

**Marketing Research and Analysis -II (Application Oriented)**  
**Prof. Jogendra Kumar Nayak**  
**Department of Management Studies**  
**Indian Institute of Technology - Roorkee**

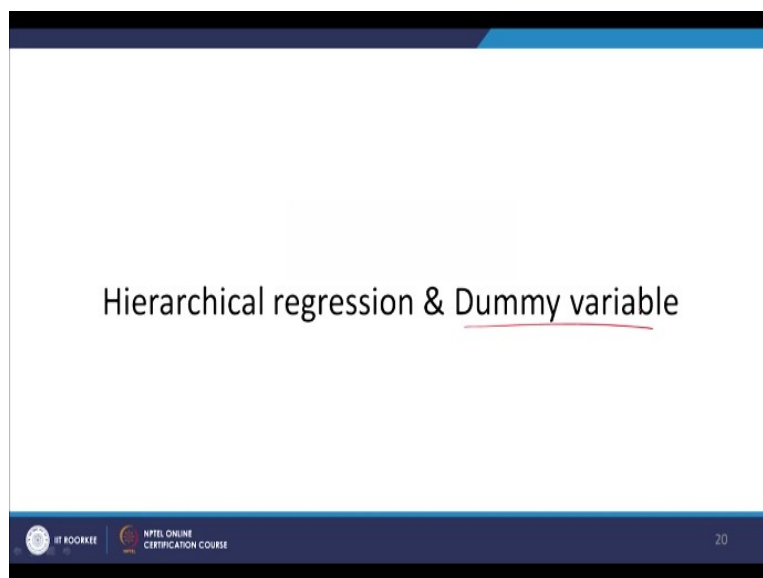
**Lecture - 48**  
**Hierarchical Regression and Dummy Variable Regression**

Welcome friends to the class of marketing research and analysis. Will continue from the last lecture where we were discussing about a special case of entering the data during a multiple regression case right. So we learnt about the effect of the order of entry of a data in case of a stepwise regression and I also explained about forward regression and backward regression right but then we realize the stepwise regression has its own problems.

That sometimes a very important variable might become unimportant and may be removed from the study which resulting in creating an overall weakness for the equation or the regression the other research study right. So to avoid that problem I said that we can use a new approach which is called the hierarchical regression right, so hierarchical regression technique is a technique which is basically used to enter the data into the regression equation in a block wise manner.

So that block manner is used on basis of is developed on basis of some theory or some logic right so that means the computer does not do it for you rather it is based on certain logic created by the researcher which is from some past studies or something right.

**(Refer Slide Time: 01:50)**



And today I will also explain you about a special case of regression called the dummy variable regression which is highly useful for researchers around who are doing some other way of research right.

**(Refer Slide Time: 02:03)**

The slide is titled "Hierarchical multiple regression process". It contains three bullet points with handwritten annotations in red:

- IVs are entered in steps (blocks)       $X_1, X_2$        $X_3, X_4$        $X_1, X_2, X_3, X_4$
- Each IV is assessed in terms of what it adds to the prediction of DV after the previous IVs have been controlled for.       $X_1$        $X_2, X_3, X_4$
- Overall model and relative contribution of each block of variables is assessed

At the bottom of the slide, there is a handwritten red "R" with a double underline, and a red arrow pointing from the underlined text in the third bullet point to this "R".

Logos for IIT ROORKEE and NPTEL ONLINE CERTIFICATION COURSE are visible at the bottom left, and the number 21 is at the bottom right.

So continue with the hierarchical multiple regression process so the IVs independent variables are the predators are entered in the steps or blocks okay. So maybe the first block maybe has got two variables  $X_1, X_2$ . The second block has got you know  $X_3, X_4$  or it could be like the first block has got only  $X_1$ , second block has got  $X_2, X_3, X_4$  or the first block has got  $X_1, X_2, X_3$  and the second block has got only  $X_4$ .

Now how this block is made completely depends on the knowledge of the researcher right. So on the basis of theory he has to do it he or she has to do it. Each IV is assessed in terms of what it adds to the prediction of the dependent variable after the previous IVs have been controlled for. So when you are using this so automatically you are controlling the other variables right.



So the overall model and the relative contribution of each block is assessed. So we will see that how. Now how do you measure this you now contribution? This contribution is measured through the R square. I had explained also in the last lecture about the concept of R square and the difference between R square and adjusted R square okay.

**(Refer Slide Time: 03:10)**

### Assumptions

The assumptions are same as standard multiple regression.

- Linearity ✓
- Normality ✓
- Homoscedasticity ✓
- No multicollinearity ✓
- No outliers ✓

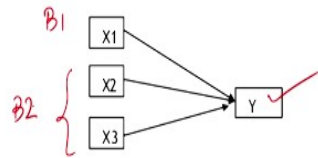
 IIT ROORKEE
  NPTEL ONLINE CERTIFICATION COURSE
 22

Some of the assumptions is normal like any multiple regression that it should be linear, normal so the homogeneity of variance, multicollinearity should not be there and outlier should be absent okay.

**(Refer Slide Time: 03:23)**

### Example

Question: A country's rate of female literacy (Y) is associated with a smaller rate of annual population increase (X1), a greater gross domestic product (X2), and a high per capita income(X3)





```

graph LR
    X1[X1] -- B1 --> Y[Y]
    X2[X2] -- B2 --> Y
    X3[X3] -- B2 --> Y
    
```

**Hierarchical Multiple Regression**

- Block 1: smaller rate of annual population increase (X1) ✓
- Block 2: a greater gross domestic product (X2), and a high per capita income(X3)

 IIT ROORKEE
  NPTEL ONLINE CERTIFICATION COURSE
 23

Now let us take this case. A country's rate of female literacy because a country like India for example, Bangladesh, India, some of the ancient economies there the female literacy rate is very low or very poor right. So we are saying that it is associated with the smaller rate of annual population increase right and a greater gross domestic product and high per capita income, so I am saying what I am thinking is.

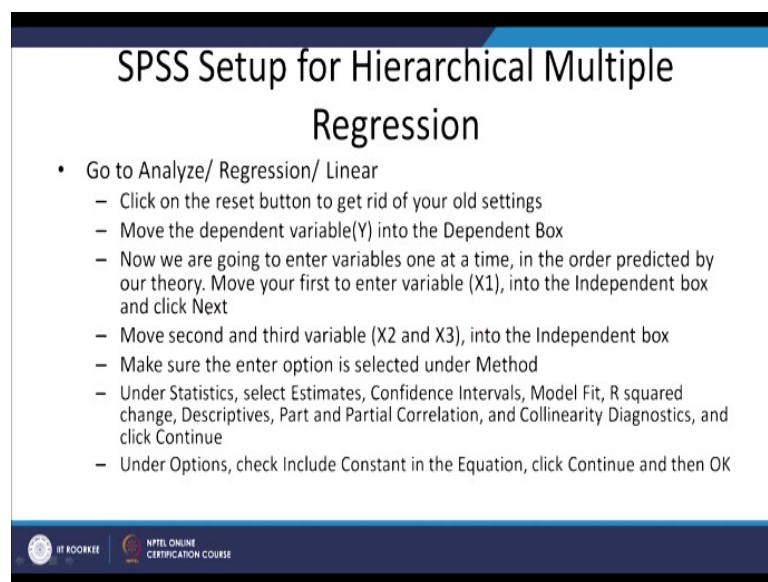
So because you know this will improve if the annual population growth is increasing or decreasing accordingly it will have an effect on the literacy rate. The GDP if it increases and

the per capita income also if it increases that also will have an impact on the Y that is my female literacy right. So these 3 variables we have taken independent variables, predictors and one dependent variable. So there are 4 variables okay.

So in the block 1 what I have done is I felt that smaller rate of annual population increase  $X_1$  is separate from these two and these two a greater gross domestic product  $X_2$  and a higher per capita income both are related to financial parameters R can be considered as 1 right. So this is block 1 I have taken and this is I have taken as block 2 right and I want to see whether my logic comes true or not.

So can I create a model which is more appropriate and more powerful and which explains my regression in a better way right.

**(Refer Slide Time: 04:59)**



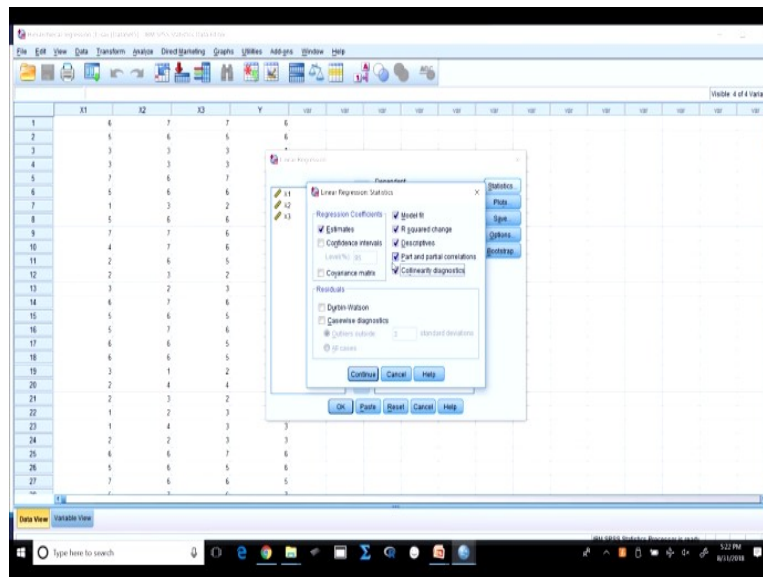
### SPSS Setup for Hierarchical Multiple Regression

- Go to Analyze/ Regression/ Linear
  - Click on the reset button to get rid of your old settings
  - Move the dependent variable(Y) into the Dependent Box
  - Now we are going to enter variables one at a time, in the order predicted by our theory. Move your first to enter variable ( $X_1$ ), into the Independent box and click Next
  - Move second and third variable ( $X_2$  and  $X_3$ ), into the Independent box
  - Make sure the enter option is selected under Method
  - Under Statistics, select Estimates, Confidence Intervals, Model Fit, R squared change, Descriptives, Part and Partial Correlation, and Collinearity Diagnostics, and click Continue
  - Under Options, check Include Constant in the Equation, click Continue and then OK

IT KOOBEE NPTEL ONLINE CERTIFICATION COURSE

So to do that so how do you do, so this is something for you, you can use it later on also. So just to show go to analyze regression linear and move the dependent variable Y into the dependent box. Now move  $X_1$  first into the independent box, click next. So let me show you and then move second  $X_2$  and  $X_3$  into the independent box. Now let me go to the data set.

**(Refer Slide Time: 05:20)**



So this is the data set which I have brought, so  $X_1$ ,  $X_2$ ,  $X_3$  and  $Y$  right. So first I will go to analyze, so during the stepwise also you are doing something similar, so reset so first is I am taking the  $Y$  as my dependent variable which is my literacy, female literacy. During if you remember in the stepwise what we were doing, we were taking all the 3 variables  $X_1$ ,  $X_2$ ,  $X_3$  and then we were giving it an approach whether we want to do take it entire.

