fair enough to proceed for a regression modeling and do the prediction and then come to the kind of management decision.

So, now, the way yesterday, we have discussed. So, we can proceed for this calculations and then in a same things.

(Refer Slide Time: 17:10)



Predicting the Market Price of Home

$$\hat{Y} = 57.4 + 0.0177 X_1 - 0.666 X_2$$

$For \; X_1 = 2500 \; and \; X_2 = 12,$

$$\hat{Y} = 57.4 + 0.0177(2500) - 0.666(12)$$

$$= 93.658 \text{ thousand dollars}$$

So, after the estimations you will get this Y estimate and the parameters. Now, this is actually b 0 and this is b 1 and this is b 2 so; that means, initially we are assuming that. So, high is the square feet; then high is the housing price and the number of year’s means; it depends upon the starting of the construction.

So, when the building will be older and older and older then the value of that particular price will be negatively connected. So, as a result the coefficient is also coming negative.

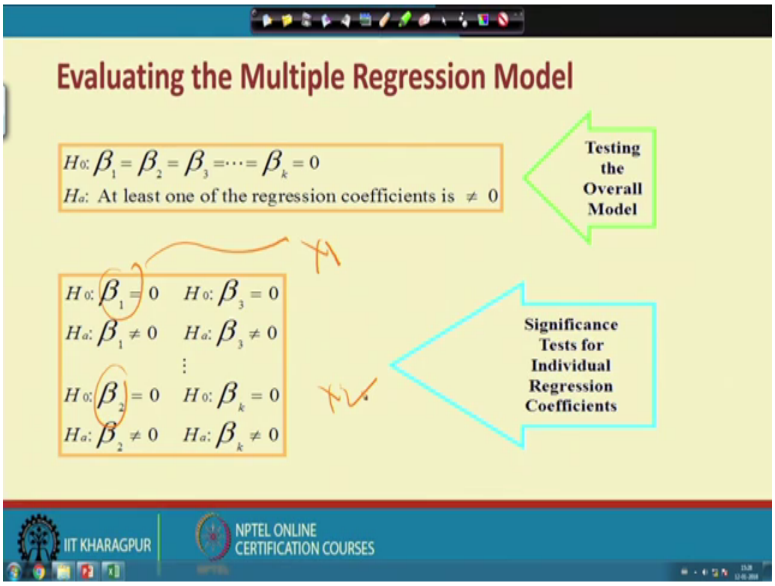
So; that means, before we start the process, you must have actually theoretical expectation, what should be the kind of link and that need to be validated, through this data set or the kind of samples. So, now, our job is actually to predict the market price of a particular house and here, there are two variables and having a X 1 know kind of fix, of the square feet and the kind of number of years of that building. So, you can actually predict, what should be the housing price. So, now, putting X 1 equal to 2500 and X 2 equal to 12.

So, you can say that the housing price for this particular X 1 to 2500 and X 2 12 will be a 93.66 thousand dollars. So, this much actually you have to predict so; that means, just to imagine, how regression can useful for doing the predictions and that to how predictive analytics can be very useful for doing the kind of management decision. So, now, only thing is to know, how these parameters are coming and in fact, we have already discussed the particular process, how to get all these parameters.

So, once you get these parameters, then by default you can do the kind of prediction. So, we can go to the excel sheet and in the spread sheet very easily get this output and just you need to connect with dependent variable and independent variable and by default, you will get this as a result. So, after getting this result; so, you to just check that, whether these coefficients are statistically significant and the goodness fit of this particular model.

So, that which we have already discussed in the previous lecture and that is same thing here that was with respect to two variable cases.

(Refer Slide Time: 19:48)



