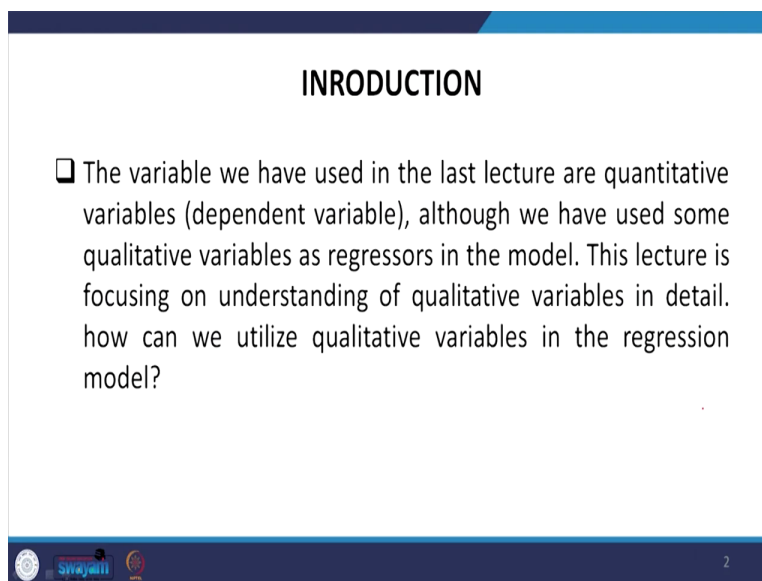


Handling Large Scale Unit Level Data Using STATA
Professor Pratap C. Mohanty
Department of Humanities and Social Sciences
Indian Institute of Technology, Roorkee
Lecture 31
Introduction to Qualitative Variables

Welcome friends once again to the NPTEL MOOC module on Handling Large Scale Unit Level Data using STATA. We are at the last but one week of our completion of the module, where we have specifically targeted the understanding of Stata through the qualitative data. So far we tried our best to enrich you, enlighten you with the background understanding of Stata through continuous variables. Now onwards it will be completely, at least for this week, it will be completely dedicated to qualitative information and qualitative data.


Without spending much time I just wanted to mention that to apply the Stata very correctly we already circulated you, already recorded you a lecture on review of Stata command, and please go through that. Hardly in 10-15 minutes you can able to wrap up all the important commands for better results. So, let us stick to the introduction of qualitative variables. The meaning itself is clarifying many things, that it is qualitative. The information that contains the variable describes many important aspects which cannot be just compounded by a quantitative variable are continuous variable.

(Refer Slide Time: 02:02)



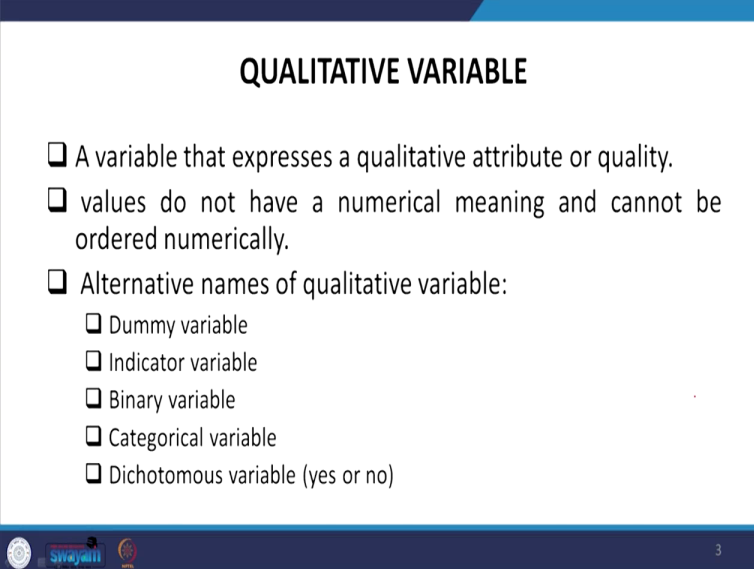
INRODUCTION

- The variable we have used in the last lecture are quantitative variables (dependent variable), although we have used some qualitative variables as regressors in the model. This lecture is focusing on understanding of qualitative variables in detail. how can we utilize qualitative variables in the regression model?

 2

So the variable used so far as I mentioned is all about quantitative, metric data or continuous data. Though we have used some qualitative variables as regressors or in the explanatory variables, this lecture is dedicated to the understanding of qualitative variables only along with quantitative variables in explanatory variables as well. We are mentioning and we will clarify all your doubts related to these issues. How can we utilize qualitative variables in the regression model is our most important challenge.

(Refer Slide Time: 02:47)



QUALITATIVE VARIABLE

- A variable that expresses a qualitative attribute or quality.
- values do not have a numerical meaning and cannot be ordered numerically.
- Alternative names of qualitative variable:
 - Dummy variable
 - Indicator variable
 - Binary variable
 - Categorical variable
 - Dichotomous variable (yes or no)

3

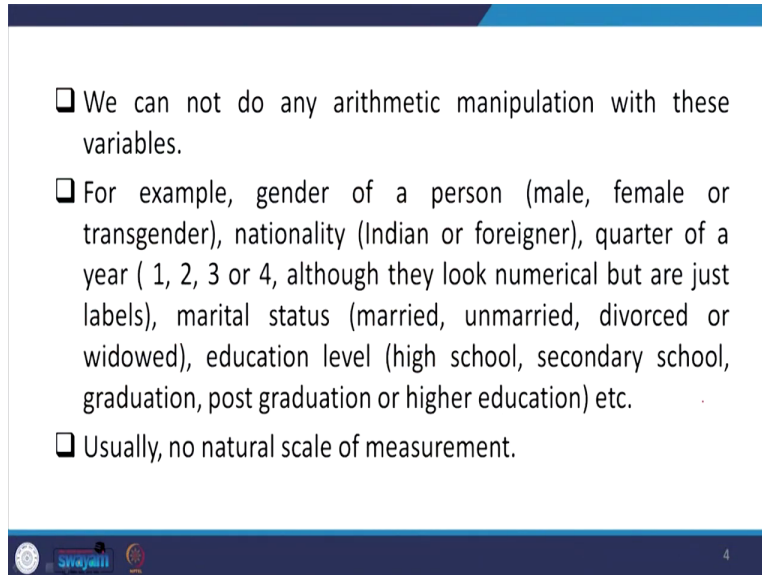
What do you mean by qualitative variable? A variable that expresses a quality attribute or quality or explanation is called qualitative variables. These are the values that do not have the numerical meaning or cannot be ordered numerically. There are some alternative approaches by which you can able to present them into numerical format.

So, there are alternative names of the qualitative variables, such as dummy variables. It is famously used. I think all might have heard. And indicator variable, binary variable, categorical variable, dichotomous variable. Some of them we already discussed during our introduction to data.

So, but these are not exactly the same. We are now clarifying bits by bits with its categories and how those can be interpreted with the help of STATA. So, when it is dichotomous it is strictly yes and no or when it is binary it is strictly 1 and 0, usually computer reads. So, dummy basically, it

considers one category as dummy as against another one. So, we are going to clarify in a short while.

