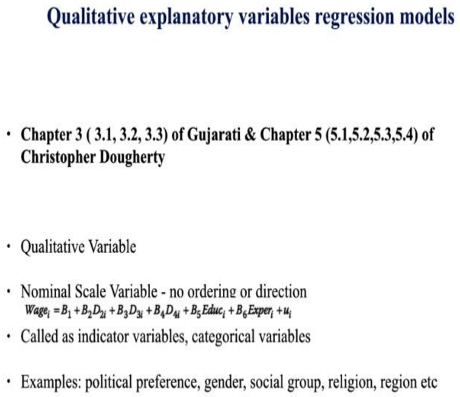
Applied Econometrics
Prof. Tutan Ahmed
Vinod Gupta School of Management
Indian Institute of Technology - Kharagpur

Module - 7
Lecture - 57
Dummy Variable

Hello and welcome back to the lecture on Applied Econometrics. So, in today's lecture, we are going to talk about dummy variable. Now, in the previous lecture where we spoke about the functional specification, we have seen the different variables and how these different variables are included in a regression equation. And the kind of data we collect in economics, in the domain of economics or public policy, we will see at the end of the day, most of the variables are basically qualitative variable or they are categorical variable. And what we mean by qualitative variable?

(Refer Slide Time: 01:02)



Qualitative explanatory variables regression models

- **Chapter 3 (3.1, 3.2, 3.3) of Gujarati & Chapter 5 (5.1,5.2,5.3,5.4) of Christopher Dougherty**
- Qualitative Variable
- Nominal Scale Variable - no ordering or direction
  $Wage_i = B_1 + B_2D_{2i} + B_3D_{3i} + B_4D_{4i} + B_5Educ_i + B_6Exper_i + u_i$
- Called as indicator variables, categorical variables
- Examples: political preference, gender, social group, religion, region etc

So, qualitative variable means the variables which really do not have any specific number assigned, or you cannot really express them quantitatively. For example, if you see a male and female, now how exactly you will value male and female? So, they are 2 categories; that is it. Or if you think about other variables, for example political preferences; so, if am a left or I am a rightist or I am a centrist, so, how do you really assign values to them, like, how do you really order them?

You cannot really order them. So, they are just different categories; that is it. Or if it is social group, like, social group in India could be like General or Scheduled Caste, Scheduled Tribe or OBC. So, you cannot really assign a specific value to each of these different categories. Or if you think about religion, Hindu, Muslim, Buddhists, Jains, Christians, Persians, and so many different religions you have.

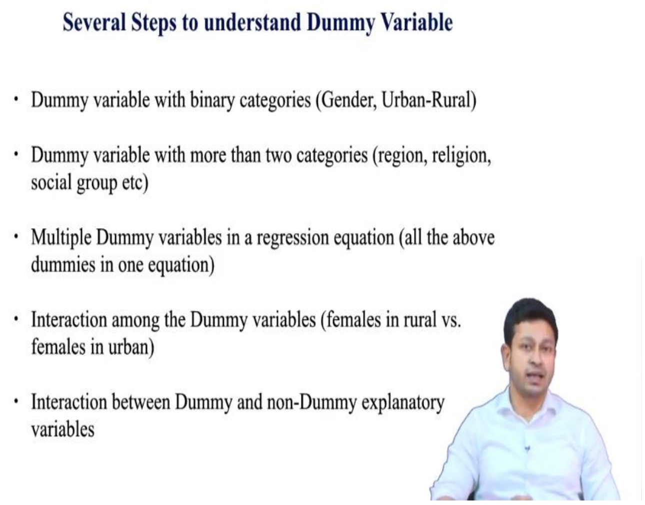
Now, how do you really assign a value to them? You cannot really assign a value to them. And interestingly, we will see in a while that most of the variables that we collect data on, from the field, are actually categorical variable. And all these different types of example that I have given of the qualitative variable, these are called dummy variable when you include them in a regression equation.