Evaluating the Multiple Regression Model

H_0: $\beta_1 = \beta_2 = \beta_3 = \cdots = \beta_k = 0$

H_a: At least one of the regression coefficients is $\neq 0$

Testing the Overall Model

H_0: $\beta_1 = 0$ $\quad$ H_0: $\beta_3 = 0$

H_a: $\beta_1 \neq 0$ $\quad$ H_a: $\beta_3 \neq 0$

$\vdots$

H_0: $\beta_2 = 0$ $\quad$ H_0: $\beta_k = 0$

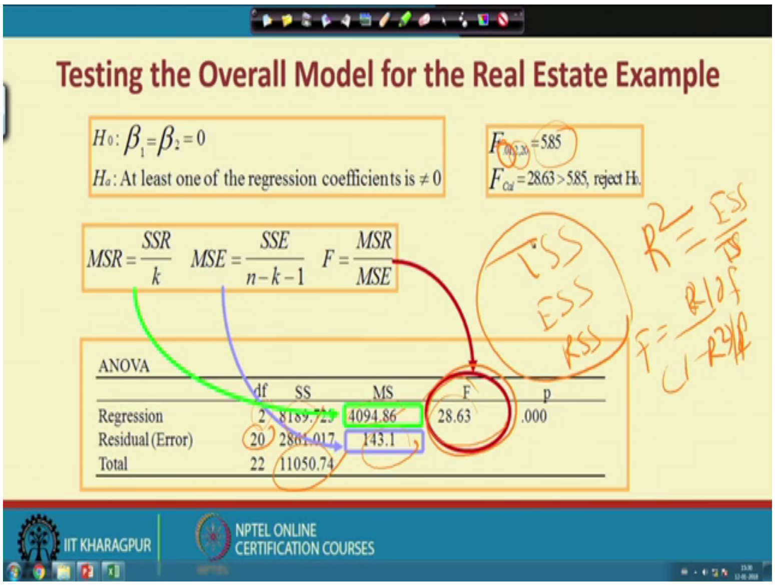
H_a: $\beta_2 \neq 0$ $\quad$ H_a: $\beta_k \neq 0$

Significance Tests for Individual Regression Coefficients

Now, it is with respect to three variable case, where one dependent variable, with the two independent variables; so, now, knowing the output, so, we can proceed to see how the kind of testing? So, this is what simple prediction. Now, in this case; so, the coefficients are actually beta 1 beta 2 beta 3. So, that needs to be equal to 0. So, the alternative hypothesis, they are not equal to 0. So, as a result; so, we like to check what is the sample estimates and then on the basis of sample estimates and we have to apply the t statistic and then check the significance of the parameters. So; that means, in this case, in this case, we have actually two parameters that is, beta 1 and beta 2, beta 1 is reflected by X 1 and beta 2 is reflected by X 2.

So; that means, we like to know, which particular variable is significant and which particular variable is not significant? We need actually both variable, should be statistically significant and then we like to check, whether both are statistically significant or they are not at all significant or at least one is significant to predict the Y that to housing price with respect to number of years and the kind of square feet right. So, this is how the process and then will check a how is the kind of testing.

(Refer Slide Time: 21:04)



Testing the Overall Model for the Real Estate Example

$H_0: \beta_1 = \beta_2 = 0$

H_a: At least one of the regression coefficients is $\neq 0$

$F_{.01,2,20} = 5.85$

$F_{Cal} = 28.63 > 5.85$, reject H_0

$MSR = \frac{SSR}{k}$ $MSE = \frac{SSE}{n-k-1}$ $F = \frac{MSR}{MSE}$

ANOVA

	df	SS	MS	F	p
Regression	2	8189.723	4094.86	28.63	.000
Residual (Error)	20	2861.017	143.1		
Total	22	11050.74			

TSS
ESS
RSS
$R^2 = \frac{ESS}{TSS}$
$F = \frac{R^2/df}{(1-R^2)/df}$
IIT KHARAGPUR
NPTEL ONLINE CERTIFICATION COURSES

So, the testing followed by two way structure, first check is the overall fitness of the model. So, that is the ANOVA, which we have already discussed in the last lecture.

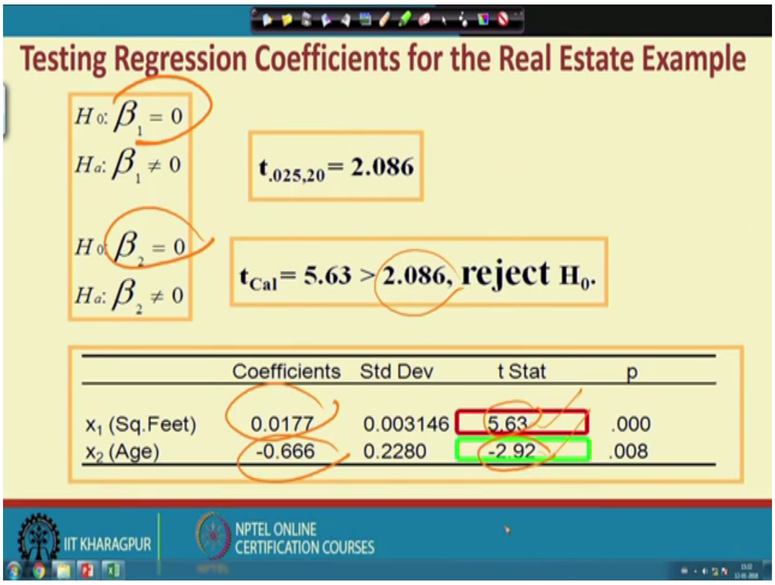
So, the ANOVA statistic depends upon three things; TSS ESS and RSS and it will reflected by two major indicators, that is R square, the coefficient of determination and which is the ratio between ESS by TSS, then f, which is reflected by a, reflected by R square by 1 minus R square and subject to degree of freedom, which we have already calculated here. So, this is actually f calculation and this is the R square, value will be just with the help of X plane sum square by total sum square. It will get R square value and like the previous discussion, higher the R square better is the model fit lower their square lower is the model fit, whether it is a higher and lower.