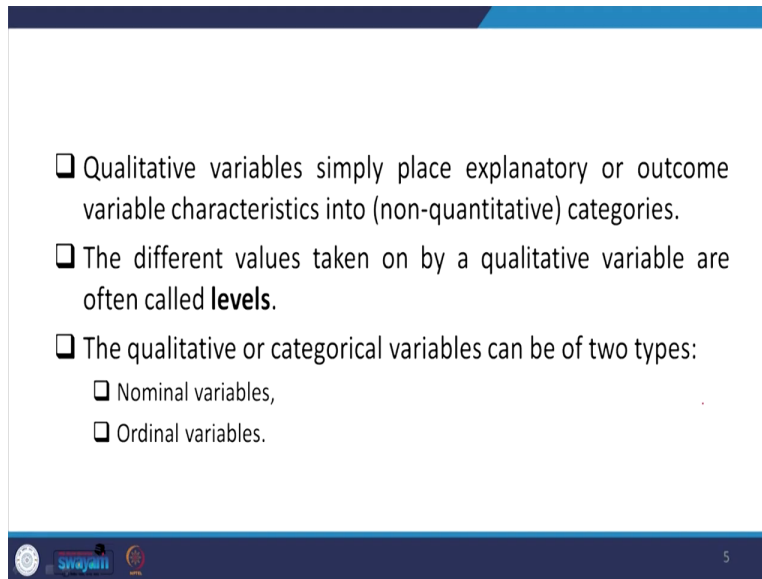
(Refer Slide Time: 04:04)



- ❑ We can not do any arithmetic manipulation with these variables.
- ❑ For example, gender of a person (male, female or transgender), nationality (Indian or foreigner), quarter of a year (1, 2, 3 or 4, although they look numerical but are just labels), marital status (married, unmarried, divorced or widowed), education level (high school, secondary school, graduation, post graduation or higher education) etc.
- ❑ Usually, no natural scale of measurement.

We cannot do any arithmetic manipulation with these variables, like mean or standard deviation is not clearly understood, just average of those categorical variables. We mentioned several times in our previous lecture that it is not suggested and that will result into spurious interpretation. So, need not be presented with so many example, but at maximum I can mention that if it is related to nationality, Indian or foreigner, average of nationality gives no answer. So, like quarters in a year or gender, male, female, transgender etc. are important. Please go through each of the example we cited and I am not mentioning again. So, usually, no natural scale of measurement possible in these variables.

(Refer Slide Time: 04:59)

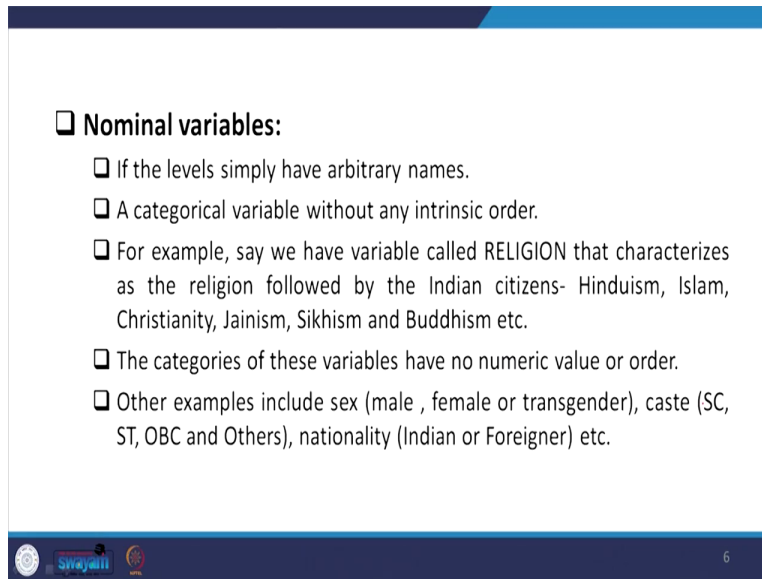


- ❑ Qualitative variables simply place explanatory or outcome variable characteristics into (non-quantitative) categories.
- ❑ The different values taken on by a qualitative variable are often called **levels**.
- ❑ The qualitative or categorical variables can be of two types:
 - ❑ Nominal variables,
 - ❑ Ordinal variables.

Qualitative variables simply place explanatory or outcome variable characteristics into categories. The different values taken on, by a qualitative variable are also called levels. I think level and its values how these are going to be attached, we already discussed, but once again we will clarify through qualitative variables only.

The qualitative or categorical variables can be of two types, broadly either nominal, where no ordering is possible, and usually there is no comparison between one category with another just by their magnitude. They are separated completely with certain reasons. Whereas ordinal has a systematic ordering and we can interpret differently.

(Refer Slide Time: 06:05)



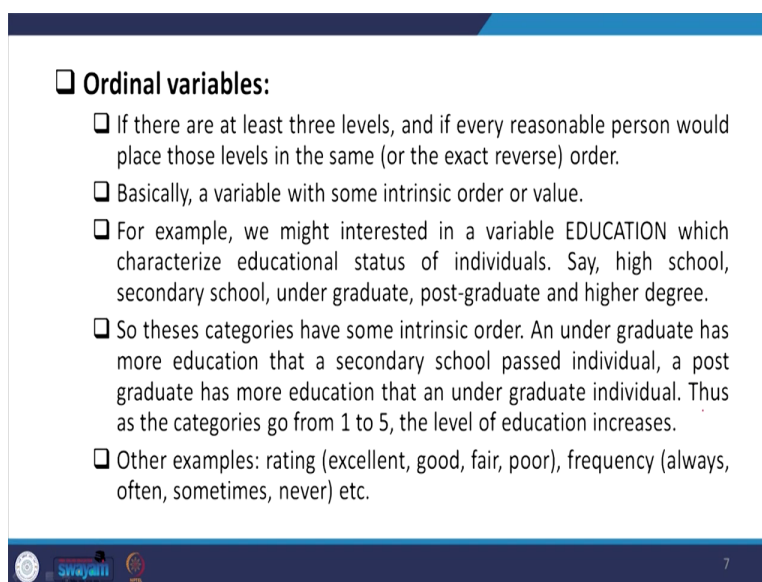
Slide 6: Nominal variables

- ❑ **Nominal variables:**
 - ❑ If the levels simply have arbitrary names.
 - ❑ A categorical variable without any intrinsic order.
 - ❑ For example, say we have variable called RELIGION that characterizes as the religion followed by the Indian citizens- Hinduism, Islam, Christianity, Jainism, Sikhism and Buddhism etc.
 - ❑ The categories of these variables have no numeric value or order.
 - ❑ Other examples include sex (male , female or transgender), caste (SC, ST, OBC and Others), nationality (Indian or Foreigner) etc.

6

So, nominal variables, I think we have said many times earlier but once again, if the labels simply have arbitrary names are also called nominal, like gender, male, female, any label you do it, it does not have any ordering. A categorical variable without any intrinsic order is called nominal variable, like citizenship we discussed. The categories of these variables have no numeric values or order. Other examples are like sex of the persons or caste and social groups etc.

(Refer Slide Time: 06:39)



Slide 7: Ordinal variables

- ❑ **Ordinal variables:**
 - ❑ If there are at least three levels, and if every reasonable person would place those levels in the same (or the exact reverse) order.
 - ❑ Basically, a variable with some intrinsic order or value.
 - ❑ For example, we might interested in a variable EDUCATION which characterize educational status of individuals. Say, high school, secondary school, under graduate, post-graduate and higher degree.
 - ❑ So theses categories have some intrinsic order. An under graduate has more education that a secondary school passed individual, a post graduate has more education that an under graduate individual. Thus as the categories go from 1 to 5, the level of education increases.
 - ❑ Other examples: rating (excellent, good, fair, poor), frequency (always, often, sometimes, never) etc.