So, in this lecture, we are going to see how to really include these variables in a regression equation and what are the different types of dummy variable we may find and what are the different tests we might have to do to understand which dummy variables to include in our regression equation. As I said, it is a qualitative variable. And also it is called a nominal scale variable, because there is no ordering done, there is no direction.

And it is also called indicator variable. And because it represents different categories, it is called categorical variable. And we will see that because of this categorical nature, because of this qualitative nature of this variable, always how we enumerate the impact of one dummy variable is with respect to one reference dummy variable. So, there is a fixed category, and you always try to measure the impact of other categories vis-a-vis the fixed category.

For example, if I include a dummy variable called gender, and if I have like my fixed category is male, then I try to see the impact of being female on the outcome variable. So, we will just see all these different concepts in all these different categories of dummy variable that we are going to see.

(Refer Slide Time: 03:31)



Several Steps to understand Dummy Variable

- Dummy variable with binary categories (Gender, Urban-Rural)
- Dummy variable with more than two categories (region, religion, social group etc)
- Multiple Dummy variables in a regression equation (all the above dummies in one equation)
- Interaction among the Dummy variables (females in rural vs. females in urban)
- Interaction between Dummy and non-Dummy explanatory variables

Now, I will divide the lecture into different small parts. So, for example, first we will talk about most basic simple, a dummy variable that is a binary dummy variable. Let us say it is only of 2 categories; so, it could be male, female; it could be rural, urban; it could be white, non-white; so, these are the categories where there is binary categories. So, you have one category and you can basically measure the impact of the other category vis-a-vis the first category.

So, you can define how you actually want to choose the first category. So, we will talk about that in detail. Then we can have more than 2 categories. The examples that I have given, if it is a social group, so, you have SC, ST, OBC, General; at least these 4 social categories we have in India. Now, then, you do not have a binary category; so, you have like 4 categories. And 4 categories essentially talking about the same population, as in like the overall same population, these are different sub-populations within the overall population.

And how do you really show the sub-population vis-a-vis same population, when I have more than 2 sub-populations? You can also think of like people living in different region. Some people can live in the western part of India; some people can stay in the southern part of India; some people can stay in the middle of the India; some people can stay in the eastern part of India; some people can stay in the northern part of India.

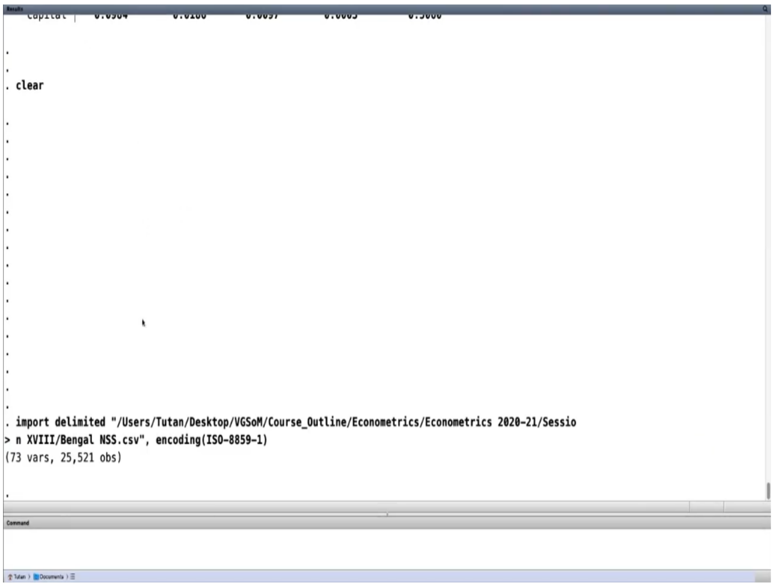
Now, all these regional aspects can actually influence your outcome. So, whatever outcome variable you have, so, these regions can influence your outcome variable. Now, how do you really represent these different regions. So, then again it comes in the category where we have more than 2 categories for the dummy variable, for the specific variable where you are actually trying to capture the impact of different regions in India, let us say.