That means all at one time or stepwise that means that computer decides which is the highest best predictor and then the next predictor or we use the backward, forward or any of these methods but here we are not doing that, what we are doing is we will use a block method. So first I am using first taking the independent variable  $X_1$  here and then I am saying next then I am going to the  $X_2$  and  $X_3$  and I am considering this to be the next block.

So there are how many blocks now? Two blocks although there are 3 variables, there are 2 blocks okay. So you could have made 3 blocks also if your logic permits but then if you have made 3 blocks then the point is what is the order that you have to decide okay. Now going to statistics I want the R square change because this is what will tell me my contribution, the descriptives, the partial correlation and the collinearity diagnostics to ensure that there is no multicollinearity problem.

**(Refer Slide Time: 06:46)**

**Correlations**

	Y	X1	X2	X3	
Pearson Correlation	Y	1.000	.711	.737	.748
	X1	.711	1.000	.801	.794
	X2	.737	.801	1.000	.807
	X3	.748	.794	.807	1.000
Sig. (1-tailed)	Y		.000	.000	.000
	X1	.000		.000	.000
	X2	.000	.000		.000
	X3	.000	.000	.000	
N	Y	389	389	389	389
	X1	389	389	389	389
	X2	389	389	389	389
	X3	389	389	389	389

**Variables Entered/Removed<sup>a</sup>**

Model	Variables Entered	Variables Removed	Method
1	X1 <sup>a</sup>		Enter
2	X2, X3 <sup>b</sup>		Enter

a. Dependent Variable: Y  
b. All requested variables entered.

**Model Summary**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.711 <sup>a</sup>	.505	.504	1.157	.505	305.185	1	387	.000
2	.787 <sup>a</sup>	.619	.615	1.038	.134	57.811	2	385	.002

a. Predictors: (Constant), X1  
b. Predictors: (Constant), X1, X2, X3

Now look at it, now if you look at this this is the descriptive statistics which I think I am sure you can understand and you can utilize. Now this is the correlation right, now let us see the model summary, now yes this is of my greatest importance. Now when I am using the  $X_1$  so there are two models, two blocks, two models. The first model for the first part of the equation when I used  $X_1$  only my R was 0.711 and I am sure this R you understand.

This is a multi-correlation right and R square is the square of it 0.505. Now adjusted R square is similar to that right 0.504. Now this is significant? Yes, it is significant. Now second when we entered the second block that is  $X_2$  and  $X_3$  together so the new correlation is 0.787 and R square is now 61.9 or 0.619 and if you look at the significance it is also significant. That means both are significant, both are explaining right.

There is no effect that null hypothesis that it has no effect has been rejected right. Now if I go back now let us go back to my PPT.

**(Refer Slide Time: 08:01)**

### Interpretation of Hierarchical Multiple Regression (MR)

- Check the R Square in the Model Summary box. Variables entered in Block 1 (X1) explained 50% of the variance ( $.505 \times 100$ ) in DV.
- After Block 2 variable (X2 and X3) has been included, the model as a whole explained 61.6% of variance in DV.
- In the column labelled R Square Change (on the line marked Model 2) – explained additional 11% of the variance in DV.
- This is significant contribution, as indicated by Sig. F Change value for this line (.000)

Model Summary									
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.711 <sup>a</sup>	.505	.504	1.157	.505	395.185	1	387	.000
2	.787 <sup>b</sup>	.619	.616	1.018	.114	57.611	2	385	.000

a. Predictors: (Constant), X1  
b. Predictors: (Constant), X1, X3, X2

IIT ROORKEE
 NPTEL ONLINE CERTIFICATION COURSE
 25

Now this is my model summary right, so check the R square and you see that it explained how much 50%, 50% of the variance right. After block 2 has been included, the model explains how much 61.6 adjusted R square I am taking right. So the additional 11.4% or 11% I am just for easy I have written 11% variance in the dependent variable is explained by this two new variables right. This is significant contribution as you can see from here right.

**(Refer Slide Time: 08:37)**

### Interpretation of Hierarchical Multiple Regression (MR)

The ANOVA table indicates that the model as a whole (which includes both blocks of variables) is significant

$F(3, 385) = 208.674, p < .0005$

ANOVA <sup>a</sup>						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	528.914	1	528.914	395.185	.000 <sup>b</sup>
	Residual	517.960	387	1.338		
	Total	1046.874	388			
2	Regression	648.222	3	216.074	208.674	.000 <sup>c</sup>
	Residual	398.652	385	1.035		
	Total	1046.874	388			

a. Dependent Variable: Y  
b. Predictors: (Constant), X1  
c. Predictors: (Constant), X1, X3, X2

IIT ROORKEE
 NPTEL ONLINE CERTIFICATION COURSE
 26

Now what is the ANOVA table tell us. The ANOVA table indicates that the model as a whole is significant. Now you see F (3,385) so what it is saying 3 and 385. So when you are having numerator and denominator and your value is how much 208.674 and it is significant okay.

**(Refer Slide Time: 08:59)**

### Interpretation of Hierarchical Multiple Regression (MR)

Check Standardized Coefficient (Beta values) and Sig. box in Model 2  
 All the variables make a unique significant contribution ( $p > .05$ )  
 The best predictor of female literacy is X3 ( $\beta = .352$ ) followed by X2 ( $\beta = .30$ ), and X1 ( $\beta = .19$ ).

$Y$        $Per$        $GDP$        $Pop$

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	1.324	.195		6.795	.000					
	X1	.758	.038	.711	19.879	.000	.711	.711	.711	1.000	1.000
2	(Constant)	.637	.183		3.478	.001					
	X1	.204	.062	.191	3.311	.001	.711	.166	.104	.296	3.379
	X2	.313	.062	.300	5.036	.000	.737	.249	.158	.279	3.583
	X3	.367	.061	.352	5.991	.000	.746	.292	.188	.287	3.482

a. Dependent Variable: Y

→ Per capita income.

Now interpretation how it is given or how you should write all the variables make a significant contribution. Yes, because it is all significant right, so  $X_1$ ,  $X_2$ ,  $X_3$  all are significant. The best predictor is of female literacy which is the Y is which one now  $X_3$ . Now let us see what is  $X_3$  doing. Now  $X_3$  is the beta value is how much 0.352 so highest right and it is not because you have put it in this order.

Sometimes it could have been different also, the orders might have changed because it is contributing and what is this? This is my per capita income of my parents or the family right. This  $X_1$  was the population right growth and this was the GDP. So if you look at it followed by  $X_2$ , so first per capita, then GDP and then finally which one, the population growth. These 3 in this order they are contributing okay.

**(Refer Slide Time: 09:57)**

### Interpretation of Hierarchical Multiple Regression (MR)

#### Reporting the hierarchical multiple regression results

Hierarchical multiple regression was performed to investigate the affect of annual population increase ( $X_1$ ), a greater gross domestic product ( $X_2$ ), and a high per capita income( $X_3$ ) to predict levels of female literacy. after controlling for  $X_2$  and  $X_3$ .

In the first step of hierarchical multiple regression, one predictor was entered: annual population increase ( $X_1$ ). This model was statistically significant  $F(1, 387) = 395.185, p < .0005$  and explained 50% of variance in Y. This factor made a significant unique contribution to the model. After entry of  $X_2$  and  $X_3$  at Step 2 the total variance explained by the model as a whole was 61% ( $F(3, 385) = 208.674, p < .001$ ). The introduction of  $X_2$  and  $X_3$  explained additional 11% of variance in Y, after controlling for  $X_2$  and  $X_3$  ( $R^2$  Change = .114;  $F(2, 385) = 57.611; p < .001$ ). In the final adjusted model all the predictor variables were statistically significant, with  $X_3$  recording a higher Beta value ( $\beta = .352, p < .001$ ) than the  $X_2$  ( $\beta = .30, p < .001$ ) and  $X_1$  ( $\beta = .191, p < .01$ ).

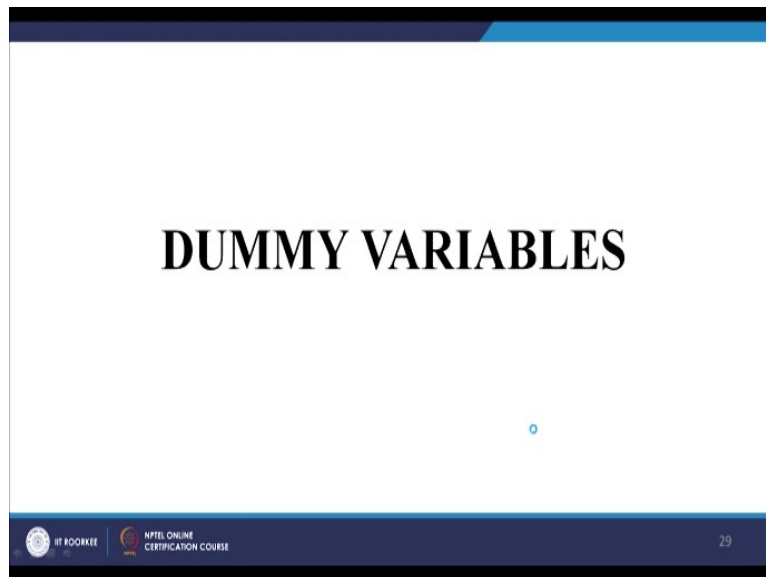


Now how do you write? Hierarchical multiple regression was performed to investigate the effect of annual population increase a greater gross domestic product and highest per capita income to predict the levels of female literacy after controlling for  $X_2$  and  $X_3$  right. So if you do not write this also no issues right but you understand for understanding. So the first step predictor was entered, one predictor was entered.

If in your case suppose another case you are doing and there are two predictors so it is a two predictors were entered right, annual population increase  $X_1$ . This was significant at see  $F(1, 387)$  right and this is significant at 005 and explained 50% of the variance. This is how you should write. This factor made a significant unique contribution to the model. After entry of  $X_2$  and  $X_3$ , the total variance explained is now 61%.

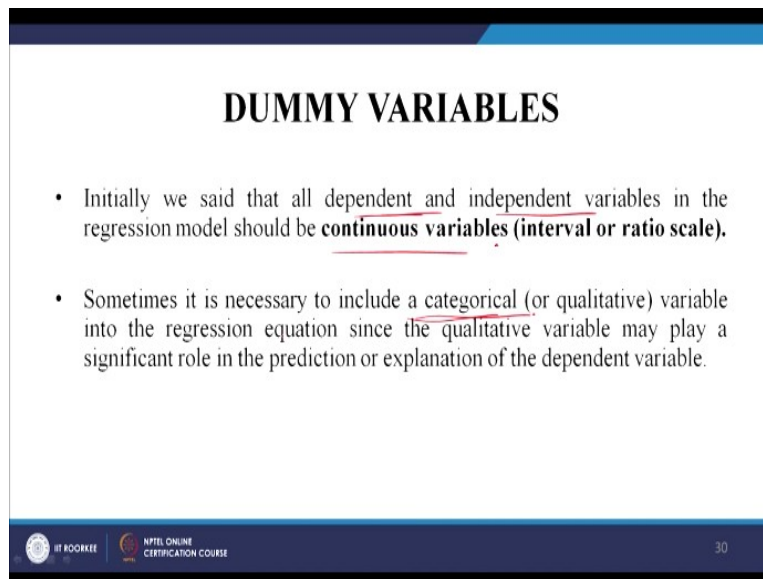
And this is how you should write and the explained additional explanation was 11% right. So in the final adjusted model all the predictors were statistically significant with  $X_3$  recording a higher beta value this much, then  $X_2$  and this. So this is how you should write the output of the hierarchical multiple regression right.