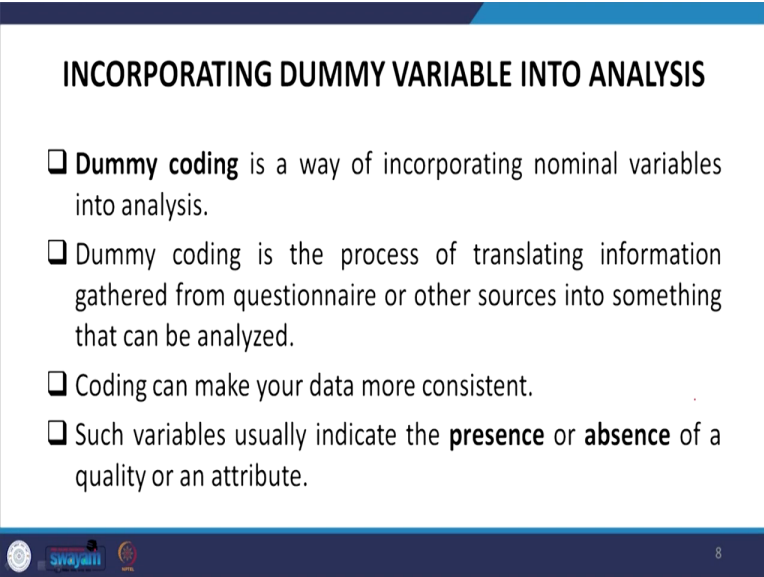
7

Then coming to the ordinal variables, as we have already been mentioned that it is a clear ordering, clear hierarchical patterns that can be understood from the very nature of the data or the variable. If there are at least three levels, and if every reasonable person would place those levels in the same or the exact reverse order, then that is called ordinal variables.

Basically variable with some intrinsic order is called ordinal variables, like variable education which characterize educational status of the individuals or standard of education, years of education. Even when we are saying years of education maybe, but it does not have a complete count, it should have a complete year. There is no fraction possible in case of years of education. And usually those are count in nature. But here we are trying to mention hierarchical standards.

So let us understand that ordinal variable have categories with intrinsic order. Then an undergraduate has more education than a secondary school passed individuals, if we can categories in an order. It is very clearly understood that higher the order higher the value or higher the value in terms of certain attributes. So, similarly, we can categories them 1 to 5 or 5 to 1, but better to give higher weightage to higher value or higher meaning or higher magnitude. Other such examples are rating like excellent, good, fair and poor, frequency etc.

(Refer Slide Time: 08:33)



INCORPORATING DUMMY VARIABLE INTO ANALYSIS

- Dummy coding** is a way of incorporating nominal variables into analysis.
- Dummy coding is the process of translating information gathered from questionnaire or other sources into something that can be analyzed.
- Coding can make your data more consistent.
- Such variables usually indicate the **presence** or **absence** of a quality or an attribute.

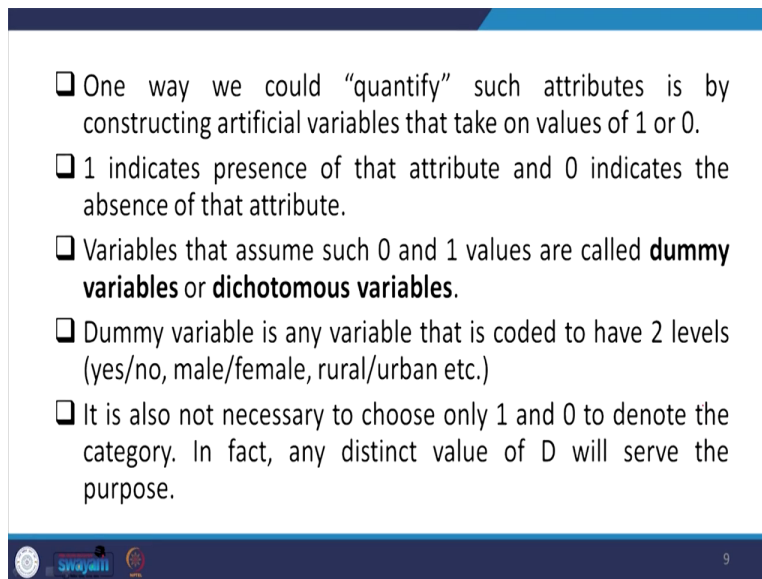
8

Coming to the dummy variable, we are trying to incorporate these variables into the model. And what is this dummy variable and how the codings are made we are going to discuss now. Dummy

coding is a way of incorporating nominal variables into analysis more or less understood as nominal variables, not as the ordinal variables.

Dummy coding is the process of translating information gathered from questionnaires or other sources into something that can be analyzed. Coding can make your data more consistent. Such variables usually indicate the presence or simply absence of a quality or an attributes.

(Refer Slide Time: 09:26)

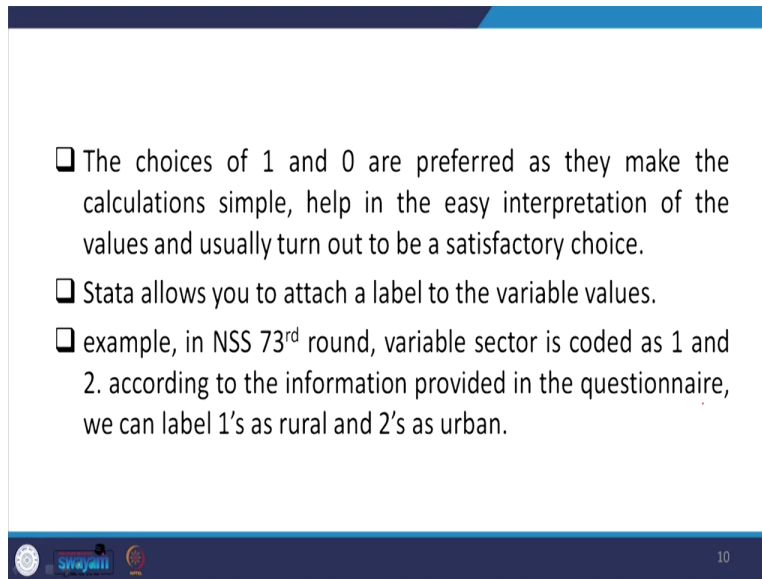


- ❑ One way we could “quantify” such attributes is by constructing artificial variables that take on values of 1 or 0.
- ❑ 1 indicates presence of that attribute and 0 indicates the absence of that attribute.
- ❑ Variables that assume such 0 and 1 values are called **dummy variables** or **dichotomous variables**.
- ❑ Dummy variable is any variable that is coded to have 2 levels (yes/no, male/female, rural/urban etc.)
- ❑ It is also not necessary to choose only 1 and 0 to denote the category. In fact, any distinct value of D will serve the purpose.

So, one way we could quantify such attributes is by constructing artificial variables that take on code like 1 and 0, like in binary form. So, 1 indicates presence, 0 indicates absence of an attributes. So, variable that assume such as 0 and 1 values are called dummy variables or dichotomous variables. So, dummy variable is any variable that is coded to have two levels basically, yes no type or male female type or rural, urban etc. So, usually it has two levels.

It is also not necessary to choose only 1 and 0. We may choose any code, but it should have two levels only. In fact, Any distinct value is going to serve the purpose. One has to be clearly separated from another level.

(Refer Slide Time: 10:39)



- ❑ The choices of 1 and 0 are preferred as they make the calculations simple, help in the easy interpretation of the values and usually turn out to be a satisfactory choice.
- ❑ Stata allows you to attach a label to the variable values.
- ❑ example, in NSS 73rd round, variable sector is coded as 1 and 2. according to the information provided in the questionnaire, we can label 1's as rural and 2's as urban.