So, we need actually, the significance of overall fitness of the model that is reflected by f statistics. We have to follow the particular of structure alpha fixes on and the kind of degree of freedom then you have to find out the critical value. So, in this case, we are putting actually an alpha equal to 1 percent and then the number of sample points corresponding to the number of sample points. So, we are getting the critical value is 5.85 that is actually, because in the f statistic, we need 2 degree of freedom, the upper part and the lower part; that means, technically for this side and for this side. So, this is reflected by 2 and 20 and then, so, this is the calculated f statistics.

So, now, we have to compare since, the calculated a value is cover taking the critical. So, we can actually reject.

So, that means; so, model is saying that the fitness of the model or goodness of the fit is actually fair enough to go for the kind of prediction. So, the other checks, which we will need to have. So, that is actually with respect to the parameters and we should know that the beta 1 beta 2 should be actually statistically significant or not.

(Refer Slide Time: 23:17)



Testing Regression Coefficients for the Real Estate Example

$H_0: \beta_1 = 0$

$H_a: \beta_1 \neq 0$

$H_0: \beta_2 = 0$

$H_a: \beta_2 \neq 0$

$t_{.025,20} = 2.086$

$t_{Cal} = 5.63 > 2.086$, reject H_0.

	Coefficients	Std Dev	t Stat	p
x_1 (Sq.Feet)	0.0177	0.003146	5.63	.000
x_2 (Age)	-0.666	0.2280	-2.92	.008

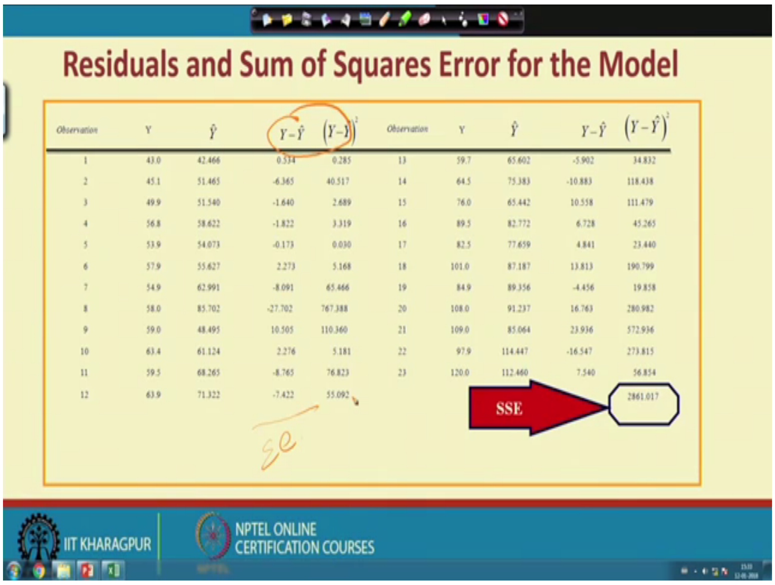
So, accordingly the null hypothesis will be. So, beta 1 equal to 0 and beta 2 equal to 0 and the counter for alternative hypothesis will be beta 1 not equal to 0 and beta 2 not equal to 0 it is also same thing. So, these are all beta coefficients. So, this is actually beta 1 coefficients and this is beta 2 coefficients; so, the intercept, which we are actually skipping here, because it is not important when we go for more and more variables to predict the Y.

So, the variable importance, more important than the kind of intercept, but when you are having less and less number of independent variable then the intercept impact may be coming high. So, now, accordingly; so, you can actually calculate the kind of structure and then you check, whether it is coming actually statistical significant or not. So, now, the same structures; so, you have to compare the calculated, with the critical and the calculated will be coefficient with the standard error and then will get the t statistics, which is calculated for X 1 and that is for beta 1 and this is for beta 2 and that will be with respect to X 2 and then critical value is 2.086.

So, in both the cases, there crossing the critical level that means, the calculate statistic is the taking the critical value as a result we are in a position to reject the null hypothesis. So, technically, so, we are in a position to predict the y now with respect to X 1 and X 2. So; that means, the problem, which we have connected. Now, to predict the housing price market price of house subject to the size of the house and the number of years the kind of with respect to their building. We have to predict, the kind of situation and then we have to take the kind of management decision, what should be future market price of that particular house subject to the square feet and kind of number of years that is the age of the house.

So, now, so, likewise, it can be also actually extent, with respect to more number of variables here.