**(Refer Slide Time: 11:12)**



So here this is what was of my interest to tell you how to understand hierarchical multiple regression that it follows logic and then how to do it and how to write the interpretation most important okay. I think you must be clear by now. If you have doubts, then you may ask during the question and answer forum. Now I will continue to a special case of regression. The special case of regression is called dummy variable regression. Now what is this dummy variable regression let me show you.

(Refer Slide Time: 11:38)



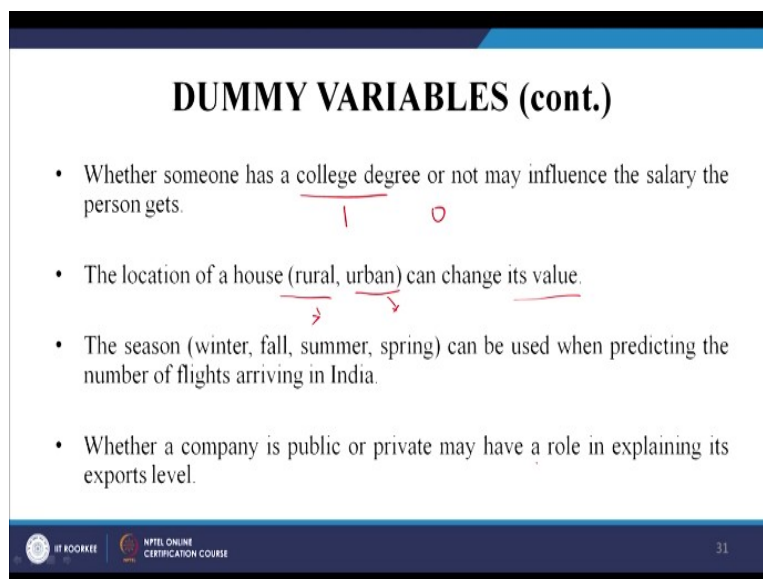
**DUMMY VARIABLES**

- Initially we said that all dependent and independent variables in the regression model should be continuous variables (interval or ratio scale).
- Sometimes it is necessary to include a categorical (or qualitative) variable into the regression equation since the qualitative variable may play a significant role in the prediction or explanation of the dependent variable.

30

If you remember initially we said that all the dependent and independent variables in the regression model should be continuous right but sometimes it happens in life that you need to include a categorical variable or a qualitative variable into the regression equation right. so how do we do deal with this case now.

(Refer Slide Time: 12:02)



**DUMMY VARIABLES (cont.)**

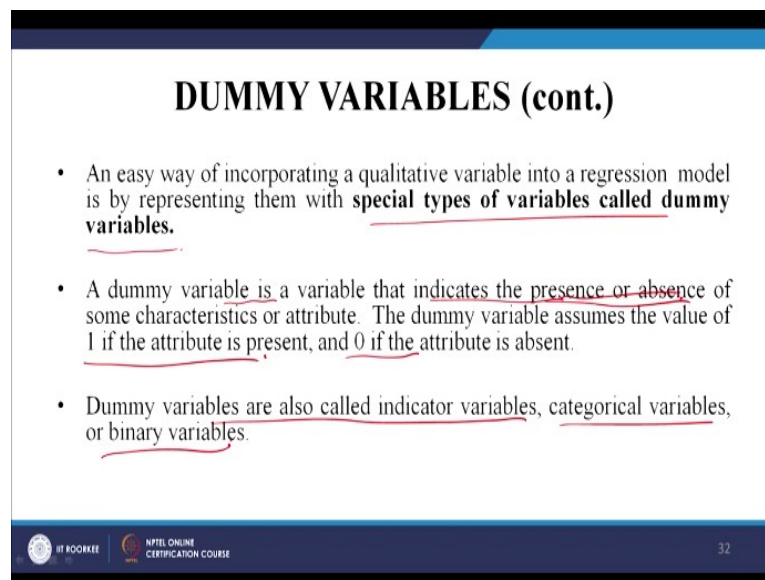
- Whether someone has a college degree or not may influence the salary the person gets.   
1 0
- The location of a house (rural, urban) can change its value.   
r u
- The season (winter, fall, summer, spring) can be used when predicting the number of flights arriving in India.
- Whether a company is public or private may have a role in explaining its exports level.

31

So whether someone has a college degree or not may influence the salary the person gets possible right. So having a degree or not having a degree let us say (1,0) now the presence or absence of (1,0) may influence the salary he gets. The location of a house can change its value. I think you all will agree if the house is in urban place the cost will be more, the price will be more than in the rural.

The season can be used when predicting the number of flights arriving in India or anywhere. Similarly, whether a company is public or private may have a role in explaining its exports level possible right. So these are some of the variables which are all categorical but they do have an impact during the regression equation. How do we include these variables into the study and explain them right let us see.

**(Refer Slide Time: 12:55)**



**DUMMY VARIABLES (cont.)**

- An easy way of incorporating a qualitative variable into a regression model is by representing them with special types of variables called dummy variables.
- A dummy variable is a variable that indicates the presence or absence of some characteristics or attribute. The dummy variable assumes the value of 1 if the attribute is present, and 0 if the attribute is absent.
- Dummy variables are also called indicator variables, categorical variables, or binary variables.

IT KOOBEE NPTEL ONLINE CERTIFICATION COURSE 32

So dummy variable is an easy way of incorporating a qualitative or a categorical variable into a regression model by representing them with special types of variables called dummy variables. So it is a dummy variable case right. A dummy variable is a variable that indicates only very simple you see the presence or absence, whether it is present or absent of some characteristics or attribute.

The dummy variable assumes the value of 1 if it is present, 0 if it is absent that is all. Dummy variables are also called indicator variables, categorical variables or binary variables because of their property. Let us say we want to predict the salary of a customer service agent within that the years of experience is one of the variables that can have an impact right.



**(Refer Slide Time: 13:46)**

## DUMMY VARIABLES

- Let's say that we want to predict the salary a customer service agent gets. We think that years of experience is one of the variables ( $X_1$ ).  
*(continuous)*
- We would also like to include whether the person is a college graduate or not. We will use a dummy variable to include this information. Therefore  $X_2$  will be  
*Categorical.*

$X_2 = 0$ , if the person is not a college graduate.

$X_2 = 1$ , if the person is a college graduate.



33

We would also like to include whether the person is a college graduate or not so what are the variables 1 that how many years of experience. So this is the continuous data right but the second one whether he is a college graduate or not that may also influence his educational level. Now this is a categorical variable right, so now I have one continuous one categorical. This I could have done the dependent variable salary so if I would have taken only this one it would have been the simple regression.

But if I am taking two that is the case of a multiple regression and the problem here is but this second one is the categorical thing.

**(Refer Slide Time: 14:27)**



## DUMMY VARIABLES

- Consider the respondent's sex. The variable sex has two categories – 1 for males and 2 for females.
- What we do is to create two dummy variables – one for males and the other for females. Here's how we do it:

$\text{-sex\_male} = 1$  if male and  $0$  if female. and  
 $\text{-sex\_female} = 1$  if female and  $0$  if male.

$$\begin{array}{r} 4 \\ M \\ 2 - \textcircled{1} = 1 \end{array} \quad \begin{array}{r} 3 \\ F \\ 2 \end{array}$$

- If there are  $k$  categories, then you use  $k - 1$  of the dummy variables in your regression analysis.
- The category that you omit becomes your comparison group. We're going to enter sex\_male into the analysis and omit sex\_female. That means that females will be the comparison group.

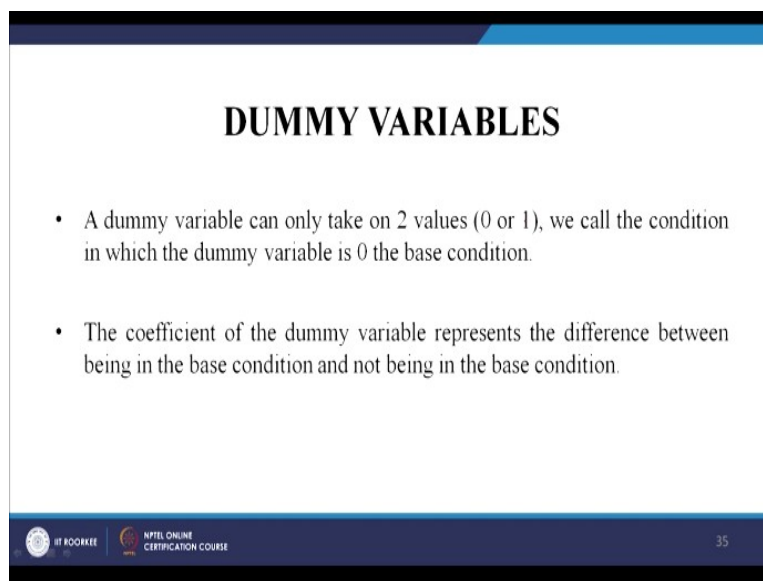


34

So consider the respondent's sex, the variable sex has two categories 1 for male and 2 for female similarly. What we do is to create two dummy variables one for males and other for

females, so how do we do it, 1 if male and 1 if female suppose somebody is a female then 1 if female 0 if male. So it is like absent present okay. Remember when you have k categories, suppose in the case of gender. How many? Male and female two categories.

So you will use k-1 dummy variable so how many here, now only one, 2-1=1 right. Suppose there would have been 3 categories then how many you will have groups in a two, 3-1 =2 right. The category that you omit the one and the third becomes your comparison group right. So this is what you compare it with. So we are going to enter let us say the sex male into the analysis and omit sex female that means that females will be the comparison group.

**(Refer Slide Time: 15:27)**



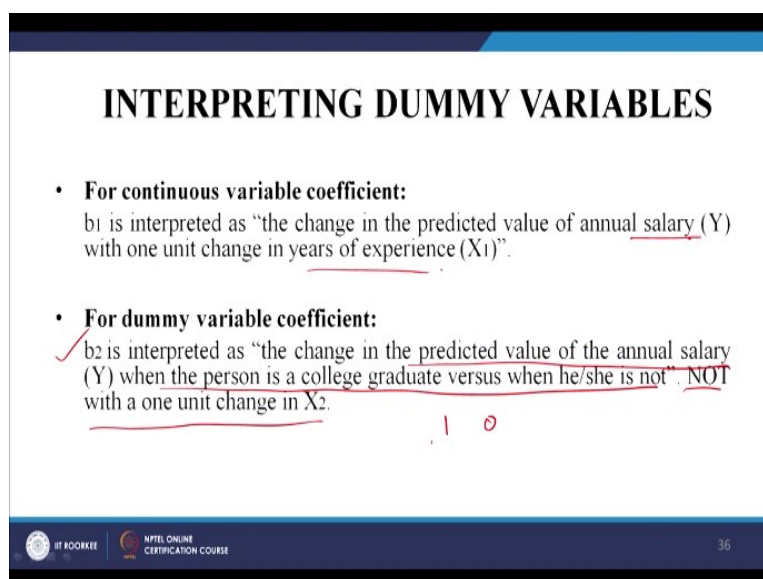
### DUMMY VARIABLES

- A dummy variable can only take on 2 values (0 or 1), we call the condition in which the dummy variable is 0 the base condition.
- The coefficient of the dummy variable represents the difference between being in the base condition and not being in the base condition.