Now, I may want to include all the category 1 and category 2 in 1 regression equation. So, when I do that, how do I really interpret the regression equation? That also, we are going to talk about. Then, we will talk about slightly different type of dummy variable, where we actually take into account the interaction of different dummy variables. For example, if I want to see the impact of being female and being in the city.

So, a female who lives in the city; so, that is an interaction dummy variable. I want to see the joint effect of females in urban India vis-a-vis females in rural India. Then I will talk about another type, where the dummy variable may interact with a non-dummy variable. So, let us say female; someone is a female, but I want to understand the impact of the female when they actually take into account the education part.

So, let us say I want to understand the education of female and how that is impacting the wage. So, female education and wage. So, I am trying to understand the joint effect of someone being female and their education, and how that is influencing the wage. Obviously, we will see this category when we talk further on the dummy variable part, different types of dummy variable. Now, let me actually show you the data or data sets, so that we can actually understand how a dummy variable looks like.

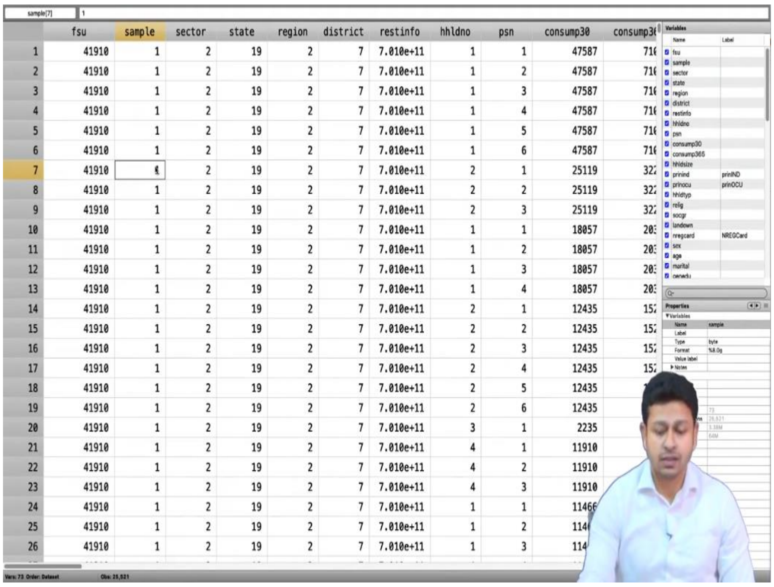
(Refer Slide Time: 06:57)



```
    capital |   0.0904      0.0180      0.0097      0.0005      0.5000

.
.
. clear

.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
. import delimited "/Users/Tutan/Desktop/VGSoM/Course_Outline/Econometrics/Econometrics 2020-21/Sessio
> n XVIII/Bengal NSS.csv", encoding(ISO-8859-1)
(73 vars, 25,521 obs)

.
```

So, I already have imported this data set. So, this is from national sample survey data, NSS; and this is for West Bengal. So, we got all the values of all the different fields from national sample survey data for the state West Bengal. And this is a sixty-eighth round national sample survey data. And let me first show you the data set and how it looks like.

(Refer Slide Time: 07:30)



