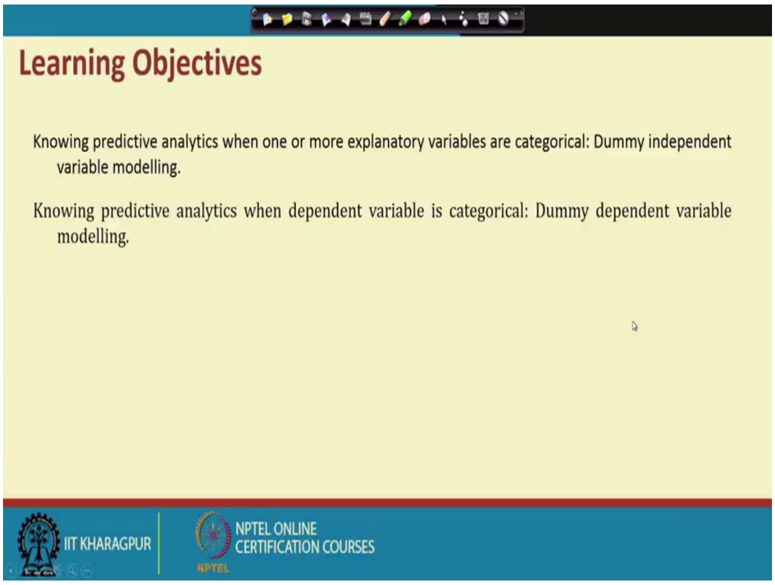
Business Analytics for Management Decision
Prof. Rudra P Pradhan
Vinod Gupta School of Management
Indian Institute of Technology, Kharagpur

Lecture – 31
Predictive Analytics : Dummy Modelling

Hello everybody and this is Rudra Pradhan here. Welcome you all to BMD lecture series. Today, we will continue with the predictive analytics, that too coverage is dummy modelling. So, this is a typical case, here either dependent variable or independent variable will be categorical. So; that means, we have already solved some of the problems, that too for bivariate structure and multivariate structure, where most of the times variables behavior is represented by quantitative way.

So; that means, we have actually numerical values that true for both dependent variable and independent variables, while doing some kind of prediction. But in the real life scenario, in the kind of day to day business environment, some of the variables are; there where we do the prediction and the variables naturally such that. So, the information related to that variables are not actually numeric. So, it is kind of categoricals or simply called as qualitative. So, as a result we have a two different structure. So, either dependent variable will be categorical or independent variable will be categorical or both will be categoricals.

(Refer Slide Time: 01:42)



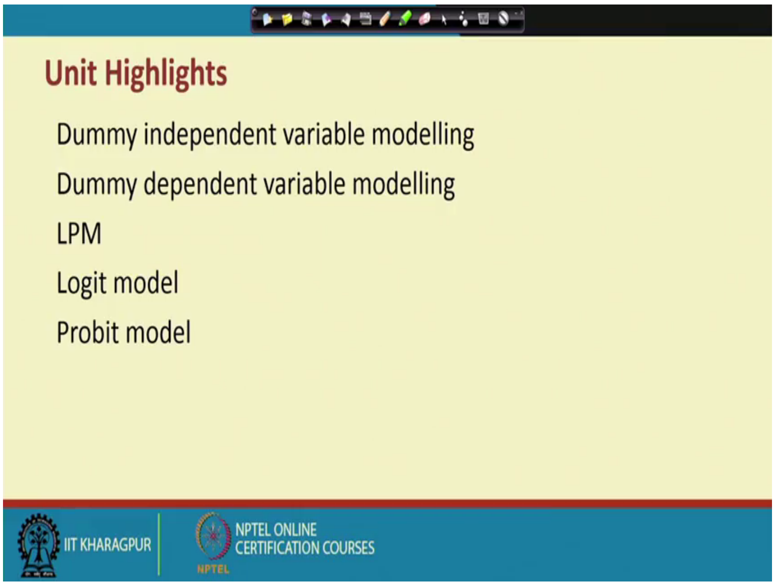
Learning Objectives

Knowing predictive analytics when one or more explanatory variables are categorical: Dummy independent variable modelling.

Knowing predictive analytics when dependent variable is categorical: Dummy dependent variable modelling.

So, as a result we have two different sets of problems. So, first problems knowing predictive analytics, when one or more explanatory variables are categorical that is actually called as a dummy independent variable modelling and again knowing predictive analytics, when dependent variable is categorical that is called as a dummy dependent variable modelling.

(Refer Slide Time: 02:07)



Unit Highlights

Dummy independent variable modelling

Dummy dependent variable modelling

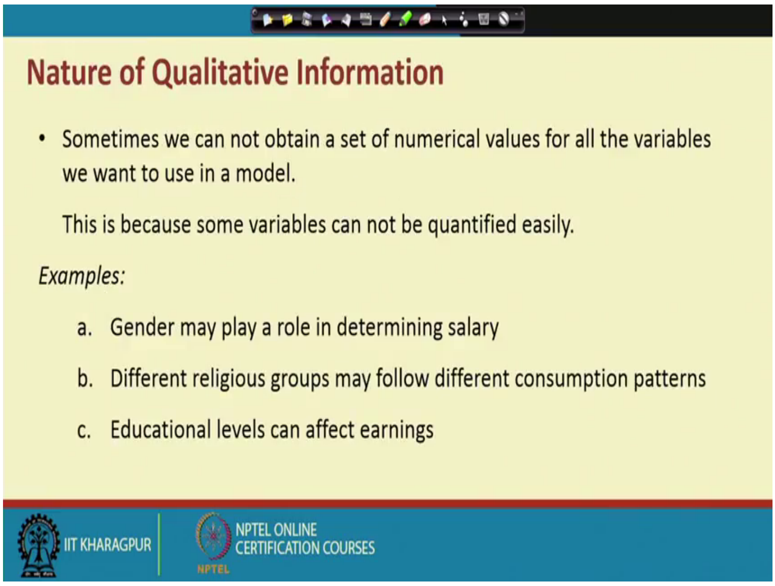
LPM

Logit model

Probit model

So, first we start with the dummy independent variable modelling. So, that is little bit, easy to understand and then we will go for dummy dependent variable modelling.

(Refer Slide Time: 02:17)



Nature of Qualitative Information

- Sometimes we can not obtain a set of numerical values for all the variables we want to use in a model.

  This is because some variables can not be quantified easily.

Examples:

a. Gender may play a role in determining salary

b. Different religious groups may follow different consumption patterns

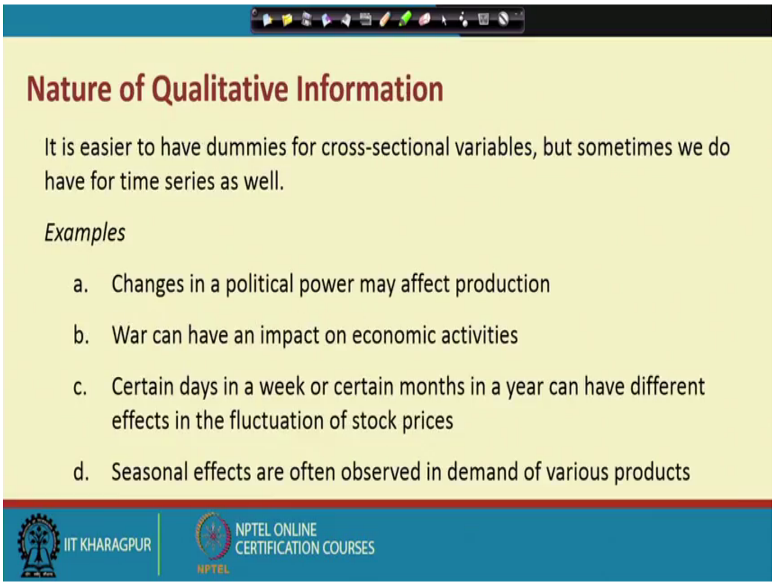
c. Educational levels can affect earnings

.

So, in the case of dummy independent variable modeling; so, first end requirement is to understand the dummy structure. So; that means, it is a variables, where the information is not actually in the form of numeric structures. So, it is in the form of some kind of, binary or some kind of categoricals or something like that.

So, we have a plenty of examples for that and like, gender is the classic examples, religious group is the another classic examples, educational levels can be also another classic examples, while studying the impact of independent variable or dependent variables. So, these variables may have actually different kind of impact.

(Refer Slide Time: 03:06)



Nature of Qualitative Information

It is easier to have dummies for cross-sectional variables, but sometimes we do have for time series as well.