35

I will show you with an example. A dummy variable can only take two values 0 or 1 right.

**(Refer Slide Time: 15:33)**



### INTERPRETING DUMMY VARIABLES

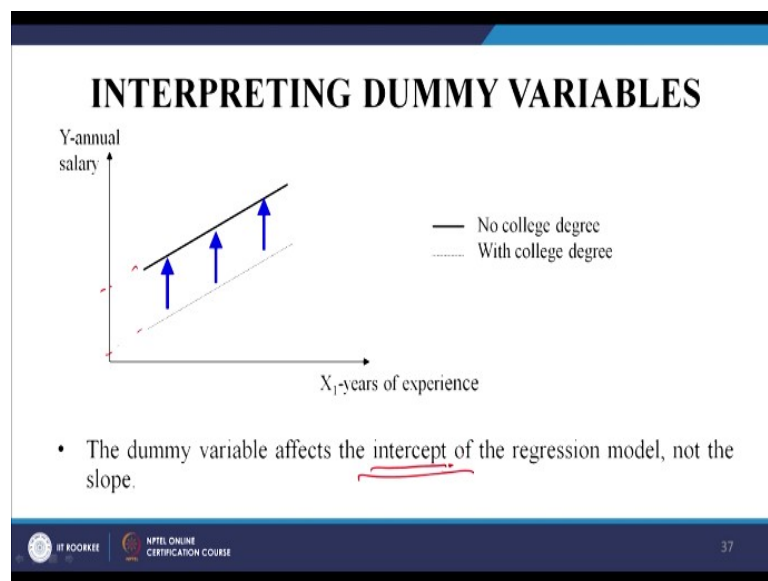
- **For continuous variable coefficient:**  
b1 is interpreted as "the change in the predicted value of annual salary (Y) with one unit change in years of experience (X1)".
- **For dummy variable coefficient:**  
b2 is interpreted as "the change in the predicted value of the annual salary (Y) when the person is a college graduate versus when he/she is not". NOT with a one unit change in X2.  
1 0

36

So for continuous variable coefficient  $b_1$  is interpreted the slope, the change in the predicted value of annual salary  $Y$  with one unit change in years of experience true right. So how much one unit change with one unit change how much you know the annual salary is changing. For dummy variable coefficient  $b_2$  which is a categorical second case the change in the predicted value of the annual salary when the person is a college graduate versus when he or she is not, with a one unit change in  $X_2$ .

How this is one unit change because there is 1 and 0 two values. So either it is 1 or it is 0 right.

**(Refer Slide Time: 16:15)**



So let us so what it says is if you draw this, the annual salary and years of experience for example with college degree without college degree, without college degree or with college degree the dummy variable affects the intercept you see. The intercept is affected but now the slope, the angle is the same. So the rate of change does not change right but the intercept changes okay.

**(Refer Slide Time: 16:41)**

## DUMMY VARIABLE EXAMPLE

Y: annual salary

X<sub>1</sub>: years of experience

X<sub>2</sub>: 1 if the person has a college degree, 0 otherwise.

Suppose the regression equation is

$$\hat{y} = \underbrace{25}_{\text{Call. } a} + \underbrace{2.5}_{\text{Exp}} x_1 + \underbrace{8}_{\text{Call. } b_2} x_2$$

Assume that the person has 5 years of experience. What would his salary be if he is not a college graduate? What would his salary be if he is a college graduate?

So let us take this regression equation,  $Y=25$  this is the intercept, it is just hypothetically  $\hat{y}=a+b_1 X_1+b_2 X_2$  okay. So 'a' is this one,  $b_1$  is my slope right for the first one first independent variable and  $b_2$  is my second right. So  $X_2$  is my categorical variable,  $X_1$  is my in this case my experience, so experience plus college graduate or not okay.

Assume that a person has 5 years of experience, so this is 5 okay. What would be salary if he is not a college graduate? If he is not a college graduate, so just or what would salary be if he is a college graduate? Now you have to understand, so if you use this, if it is 1 or 0 depending right so you can always compare whether if it is 0 then but you should know what is 1 and what is 0, 1 is for male, 0 is for female or whatever it is right.

**(Refer Slide Time: 17:47)**

The salary of a person with 5 years of experience:

- If he is not a college graduate: *putting the values in the regression equation*

$$\hat{y} = 25 + 2.5x_1 + 8x_2$$

$$= 25 + 2.5(5) + 8(0)$$

$$= 37.5$$

- If he is a college graduate:

$$= 25 + 2.5(5) + 8(1)$$

$$= 45.5$$

So let us see this. So if he is not a college graduate, so what it is becoming  $25+2.5(5)+8(0)$  so my salary is 37.5 but if it is a college graduate so this is the difference okay.

(Refer Slide Time: 18:04)

### MULTI-CATEGORY DUMMY VARIABLES

- What if we want to use a categorical variable that has more than 2 levels? For example, how do we use dummy variables for a "season" variable?
- We cannot assign numbers 1, 2, 3, 4... because a dummy variable can only take on values 0 and 1.
- Instead we use multiple dummy variables to code the multi-category variable.
- When a categorical variable has **d** levels, **d-1** number of dummy variables are used to code this categorical variable.
- You take one level to be the base condition where all of the dummy variables are 0.

40

Now multi-category, when we have more than two levels right so let us say for season, we cannot assign numbers 1, 2, 3, 4 because a dummy variable can only take two values 0 and 1. So instead we use multiple dummy variables to code it right. So again the levels will be **d-1** so **k-1** or **d-1** however you may show right and the base condition take one level to be the base condition where all the dummy variables are 0 okay.

(Refer Slide Time: 18:33)

### MULTI-CATEGORY DUMMY VARIABLES

- **For example** to code seasons, we need  $4 - 1 = 3$  dummy variables ( $X_1, X_2, X_3$ ).
- Let's take winter as our base case. We designate  $X_1$  to represent Spring,  $X_2$  to represent Summer,  $X_3$  to represent Fall. Only one of the dummy variables can be 1 at a time.

	X1	X2	X3	
Winter	0	0	0	(base case)
Spring	1	0	0	$X_1=1$ when spring
Summer	0	1	0	$X_2=1$ when summer
Fall	0	0	1	$X_3=1$ when fall

- Winter:  $\underline{0,0,0}$       Spring: 1,0,0
- Summer: 0,1,0      Fall: 0,0,1

41

Now let us see this. To code seasons we need how many dummy variables, how many seasons in a year, 4. So  $4-1=3$  dummy variables you require  $X_1, X_2, X_3$ . So let us say winter, so 0 it is not a winter,  $X_2$  it is not even this one right. So this is a base case spring. If it is  $X_1$



let us say  $X_1$  is 1 and this is 0 right. So the third one also will be, fourth one will also be 0. Summer 0 1 0 0 1.

So if you see it is either only one of it or it is not there right. So winter is 0 0 0 right. It saying in the case of winter 0 0 0 so we designate as you can see  $X_1$  to represent spring,  $X_2$  to represent summer,  $X_3$  fall right. So these 3 it has been taken right. Only one of the dummy variables can be at one at a time. So winter, summer and fall we have taken, winter, spring summer and fall and this is comparison group right.

So when it is a case of winter all are 0 obviously. When it is spring, it is 1 0 0 right, so 1 0 0. When it is summer, just 0 1 0 so you can do that right.

**(Refer Slide Time: 19:50)**

- Hence, we need only three dummy variables to capture all information about season variable.
- We do not need to create a fourth dummy variable for **“Winter”** because this category is already represented by the other three dummy variables—when each of the dummy variables is 0, then it must be the case that the season is “Winter.”
- This category is the excluded or reference category, and the  $b$  coefficient for each dummy variable is compared against it.

0 comp / ref

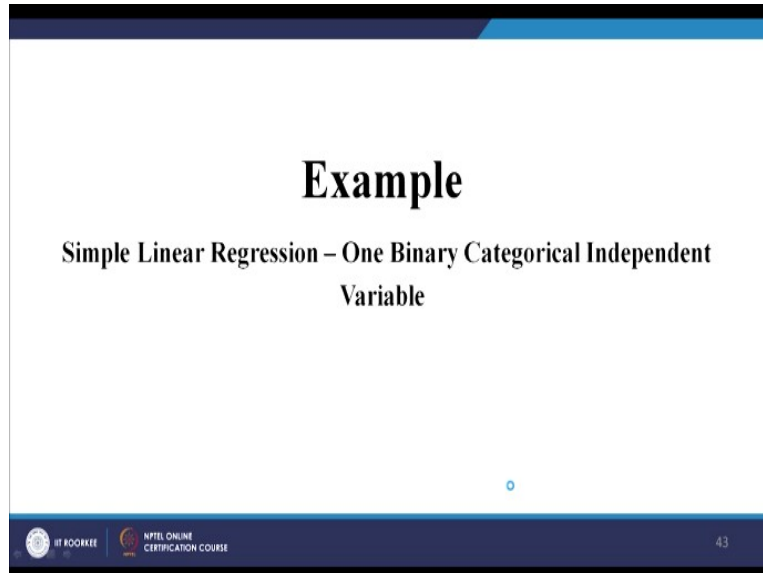
IIT ROORKEE NPTEL ONLINE CERTIFICATION COURSE 42

Hence, we need only 3 dummy variables to capture our information. We do not want to create a fourth dummy variable because this is it will be redundant unnecessary because the category is already represented by the other 3 dummy variables when each of the dummy variable is 0, it must be the fourth one which is winter in this case right. If it is all 0, 0 0 then it is a case of a winter right.

If it is 1 for you know it is  $X_1$  this is spring right  $X_1$  is spring so if it is 1 then it cannot be winter or summer or fall same thing right. The category is this one category is excluded or called the reference category and the beta coefficient for each dummy variable is compared against it so that means what you have one comparison group or reference category which is comparison or reference right.

And everybody every other variable is compared against this reference category okay. Let us see that.