sample[7] 1

	fsu	sample	sector	state	region	district	restinfo	hhldno	psn	consump30	consump3
1	41910	1	2	19	2	7	7.010e+11	1	1	47587	71
2	41910	1	2	19	2	7	7.010e+11	1	2	47587	71
3	41910	1	2	19	2	7	7.010e+11	1	3	47587	71
4	41910	1	2	19	2	7	7.010e+11	1	4	47587	71
5	41910	1	2	19	2	7	7.010e+11	1	5	47587	71
6	41910	1	2	19	2	7	7.010e+11	1	6	47587	71
7	41910	1	2	19	2	7	7.010e+11	2	1	25119	32
8	41910	1	2	19	2	7	7.010e+11	2	2	25119	32
9	41910	1	2	19	2	7	7.010e+11	2	3	25119	32
10	41910	1	2	19	2	7	7.010e+11	1	1	18057	20
11	41910	1	2	19	2	7	7.010e+11	1	2	18057	20
12	41910	1	2	19	2	7	7.010e+11	1	3	18057	20
13	41910	1	2	19	2	7	7.010e+11	1	4	18057	20
14	41910	1	2	19	2	7	7.010e+11	2	1	12435	15
15	41910	1	2	19	2	7	7.010e+11	2	2	12435	15
16	41910	1	2	19	2	7	7.010e+11	2	3	12435	15
17	41910	1	2	19	2	7	7.010e+11	2	4	12435	15
18	41910	1	2	19	2	7	7.010e+11	2	5	12435	
19	41910	1	2	19	2	7	7.010e+11	2	6	12435	
20	41910	1	2	19	2	7	7.010e+11	3	1	2235	
21	41910	1	2	19	2	7	7.010e+11	4	1	11910	
22	41910	1	2	19	2	7	7.010e+11	4	2	11910	
23	41910	1	2	19	2	7	7.010e+11	4	3	11910	
24	41910	1	2	19	2	7	7.010e+11	1	1	11466	
25	41910	1	2	19	2	7	7.010e+11	1	2	114	
26	41910	1	2	19	2	7	7.010e+11	1	3	114	

Vars: 73 Order: Dataset Obs: 25,521

Name	Label
fsu	
sample	
sector	
state	
region	
district	
restinfo	
hhldno	
psn	
consump30	
consump365	
hhldsize	
prinind	prinIND
prinocu	prinOCU
hhldtyp	
relig	
socgr	
landown	
nregcard	NREGCard
sex	
age	
marital	
genedu	

Name sample
Label
Type byte
Format %8.0g
Value label

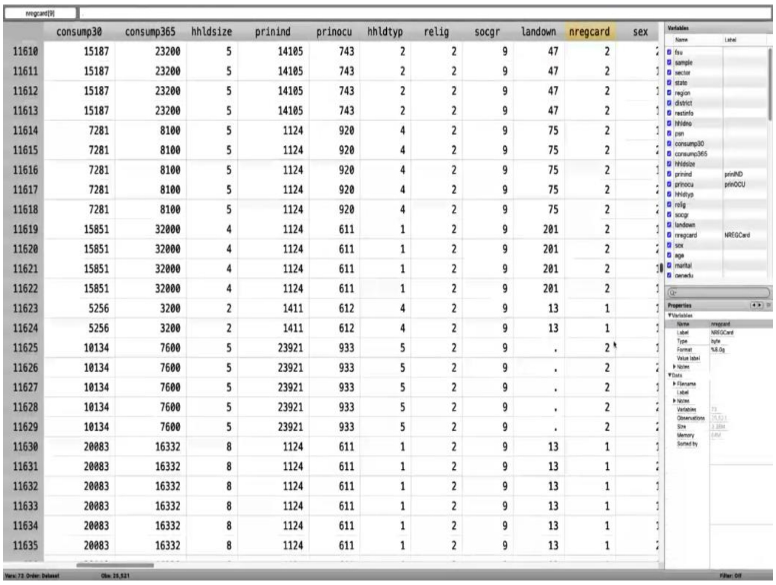
This is just to give you an idea how important dummy variable is when we actually collect data from the field. So, these are the different fields I have. And I will explain these fields to give you an idea about the importance of dummy variable. So, these are all these different fields; maybe there are 70, 80 fields that you have when you actually collect data from the field. So, this fsu, sample, these are all some numbers that we get to identify an individual; so, we do not need to bother about that.

Sector: Sector is basically representing rural and urban, and it has a value and NSS will show you 2 and 1. So, it is showing rural and urban. So, either it is 2 or it is 1. See, here it is all 1. It was 2 previously. State is 19. So, for all the different states, we have like all these qualitative numbers. So, 19 means West Bengal. So, the states will vary from 1 to 34, 35. This is again a qualitative variable. Then you have region.