Examples

a. Changes in a political power may affect production

b. War can have an impact on economic activities

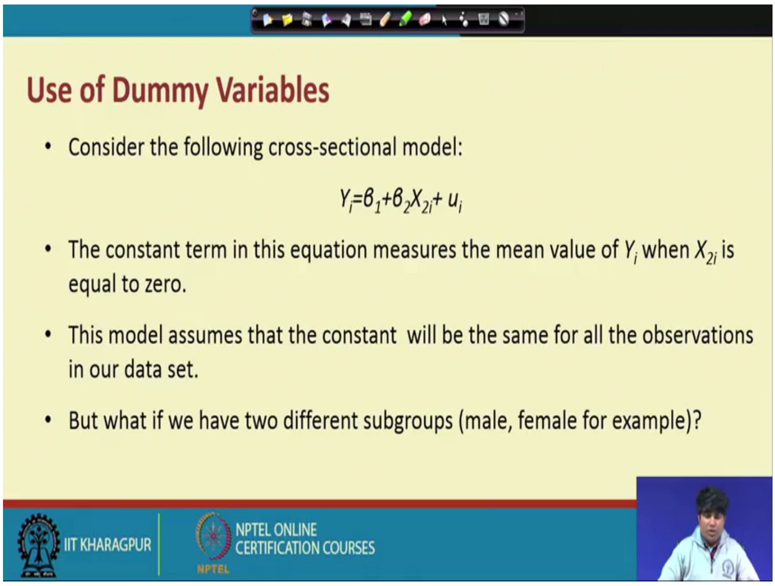
c. Certain days in a week or certain months in a year can have different effects in the fluctuation of stock prices

d. Seasonal effects are often observed in demand of various products

So, now with the help of, dummy modelling, we can address all these issues and by the way, dummy modelling can be applied to cross sectional modelling and it can be applied to time series modeling. In the cross sectional kind of environment we can find out like gender impact, then educational impact in a religion impact, but in the kind of time series, modelling dummy variable can be used to check the structural break to find out the aftershock effect, before shock effect, then , kind of , where effect before the situation, after the situation, sometime we like to predict the productions or growth or when we change the kind of political power 1.5, one point of time to another point of time.

Sometimes, we can study the seasonality, we can study the day of the week effect. So, many ways, actually we can use dummy modeling and that to in the kind of, predictive kind of environment.

(Refer Slide Time: 04:03)



Use of Dummy Variables

- Consider the following cross-sectional model:

$$Y_i = \beta_1 + \beta_2 X_{2i} + u_i$$

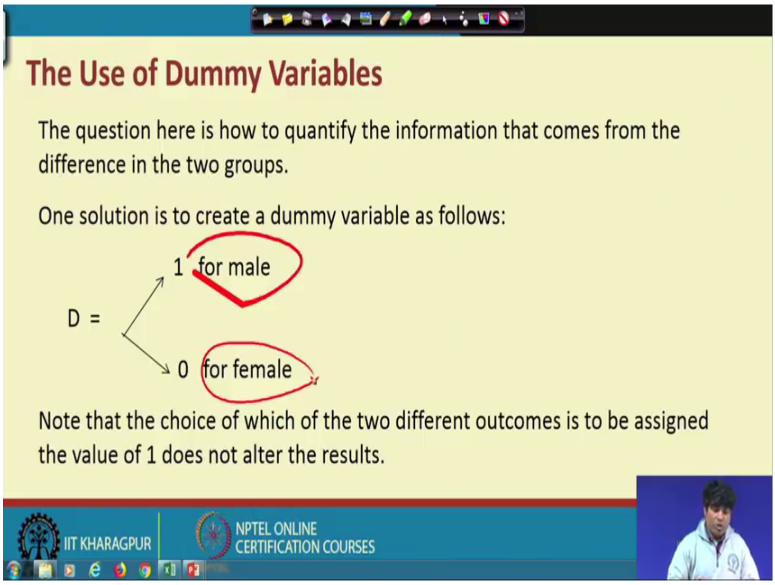
- The constant term in this equation measures the mean value of Y_i when X_{2i} is equal to zero.
- This model assumes that the constant will be the same for all the observations in our data set.
- But what if we have two different subgroups (male, female for example)?

.

Let us start, with the kind of structure, what is exactly the dummy concept and how we have to address this. So, we can start with a simple model here. So, Y with respect to X and here, the model is actually Y beta 1 and beta 2 X 2 and here the constant terms in this equation measures the mean value of Y a, when X 2, equal to 0. So; that means, the mod, this model assumes that the constant will be, this will be same for all the observation in our data set, but actually in real life scenario. So, there is a high chance, that they have the impact actually of two different subgroups. For instance, we have actually 10000 data points and some data points are with respect to male and some data points with respect to female.

So, there is a high chance that the male female impact will be, you are having, different while connecting Y and X. So, dummy variable can be used to find out the typical difference between male impact and the female impact. So, let us see how? How is this particular structure and then we will be moving to this, discussions.

(Refer Slide Time: 05:30)



The Use of Dummy Variables

The question here is how to quantify the information that comes from the difference in the two groups.

One solution is to create a dummy variable as follows:

D = 1 for male

D = 0 for female

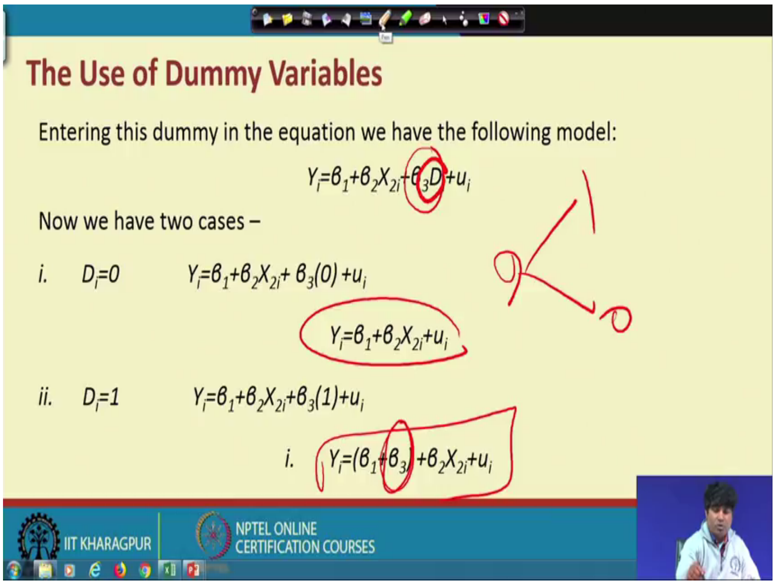
Note that the choice of which of the two different outcomes is to be assigned the value of 1 does not alter the results.

So, in the kind of dummy modelling. So, we usually use the concept called as D and in the case of gender impact. So, we put actually D equal to dummy variable, which is equal to 1. for one indication and let us say male and when it is not actually male then it will be female, then that will be 0 indication.

So; that means, this is the case, we can have here binary representation, but sometimes depending upon a particular, variable. So, it can have a different cluster altogether. For instance, in the religions, it may have actually different groupings, but still dummy variable can be used to find out the particular impact.

So, now in order to better understand you can actually, let us move to the particular structure. So, this is what actually original, our original model is this one. So, Y equal to beta 1 beta 2 X 2 and the error term and then again with a dummy in part; that means, we try to establish whether gender is having, linked between Y and X.

(Refer Slide Time: 06:40)



The Use of Dummy Variables

Entering this dummy in the equation we have the following model:

$$Y_i=\beta_1+\beta_2X_{2i}+\beta_3D+u_i$$

Now we have two cases –

i. $D_i=0$ $\quad Y_i=\beta_1+\beta_2X_{2i}+\beta_3(0)+u_i$

$$\boxed{Y_i=\beta_1+\beta_2X_{2i}+u_i}$$

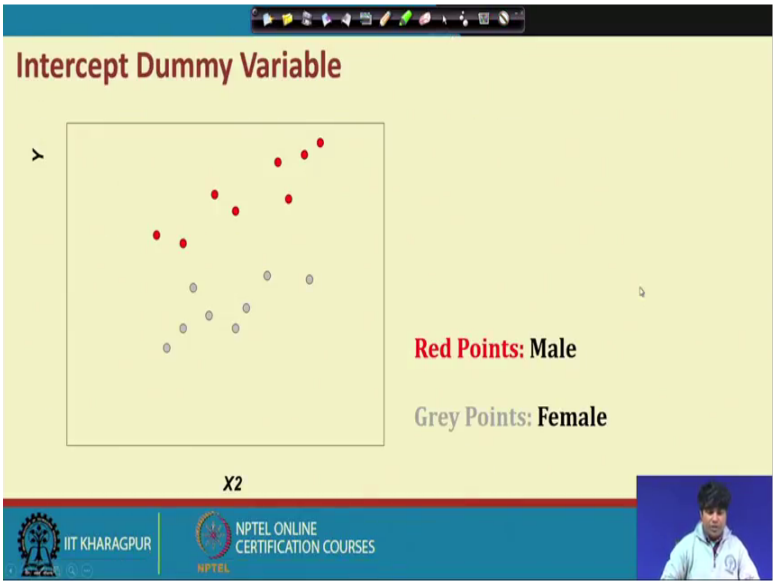
ii. $D_i=1$ $\quad Y_i=\beta_1+\beta_2X_{2i}+\beta_3(1)+u_i$

i. $\boxed{Y_i=(\beta_1+\beta_3)+\beta_2X_{2i}+u_i}$

So, as a result we allow the dummy modelling here, and in this case. So, D is the kind of dummy structure and here in the D, is the kind of structure. So, it D can be equal to 1 and D can be equal to 0. So, now when D equal to 0, then the model will be restricted to this much only. So, in that case this particular impact will be vanished, and when D equal to 1, then the model will be moved to this ones; that means, beta 3. This is the extra component, will be added into the process. So; that means, it will give you two different structure altogether. So, the mean impact while studying in the impact of X on Y.