**(Refer Slide Time: 20:46)**



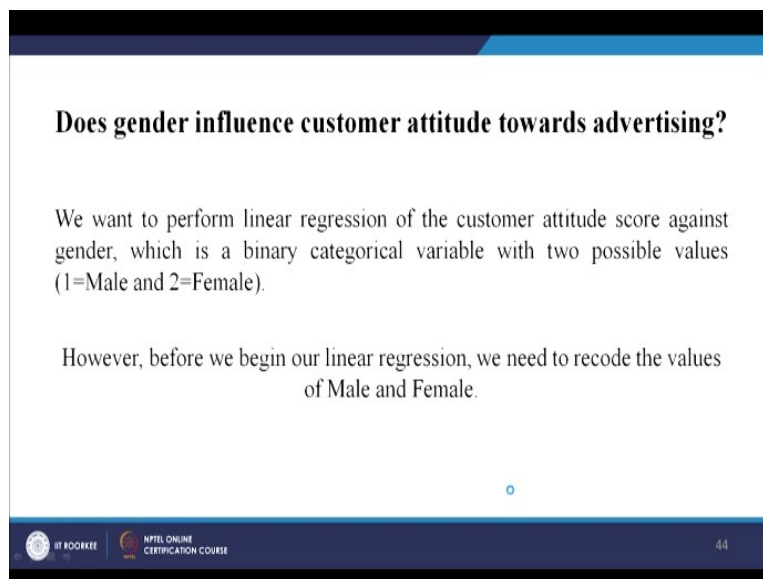
**Example**

Simple Linear Regression – One Binary Categorical Independent Variable

43

The slide features a dark blue header and footer. The footer contains the IIT Kharagpur logo on the left, the NPTEL ONLINE CERTIFICATION COURSE logo in the center, and the number 43 on the right. A small blue circle is positioned in the lower right area of the slide content.

**(Refer Slide Time: 20:50)**



**Does gender influence customer attitude towards advertising?**

We want to perform linear regression of the customer attitude score against gender, which is a binary categorical variable with two possible values (1=Male and 2=Female).

However, before we begin our linear regression, we need to recode the values of Male and Female.

44



The slide features a dark blue header and footer. The footer contains the IIT Kharagpur logo on the left, the NPTEL ONLINE CERTIFICATION COURSE logo in the center, and the number 44 on the right. A small blue circle is positioned in the lower right area of the slide content.

**(Refer Slide Time: 20:55)**

## Why must we do this?

- The codes 1 and 2 are assigned to each gender simply to represent which distinct place each category occupies in the variable **gender**.
- However, linear regression assumes that the numerical amounts in all independent, or explanatory, variables are meaningful data points. So, if we were to enter the variable **gender** into a linear regression model, the coded values of the two gender categories would be interpreted as the numerical values of each category.
- This would provide us with results that would not make sense, because for example, the gender **Female** does not have a value of 2.

**We can avoid this error in analysis by creating dummy variables.**



45



Now how do you do it? Simple in a regression, so will begin this with a coding of male and female gender right. So how do you code? Why we must do? I think I have explained so I am not getting into it.

**(Refer Slide Time: 21:01)**

Because our **gender** variable only has two categories, turning it into a dummy variable is as simple as recoding the values of **Male** and **Female** in **gender** from 1=Male and 2=Female to 0=Male and 1=Female.

### STEPS TO BE FOLLOWED:

- To begin, select **Transform** and **Recode into Different Variables**.
- Find the variable **gender** in the variable list and move it to the **Numeric Variable -> Output Variable** text box.
- Next, under the **Output Variable** header, enter in the name and label for the new gender variable.
- Click **Change**, to move new output variable into the **Numeric Variable -> Output Variable** text box in the center of the dialogue box.



46

So steps to be followed. To begin, select transform and recode into different variables right. Let me show you instead of telling you. So let us go to the dummy variable the simple one.

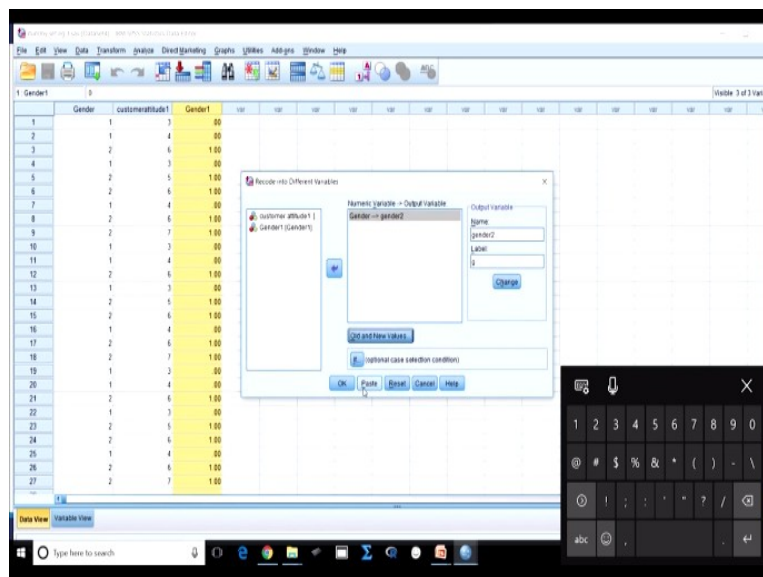
**(Refer Slide Time: 21:11)**

**STEPS TO BE FOLLOWED (continued) :**

- Then, select **Old and New Values**.
- Enter **1** under the **Old Value** header and **0** under the **New Value** header.
- Click **Add**. You should see **1 → 0** in the **Old → New** text box. Now enter **2** under the **Old Value** header and **1** under the **New Value** header. ○
- Click **Add**, and then **Continue**.
- Finally, click **OK** in the original **Recode into Different Variables** dialogue box. Scroll down to the very end of the variables list in **Variable View**.
- You should see your new dummy variable **gender1** at the end of the list, as it's the last variable to be created.

47

(Refer Slide Time: 21:17)

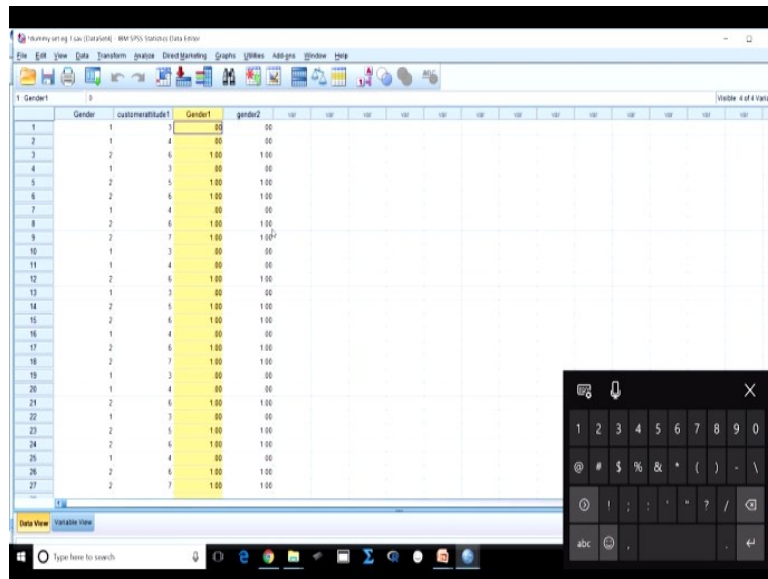


This is the one which I had already done, I may delete it also, I can delete or if you can see I can again do it in a different name maybe let it this be transform. Now go to this recode into different variables. Now what do we want to change, now gender. Gender I want to change. Now what do you want to name it, you have to give a label okay. So the label is let us say I am giving you this time gender okay.

Some you know just to differentiate 2 okay and you just label it as let us say gender g okay, so change. Now go to the old and new values and now put in the value. For example, in your case, in this case for example 1 is let us say male and 2 is female let us say, so value 1. What is the value? 1 and what is the new value? The new value is 0 right add and all other values

will be how much now? 1. So you add this, so 1 is 0 and others are all 1 continue right and okay.

**(Refer Slide Time: 22:40)**



Now if you see we have created a similar same this, this is same thing I had done it earlier, I want to show you how to do it. Now we have created a variable right. Now let us go back to the steps right.

**(Refer Slide Time: 22:53)**

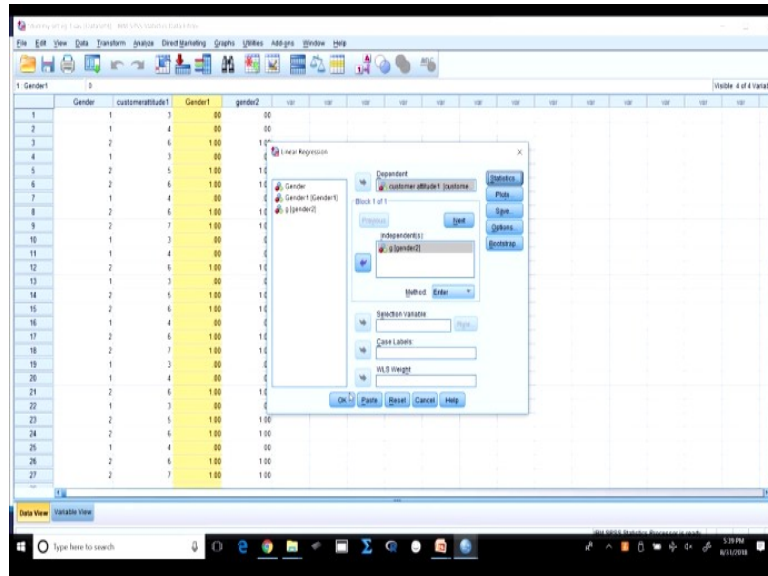
**STEPS TO BE FOLLOWED (continued) :**

- To perform simple linear regression, select **Analyze, Regression, and Linear...**
- Find **customer attitude1** in the variable list and move it to the **Dependent** box at the top of the dialogue box. Find **gender1** in the variable list and move it to the **Independent(s)** box in the center of the dialogue box.
- Click **OK**.
- The following output in the SPSS Output window will be generated

48

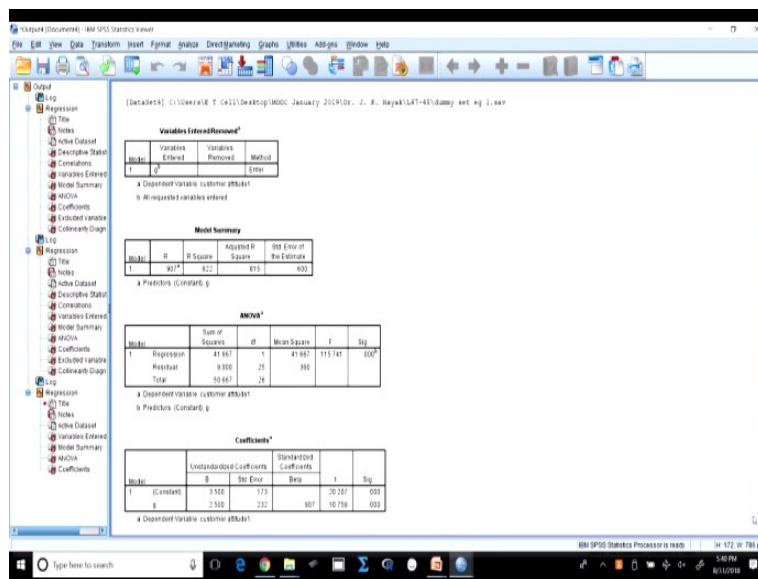
So the variable is created and then you can use it for regression. Suppose let us go back.