(Refer Slide Time: 25:18)



Residuals and Sum of Squares Error for the Model

Observation	Y	$\hat{Y}$	$Y-\hat{Y}$	$(Y-\hat{Y})^2$	Observation	Y	$\hat{Y}$	$Y-\hat{Y}$	$(Y-\hat{Y})^2$
1	43.0	42.466	0.534	0.285	13	59.7	65.602	-5.902	34.832
2	45.1	51.465	-6.365	40.517	14	64.5	75.383	-10.883	118.438
3	49.9	51.540	-1.640	2.689	15	76.0	65.442	10.558	111.479
4	56.8	58.622	-1.822	3.319	16	89.5	82.772	6.728	45.265
5	53.9	54.073	-0.173	0.030	17	82.5	77.659	4.841	23.440
6	57.9	55.627	2.273	5.168	18	101.0	87.187	13.813	190.799
7	54.9	62.991	-8.091	65.466	19	84.9	89.356	-4.456	19.858
8	58.0	85.702	-27.702	767.388	20	108.0	91.237	16.763	280.982
9	59.0	48.495	10.505	110.360	21	109.0	85.064	23.936	572.936
10	63.4	61.124	2.276	5.181	22	97.9	114.447	-16.547	273.815
11	59.5	68.265	-8.765	76.823	23	120.0	112.460	7.540	56.854
12	63.9	71.322	-7.422	55.092					

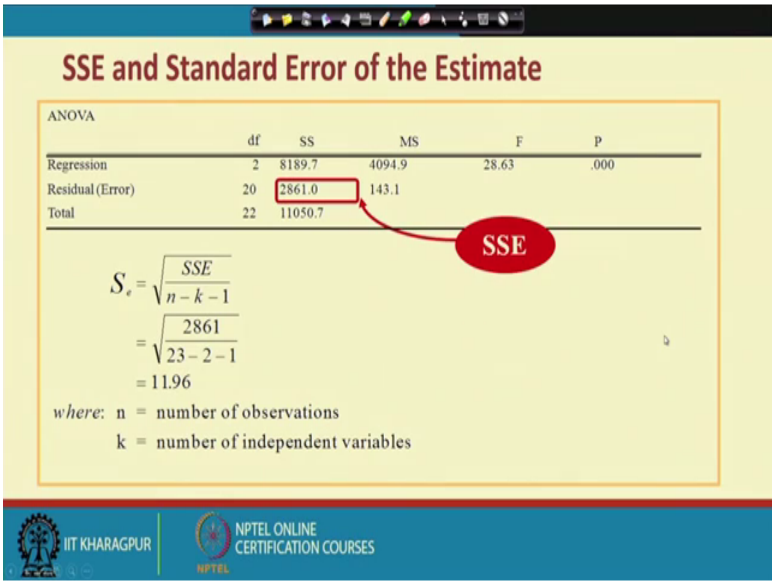
SSE 2861.017

So, the same structure; so, we once, we get the estimated model. So, we will be dealing with residuals.

So, first thing you to find out the residuals; so, which is actually, the difference between Y minus Y head and then we have to see that the sum of the estimated error sum should be equal to 0 so; that means, technically, this error sum should be equal to 0 and sum of the error square is nothing, but actually squaring, all the items and then you take the summation that is actually sum of the error sum of the squares and then once, you get the error sum of squares. So, with the help of this error sum squares.

So, you can actually do it entire testing's that is the kind of the R, the goodness fit check and the kind of specification check that is with respect to the parameters beta 1 and beta 2 in means particularly with respect to this problem.

(Refer Slide Time: 26:20)



SSE and Standard Error of the Estimate

ANOVA

	df	SS	MS	F	P
Regression	2	8189.7	4094.9	28.63	.000
Residual (Error)	20	2861.0	143.1		
Total	22	11050.7			

$$S_e = \sqrt{\frac{SSE}{n-k-1}}$$
$$= \sqrt{\frac{2861}{23-2-1}}$$
$$= 11.96$$

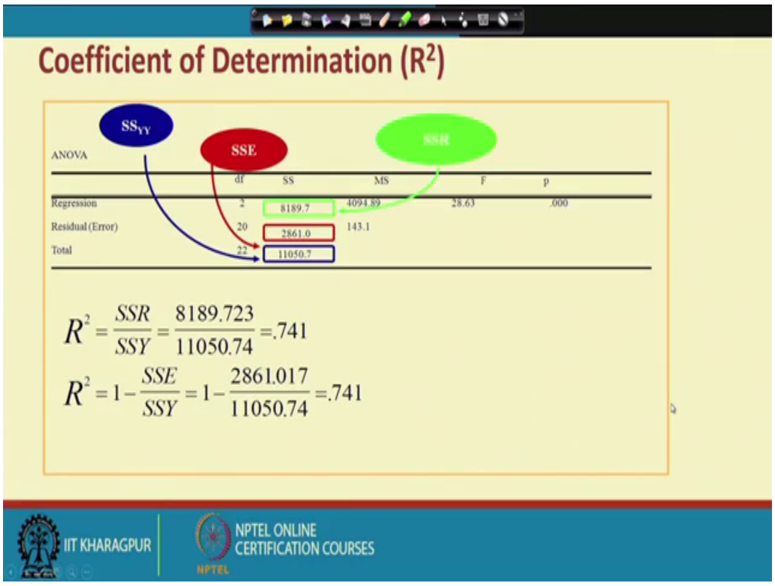
where: n = number of observations

k = number of independent variables

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, now this is how the particular structure and we have already discussed. So, this is what actually the calculated f value and it is actually statistically significant right.