So, in order to understand much better. So, let us, go to this particular plotting and you will see here.

(Refer Slide Time: 07:30)



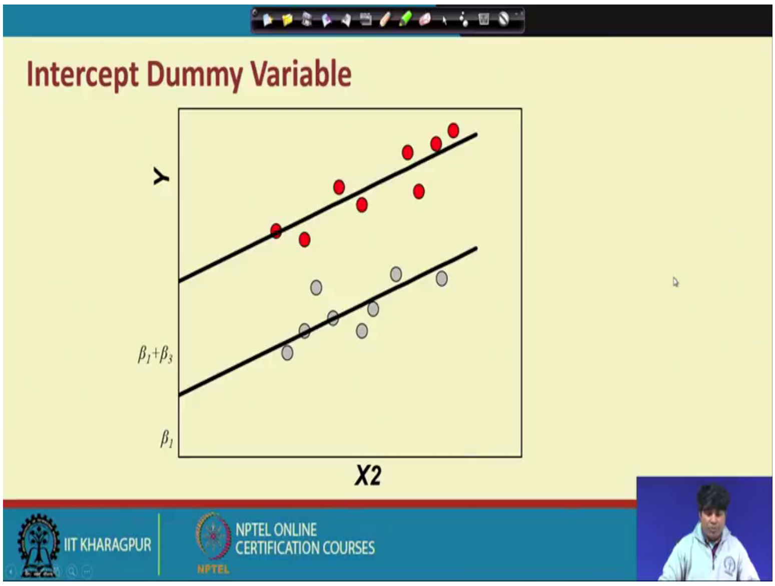
Intercept Dummy Variable

Red Points: Male

Grey Points: Female

So, the red plotting will give you the male signal and the gray gives the kind of female clustering. So; that means, by default dummy will give you two different cluster all together, while studying the relationship between Y and X.

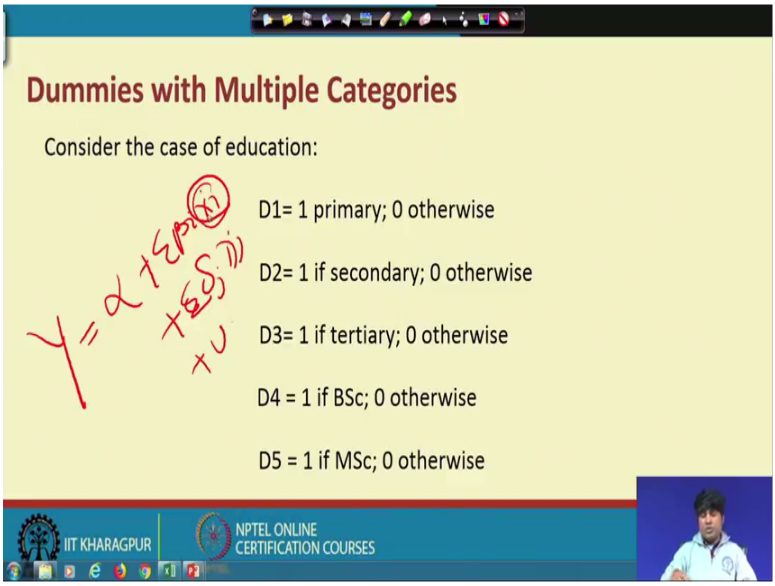
(Refer Slide Time: 07:49)



Intercept Dummy Variable

Now,. so, corresponding to the male groups. So, we can have a kind of predicted lines and against with female groups. So, we can have actually predicted line. So, the difference between the two predicted lines will be, with respect to beta 3, that is the dummy impact right, which we have already highlighted from this particular equation, this is a beta 3 component. So, the beta 3 is the kind of , this is the difference between this beta 3 component. So, this is the beta 1 component and then. So, the beta 3 is added and this should give you the another kind of indication.

(Refer Slide Time: 08:28)



Dummies with Multiple Categories

Consider the case of education:

$Y = \alpha + \Sigma\beta_i X_i + \Sigma\delta_j D_j + u$

D1= 1 primary; 0 otherwise

D2= 1 if secondary; 0 otherwise

D3= 1 if tertiary; 0 otherwise

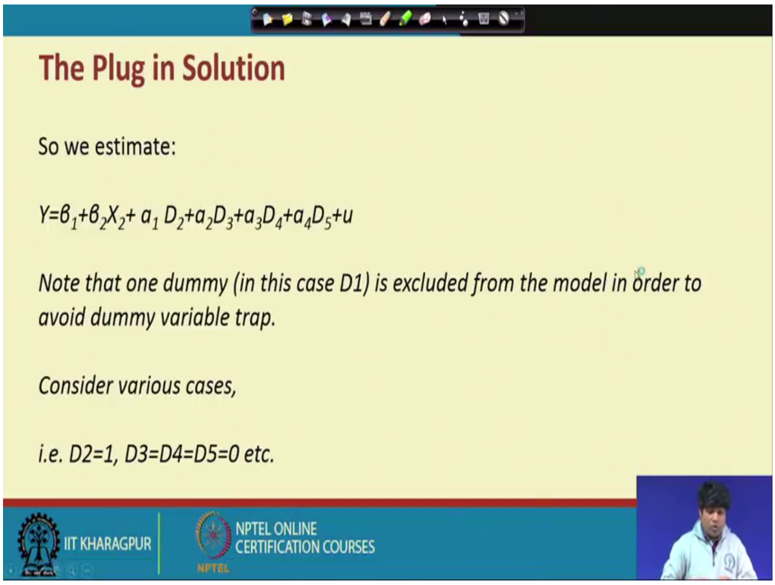
D4 = 1 if BSc; 0 otherwise

D5 = 1 if MSc; 0 otherwise

So, that is the beauty of the particular dummy structure and dummy can be used for a single variables and it can be used for multiple categories. So, in a particular models. So, this kind of situation, we have a 5 different dummies. So; that means, technically. So, usually the model can be written like this Y equal to alpha plus summation beta i X i and plus summation summation delta j Dj plus error terms right. So, this is the general form of dummy modelling and here X X is a in this is X, is the independent variable clusters and which may have actually values in a kind of quantitative structure and Dj is a kind of variable, , which information will be in the form of some kind of categoricals right. So; that means, technically qualitative in nature.

So, this is the classic examples, in this case we have actually a 5 dummies and then we like to check how primary impact can affect a link. Again secondary impact, how secondary impact will be affecting the link like this . So, it is. So, many process, we have to, or go for the kind of investigation.

(Refer Slide Time: 09:49)



The Plug in Solution

So we estimate:

$Y=\beta_1+\beta_2X_2+ a_1 D_2+a_2D_3+a_3D_4+a_4D_5+u$

Note that one dummy (in this case D1) is excluded from the model in order to avoid dummy variable trap.

Consider various cases,

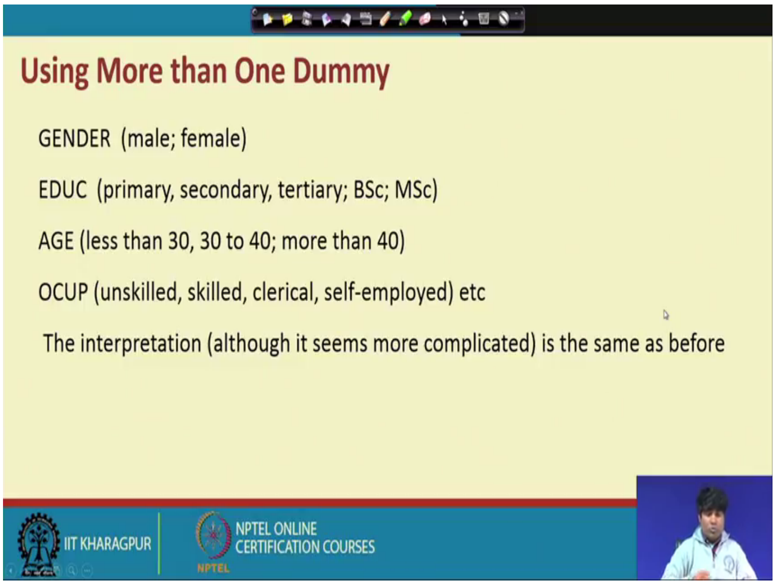
i.e. D2=1, D3=D4=D5=0 etc.

.

So, let us link them to a particular model and corresponding to the previous structure. So, the same structure here. So, we start with, this is actually original model, then this is the representation of single dummy and this is the case of multiple dummy. So, we have here multiple dummies, and in fact, we have a 5 dummies, but in the models, which we are, presenting here. So, it is having actually a for dummies. So; that means, technically when all dummies will be 0, then . So, it will take care the first dummy by default. So, that is nothing, but actually beta 1 plus beta 2 X 2, when we go for the kind of estimations right.

So, this is how the particular structure through which, actually you can go for the understanding of dummy modelling and then we can actually move to further discussion.