The choices of 1 and 0 are preferred as they make the calculation simple and help in easy interpretation of the values and usually turns out to be satisfactory choice. Stata allows you to attach a label to the variable and their values if it is dummy. For example, in the National Sample Survey 73rd round on unorganized enterprises, we discussed variables such as sector with code 1 and 2 that represent rural and urban.

And according to the information provided in the questionnaire, we are saying that the level 1 stands for rural and 2 stands for urban. So, our dummy here, we can make any one of them as dummy. So, if coding is defined very correctly our interpretation for the dummy will be very good. We are going to discuss very clearly.

(Refer Slide Time: 12:22)

Without Label

Sector	Freq.	Percent	Cum.
1	35,766	49.31	49.31
2	36,762	50.69	100.00
Total	72,528	100.00	

With Label

Sector	Freq.	Percent	Cum.
rural	35,766	49.31	49.31
urban	36,762	50.69	100.00
Total	72,528	100.00	

Coming to the labeling and in the dummy if it is defined, if the variable gives information about the dummy in two categories so without label it is just coming, in the sector example case it is, if you derive the result, it only gives 1 or 2. But after labeling, like value labeling we already mentioned, value of the variable and their labeling we already defined. So, then if 1 and 2, if you label it, with 1 stands for rural and 2 stands for urban then accordingly we can have different interpretation. But the result remains same. Interpretation becomes easy if you label in that particular category of the dummy.

(Refer Slide Time: 12:37)

Coding for Nominal data with Multiple categories:

- ❑ Coding process is similar with other categorical variables. Such as nominal with more than two categories.
- ❑ For example, in NSS 73rd round, variable ownership type is coded as 1,2,3,4,5,6, and 7. according to the questionnaire, we can label 1's as male proprietary, 2's as female proprietary, 3's as transgender proprietary, 4's as enterprise with partners from the same household, 5's as enterprise with partners from different households, 6's as SHGs and 7's as trusts.

So, coding for nominal data with multiple categories if any, coding process is similar with other categorical variables such as nominal with more than two categories. In our example, variable ownership type is with the code available as 1, 2, 3, 4, 5, 6, 7. According to the questionnaire, we can label the values 1 for male proprietary, then 2 for female proprietary, 3 for transgender and so on. We have already discussed. You just go and find out.

(Refer Slide Time: 13:17)

Without Label

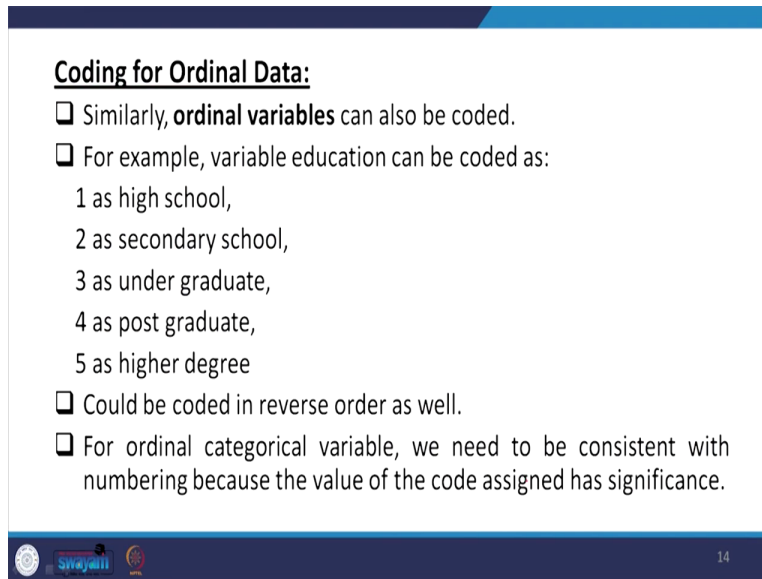
Type of ownership code	Freq.	Percent	Cum.
1	60,350	83.37	83.37
2	8,804	12.16	95.53
3	9	0.01	95.54
4	1,346	1.86	97.40
5	938	1.30	98.70
6	535	0.74	99.44
7	406	0.56	100.00
Total	72,388	100.00	

With Label

Type of ownership code	Freq.	Percent	Cum.
male_prop	60,350	83.37	83.37
female_prop	8,804	12.16	95.53
transgender_prop	9	0.01	95.54
samehhpartners	1,346	1.86	97.40
differnthpartners	938	1.30	98.70
SHG	535	0.74	99.44
trusts	406	0.56	100.00
Total	72,388	100.00	

Then I think these are the labeling we just discussed. Then 1 stands here for male, then 2 stands for female, then 3 transgender, and accordingly, 7 stands for trust, who defines to be the owner of the enterprise.

(Refer Slide Time: 13:33)



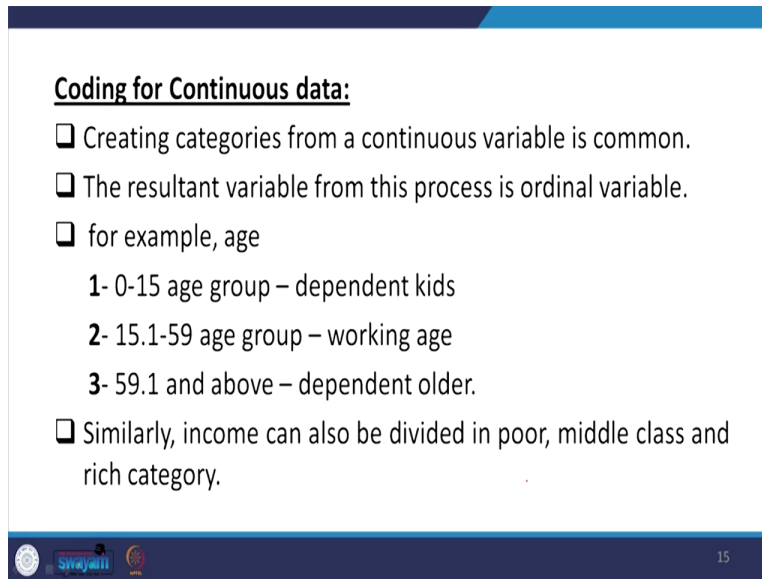
Coding for Ordinal Data:

- Similarly, **ordinal variables** can also be coded.
- For example, variable education can be coded as:
 - 1 as high school,
 - 2 as secondary school,
 - 3 as under graduate,
 - 4 as post graduate,
 - 5 as higher degree
- Could be coded in reverse order as well.
- For ordinal categorical variable, we need to be consistent with numbering because the value of the code assigned has significance.

14

Coming to the coding of ordinal data which is similarly ordinal variables can also be coded the way we coded the binary, different variable or the dummy variables. For example, like education case we discussed 1 can be for high school, 2 can be for secondary school and so on, 5 for the higher degree education achievement. So, the ordinal variable could be coded in reverse order as well depending upon the requirement of interpretation, so that can be coded in a reverse order. For ordinal categorical variable, we need to be consistent with numbering, because the value of the code assigned has certain significance. Higher the value, higher ordering is more justified.