So, regions are basically bigger unit than a district and a smaller unit than a state. And you have, within a state, there are several regions. So, these again are categorical variables; 2, 3, you will see; the values of region would be 2, 3, 5 and so forth. District: Definitely it is a categorical variable, because I have say 18 or 19 districts in West Bengal. So, you have all the numbers assigned to the district 16, 17, 19, 18, 7, 4, 1, 2.

Then you have some identification number. You do not really have to bother about that. Household number: It is again just a number. So, it is a only category. Personal serial number: This is again a category. We do not really need to bother about that, because we essentially are trying to get an aggregate effect; so, not really at household level or something.

(Refer Slide Time: 09:34)



	consump30	consump365	hhldsize	prinind	prinocu	hhldtyp	relig	socgr	landown	nregcard	sex
11610	15187	23200	5	14105	743	2	2	9	47	2	2
11611	15187	23200	5	14105	743	2	2	9	47	2	1
11612	15187	23200	5	14105	743	2	2	9	47	2	1
11613	15187	23200	5	14105	743	2	2	9	47	2	1
11614	7281	8100	5	1124	920	4	2	9	75	2	1
11615	7281	8100	5	1124	920	4	2	9	75	2	2
11616	7281	8100	5	1124	920	4	2	9	75	2	1
11617	7281	8100	5	1124	920	4	2	9	75	2	2
11618	7281	8100	5	1124	920	4	2	9	75	2	2
11619	15851	32000	4	1124	611	1	2	9	201	2	1
11620	15851	32000	4	1124	611	1	2	9	201	2	2
11621	15851	32000	4	1124	611	1	2	9	201	2	1
11622	15851	32000	4	1124	611	1	2	9	201	2	1
11623	5256	3200	2	1411	612	4	2	9	13	1	1
11624	5256	3200	2	1411	612	4	2	9	13	1	1
11625	10134	7600	5	23921	933	5	2	9	.	2	1
11626	10134	7600	5	23921	933	5	2	9	.	2	2
11627	10134	7600	5	23921	933	5	2	9	.	2	1
11628	10134	7600	5	23921	933	5	2	9	.	2	2
11629	10134	7600	5	23921	933	5	2	9	.	2	2
11630	20083	16332	8	1124	611	1	2	9	13	1	1
11631	20083	16332	8	1124	611	1	2	9	13	1	2
11632	20083	16332	8	1124	611	1	2	9	13	1	1
11633	20083	16332	8	1124	611	1	2	9	13	1	1
11634	20083	16332	8	1124	611	1	2	9	13	1	1
11635	20083	16332	8	1124	611	1	2	9	13	1	2

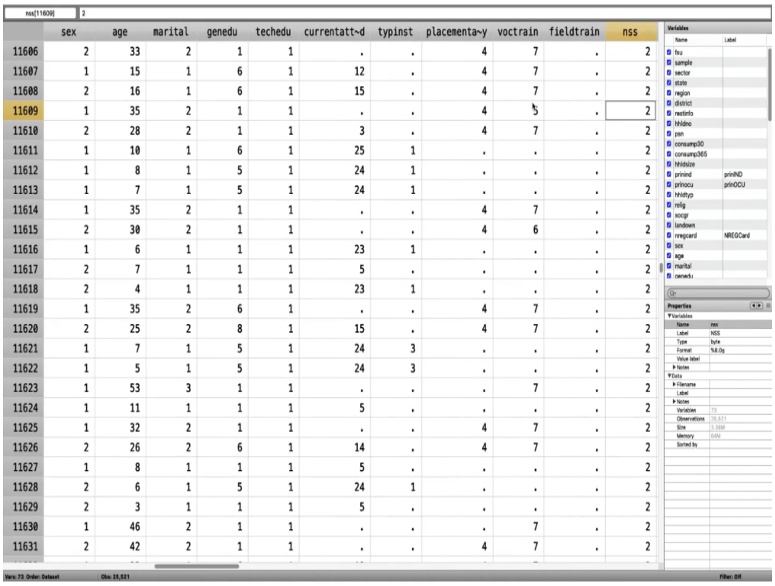
Consumption for 30 days and consumption for 365 days: These are numerical variables; these are quantitative variables. So, for the first time, we are seeing something which is a quantitative variable. So, it represents some, in monetary term; so, how much consumption was done for 30 days or how much consumption was done for 365 days. Household size: That is a quantitative variable; number of people living in a household.