(Refer Slide Time: 26:32)



Coefficient of Determination (R²)

ANOVA

	df	SS	MS	F	p
Regression	2	8189.7	4094.89	28.63	.000
Residual (Error)	20	2861.0	143.1		
Total	22	11050.7			

$$R^2 = \frac{SSR}{SSY} = \frac{8189.723}{11050.74} = .741$$

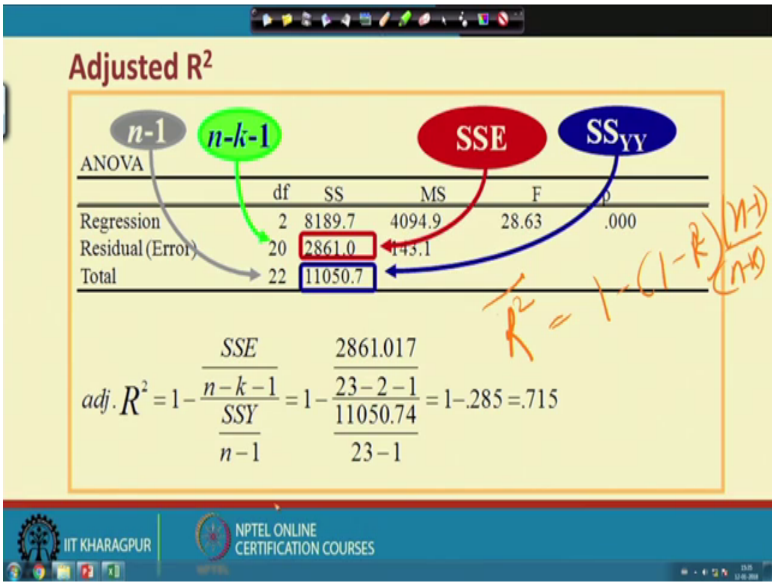
$$R^2 = 1 - \frac{SSE}{SSY} = 1 - \frac{2861.017}{11050.74} = .741$$

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Now, so, the R square value is coming here, in this particular structure that is the ratio between ESS to TSS. So, that is coming actually 7 is 0.741 so; that means, technically. So, it is actually coming 70 percent, what we called 74 four percent. So, the impact of independent variable to dependent variable is reflected by 74 percent so; that means, it is a high accuracy to predict the kind of housing price with respect to a size of the house and the kind of age of the house.

So, this is actually very interesting to know the kind of the structure and then you to take the management decision as per the particular requirement and knowing all these things. So, it is easy to easy to check the kind of process.

(Refer Slide Time: 27:25)



Adjusted R²

ANOVA

	df	SS	MS	F	p
Regression	2	8189.7	4094.9	28.63	.000
Residual (Error)	20	2861.0	143.1		
Total	22	11050.7			

$$\bar{R}^2 = 1-(1-R)\frac{(n-1)}{(n-k)}$$

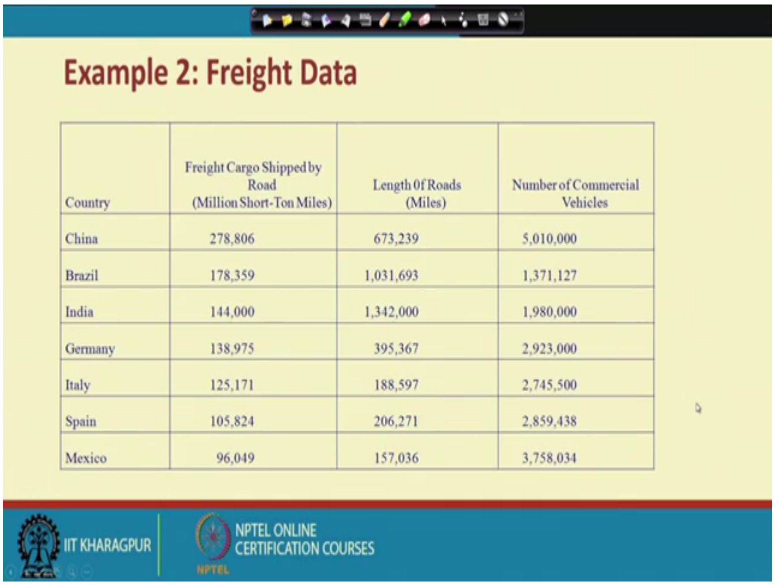
$$adj.R^2 = 1-\frac{\frac{SSE}{n-k-1}}{\frac{SSY}{n-1}} = 1-\frac{\frac{2861.017}{23-2-1}}{\frac{11050.74}{23-1}} = 1-.285 = .715$$