(Refer Slide Time: 14:25)



Coding for Continuous data:

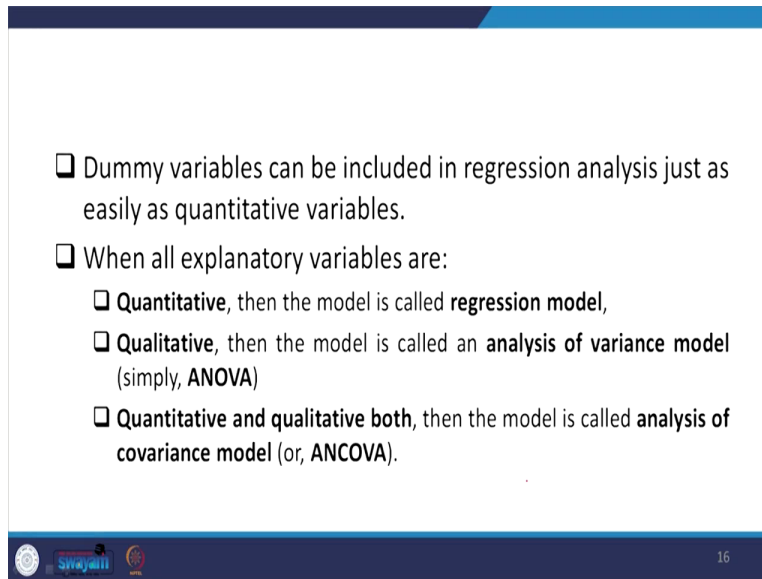
- ❑ Creating categories from a continuous variable is common.
- ❑ The resultant variable from this process is ordinal variable.
- ❑ for example, age
 - 1- 0-15 age group – dependent kids
 - 2- 15.1-59 age group – working age
 - 3- 59.1 and above – dependent older.
- ❑ Similarly, income can also be divided in poor, middle class and rich category.

15

So, in case of continuous data, creating categories from the continuous variable is common, very common because like expenditure of the person or age of the person we can categories into different groups, standard of living of the person that can be categorized into groups.

Here, for example, our continuous data is 0 to 15 age group, dependent kids or working age group. We can make them in different brackets. Like among all persons we categorized into three here from the continuous series as well. So, continuous can be categorized as well. That we dealt earlier also. Coming to the income, similarly, income can also be divided into poor, middle class and rich category.

(Refer Slide Time: 15:24)



❑ Dummy variables can be included in regression analysis just as easily as quantitative variables.

❑ When all explanatory variables are:

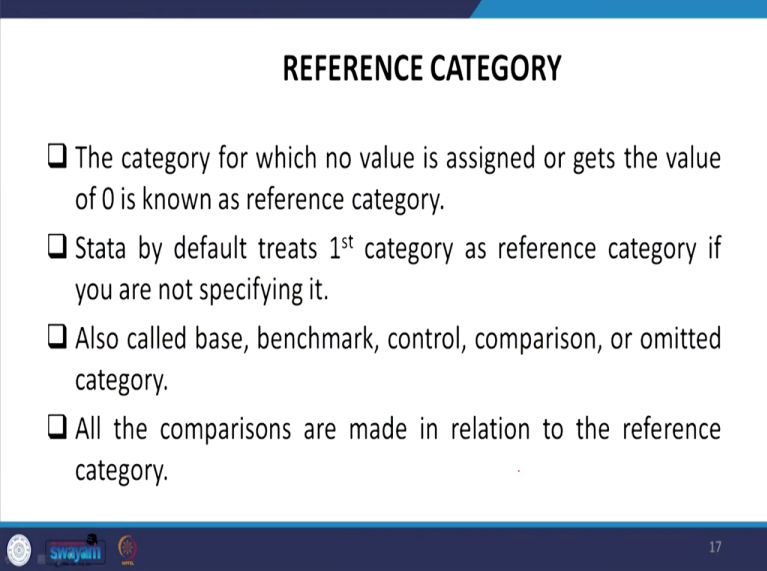
- ❑ **Quantitative**, then the model is called **regression model**,
- ❑ **Qualitative**, then the model is called an **analysis of variance model** (simply, **ANOVA**)
- ❑ **Quantitative and qualitative both**, then the model is called **analysis of covariance model** (or, **ANCOVA**).

16

Dummy variables can be included in regression analysis just as easily as quantitative variables. The categorical variables can also be included in regression analysis also as part of the explanatory variable. When all explanatory variables are quantitative, we are coming to a little new addition of the existing knowledge so far related to Stata and their model specification.

When all explanatory variables the regressors when they are all quantitative in nature those type of regression models are called regression models, where if it is full of qualitative variables then the analysis of those type of model dealing with the qualitative variables are called analysis of variance model or ANOVA. When it is of mix then mix of qualitative as well as quantitative analysis are called analysis of covariance, called ANCOVA model.

(Refer Slide Time: 17:14)



REFERENCE CATEGORY

- ❑ The category for which no value is assigned or gets the value of 0 is known as reference category.
- ❑ Stata by default treats 1st category as reference category if you are not specifying it.
- ❑ Also called base, benchmark, control, comparison, or omitted category.
- ❑ All the comparisons are made in relation to the reference category.

17

When we have defined different categories, different labeling of the variable, maybe it is dummy, categorical, ordered categorical or continuous then categorical, if you have made all those things, next question comes which one to be considered to be a reference category? Which should be considered as a base category so that the another category can be interpreted very correctly, like the category for which no value as assigned or gets the value of 0 is known as reference category.

So, Stata by default treats first category in the categorical variable as the reference category if it is not assigned in a different way. If you do not specifically mention the another category by our command, it by default considers first category as the base. So, this reference category is also called base category, benchmark category, control, comparison or omitted category. So, all the comparisons are made in relation to the reference category. We are going to discuss everything with the help of our data in our next class. In this lecture, I am specifically mentioning everything as a background to the understanding of qualitative variables.

(Refer Slide Time: 18:11)

- ❑ Must keep track of the reference category, if you have several dummy variables for better and clearer interpretation.
- ❑ For example, if we are interested in analysing women entrepreneurs' performance in different sectors (rural and urban). Or, more specifically interested in analysing how urban women entrepreneurs differ from rural women entrepreneurs. The sector variable is coded as rural and urban, we would select "rural" as the reference category on the nonmetric "sector" variable.

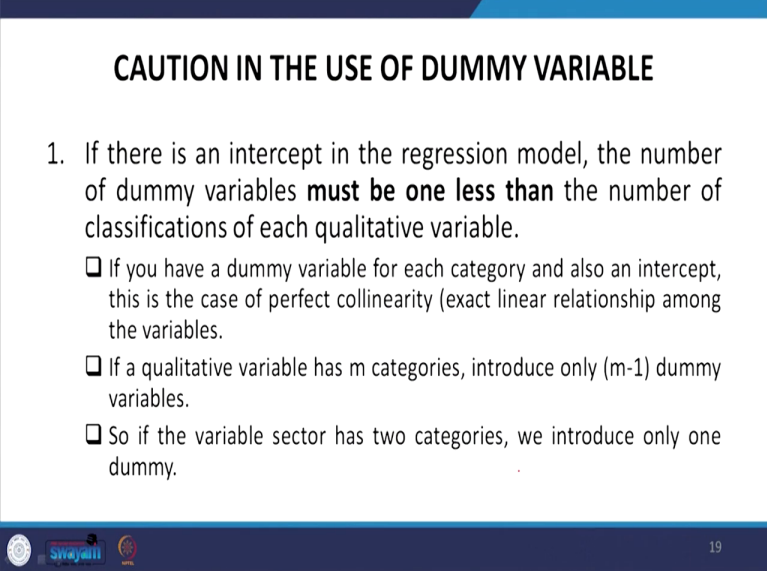


18

We must keep track of the reference category if you have several dummy variables for better and clearer interpretation. If there are so many dummy variables, reference category would be very clearly marked. Otherwise, the interpretation might be very confusing and misleading.