(Refer Slide Time: 10:40)



Using More than One Dummy

GENDER (male; female)

EDUC (primary, secondary, tertiary; BSc; MSc)

AGE (less than 30, 30 to 40; more than 40)

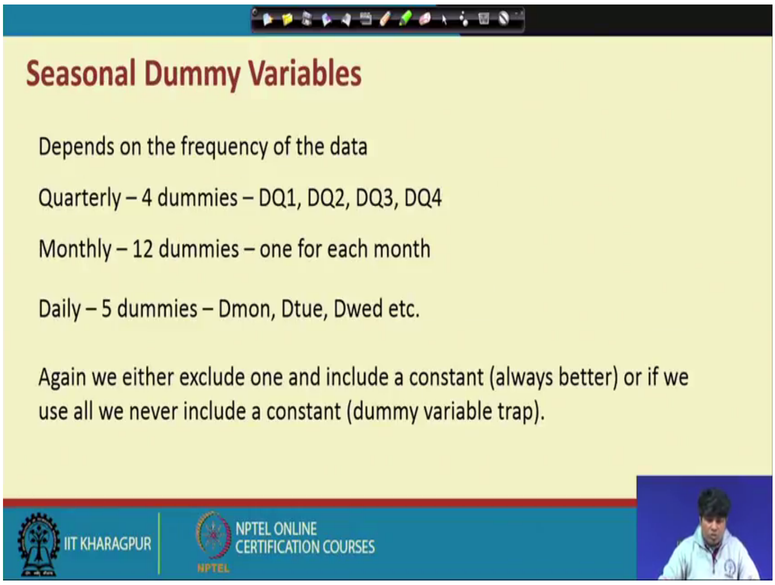
OCUP (unskilled, skilled, clerical, self-employed) etc

The interpretation (although it seems more complicated) is the same as before

.

So, the dummy can be with respect to gender with respect to education, with respect to age, with respect to occupations and then with respect to kind of religion. So, many things can be incorporated to study the kind of impact, Then against we can look for the structural break after certain percept. So, many things you can go for quarterly effect . So, there are many different ways, you can actually handle, this particular problem. This is very interesting technique, through which you can actually solve some of the problems, business problems and then it will come with very fantastic results as per the management requirement.

(Refer Slide Time: 11:25)



Seasonal Dummy Variables

Depends on the frequency of the data

Quarterly – 4 dummies – DQ1, DQ2, DQ3, DQ4

Monthly – 12 dummies – one for each month

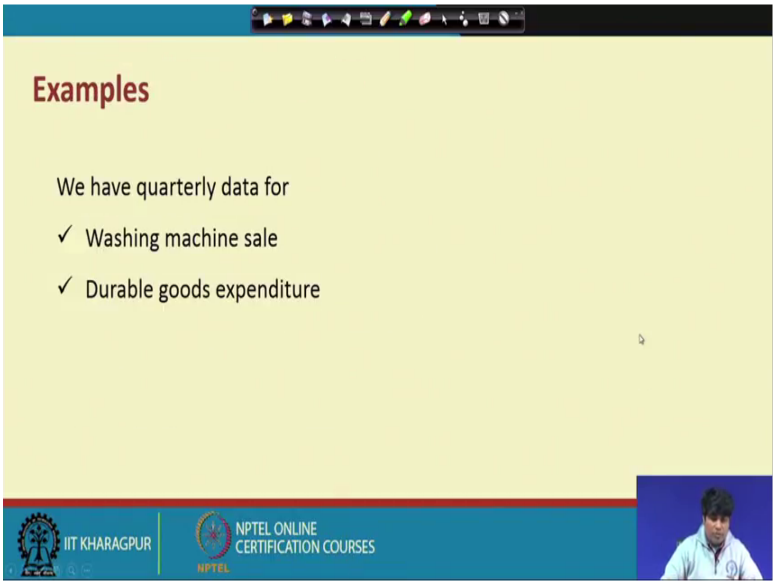
Daily – 5 dummies – Dmon, Dtue, Dwed etc.

Again we either exclude one and include a constant (always better) or if we use all we never include a constant (dummy variable trap).

.

So, in order to, go, something, in depth. So, in the time series contest, we can study the quarterly impact. We can go for monthly, for a monthly effect, we can go for weekly paid, we can go for a daily effect. So, there are many different instance. It is a very interesting actually so; that means, technically, we may have a 2 variable, but if you are, some data points are very high. So, ultimately you can actually with the help of dummy we can extend this model from bivariate to a multivariate. So, by adding one after another dummy depending upon the particular requirement and the kind of particular objective.

(Refer Slide Time: 12:04)



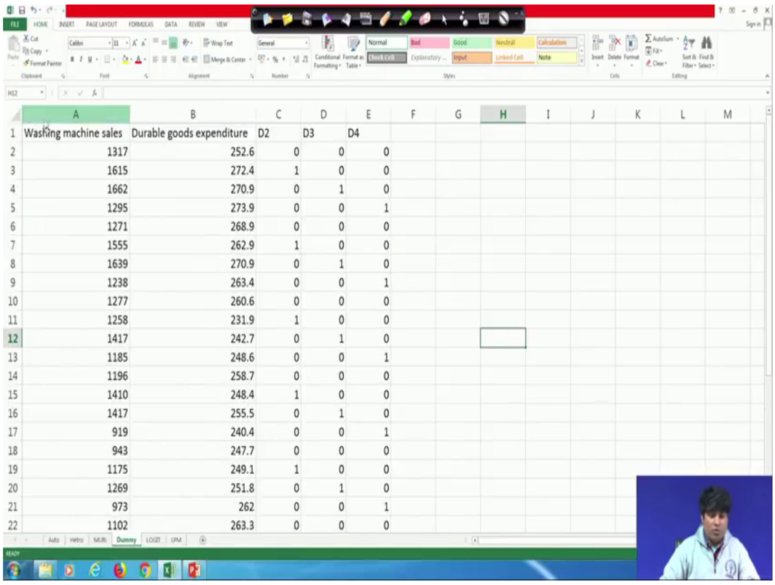
Examples

We have quarterly data for

- ✓ Washing machine sale
- ✓ Durable goods expenditure

So, now in order to better understand the particular structures; that means, technically, this is actually complete structure of multivariate structure appear on dummy modelling and we like to check. How it is actually happening in the real life scenario? When we will be dealing with data and the kind of problem. So, in order to better understand. So, I will take you to a spreadsheet, where we have a 2 variable washing machines. Sales of washing machines and durable goods expenditure with the intention that the selling of washing machine depends upon the durable goods expenditure. And it is actually quarterly data and we like to check whether there is actually the kind of what we can called as quarterly impact in between washing machine sales and the durable goods expenditure

(Refer Slide Time: 12:54)



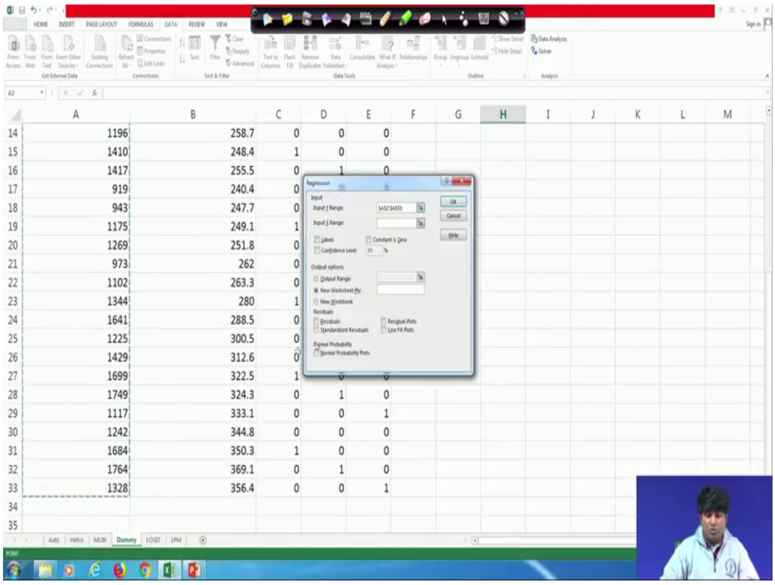
	A	B	C	D	E
1	Washing machine sales	Durable goods expenditure	D2	D3	D4
2	1317	252.6	0	0	0
3	1615	272.4	1	0	0
4	1662	270.9	0	1	0
5	1295	273.9	0	0	1
6	1271	268.9	0	0	0
7	1555	262.9	1	0	0
8	1639	270.9	0	1	0
9	1238	263.4	0	0	1
10	1277	260.6	0	0	0
11	1258	231.9	1	0	0
12	1417	242.7	0	1	0
13	1185	248.6	0	0	1
14	1196	258.7	0	0	0
15	1410	248.4	1	0	0
16	1417	255.5	0	1	0
17	919	240.4	0	0	1
18	943	247.7	0	0	0
19	1175	249.1	1	0	0
20	1269	251.8	0	1	0
21	973	262	0	0	1
22	1102	263.3	0	0	0

.

So; that means, actually the model is you with respect to Y beta 1 beta 2 o X ones then delta 1 D 1 delta 2 D 2 delta 2 D 3 delta for divorce, but here, because of a dummy variable trap. So, we are using only 3 dummy so; that means, technically, when D 2 D 3 D 4 will be 0 by default that we will take care the impact of the first quarter and again. So, when D 2 equal to 1 and others will be 0. So, this will take care the impact of second quarter and again. So, when D 3 equal to 1 and others are 0.