Then we have this industry code and then occupation code. These are only just codes. So, these are all categorical variables. Then household types: What type of household based, depending on the major occupation of the household owner. We create different categories for household types. So, this again is a categorical variable. Then we have religion. So, religion, we have like values 1, 2, 3, so forth.

So, this is basically Hindu, Muslim and other religions. Then you have social group. Social group, as we said, this is a categorical variable. So, it could be 1, 2, 3, 9. So, there are 4 categories in national sample survey. So, it is basically General, Scheduled Caste, Scheduled Tribe and OBC. Then land ownership is a quantitative variable. So, how many hectares or acres of land you have.

MN rega card: If someone is having MN rega card; so, it will either be 0 or 1; or basically it has 2 values, 1 or 2 here. So, this again is a dummy variable. Then we have gender or sex. That is again a dummy variable; so, 1 and 2.

(Refer Slide Time: 11:11)

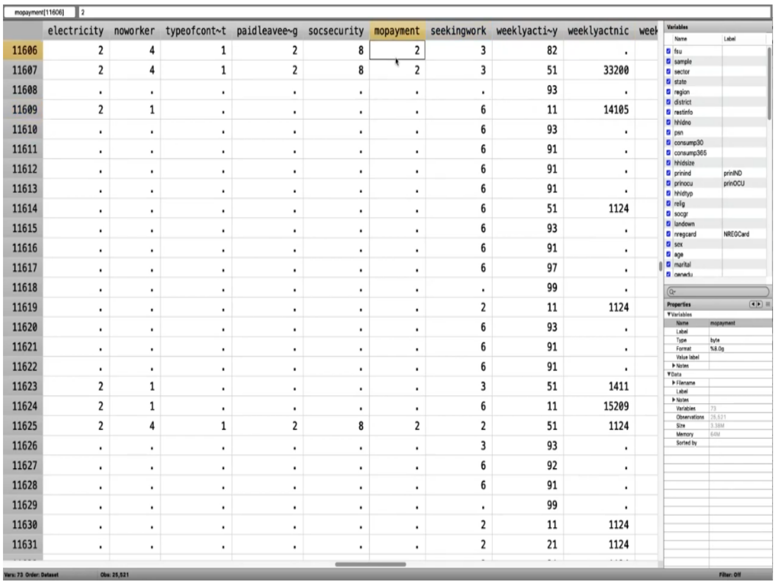


	sex	age	marital	genedu	techedu	currentatt~d	typinst	placementa~y	voctrain	fieldtrain	nss
11606	2	33	2	1	1	.	.	4	7	.	2
11607	1	15	1	6	1	12	.	4	7	.	2
11608	2	16	1	6	1	15	.	4	7	.	2
11609	1	35	2	1	1	.	.	4	5	.	2
11610	2	28	2	1	1	3	.	4	7	.	2
11611	1	10	1	6	1	25	1	.	.	.	2
11612	1	8	1	5	1	24	1	.	.	.	2
11613	1	7	1	5	1	24	1	.	.	.	2
11614	1	35	2	1	1	.	.	4	7	.	2
11615	2	30	2	1	1	.	.	4	6	.	2
11616	1	6	1	1	1	23	1	.	.	.	2
11617	2	7	1	1	1	5	.	.	.	.	2
11618	2	4	1	1	1	23	1	.	.	.	2
11619	1	35	2	6	1	.	.	4	7	.	2
11620	2	25	2	8	1	15	.	4	7	.	2
11621	1	7	1	5	1	24	3	.	.	.	2
11622	1	5	1	5	1	24	3	.	.	.	2
11623	1	53	3	1	1	.	.	.	7	.	2
11624	1	11	1	1	1	5	.	.	.	.	2
11625	1	32	2	1	1	.	.	4	7	.	2
11626	2	26	2	6	1	14	.	4	7	.	2
11627	1	8	1	1	1	5	.	.	.	.	2
11628	2	6	1	5	1	24	1	.	.	.	2
11629	2	3	1	1	1	5	.	.	.	.	2
11630	1	46	2	1	1	.	.	.	7	.	2
11631	2	42	2	1	1	.	.	4	7	.	2