For example, if you are interested in analyzing women entrepreneurs' performance in different sector, in rural or urban or more specifically if you are interested in analyzing how urban women entrepreneurs differ from the rural women entrepreneurs, the sector variable is coded as rural and urban and we would select rural as the reference category on the non-metric sector variable.

(Refer Slide Time: 18:56)



CAUTION IN THE USE OF DUMMY VARIABLE

1. If there is an intercept in the regression model, the number of dummy variables **must be one less than** the number of classifications of each qualitative variable.
 - If you have a dummy variable for each category and also an intercept, this is the case of perfect collinearity (exact linear relationship among the variables).
 - If a qualitative variable has m categories, introduce only $(m-1)$ dummy variables.
 - So if the variable sector has two categories, we introduce only one dummy.

19

The caution here in the dummy variable case is that this should be very clearly noticed and clearly understood. This is quite intriguing. If there is an intercept in the regression model, a fixed portion, fixed coefficient in the model, the number of dummy variables must be less than the number of the classification of each qualitative variables. So, if the qualitative variable is classified into two categories, the dummy should not be two, it should be less than two that should be one dummy must be there. Otherwise, there will be some problems in the model. We are going to discuss that in our next slide.

In between, I just wanted to mention that, if you have a dummy variable for each category, also an intercept, this is the case of perfect collinearity. And why collinearity, we will explain with the help of example in our next couple of slides. So, if qualitative variable has m categories and we introduce m minus 1 dummy variables then that is I think perfectly fine. If the variable sector has two categories, we introduce only one dummy for our better analysis.

(Refer Slide Time: 20:50)

□ If we violate this rule, we will fall into what is called the **Dummy variable trap**.

Consider a model,

$$Y = \beta_0 + \beta_1 X + \beta_2 D_2 + \beta_3 D_3 + \epsilon$$

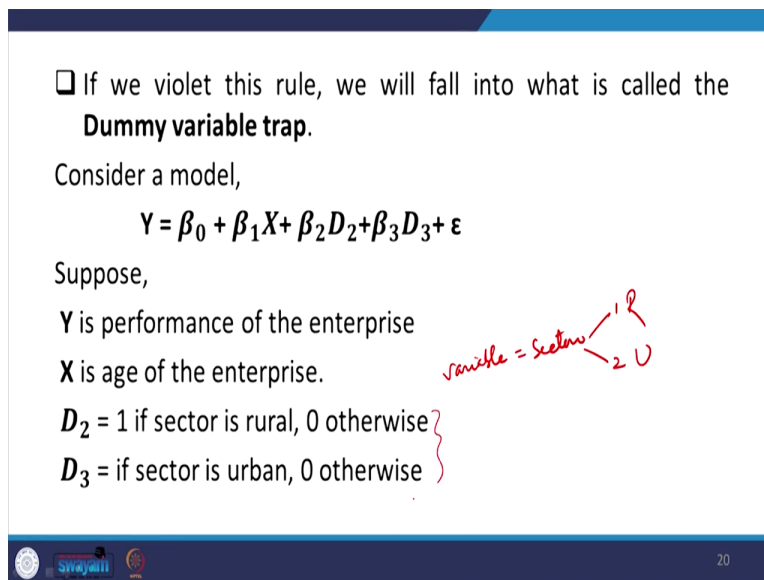
Suppose,

Y is performance of the enterprise

X is age of the enterprise.

$D_2 = 1$ if sector is rural, 0 otherwise }
 $D_3 = 1$ if sector is urban, 0 otherwise }

variable = Sector
1 R
2 U



Let us explain what is the problem of defining the exact number of dummy equivalent to the number of or equalized with the number of categories in the qualitative variable. We will be facing a problem that is called dummy variable trap. So, it is very famously understood in econometrics model.

What do you mean by dummy variable trap? Let us consider this model. This model suggests that any model of dependent and independent variable, if you have a dummy and dummy having two categories and we have defined two different dummies for the two categories of that variable and first dummy is D_2 and second dummy is D_3 , along with another explanatory variable and also intercept. And how it is going to create problem?

Here our explanatory variable, our dependent variable is performance of the enterprise depends on X is our age of the enterprise, how old the enterprise is. Then dummy variables are like rural and urban sector. Variable name here is sector. So, that is categorized rural and urban. So, this is rural, this is urban.

If you define that D_2 is equal to 1 if sector is rural. If the sector is rural then you define 1 or 0 otherwise. So, 1, 0 if you define like this, and in D_3 another dummy variable you defined as 1 if it is urban or 0 otherwise. It is just the reverse to each other and how these are creating problems of perfect multicollinearity we are going to explain.

(Refer Slide Time: 22:20)

The above model cannot be estimated because of perfect collinearity between D_2 and D_3 .

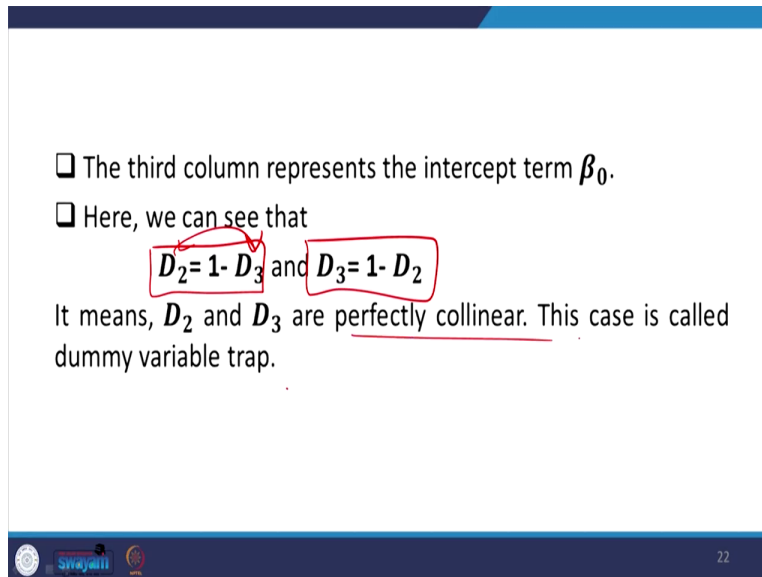
- Suppose we have a sample of 2 rural enterprise and 3 urban. The matrix can be designed as:

		β_0	D_2	D_3	X
rural	Y_1	1	1	0	X_1
urban	Y_2	1	0	1	X_2
rural	Y_3	1	1	0	X_3
urban	Y_4	1	0	1	X_4
urban	Y_5	1	0	1	X_5

So, like here these two is our matter of concern. So, either like if it is rural then we have said 1, otherwise it is 0. Then if it is urban it will be 0. Then it will be, in case of D3 it is 1. So, basically, in all the cases it is simply the opposite, then, along with the explanatory variables we have. So, basically, we are explaining the same thing twice. We are explaining the same thing, same information in a reverse manner.

So, that means there is a perfect linearity between two variables. Two variables are perfectly correlated to each other. So, on the average these are same. So, these two are correlated to each other and Stata result is going to be misleading. So, we will be in a trap.

(Refer Slide Time: 23:33)



□ The third column represents the intercept term β_0 .