**(Refer Slide Time: 22:59)**



Now I want to do a regression to study the effect of gender and customer attitude. Now what I am doing is go to regression linear so my customer attitude is my dependent variable. This should have been in scale but do not worry you can change that and gender 2 now, imagine this is the new one which I have built is my independent variable. Now what I am doing is I am trying to measure it.

**(Refer Slide Time: 23:31)**



So I do not want anything else, I just want to run it and see the effect. So how much impact is gender making on the customer attitude or the product you know the attitude towards the product. Now R square is 82%, so gender explains 82% of the study okay and it is significant right. So that means what in this case we can say gender is a very important variable which affects the consumer attitude or the customer attitude okay. Now let us go back. Now this was you see this is how it comes right. So the coefficient I have written this all for you.

(Refer Slide Time: 24:04)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	3.500	.173		20.207	.000
	Gender1	2.500	.232	.907	10.758	.000

a. Dependent Variable: customer attitude1

$$Y = a + bX$$
$$3.5 + 2.5 \text{ (Gender)}$$

IT ROORKEE NPTEL ONLINE CERTIFICATION COURSE 50

(Refer Slide Time: 24:07)

### Interpretation

Using the output from SPSS, we can calculate the mean customer attitude towards advertisement for men and women using the following regression equation:

$$Y = a + bX$$

where Y is equal to our dependent variable and X is equal to our independent variable.

Into this equation, we will substitute **a** and **b** with the statistics provided in the **Coefficients** output table, **a** being the **constant coefficient** and **b** being the **coefficient associated with gender** (our explanatory variable).

In this example, our equation should look like this:

$$\text{Customer attitude1} = 3.5 + (2.5 \times \text{gender})$$

IT ROORKEE NPTEL ONLINE CERTIFICATION COURSE 51

So the output we can calculate the mean customer attitude towards advertisement for men and women maybe it was customer attitude towards advertisement right. So  $Y=a+bX$  where Y is equal to is the dependent variable, X is our independent variable which is gender okay. So now if you look let us go back to the same data set. Now this is my a right so we know  $Y = a+bX$  right.

$Y=3.5+2.5*(\text{Gender})$ . So if you see what we have done male was 0, female was 1.

(Refer Slide Time: 24:50)

## Interpretation

Since **gender** takes on the value of 1 for female and 0 for male, the predicted scores are as follows:

**Customer attitude** =  $3.5 + (2.5 \times 1) = \underline{6}$  (Females)  
**Customer attitude** =  $3.5 + (2.5 \times 0) = \underline{3.5}$  (Males)

So, on average, female respondents reported a customer attitude score that is 2.5 points *higher* than male respondents.

© IIT ROORKEE NPTEL ONLINE CERTIFICATION COURSE 52

So if you see what we have done, male was 0, female was 1. So if you do that so the customer attitude for females is coming 6, for males it is coming 3.5. So on an average female respondents reported a customer attitude score that is 2.5 points higher than the male respondents. This is how you interpret okay.

**(Refer Slide Time: 25:13)**

## Interpretation

- Remember we can use the  $r^2$  statistic (which is calculated in the **Model summary** output table) to gauge how much variation in the dependent variable is explained by the independent variable.
- In our example, the  $r^2$  is 0.822 82.2 % of the variation in customer attitude score is explained by gender.

© IIT ROORKEE NPTEL ONLINE CERTIFICATION COURSE 53

And the R square was 82.2% which is explained by the gender, this is also you should mention right. So this is I will show you a new example now.

**(Refer Slide Time: 25:18)**



## Reporting the results

A simple linear regression was calculated to predict customer attitude towards advertisement based on gender. A significant regression equation was found ( $F(1, 25) = 115.741, p < .000$ ), with an  $R^2$  of .822. Customers' predicted attitude towards advertisement is equal to  $3.500 + 2.500$  (gender).

◦

IT ROORKEE NPTEL ONLINE CERTIFICATION COURSE 54

**(Refer Slide Time: 25:22)**

## Example

Simple Linear Regression: One Categorical Independent Variable  
with Several Categories

IT ROORKEE NPTEL ONLINE CERTIFICATION COURSE 55

Now this is a one categorical independent variable with several categories okay.

**(Refer Slide Time: 25:28)**

## Does ethnicity influence customer attitude score?

Suppose you want to see how the customer attitude score of the respondents towards advertisement is related to their ethnic group. In this dataset, the variable **ethngrp2** has 5 categories (1=Kashmiris, 2= Bengalis, 3= Punjabis, 4= Tamil, and 5=Other).

However, because linear regression assumes all independent variables are numerical, if we were to enter the variable **ethngrp2** into a linear regression model, the coded values of the five categories would be interpreted as numerical values of each category. Using **ethngrp2** in a linear regression without changing the coded values of the categories would give us results that would not make sense.

To avoid error, we're going to create dummy variables for **ethngrp2**.

Suppose this is interesting, you want to see how the customer attitude score of the respondents towards advertisement is related to their ethnic group. So now we have taken various ethnic groups for example 1 Kashmiris. In India, these are the different states and people from these states are called like this. West Bengal Bengalis 2, Punjab Punjabis, Tamil Nadu Tamil and rest are others okay.

However, because linear regression assumes all independent variables are numerical. So we have to now convert it into since these are all categorical variables we have to do a dummy variable right.

(Refer Slide Time: 26:09)

## Dummy Variable

Each dummy variable represents one category of the explanatory variable and is coded with 1 if the case falls in that category and with 0 if not.

**For example**, in the dummy variable for **Bengalis** ethnicity, all cases in which the young person's ethnicity is **Bengalis** will be coded as 1 and all other cases are coded as 0.

In the dummy variable for **Kashmiris**, all cases in which the young person's ethnicity is **Kashmiris** will be coded as 1 and all other cases are coded as 0.

The same will be done in the **Tamil**, in the **Punjabis**, and in the **Other** ethnicity dummy variables. This allows us to enter in the ethnicity values as numerical.

So what we will do, you see here. Each dummy variable represents one category of the explanatory variable and is coded with 1. We have done earlier also. If the case falls in that

category and with 0 if not okay. For example, if the dummy variable for Bengalis ethnicity, all cases in which the young person's ethnicity is Bengali will be coded as 1 and others as 0. Similarly, for Kashmiri, Tamil and others okay.

**(Refer Slide Time: 26:32)**

**STEPS TO BE FOLLOWED:**

- To begin creating five dummy variables (one for each of the categories in **ethngrp2**), select **Transform** and then **Recode into Different Variables**.
- Find **ethngrp2** in the variable list and move it to the **Numeric Variable -> Output Variable** text box.
- Under the **Output Variable** header, type in the name and label of the first dummy variable. Because 1= Kashmiris in the **ethngrp2** variable values, we can start by creating the **KASHMIRIS** dummy variable. We can label it "Kashmiris Ethnic Group".
- Click **Change**.
- Next, click **Old and New Values**.
- Because 1= Kashmiris in **ethngrp2**, enter 1 under the **Old Value** header and 1 under the **New Value** header.
- Click **Add**.

58

So how do you do? First recode into different variables same thing. Now move to the numerical variable output and here type in the name and label of the first dummy variable. Suppose 1 is Kashmiri in the ethnic group two variable clues we can start by creating the Kashmiri. Suppose in our case in the original data which I will show you now. Let us say I have brought the data set.

**(Refer Slide Time: 26:55)**

The screenshot shows the SPSS Variable View window. The variable list includes:

Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1 ethnicity2	Numeric	8	2		None	None	8	Right	Nominal	Input
2 customarr	Numeric	8	2		None	None	8	Right	Scale	Input
3 price	Numeric	8	2		None	None	8	Right	Nominal	Input

Now let us go to the variable view. Now you see this. I already know that 1 is Kashmiri, 2 is Bengali, 3 is Punjabi is a coding just a coding right, 4 is Tamil, 5 is other okay. So I know it okay.

**(Refer Slide Time: 27:11)**

The screenshot shows the SPSS Data Editor window with a dataset named 't\_bengal'. The 'Data View' tab is active, displaying a table with 29 rows and 13 columns. The columns are: ethnicity2, cost, price, kashmir, bengal, punjab, tamil, other, and several unlabeled columns. The 'ethnicity2' column contains values 1 through 5, representing different ethnicities. The 'kashmir' column has a value of 1 for ethnicity 1 and 0 for others. The 'bengal' column has a value of 1 for ethnicity 2 and 0 for others. The 'punjab' column has a value of 1 for ethnicity 3 and 0 for others. The 'tamil' column has a value of 1 for ethnicity 4 and 0 for others. The 'other' column has a value of 1 for ethnicity 5 and 0 for others. A virtual keyboard is visible in the bottom right corner of the screenshot.

ethnicity2	cost	price	kashmir	bengal	punjab	tamil	other
1	1.00	10.00	1.00	0.00	0.00	0.00	0.00
2	1.00	9.50	0.00	1.00	0.00	0.00	0.00
3	2.00	7.50	0.00	0.00	1.00	0.00	0.00
4	2.00	7.00	0.00	0.00	0.00	1.00	0.00
5	3.00	5.50	0.00	0.00	0.00	0.00	1.00
6	3.00	5.00	0.00	0.00	0.00	1.00	0.00
7	4.00	3.50	0.00	0.00	0.00	0.00	1.00
8	4.00	3.00	0.00	0.00	0.00	0.00	1.00
9	5.00	1.50	0.00	0.00	0.00	0.00	1.00
10	5.00	1.00	0.00	0.00	0.00	0.00	1.00
11	1.00	9.75	0.00	0.00	0.00	0.00	0.00
12	1.00	9.50	0.00	0.00	0.00	0.00	0.00
13	2.00	7.50	0.00	0.00	0.00	0.00	0.00
14	2.00	6.75	0.00	0.00	0.00	0.00	0.00
15	3.00	5.50	0.00	0.00	0.00	0.00	0.00
16	3.00	4.75	0.00	0.00	0.00	0.00	0.00
17	4.00	3.25	0.00	0.00	0.00	0.00	0.00
18	4.00	3.00	0.00	0.00	0.00	0.00	0.00
19	5.00	1.25	0.00	0.00	0.00	0.00	0.00
20	5.00	1.00	0.00	0.00	0.00	0.00	0.00
21	1.00	10.00	0.00	0.00	0.00	0.00	0.00
22	1.00	9.50	0.00	0.00	0.00	0.00	0.00
23	2.00	7.50	0.00	0.00	0.00	0.00	0.00
24	2.00	7.00	0.00	0.00	0.00	0.00	0.00
25	3.00	5.50	0.00	0.00	0.00	0.00	0.00
26	3.00	5.00	0.00	0.00	0.00	0.00	0.00
27	4.00	3.50	0.00	0.00	0.00	0.00	0.00
28	4.00	3.00	0.00	0.00	0.00	0.00	0.00
29	5.00	1.50	0.00	0.00	0.00	0.00	0.00

So now how do I create my dummy variables? So let us do it let us transform. Now we go to you know recode into different variables. So we want ethnicity so I have already taken, so let us reset it okay ethnicity. Now first one I am doing it for let us code the key. So Kashmir right Kashmir and I want to label it as only K it is for just change easy right, old and new values. What was the old value for Kashmir?