And then we will see how the particular structure? So, R square will be again a there is a kind of adjusted R square and since, when you go for more and more independent variable then R square will start increasing indefinitely. So, as a result; so, when we are dealing with multiple regression or multi multivariate regression that time, it is the adjusted R square, which can reflect the kind of predictions rather than simple R square.

So, usually R adjusted R square, it can be written as this is adjusted R square, this is nothing, but 1 minus R square were into n minus 1 by n minus k that is the degree of freedom and simply you can calculate it, by this particular process and if the sample size is very strong enough, then the difference between R square and adjusted R square will not be higher, but if, but in the case of small sample. So, there will be difference and that may affect the kind of predictions, but by the way.

So, whether it is R square or adjusted R square, both are the kind of indicator, through which will do the predictions and that to predict the housing price, with respect to size of the house and the age of the house.

(Refer Slide Time: 28:51)



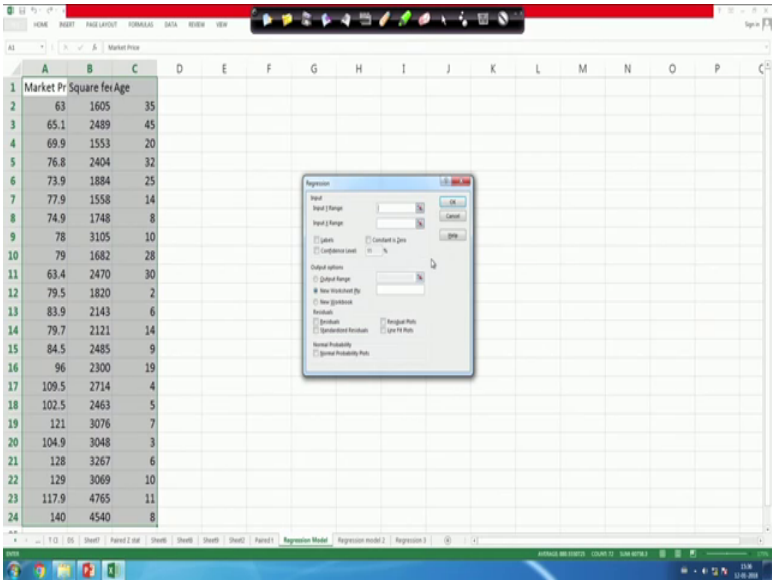
Example 2: Freight Data

Country	Freight Cargo Shipped by Road (Million Short-Ton Miles)	Length Of Roads (Miles)	Number of Commercial Vehicles
China	278,806	673,239	5,010,000
Brazil	178,359	1,031,693	1,371,127
India	144,000	1,342,000	1,980,000
Germany	138,975	395,367	2,923,000
Italy	125,171	188,597	2,745,500
Spain	105,824	206,271	2,859,438
Mexico	96,049	157,036	3,758,034

IIT KHARAGPUR
NPTEL ONLINE CERTIFICATION COURSES

So, this is another way to check the consistency of the models or the accuracy of the models and this is another kind of example. Let us, I take you to excel sheet then I will show you.

(Refer Slide Time: 29:04)



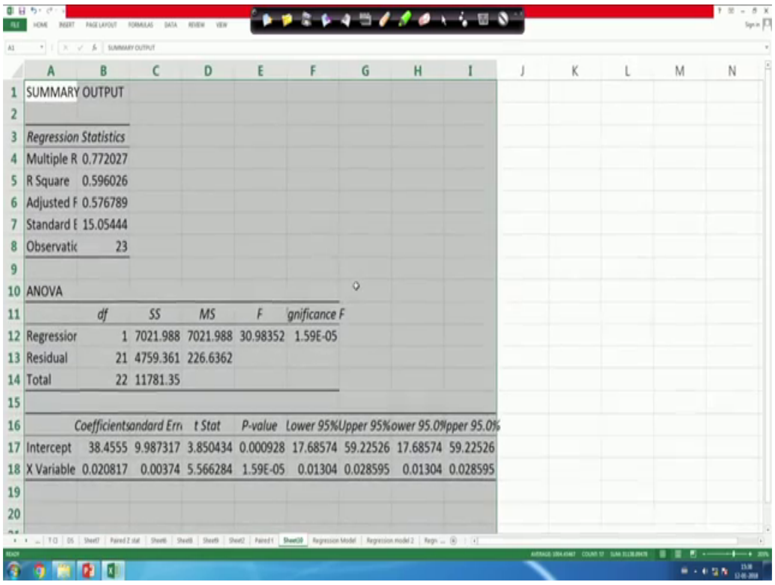
	A	B	C
1	Market Pr	Square fe	Age
2	63	1605	35
3	65.1	2489	45
4	69.9	1553	20
5	76.8	2404	32
6	73.9	1884	25
7	77.9	1558	14
8	74.9	1748	8
9	78	3105	10
10	79	1682	28
11	63.4	2470	30
12	79.5	1820	2
13	83.9	2143	6
14	79.7	2121	14
15	84.5	2485	9
16	96	2300	19
17	109.5	2714	4
18	102.5	2463	5
19	121	3076	7
20	104.9	3048	3
21	128	3267	6
22	129	3069	10
23	117.9	4765	11
24	140	4540	8