□ Here, we can see that

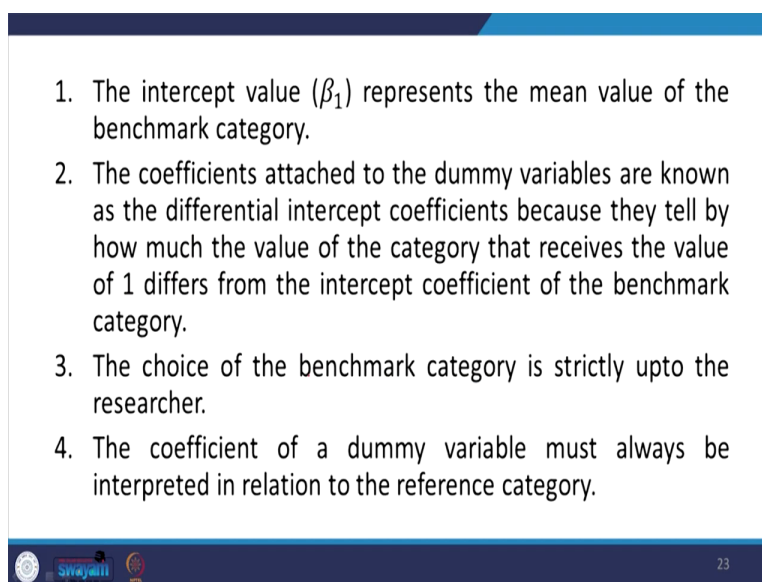
$D_2 = 1 - D_3$ and $D_3 = 1 - D_2$

It means, D_2 and D_3 are perfectly collinear. This case is called dummy variable trap.

22

How to mention that? Basically, what is this? How they are related? D_3 nothing but 1 minus D_2 , D_2 is nothing but 1 minus D_3 , which we just said. 1 is basically 1 minus 0 or 0 is equal to 1 minus 1. So, this is just the opposite. So, that means these two are linearly related to each other. These are D_3 and D_2 are perfectly linearly related to each other. So, that is the reason why the estimation is going to be problematic and due to perfect collinearity and this is called the dummy variable trap.

(Refer Slide Time: 24:13)



1. The intercept value (β_1) represents the mean value of the benchmark category.
2. The coefficients attached to the dummy variables are known as the differential intercept coefficients because they tell by how much the value of the category that receives the value of 1 differs from the intercept coefficient of the benchmark category.
3. The choice of the benchmark category is strictly upto the researcher.
4. The coefficient of a dummy variable must always be interpreted in relation to the reference category.

23

The intercept value represents the mean value of the benchmark category. The coefficients attached to the dummy variables are known as the differential intercept coefficients because they tell by how much the value of the category that receives the value of 1 differs from the intercept category, coefficient of the benchmark category. The choice of the benchmark category is strictly up to the researcher. We already mentioned. The coefficient of a dummy variable must always be interpreted in relation to reference category.

(Refer Slide Time: 24:49)

ESTIMATION WITH DUMMY VARIABLES

Intercept Dummy Variable

□ Consider the following model with X_1 as quantitative and D_2 as an indicator variable

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 D_2 + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$$

$$D_2 = \begin{cases} 0 & \text{if an observation belongs to group A} \\ 1 & \text{if an observation belongs to group B} \end{cases}$$

If $D_2 = 0$, then

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 \cdot 0 + \varepsilon$$

Let us understand some estimation with dummy variables. How we can be able to understand the estimations, dummy variable models? So, we are now presenting here the intercept dummy variables. What is this? The model with X_1 as the quantity variable and D_2 is an indicator variable. In this case, this is our dummy variable and we have not included both as dummy because it has two categories, so only one dummy is mentioned correctly.

So, in this case what the dummy variable and how it is interpreted, it is in 0 and 1 form. It is equal to 0 if the observation belongs to that group A or 1 if it belongs to another group. So, if D_2 is equal to 0 either of the option in the model, if it is equal to 0 then what is left, it is basically, this is left with the error term.

Then otherwise if it is 1 then only another, like the slope remains, if this equal to 1 then this one, this will be 1, this one plus β_2 , only intercept shifted by β_2 , intercept increased by β_2 , but the

slope remains constant. If we exercise by D_2 as the dummy having 0 and 1 what is the result here? Only intercept is changed by the extra coefficient of the dummy and the slope remain constant. This is what we are explaining.

(Refer Slide Time: 26:30)

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

$$E(Y | D_2 = 0) = \beta_0 + \beta_1 X_1$$
 Straight line relationship between β_0 and β_1 .
 If $D_2 = 1$, then

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 \cdot 1 + \varepsilon$$

$$= (\beta_0 + \beta_2) + \beta_1 X_1 + \varepsilon$$

$$E(Y | D_2 = 1) = (\beta_0 + \beta_2) + \beta_1 X_1$$
 straight-line relationship with intercept $(\beta_0 + \beta_2)$ and slope β_1 .

So, Y given the expected value of the dependent variable with respect to dummy variable or conditional upon the dummy variable if it is 0 then we get the expected value as $\beta_0 + \beta_1 X_1$. So, there exists a straight line relationship between β_0 and β_1 .

If it is 1, there still exists, as I just said, still exists, this is the one we inserted, still exists a linearly relationship, but only difference is that our intercept is increased by β_2 but slope remain constant. So, that is what the difference when we go by a dummy variable model.

(Refer Slide Time: 27:17)

The quantities $E(Y|D_2 = 0)$ and $E(Y|D_2 = 1)$ are the average responses when an observation belongs to group A and group B, respectively. Thus

$$\beta_2 = E(Y|D_2 = 1) - E(Y|D_2 = 0)$$

Difference between the average values of Y with $D_2 = 0$ and $D_2 = 1$.

This type of regression is called parallel regression because the slope remains the same but intercept changes.



26

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

$$E(Y|D_2 = 0) = \beta_0 + \beta_1 X_1$$

Straight line relationship between β_0 and β_1 .

If $D_2 = 1$, then

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 \cdot 1 + \varepsilon$$

$$= (\beta_0 + \beta_2) + \beta_1 X_1 + \varepsilon$$

$$E(Y|D_2 = 1) = (\beta_0 + \beta_2) + \beta_1 X_1$$

straight-line relationship with intercept $(\beta_0 + \beta_2)$ and slope β_1 .



25

ESTIMATION WITH DUMMY VARIABLES

Intercept Dummy Variable

- Consider the following model with X_1 as quantitative and D_2 as an indicator variable

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 D_2 + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$$

$$D_2 = \begin{cases} 0 & \text{if an observation belongs to group A} \\ 1 & \text{if an observation belongs to group B} \end{cases}$$

If $D_2 = 0$, then

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 \cdot 0 + \varepsilon$$

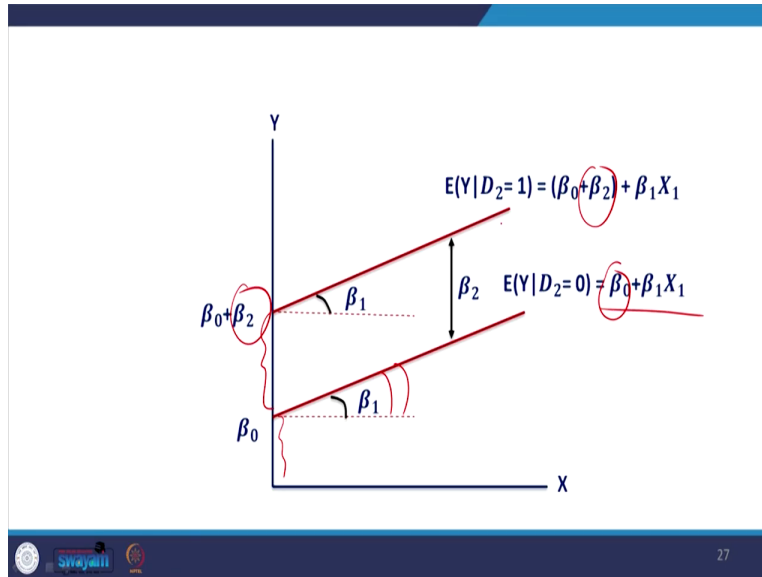


So how to estimate that beta 2 then, beta 2, the slope how to get it basically? how to get the beta 2 in this case because all are same? So, that means the expected value of Y given D2 is equal to 1 minus expected value of Y given D2 is equal to 0. So, once we subtract, it we will get the value of beta 2, isn't it?