I have shown you earlier, it was 1. So the new value is also 1. So I am adding it. Now you see coming to here, all other values I will put it as 0 right, so all other values are 0. Now add continue and you just give it okay. So if you go to the file you see this file not this file the other file right, so one label has been created for Kashmir. Similarly, again we will go it for again for Bengalis correct.

So here I will change it to Bengal and I will mark it as B okay so Bengal was 2 right so here I am changing old and new values so my this is old value is 2 and new value is 1 add. Now let us do it for the others right. Transform, recode, so now we have taken already Bengal so Bengal I have given B right and then I will go to the values so let us say if you have I will just show you.

Again, I will show you, so the old value for Bengal was what, 2 right. So old value for Bengal was 2 so but now the present value will be because dummy it is 1 add. So all other values will be marked as 0 right, 0 so continue right and okay. Now let us go back to the model, so you have created one file for Bengal right. Now again go for the third, the third was if I remember let us check.

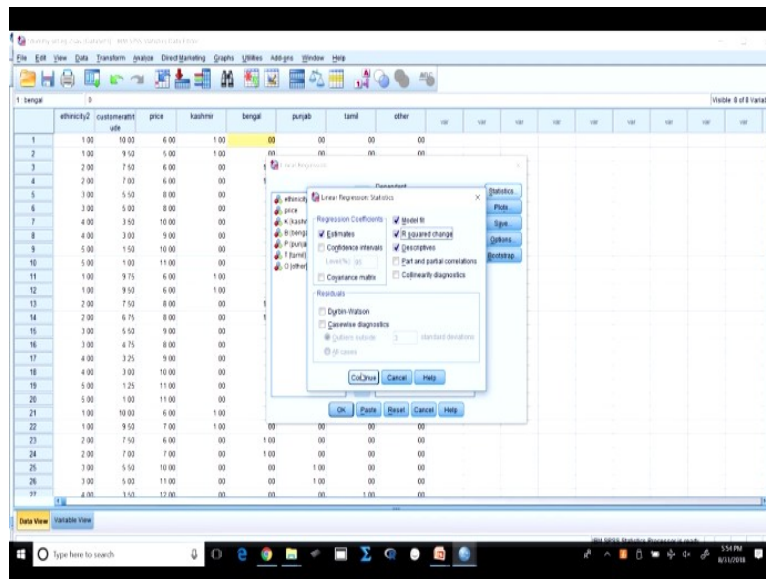
What was the third? Third was Punjabi then Tamil then other okay. So let us go to the third one. So transform right so this one we will change it now, so this is Punjabi let us say Punjab right Punjab and this is we are giving it a big P okay change, so it has changed here. Now old value, old value for Punjab was 3 okay and you just have to remove these 2 okay earlier data. So again let us do it, 3 so Punjabi was 3, now it is 1.

All other values will be marked as so value is 3, Punjab was 3, now we are marking it 1, we need to add it and then go to all other values and mark it 0 right. So again add it, so now continue right okay. So you will see that a new variable has been created. Similarly, we will create it for Tamil and others also. So let us make it fast different variables, so ethnicity is Tamil and this is T, change old and new values.

First remove these things otherwise they will confuse you. So this is 4 right 4, now it is 1 add it, all other values kindly make it 0 add it again, continue okay right. So let us see so again Tamil has been created and finally we need to go for the last which is for others right. So this is others and this is let us say O, change old and new values, so this is 5 right this is 5 so 5 here we get 5 and here it is 1 right, add and all other values you make it 0 right, add okay.

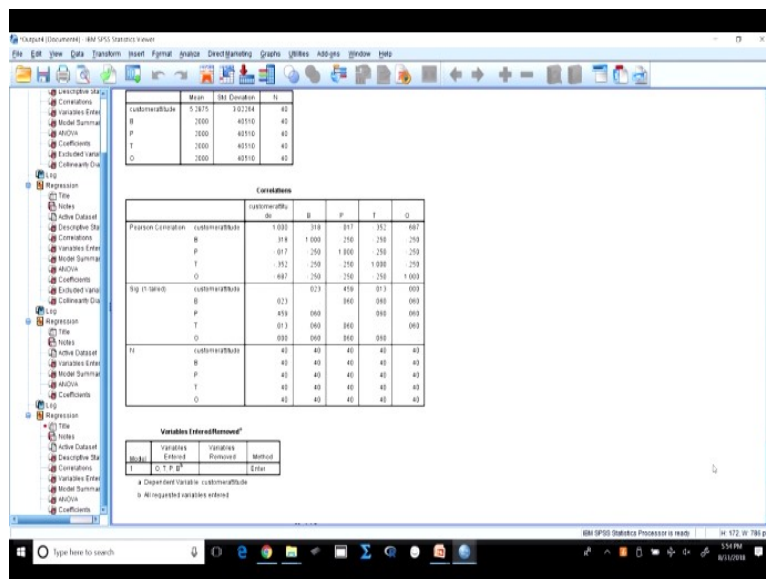
So now you see you have got 5 factors right, the 5 dummy variables have been created but when there are 5 we said we always need 4 right  $k-1$ . So we need  $5-1$  to see the effect okay and the last one will be the reference or the comparison group.

**(Refer Slide Time: 32:28)**



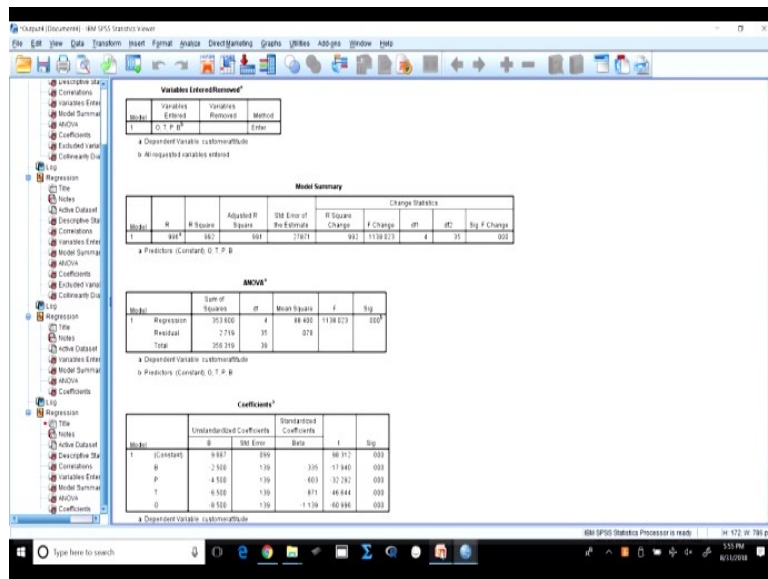
Now I want to see the effect of this on the customer attitude by the ethnicity right. So let us go to regression linear so my customer attitude is my dependent variable and now what I am doing is I may take I am taking Kashmir as my reference group so I am leaving it and Punjabi, Tamil, others as my independent variables. Now statistics I just want to see the descriptives the R square that is all.

**(Refer Slide Time: 32:54)**



Even that also I do not require, I just need to check the model. Now you see the mean and standard deviation has been given to you right.

**(Refer Slide Time: 33:02)**





Now if you look at the model, so this is the correlation right and go down the model summary, so this model is explaining almost 99.2%. That means ethnicity affects, the effect of ethnicity on consumer attitude of advertisements is 99% in this case. It is highly significant and look at the coefficients. So if you look at the coefficients what it will be?

Now B is the constant intercept is this much and it is all significant so let me and the slopes are also given to you right. So let us go to the here I have used the same data and I have explained it right.

(Refer Slide Time: 33:44)

**STEPS TO BE FOLLOWED (continued) :**


- You should see 1->1 in the **Old -> New** text box. Now, because in this dummy variable we want all the other values to be 0, click **All other values** under the **Old Value** header and enter **0** under the **New Value** header.
- Click **Add**.
- Click **Continue**, and then **OK** in the original **Recode into Different Variables** dialogue box.
- To check that you have successfully created a dummy variable called **KASHMIRIS**, scroll down to the end of the variable list in Variable View. **KASHMIRIS** should be the last variable in the list, as it is the latest variable to be created.
- Let's recode one more dummy variable together. Go back to **Transform** and **Recode into Different Variables**.

 IIT KOOHKEE
  NPTEL ONLINE CERTIFICATION COURSE
 59

(Refer Slide Time: 33:46)

**STEPS TO BE FOLLOWED (continued) :**


- All of our previous information will still be entered. Click to highlight **ethngrp2** -> **KASHMIRIS** and then click the blue arrow to remove this from the **Numeric Variable** -> **Output Variable** text box.
- Find **ethngrp2** again and move it to the **Numeric Variable** -> **Output Variable** box. Under the **Output Variable** header, enter in **BENGALIS** as the output variable name and **Bengalis Ethnic Group** as the output variable label. Click **Change**.
- Click **Old and New Values**. All of our previous work will be saved here and we no longer need it. Highlight the commands in the **Old** -> **New** box and click **Remove**.
- Because 2= Bengalis, enter 2 under the **Old Value** header and 1 under the **New Value** header (as in this dummy variable, we want Bengalis to have a value of 1). Click **Add**. Under the **Old Value** header, select All other values and enter 0 under the **New Value** header.
- Click **Add**.
- Click **Continue** and then **OK** in the original Recode into Different Variables dialogue box.



IT KOOKEE NPTEL ONLINE CERTIFICATION COURSE 60

(Refer Slide Time: 33:47)

**STEPS TO BE FOLLOWED (continued) :**

- Repeat the above steps for the **Punjabis**, **Tamil**, and **OTHER** ethnicity categories.
- Remember when entering these following categories, you must use their corresponding values when recoding: 3= Punjabis, 4= Tamil, and 5=Other.
- **For example**, when recoding **ethngrp2** into the **PUNJABIS** dummy variable, you will use **3** as the **Old Value** and **1** as the **New Value** for **PUNJABIS**, while recoding all other values to **0**. You'll use **4** as the **Old Value** and **1** as the **New Value** for **TAMIL**, while recoding all other values to **0**. And, finally, you'll use **5** as the **Old Value** and **1** as the **New Value** for **OTHER**, while recoding all other values to **0**.
- When you are finished, you should have five new dummy variables at the end of your variable list in **Variable View**. 

IT KOOKEE NPTEL ONLINE CERTIFICATION COURSE 61

(Refer Slide Time: 33:48)



#### STEPS TO BE FOLLOWED (continued) :

**Before we begin:** when we fit our model in SPSS, we need to select one dummy variable as the baseline category (the category against which we compare all the other categories).

In this example, we will use **KASHMIRIS** as the baseline category. As the **KASHMIRIS** variable is now our baseline, we don't have to include it in the linear regression model. *(Using the reference category makes all interpretation in reference to that category. The reference category is usually chosen based on how you want to interpret the result)*

We will, however, need to include all of the other dummy variables for ethnicity in the model. Basically, this means we are comparing all the ethnicities to the **KASHMIRIS** ethnicity.

So now after doing this, in this example we have used Kashmiris as the baseline category so that is why we did not take it inside. As the Kashmiris, variable is now a baseline, we do not have to include in the linear regression model right. We will, however, need to include all the other dummy variables for ethnicity right like for example Punjabi, Tamil, others we have done it.