So, this is actually the problem, which has taken, is like this, we like to know how to feed the regression model in order to predict the housing price subject to square feet and the age of the house. So, what will you do, go to the data analysis and then select the data analysis package and then check the regressions.

So; that means, see the idea is here in the predictive analytics, you should know the kind of techniques, the kind of procedures, then you can use any software's, there is no hard and fast rules. So, you can use excel spread sheets, you can use SPSS, you can use e views. You can use R software, you can use SAAS, you can use STATA. So, there are so, many software's are there, you can use MATLAB. So, but the thing is that so, we need to get the estimated equations, in this particular case Y, with respect to X 1 and X 2 since, we know the model. So, we have to just connect and then the software will, by default, will give you the result, because when we have actually more number of data and the problem is very complicated.

So, getting the result manually very difficult; so, that is why we need actually the kind of the software, through which will get the results, ones you get the result then you will go for the kind of interpretation and then we will go for the predictions and then we will go for some kind of management decision. So, now, corresponding to this problem; so, what will you do? So, first you give the input range. So, in this case, you put actually input range here. So, then this is actually a Y, then you put actually X 1. So, this is same way, same we have to find out and then just put; this is same way you are getting the results. So, this is very easy actually. So, getting one after another variable.

(Refer Slide Time: 30:58)



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.772027
R Square	0.596026
Adjusted F	0.576789
Standard E	15.05444
Observatic	23

ANOVA

	df	*SS*	*MS*	*F*	*gnificance F*
Regressior	1	7021.988	7021.988	30.98352	1.59E-05
Residual	21	4759.361	226.6362		
Total	22	11781.35			

	Coefficients	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	38.4555	9.987317	3.850434	0.000928	17.68574	59.22526	17.68574	59.22526
X Variable	0.020817	0.00374	5.566284	1.59E-05	0.01304	0.028595	0.01304	0.028595

So, then it will give you the output of the full models and this is how the kind of f statistic and then the R square adjusted, R square and the number of observations counting and the ESS RSS TSS and the intercepts the coefficient of all these parameters.

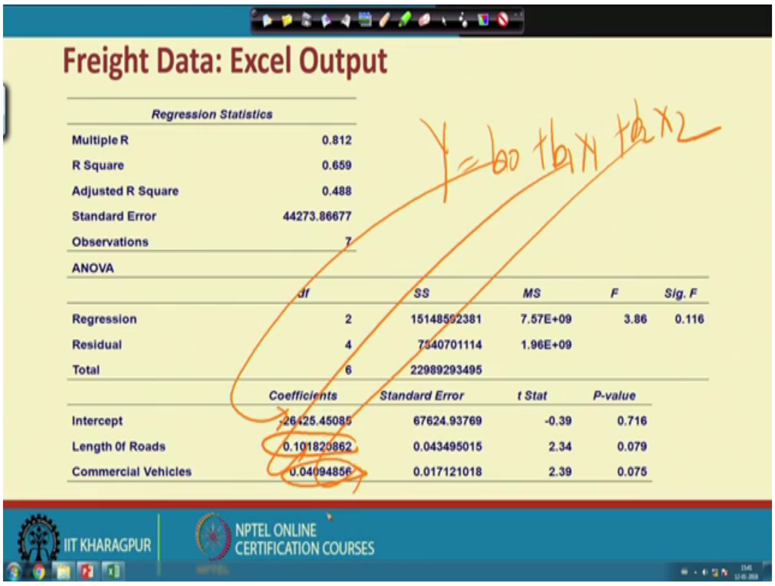
So; that means, actually. So, we can also have the kind of lower bound, upper bound so; that means, software will give you all kinds of results. So, it will help you to go for the kind of prediction, but only thing that only requirement is that you have to know the technique, understand the technique then point out the right dependent variable, point out the right independent variables, fix the kind of structure, check the data, then will go for the kind of processing.