So, this will take you to the impact of third quarter. And finally, when D 4 equal to 1, this will give you the impact of fourth quarters. So, as a result. So, what will you do in this particular process? Same way, you go to the data analysis package and again the same standard technique which you like to apply, here is a regression analysis and you choose the regression and this is as usual simple regression modelling and that to multiple regression modelling.

(Refer Slide Time: 14:09)



	A	B	C	D	E
14	1196	258.7	0	0	0
15	1410	248.4	1	0	0
16	1417	255.5	0	1	0
17	919	240.4	0		
18	943	247.7	0		
19	1175	249.1	1		
20	1269	251.8	0		
21	973	262	0		
22	1102	263.3	0		
23	1344	280	1		
24	1641	288.5	0		
25	1225	300.5	0		
26	1429	312.6	0		
27	1699	322.5	1	0	0
28	1749	324.3	0	1	0
29	1117	333.1	0	0	1
30	1242	344.8	0	0	0
31	1684	350.3	1	0	0
32	1764	369.1	0	1	0
33	1328	356.4	0	0	1
34					
35					

Regression

Input

Input Y Range: $A$2:$A$33

Input X Range:

☐ Labels ☐ Constant is Zero

☐ Confidence Level 95 %

Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

Residuals

☐ Residuals ☐ Residual Plots

☐ Standardized Residuals ☐ Line Fit Plots

Normal Probability

☐ Normal Probability Plots

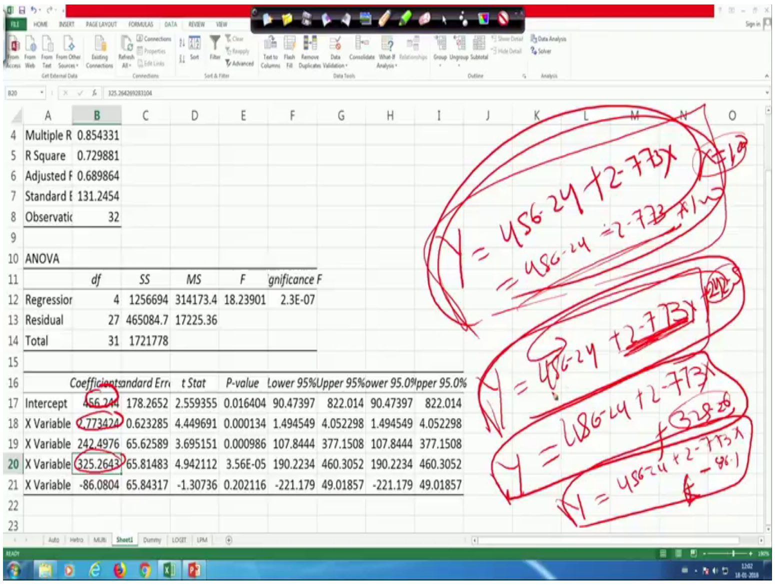
.

So, the first end requirement is here to give the indication about the data points that to with respect to dependent variables and then again. So, we need to indicate the structure of in independent variables and that true with structure.

So, now after highlighting this ones, then software is ready to operate and you see here the kind of the representation of D 2 D 3 D 4. It is actually initially by default, it will not be there, you have to create it. So, the creation is like this, this is by default, you see here second quarter data points and this is actually third data point third quarter data point, then this is fourth quarter data point.

So, similarly you have to operate, this is first and this is again second, this is third and this is fourth. So, by default actually, it is actually artificially created with the structure of this particular data in the game between washing machine sales and durables durable goods expenditure.

(Refer Slide Time: 15:19)



Multiple R	0.854331
R Square	0.729881
Adjusted F	0.689864
Standard E	131.2454
Observatic	32

ANOVA

	df	SS	MS	F	gnificance F
Regressior	4	1256694	314173.4	18.23901	2.3E-07
Residual	27	465084.7	17225.36		
Total	31	1721778			

	Coefficients	andard Err	t Stat	P-value	Lower 95%	Upper 95%	ower 95.0%	pper 95.0%
Intercept	456.244	178.2652	2.559355	0.016404	90.47397	822.014	90.47397	822.014
X Variable	2.773424	0.623285	4.449691	0.000134	1.494549	4.052298	1.494549	4.052298
X Variable	242.4976	65.62589	3.695151	0.000986	107.8444	377.1508	107.8444	377.1508
X Variable	325.2643	65.81483	4.942112	3.56E-05	190.2234	460.3052	190.2234	460.3052
X Variable	-86.0804	65.84317	-1.30736	0.202116	-221.179	49.01857	-221.179	49.01857

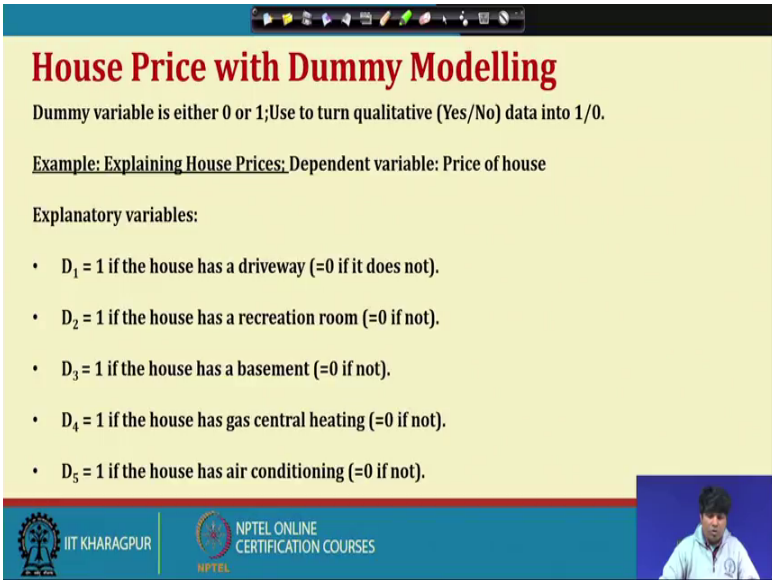
So, after arranging put on, and then you will find a regression output here, and this is what actually the intercept and this is by default, actually variable durable goods expenditure that is X 1 and this is D 2 D 3 D 4.

So; that means, technically. So, the model is now interesting. So, we like to know what is the impress of, all are actually positively linked except the last variable, which is having negative impact and probably that particular quarter, this the particular link is not actually so positive, because of some kind of seasonality issue and. In fact, how you will be find the particular quarterly impact. So, this is actually the dependent variable is here, the dependent variable is here, washing machine sales and then the kind of output is with respect to durable goods expenditure, that is a, this 2.773. So; that means,. So, the first quarter impact will be intercept plus this one. So, 456.24 and plus 2.77 and this is actually the first, first quarter impact.

And where D D 2 D 3 and D 2 D 3 D 4 equal to 0. Now, for second quarter. So, the impact will be 4 4 56.24 plus 2.773 X X plus this particular impact 2 2 42.5. So, this 240.0 42.5 by default will be added into the intercept, because this is a constant number, when you put actually D equal to 1 then; that means, the particular item will be only left 2, adding in the case of interceptand again. So, when the third quarter is concerned. So, D 2 will be 0 and D 4 will be 0 and as a result. So; that means, technically I will, let me give you the kind of structure here. So, here is. So, this is actually the intercept. So, intercept plus X variable impact. So, this is the first quarter impact so; that means,. So, Y equal to 456.24 applause 2.773 X. This is the first quarter impact, when you put X, something like , let us say expenditure is actually say 1000 then why equal to 456.24 plus 2.773 into 1000. So, this will give you the quarterly impact right and that to first quarter impact and for the second quarter impact.So, it will be 456.24 last 2.773 X plus these 242.5 plus 242.5 so; that means, by default this will be added into the intercept and that is the extra impact will be happening in the second quarter and by default, this is the X variable impact and again putting the X value here, then we can get the quantitative pictures like the, like this case right, similarly, in the third quarters. So, it will be a 456.24 plus 2.7773 X plus this variable impact already. So, plus 325.26.

Similarly, in the fourth quarter case. So, it will be 456.24 plus 2.377 X plus it will be actually minus, because the coefficient is coming so; that means, it will be minus 86.1. So, that will be actually less impact so; that means, if you go by kind of comparative structures, which particular quarter is more effective then you can find here. So, 324 25.26 so; that means, the third quarter is having high impact followed by second quarter then third quarter as the sorry, first quarter and finally, it will be in the fourth quarter. So, this is how the impact can be studied to give, to go for the prediction of washing machine sales subject to durable goods expenditure. So, you as per the business requirement, to precise to , go a little bit more.

(Refer Slide Time: 19:56)



House Price with Dummy Modelling

Dummy variable is either 0 or 1;Use to turn qualitative (Yes/No) data into 1/0.