**(Refer Slide Time: 34:07)**

#### STEPS TO BE FOLLOWED (continued) :

- To perform simple linear regression, select **Analyze, Regression,** and then **Linear...**
- In the dialogue box that appears, move **customer attitude1** to the **Dependent** box and **BENGALIS, PUNJABIS, TAMIL,** and **OTHER** to the **Independent(s)** box.
- Click **OK.**
- The following output in the SPSS Output window will be generated

**(Refer Slide Time: 34:09)**

**Model Summary**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.996 <sup>a</sup>	.992	.991	27871

a. Predictors: (Constant), o, t, p, b

**ANOVA<sup>a</sup>**

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	353.600	4	88.400	1138.023	.000 <sup>b</sup>
	Residual	2.719	35	.078		
	Total	356.319	39			

a. Dependent Variable: customerattitude  
b. Predictors: (Constant), o, t, p, b

IIT ROORKEE NPTEL ONLINE CERTIFICATION COURSE 64

And then we have taken this output. So this output you remember, R square was 99.2, adjusted R square is this much, it is significant and just ask what is this much, it is significant. **(Refer Slide Time: 34:17)**

**Coefficients<sup>a</sup>**

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	9.687	.099		98.312	.000
	b	-2.500	.139	-.335	-17.940	.000
	p	-4.500	.139	-.603	-32.292	.000
	t	-6.500	.139	-.871	-46.644	.000
	o	-8.500	.139	-1.139	-60.996	.000

a. Dependent Variable: customerattitude



IIT ROORKEE NPTEL ONLINE CERTIFICATION COURSE 65

**(Refer Slide Time: 34:19)**

## Interpretation

- We can use our SPSS results to write out the fitted regression equation for this model and use it to predict values of **customer attitude1** for given certain values of **ethngrp2**.
- In this case, **KASHMIRIS** is our baseline, and therefore the **Constant** coefficient value of 9.687 represents the predicted customer attitude score of a respondent in that category.
- Remember that the dummy variables used in this regression model are coded as Bengal=1, Punjabi=1, Tamil=1, and Other=1. This means that we will enter in 1 as the value for X in the regression equation

$$Y = a + bX$$

 IIT ROORKEE
  NPTEL ONLINE CERTIFICATION COURSE
 66

And this is what is of greater importance to us. Now if you look at it, so in this case Kashmiris is our baseline and therefore the constant coefficient of 0.9687 this one right represents the predicted consumer attitude score of a respondent in this category. That means this is the intercept actually. When the X is not there, no X is there, the independent predictors are not there, still it is 9.687.

Remember that the dummy variables are coded as Bengal=1, Punjabi=1, Tamil=1, others=1. This means that we will enter in 1 as the value for X right.

**(Refer Slide Time: 34:54)**

## Interpretation

The predicted scores are as follows:

$$a + b(x)$$



customer attitude1 = $9.687 + (-2.500 \times 1) = 7.187$ (Bengalis)	$\begin{array}{r} 9.687 \\ 2.187 \\ \hline 2.500 \end{array}$
customer attitude1 = $9.687 + (-4.500 \times 1) = 5.187$ (Punjabis)	
customer attitude1 = $9.687 + (-6.500 \times 1) = 3.187$ (Tamil)	
customer attitude1 = $9.687 + (-8.500 \times 1) = 1.187$ (Other)	

So, on average, Bengal respondents report a customer attitude score that is 2.500 points **lower** than Kashmir respondents.

On average, Punjabi respondents report a customer attitude score that is 4.500 points **lower** than Kashmir respondents.

On average, Tamil respondents report a customer attitude score that is 6.500 points **lower** than Kashmir respondents.

On average, respondents in the Other ethnic group category report a customer attitude score that is 8.500 points **lower** than Kashmir respondents.

 IIT ROORKEE
  NPTEL ONLINE CERTIFICATION COURSE
 67

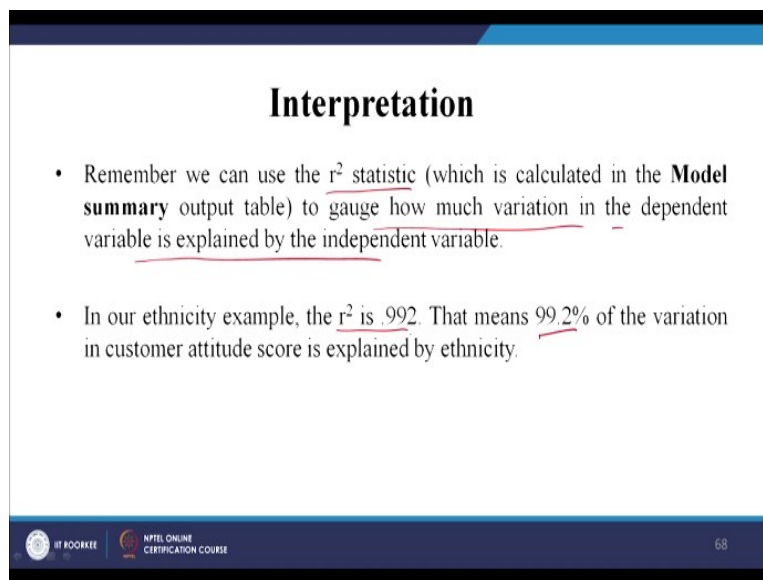
So this is how it looks like. So 9.687 a+bx okay so x is for example for the 1. Now for the first case Bengal case. So what is the comparison group? Kashmiris. So Bengal are being compared to Kashmiris, Punjabis are being compared to Kashmiris, Tamil are being

compared to Kashmiris, others are being compared to Kashmiris. So if you look at the first one -2.5, -4.5, -6.5 Bengali, -2.5 Punjabi, -4.5 it goes on right.

So when you multiple this is the total score you are getting individually for Bengalis separately. So it means that if let us say on an average Bengalis respondents report a consumer attitude score that is 2.5 times points lower than the Kashmiri respondents right. So in this case, our Kashmiri respondents is 0 right. So that means what it is only 9.687, so  $9.687 - 7.187 = 2.500$  right so this is what.

So this is 5.187 so all the time reference category is 0 so that is why we have when we compare we can say that this is 4.5 times, this is 6.5 times and this is 8.5 times.

**(Refer Slide Time: 36:21)**



**Interpretation**

- Remember we can use the  $r^2$  statistic (which is calculated in the **Model summary** output table) to gauge how much variation in the dependent variable is explained by the independent variable.
- In our ethnicity example, the  $r^2$  is .992. That means 99.2% of the variation in customer attitude score is explained by ethnicity.

IIIT ROORKEE | NPTEL ONLINE CERTIFICATION COURSE | 68

So basically now it has helped you to understand the effect of ethnicity a categorical variable on the dependent variable consumer attitude towards the advertisement and you see how nicely it has explained you and how well you can you know maybe use it in the research work. So remember we can use this r square statistics to gauge how much variation in the dependent variable is explained by the independent variable right, r square is 992, 99.2 that means which is being explained by ethnicity.

**(Refer Slide Time: 36:48)**

## Reporting the results

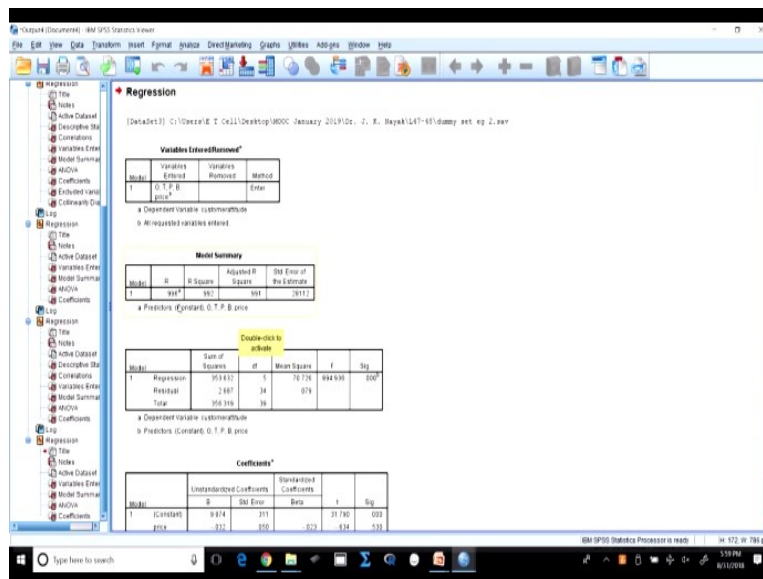
A simple linear regression was calculated to predict customer attitude towards advertisement based on ethnicity. A significant regression equation was found ( $F(4, 35) = 1138.023, p < .000$ ), with an  $R^2$  of .992.

$$y = a + b_1 X_1 + b_2 X_2 + b_3 X_3 + b_4 X_4 + b_5 X_5$$

*price*      *B*      *T*  
*price*      *price*      *price*      *price*      *price*

How do you report? A simple linear regression was calculated to predict consume attitude towards advertisement based on ethnicity and the regression equation was found at  $F(4,35)$  was this much and significant with an R square of this much. So this is what you write but I will do one more thing, I will show you how to do when you have you know a dummy variable and also let us say a continuous variable.

(Refer Slide Time: 37:12)



Now in this case you see I have a continuous variable price, the price let us say of the product and the ethnicity of the people. Now again I will do this and what I will do, I will take linear and I will take this price also into here, price actually let me change it, the variable view so price is my scale, it is not you know others are nominal that is fine right. So now let us do this, so analyze regression linear reset okay.

So dependent variable is attitude, price is my independent variable and here I am taking 4 of it right and let us run this model right. So okay so if you look at it nothing much has changed okay, so only price has been introduced into this model. So price has got a beta the beta the slope the  $b_1$  of price is 0.032 that means what? A one unit in price of the product will decrease the consumer attitude towards the product in a negative way right.

This is because it is -0.032 and other things remain the same. Then, whether it is significant, whether price has got a significant effect, well no, it is not a significant effect. So we cannot say that price has you know has any effect or not, we cannot say that. So we will conclude that price has no effect on the consume attitude but had it been significant then you could have used this value into the regression equation and taken all other dummy variables value into as the categorical variable in second equation into the second you know variable.

And measure the Y and you could see the change when you have introduced now a variable, now it will look like something like this. For example, here let us go. So if  $Y = a + b_1 X_1 + b_2 X_2 + b_3 X_3 + b_4 X_4 + b_5 X_5$  so now the  $X_2$  is my let us say the Bengali, Punjabi right the Tamil and which is my others. So this is a recent equation the latest equation we will get.

So by using this method you have understood by using dummy variable you can do so many wonderful researches and you can interpret them so nicely and you can you know you can report it and maybe it will help you in publishing your paper. So it is not only the dummy variable but also dummy variable with the continuous variable together you have taken into the equation and interpreted the result okay.

Well, I think it is clear to you and you have understood what does dummy variable mean, it is a very special case of regression, very interesting, very useful and you can use it for your benefit and you can write several research papers on that. Well, thank you so much. Have a nice day.