So, once everything is so, just to give the command to the software and software will give you just within few minutes, it will give you the result and once, you get the results, with the help of results, you will get the kind of prediction structure and you will do the predictions and then you can go for the kind of management decision. So, this is how the kind of structure, which you can have the process and; so, similarly, there is another kind of data.

So, this is again cargo data and then it is reflected by length of the rods and number of commercial vehicles. So, we are assuming that the cargo reflection is reflected by length of rods and number of commercial vehicles. So, we are expecting that there may be a positive relationship between the freight cargo and the length of the rods and then the number of commercial vehicles right.

So, this is actually the cross sectional data, because the reflection is country wise information and again, we have to check, whether the variables are statistical significant to predict a cargo the requirement. So, this is the same way again, it is with respect to dependent variable and that to with respect to two independent variables X 1 and X 2.

(Refer Slide Time: 33:07)



Freight Data: Excel Output

Regression Statistics	
Multiple R	0.812
R Square	0.659
Adjusted R Square	0.488
Standard Error	44273.86677
Observations	7

ANOVA

	df	SS	MS	F	Sig. F
Regression	2	15148592381	7.57E+09	3.86	0.116
Residual	4	7540701114	1.96E+09		
Total	6	22989293495			

	Coefficients	Standard Error	t Stat	P-value
Intercept	-26425.45085	67624.93769	-0.39	0.716
Length 0f Roads	0.101820862	0.043495015	2.34	0.079
Commercial Vehicles	0.04094856	0.017121018	2.39	0.075

$Y = b_0 + b_1 x_1 + b_2 x_2$

And again after simplifications, the excel output will give you the typical structures and again. So, once you go to the excel output, the excel output will give you the so, these are all actually intercept and this is the X 1 indications and. So, this is actually intercept and this is X 1 indications and this is X 2 indications so; that means, technically our model will be Y equal to beta 0 sorry, b 0 b 1 X 1 plus b 2 X 2 and the b 0 is this much b 1, is this much and b 2 is a this much. So, accordingly you have to just fit the kind of model and then you go for the kind of predictions.

So, now, before you go for predictions, two minimum, check you are supposed to do, whether it is a bivariate or trivariate means; this is the problem, we are dealing with a trvariate, where we have a two independent variables and one dependent variable and even if you, for multivariate structure. So, the rule is almost all same.

So, the first requirement or first check, before the prediction subject to the regression outputs; so, that you to see, whether actually the variables are statistically significant or not and then the overall fitness of the model should be also statistically reliable and also statistically significant that is actually a just through R square, adjusted R square and the f statistics so.

So, technically you will go for the right predictions and the kind of right management decisions provided all the parameters, will be statistically significant and R square will be high adjusted R square will be high and f will be also statistically significant. So, if all these are in the right track then; obviously, there will be perfect predictions and then the management decision will be very perfect, but in reality you will find, when you will be dealing with this kind of problem, with respect to a particular sample. So, you may not a get the results as per the particular expectation or the particular requirement.

So, now, if you are now, getting the result as per the expectation then there is no more complexity in the prediction process so; that means, the predictive analytics rule is a minimal there is, but when you will find some kind of complexity. For instance, if all the variables are not statistically significant or few variables are statistically significant then R square is not actually coming in to a right kind of percentage.

For instance, it is coming close to 0.01 or 0.02 like this. So, in this kind of situations; so, the signal is that the predictions cannot be actually perfect one; that means the accuracy of production; I mean prediction is a question mark. So, what we are supposed to do that, so, we have to re estimate the process; that means, the signal may means the wrong signal may be with respect to functional form or with respect to the kind of sample size or the identification of variables.

So, there may be many other issues, once you get first end output, for these particular problems, for the prediction and the kind of management decision. So, it will give you the kind of snapshot or then you will try to find out the right choice and the right combination and then once you fix choice and right combination that will be good enough to predict and then to go for the kind of management decision, but it is easy say, or easy to fix up, but in reality.

So, with the kind of dynamic world the dynamics of business, you will find lots of complexity, some variable will be significant, others will not and variables are significant

R square is not supporting, while R square is supporting, variables are not supporting. So, there are lots of issues are there.

So, now, if your theory is very good enough and the logical is very good enough and the kind of foundation or the linkage or the association is actually very perfect then obviously; so, the output regression output should suffered, but any kind of fault or any kind of issue will minimize.

The particular process and in that case, this will again challenge for the predictive analytics to find out the issues and then correct the particular procedures and you bring the kind of right estimated model, through which you can do the prediction and go for the kind of perfect management decisions. So, what are the issues and how you can deal with all these issue? We will discuss in the next lecture. So, we will stop here and.

Thank you very much have a nice day.