Example: Explaining House Prices; Dependent variable: Price of house

Explanatory variables:

- D_1 = 1 if the house has a driveway (=0 if it does not).
- D_2 = 1 if the house has a recreation room (=0 if not).
- D_3 = 1 if the house has a basement (=0 if not).
- D_4 = 1 if the house has gas central heating (=0 if not).
- D_5 = 1 if the house has air conditioning (=0 if not).

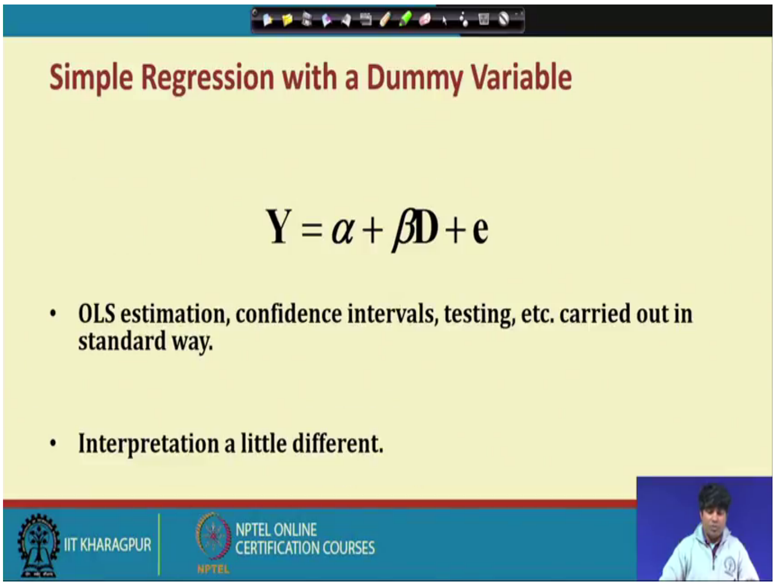
.

So, I will highlight another problem here. So, here the problem is with respect to explaining housing price, because this is a kind of real state problems where, we like to attract the customers by giving flexible price, but using the dummy modelling, you can actually predict the different price for different kind of features.

So, here is, our dependent variable is the price of house and then some of the independent variables and out up with some are qualitative in nature and some are numeric in nature and here is, we have a five different qualitative variables and these are the futures through which price house of, house price will be actually increasing so; that means, technically. So, this is the first dummy variables, it is the facility of driveway so; that means, technically. So, with a best price, if you need a driveway then that dummy will be represented by 1 and then it will add that particular impact, if you do not need this feature, then by default this D 1 impact will be reduced.

So, as a result housing price will be low against a like, D 2 or D 3. So, it is here in the case of basement. So, if you need the basement facility, then D 3 impact will go to the best price, if not then this the price will be again lower. Similarly, if you need central heating, then D 4 impact will go to the housing price and if you again need air conditioning facilities, then this will go to the impact of, in house that means technically, we have actually plenty of options to , go for the kind of housing price predictions.

(Refer Slide Time: 21:39)



Simple Regression with a Dummy Variable

$$Y = \alpha + \beta D + e$$

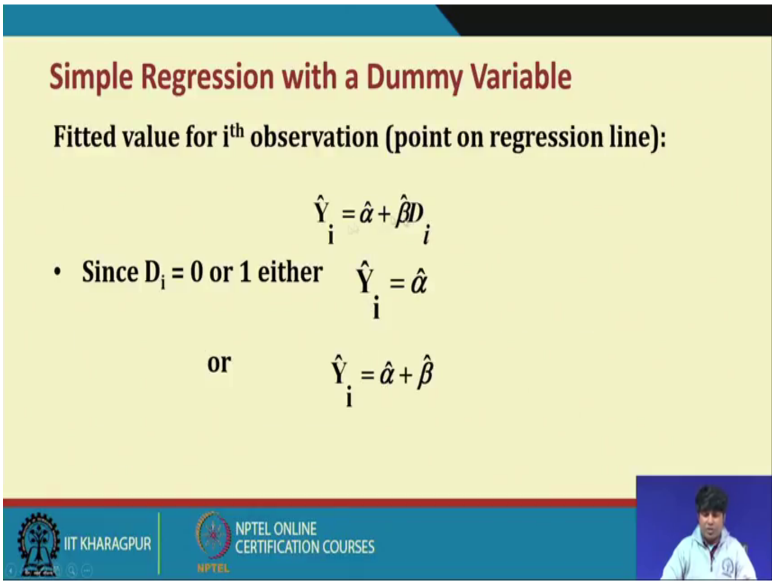
- OLS estimation, confidence intervals, testing, etc. carried out in standard way.
- Interpretation a little different.

.

So, now in order to understand the particular user, in every case actually it is the situation of yes no so; that means, this facility is there, if a solution is not there. See, if it is there then what should be the housing price, if not then what should be the housing price so; that means, it will be having actually different kind of options. So, we try to find out what are these options altogether.

So, for modelling is concerned, we start with a 1 dummy and in this case. So, Y equal to alpha plus beta D plus C. So, this is standard model and here Y is the housing price and D is the dummy modelling and that D is a dummy variables which represents a particular future.

(Refer Slide Time: 22:32)



Simple Regression with a Dummy Variable

Fitted value for i^th observation (point on regression line):

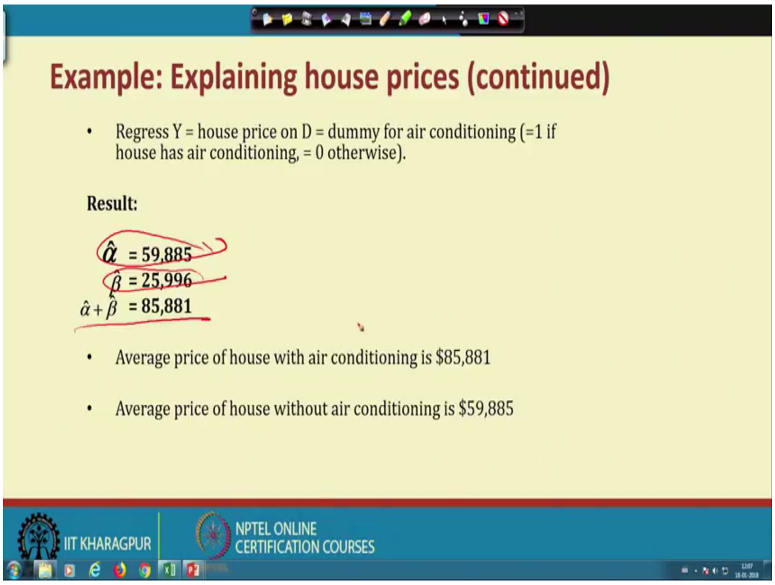
$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} D_i$$

- Since D_i = 0 or 1 either $\hat{Y}_i = \hat{\alpha}$

or $\hat{Y}_i = \hat{\alpha} + \hat{\beta}$

So, then as usual you can estimate the process, which we have already done through the excel spreadsheet then finally, you have a model equal to Y equal to alpha hat plus beta hat D only. So, you will be removed in the process as a result. So, this is your estimated model. So, this is your estimated model and by default. So, what will you do. So, you can actually predict the particular situations Y hat equal to alpha hat and where the facility is not required and the price will be again Y hat equal to alpha hat plus beta hat, where the price is actually needed so; that means, by default. So, there will be two different price for a particular dummy.

(Refer Slide Time: 23:10)



Example: Explaining house prices (continued)

- Regress Y = house price on D = dummy for air conditioning (=1 if house has air conditioning, = 0 otherwise).

Result:

$\hat{\alpha}$ = 59,885

$\hat{\beta}$ = 25,996

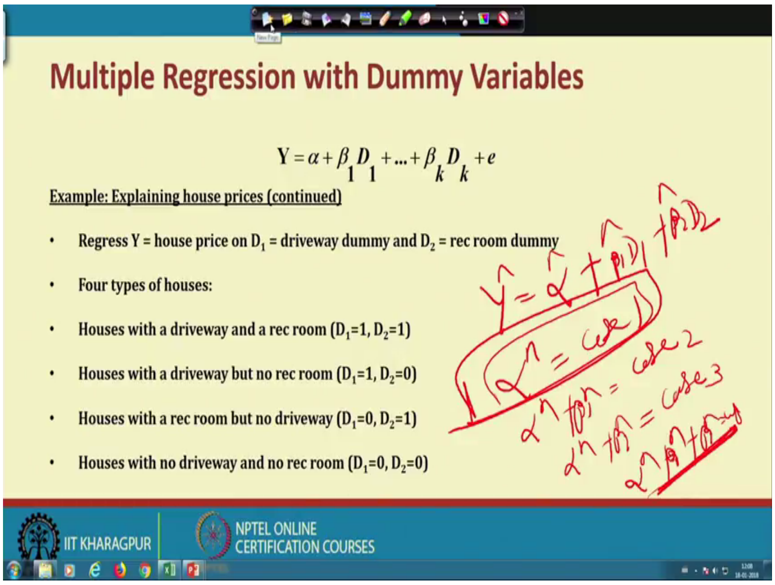
$\hat{\alpha}+\hat{\beta}$ = 85,881

- Average price of house with air conditioning is $85,881
- Average price of house without air conditioning is $59,885

So, again to adjust the particular process to adjust the particular process. Yes, to adjust this particular process, let us say here. So, this is what actually two different price. So, now, let us assume that alpha hat by estimation is coming 59,885 and beta hat is coming 25,996 so; that means, you have now two different price. So, in one case, if you need this particular, this is actually with respect to air conditioning. So, what is the house pricing with air conditioning. So, then it will be having actually alpha plus beta pack and if you do not need actually air conditioning so; that means, the beta hat impact will be removed. So, only alpha hat impact will be there, that is the best price.