So, in our original model the beta 2 can be estimated very clearly. Beta 2 if you wanted to estimate, the dummy variable and its coefficient if you want to estimate, only we need to subtract these two we will get the value of beta 2.

This type of regression analysis is called parallel regression, because the projected trend line of both the models with the dummy, different type of dummies and its coefficients result in two parallel regression lines and because the slope remains the same but intercept changes. This is what in the picture.

(Refer Slide Time: 28:39)



So, in the first one this is what beta not then beta 1 is the slope, the angle and this is changed by beta 2, so the vertical intercept is changed by beta 2, rest are same. Since these are same, so there will be parallel distance or the vertical distance remains constant.

(Refer Slide Time: 29:04)

Slope Dummy Variable

- Consider the following model with X_1 as quantitative and D_2 as an indicator variable.
- Suppose both **interact** and an explanatory variable as the interaction of them is added to the model.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 D_2 + \beta_3 X_1 D_2 + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$$
$$D_2 = \begin{cases} 0 & \text{if an observation belongs to group A} \\ 1 & \text{if an observation belongs to group B} \end{cases}$$

The slide number 28 is in the bottom right corner.

First we discussed intercept dummy variables, where the dummy variable changes the intercept and that type of regression is called parallel regression model. We are going to discuss a

dissimilar regression model. Dissimilar regression model which introduced the slope coefficient to be different and slope coefficient gets differentiated if you introduce some interact dummy.

Interact dummy, what do you mean by that? Like here X_1 is the quantity variable and D_2 is indicator variable. Suppose both interact and explanatory variable as the interaction of them added to the model. For example, here we wanted to say that this one was already there, till this it was there, this one is also there. We are only introducing a new term having interaction with both of them. Dummy is interacted with the explanatory variables. There is some relationship between these two. So, we are cleverly mentioning them together.

If we estimate the expected value of Y given the dummy or beta 2, if you wanted to estimate the beta 2, then how to do that? We simply have to put the value of D_2 dummy as 0 and 1.

(Refer Slide Time: 30:29)

□ Then,

$$E(Y|D_2=0) = Y = \beta_0 + \beta_1 X_1 + \beta_2 \cdot 0 + \beta_3 X_1 \cdot 0 + \varepsilon$$
$$= \beta_0 + \beta_1 X_1$$

The straight line with intercept β_0 and slope β_1 .

If $E(Y|D_2=1) = Y = \beta_0 + \beta_1 X_1 + \beta_2 \cdot 1 + \beta_3 X_1 \cdot 1 + \varepsilon$

$$= (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_1$$

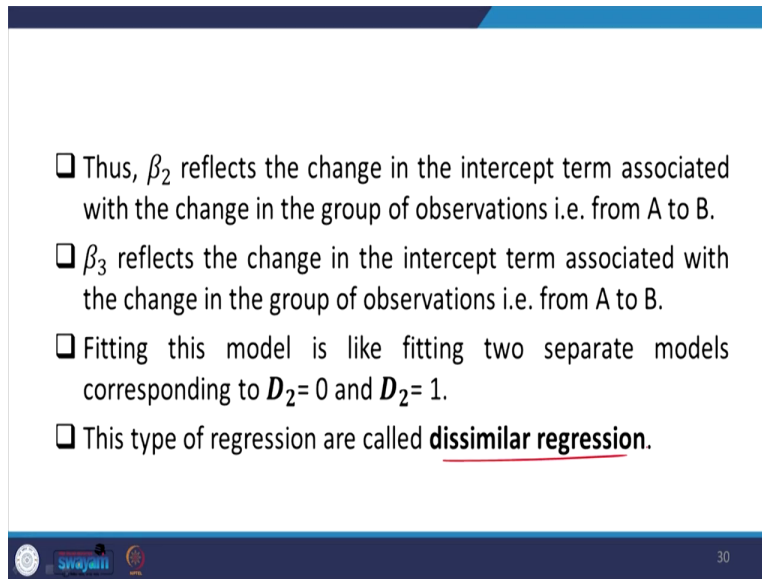
The straight line with intercept $(\beta_0 + \beta_2)$ and slope $(\beta_1 + \beta_3)$.

□ So the model has different slopes and different intercept terms.

29

In this case only when it is 0, again the same expected equation we are getting, but when it is 1, there is a change in the slope coefficient also. These are same as per the earlier model. Slope coefficient changed. So, that means our two regression lines are going to be different and they are not going to be parallel.

(Refer Slide Time: 30:53)

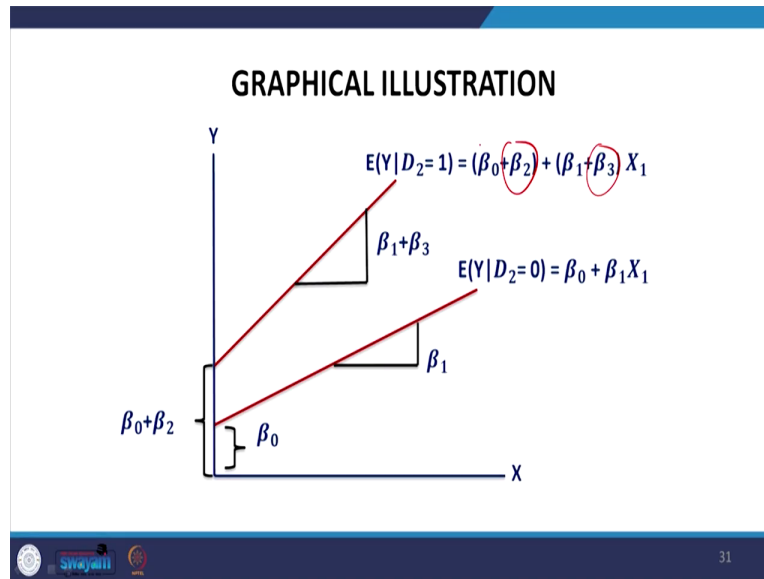


- ❑ Thus, β_2 reflects the change in the intercept term associated with the change in the group of observations i.e. from A to B.
- ❑ β_3 reflects the change in the intercept term associated with the change in the group of observations i.e. from A to B.
- ❑ Fitting this model is like fitting two separate models corresponding to $D_2=0$ and $D_2=1$.
- ❑ This type of regression are called dissimilar regression.

And so that is the reason why they are called dissimilar regression models. And so, beta 2 reflects the change in intercept term associated with the change in the group of observations. Beta 3 also reflects the change in the intercept term associated with the change in the group of observations from A to B.

So, fitting this model is like fitting two separate models corresponding to D_2 is equal to 0 and D_2 is equal to 1, and therefore, we end up with dissimilar trend lines. That is why this is called dissimilar regression.

(Refer Slide Time: 31:26)



So, slope is also different and different by this. Intercept was already proved to be different due to interaction dummies. So, that is all about the discussion in this lecture. So, we carry forward our discussion of the qualitative variables to the next class, especially here we introduce the qualitative variables in the model, in the explanatory model, explanatory variable case. But later from the next class onwards, we are going to discuss the dependent variable to be qualitative. So, there are four models broadly we will introduce here and we will largely stick to the binary kind of dependent variables.

So, let me stop here. We will continue in the next class. Thank you.