So; that means, say. So, having air conditioning facility. So, housing price will be best price, that is 59,885 plus the kind of air conditioning charge, that is reflected by beta coefficient, that is 25,996. So, as a resultthis will have a two different price through which you can predict the particular situations.

(Refer Slide Time: 24:35)



Multiple Regression with Dummy Variables

$$Y = \alpha + \beta_1 D_1 + \ldots + \beta_k D_k + e$$

Example: Explaining house prices (continued)

- Regress Y = house price on D_1 = driveway dummy and D_2 = rec room dummy
- Four types of houses:
- Houses with a driveway and a rec room (D_1=1, D_2=1)
- Houses with a driveway but no rec room (D_1=1, D_2=0)
- Houses with a rec room but no driveway (D_1=0, D_2=1)
- Houses with no driveway and no rec room (D_1=0, D_2=0)

$\hat{Y} = \hat{\alpha} + \hat{\beta}_1 D_1 + \hat{\beta}_2 D_2$

$\hat{\alpha} = \text{case}$

$\hat{\alpha} + \hat{\beta}_1 = \text{case } 2$

$\hat{\alpha} + \hat{\beta}_2 = \text{case } 3$

$\hat{\alpha} + \hat{\beta}_1 + \hat{\beta}_2 = 4$

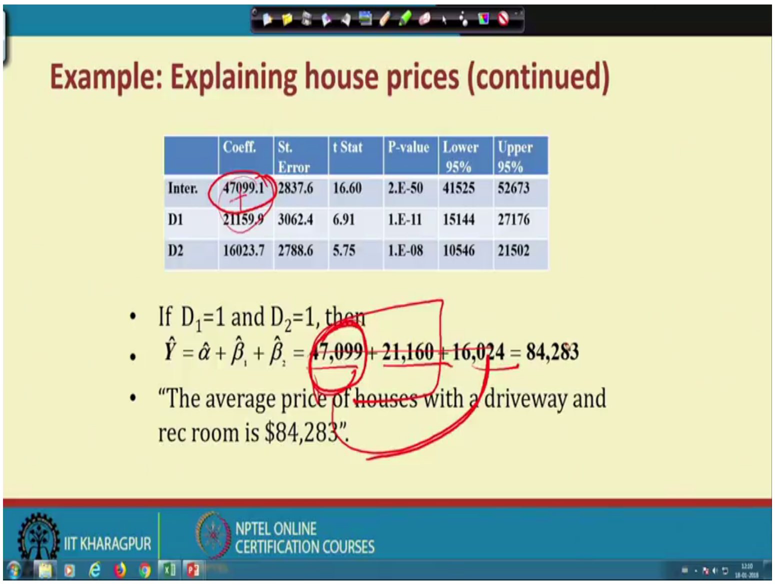
IIT KHARAGPUR

NPTEL ONLINE CERTIFICATION COURSES

So, now again. So, you can go with multiple dummies and again for simplicity. We are allowing 2 D 2 dummies all together so; that means, how many a housing price, we can actually predict here, when we have two different features all together, so; that means, technically here, the model will be Y equal to alpha, that will be represent the best price and then beta 1 D 1 and beta 2 D 2. So, let us assume that this is the estimated model and then by default error term will be removed then finally, if you do not need driveway facilities and let us say the D 2 facilities.

So, we have a four different price options, in one case you have only alpha hats that is the best price and this is the case 1 and then alpha hat plus beta 1 hat. So, this is case 2 and then alpha hat plus beta 2 hat. This is for case 3 and then finally, alpha hat plus beta 1 hat beta 1 hat plus beta 2 hat is the case for. So, this will be a substantially high price, because it, it includes two different features and this is actually having low price, because it is, it does not includes any features and the rest of the two cases, one feature is there and then what should be the kind of housing price. So, likewise having two different by these are the various options. So, D 1 1 D 1 o 1 and D 1 0 D D D 2 0 and then D 1 yes D 2 no again D D 2 S D D 1 no.

(Refer Slide Time: 26:28)



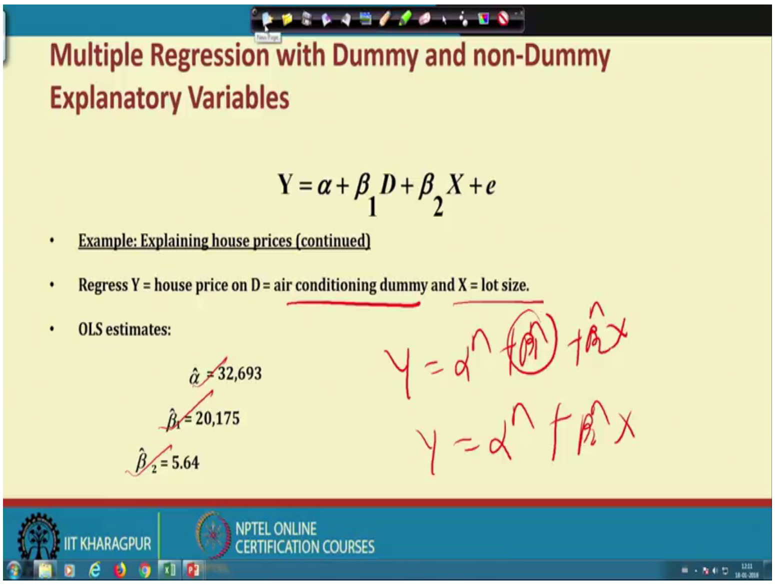
Example: Explaining house prices (continued)

	Coeff.	St. Error	t Stat	P-value	Lower 95%	Upper 95%
Inter.	47099.1	2837.6	16.60	2.E-50	41525	52673
D1	21159.9	3062.4	6.91	1.E-11	15144	27176
D2	16023.7	2788.6	5.75	1.E-08	10546	21502

- If $D_1=1$ and $D_2=1$, then
- $\hat{Y} = \hat{\alpha} + \hat{\beta}_1 + \hat{\beta}_2 = 47{,}099 + 21{,}160 + 16{,}024 = 84{,}283$
- "The average price of houses with a driveway and rec room is $84,283".

So, these are the kind of , for different option and accordingly you will find, different price pack . So, now, the estimated results are here and accordingly. So, your first price pack will be, this is the best price 4799 and against having one particular features, let us say D 1. So, then the house price will be this much against having second feature only, then 4799 a plus 161 60 1623. So, this will be the second price means third price, here connecting to alpha and alpha beta 1 and now fourth option will be. So, alpha 1 plus beta 1 plus beta 2 like this. So, this is actually the high price pack and this is the low price pack, because this is the best price. Now, this is another options against this, with this is another option and finally, all these 3 only have a 1 of sums. So; that means, having two different domains. So, we have actually four different flexible price and suffer as a prediction is concerned so; that means, this is a very attractive kind of structure, through which you can do the predictions and you come with different options for the kind of management requirement or the kind of business requirement against I am showing another kind of situation where 1 dummy and 1 numeric variables right.

(Refer Slide Time: 27:51)



Multiple Regression with Dummy and non-Dummy Explanatory Variables

$$Y = \alpha + \beta_1 D + \beta_2 X + e$$

- Example: Explaining house prices (continued)
- Regress Y = house price on D = air conditioning dummy and X = lot size.
- OLS estimates:

$\hat{\alpha} = 32,693$

$\hat{\beta}_1 = 20,175$

$\hat{\beta}_2 = 5.64$

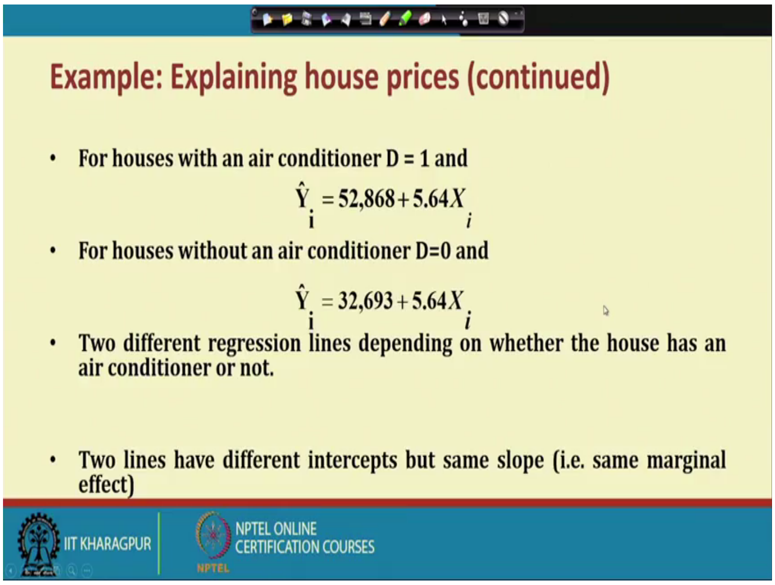
$Y = \hat{\alpha} + \boxed{\hat{\beta}_1} + \hat{\beta}_2 x$

$Y = \hat{\alpha} + \hat{\beta}_2 X$

So, D is actually having a air conditioning facility and that is actually yes no type of situation and then another kind of situation. So, this is air conditioning facilities with dummy yes no and then X is a lot size. So, if you go for a lot booking then; obviously. So, the price pack will be different.

So, now accordingly in the estimation process, assume that alpha hat is coming this much, beta hat is coming this much and beta 2 is coming this much so; that means, technically now, with 2 D to dom. with the 2 variables and 1 dummy in that 1 dummy. So, your price pack will be Y equal to alpha hat plus beta 1 hat and beta 2 X. So, this is one pack, another pack will be Y alpha hat plus beta 2 hat X. So, beta 1 will be removed, because when we have no such facility, then that price pack will be lower compared to the previous one. So, this is, this way you can actually find out two different price of zones as per the particular requirement.

(Refer Slide Time: 29:12)



Example: Explaining house prices (continued)

- For houses with an air conditioner D = 1 and

$$\hat{Y}_i = 52,868 + 5.64X_i$$

- For houses without an air conditioner D=0 and

$$\hat{Y}_i = 32,693 + 5.64X_i$$

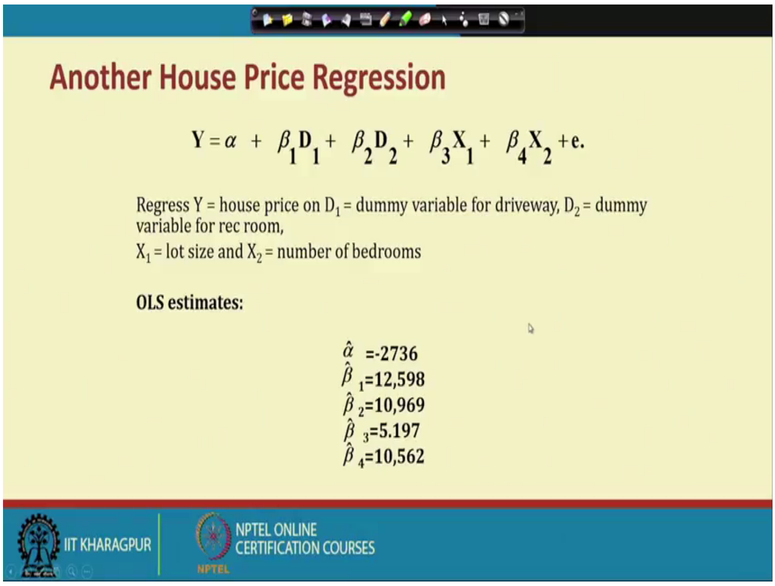
- Two different regression lines depending on whether the house has an air conditioner or not.
- Two lines have different intercepts but same slope (i.e. same marginal effect)

IIT KHARAGPUR

NPTEL ONLINE CERTIFICATION COURSES

And again, so let us assume that this is the estimated model and corresponding D 0. So, the model will be restricted to only intercept and the kind of X impact and if having that facility, then the D 1 impact will be also added into the particular process.

(Refer Slide Time: 29:31)



Another House Price Regression

$$Y = \alpha + \beta_1 D_1 + \beta_2 D_2 + \beta_3 X_1 + \beta_4 X_2 + e.$$

Regress Y = house price on D_1 = dummy variable for driveway, D_2 = dummy variable for rec room,
X_1 = lot size and X_2 = number of bedrooms

OLS estimates:

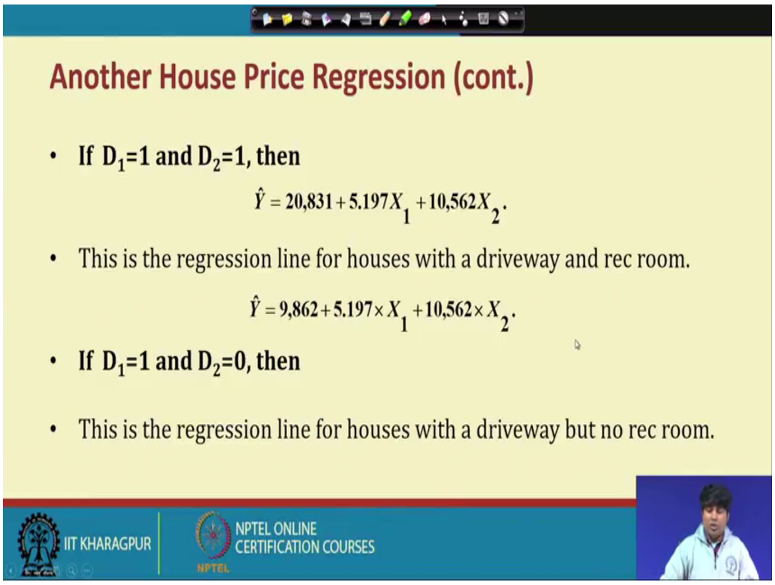
$\hat{\alpha}$ =-2736
$\hat{\beta}_1$=12,598
$\hat{\beta}_2$=10,969
$\hat{\beta}_3$=5.197
$\hat{\beta}_4$=10,562

IIT KHARAGPUR

NPTEL ONLINE CERTIFICATION COURSES

And this is again with the 2 dummies, with the 2 numeric variables. So, earlier case, we have taken X 1 is the lot size and here number of bedrooms and again. So, two different dummy structure and D 1 and D 2. So; that means, again. So, with two dummies, we have actually four different price pack.

(Refer Slide Time: 29:59)



Another House Price Regression (cont.)

- If D_1=1 and D_2=1, then

$$\hat{Y} = 20{,}831 + 5.197X_1 + 10{,}562X_2.$$

- This is the regression line for houses with a driveway and rec room.

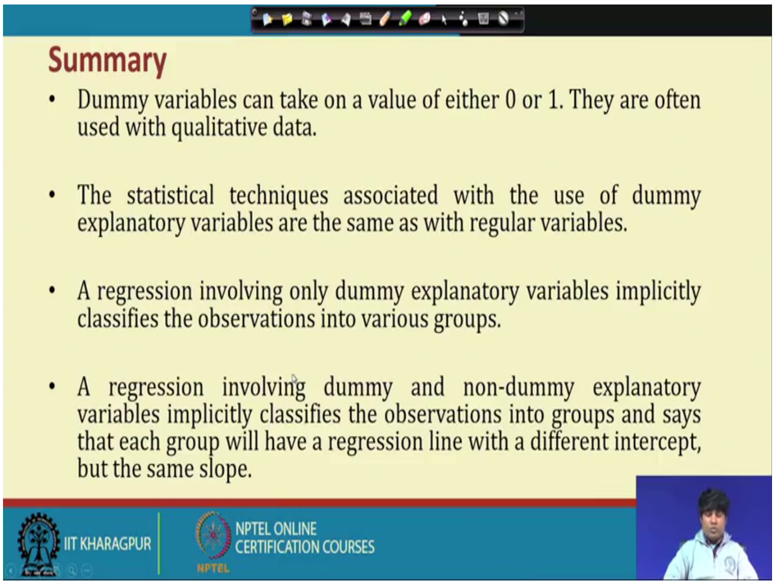
$$\hat{Y} = 9{,}862 + 5.197 \times X_1 + 10{,}562 \times X_2.$$

- If D_1=1 and D_2=0, then
- This is the regression line for houses with a driveway but no rec room.

Now, depending upon X 1 and X 2 value. So, you will be find different option altogether. So; that means, actually through dummy you will get plenty of option to go for the kind of predictions and the kind of alternatives. So, this is the case, where this is the maximum case where what the features are there and this is the second case, where one feature will be there.

And similarly, you can have actually D 2 equal to 1, D 1 equal to 0 and against you can have D 1 0 and D 2 0 simultaneously, in that case only intercept and X 1 and X 2 impact will be there the remaining D 1 impact into D 2 rema, impact will be removed in the kind of estimation process.

(Refer Slide Time: 30:37)



Summary

- Dummy variables can take on a value of either 0 or 1. They are often used with qualitative data.
- The statistical techniques associated with the use of dummy explanatory variables are the same as with regular variables.
- A regression involving only dummy explanatory variables implicitly classifies the observations into various groups.
- A regression involving dummy and non-dummy explanatory variables implicitly classifies the observations into groups and says that each group will have a regression line with a different intercept, but the same slope.

So; that means, technically whatever we have discussed here that . We are targeting the predictive analytic structures. We are predicting a kind of dependent variable, like in this context housing price with respect to independent variables, while doing all these things, see if the, if there is a chance of dummy structures, then it will give you lots of flexibility to get plenty of option or flexibility to attract the costumer and make the management decision more attractive.

So, it is a very interesting technique. So, the kind of requirement is to understand and as per the kind of need, we have to incur incorporate the dummy in the kind of predictive structure. So, having the knowledge of dummy, you can actually or having the kind of dummy modelling. So, you will be find plenty of options as per the kind of customer requirement and the kind of management requirement. So, with this we will stop you here.

Thank you very much have a nice day.