So, it represents male and female. Age is a quantitative variable. You have age. Marital status: Married, unmarried, widowed, divorced; so, these are all categorical variable. General education: You have number of years of education; so, that is a quantitative variable. And here you have technical education. Technical education: One can have Polytechnic degree, one can have engineering degree, one can have medical degree; so, these are all again categorical variable.

So, these are current attendance status. So, that is basically, if someone is actually attending school or not attending school. These are again a categorical variable. So, actually, even these variables, I am not explaining all these different variables, but most of the variables, you will see, these are all categorical variable.

(Refer Slide Time: 12:02)



mopayment[11606] 2

	electricity	noworker	typeofcont~t	paidleavee~g	socsecurity	mopayment	seekingwork	weeklyacti~y	weeklyactnic	week
11606	2	4	1	2	8	2	3	82	.	
11607	2	4	1	2	8	2	3	51	33200	
11608	.	.	.	.	.	.	.	93	.	
11609	2	1	.	.	.	.	6	11	14105	
11610	.	.	.	.	.	.	6	93	.	
11611	.	.	.	.	.	.	6	91	.	
11612	.	.	.	.	.	.	6	91	.	
11613	.	.	.	.	.	.	6	91	.	
11614	.	.	.	.	.	.	6	51	1124	
11615	.	.	.	.	.	.	6	93	.	
11616	.	.	.	.	.	.	6	91	.	
11617	.	.	.	.	.	.	6	97	.	
11618	.	.	.	.	.	.	.	99	.	
11619	.	.	.	.	.	.	2	11	1124	
11620	.	.	.	.	.	.	6	93	.	
11621	.	.	.	.	.	.	6	91	.	
11622	.	.	.	.	.	.	6	91	.	
11623	2	1	.	.	.	.	3	51	1411	
11624	2	1	.	.	.	.	6	11	15209	
11625	2	4	1	2	8	2	2	51	1124	
11626	.	.	.	.	.	.	3	93	.	
11627	.	.	.	.	.	.	6	92	.	
11628	.	.	.	.	.	.	6	91	.	
11629	.	.	.	.	.	.	.	99	.	
11630	.	.	.	.	.	.	2	11	1124	
11631	.	.	.	.	.	.	2	21	1124	

Variables

Name	Label
☑ fsu	
☑ sample	
☑ sector	
☑ state	
☑ region	
☑ district	
☑ restinfo	
☑ hhldno	
☑ psn	
☑ consump30	
☑ consump365	
☑ hhldsize	
☑ prinind	prinIND
☑ prinocu	prinOCU
☑ hhldtyp	
☑ relig	
☑ socgr	
☑ landown	
☑ nregcard	NREGCard
☑ sex	
☑ age	
☑ marital	
☑ genedu	

Properties

▼Variables	
Name	mopayment
Label	
Type	byte
Format	%8.0g
Value label	
▶ Notes	
▼Data	
▶ Filename	
Label	
▶ Notes	
Variables	73
Observations	25,521
Size	3.38M
Memory	64M
Sorted by	

Vars: 73 Order: Dataset Obs: 25,521 Filter: Off

So, for example, mode of payment. If someone is working in a factory, how the payment is made? Is it made weekly basis? Is it daily basis? Is it monthly basis? So, these are all again categorical variables. So, all these variables that we see, they are actually categorical variables. So, essentially, we will see, most of the variables that we use in the regression equation are categorical variables.

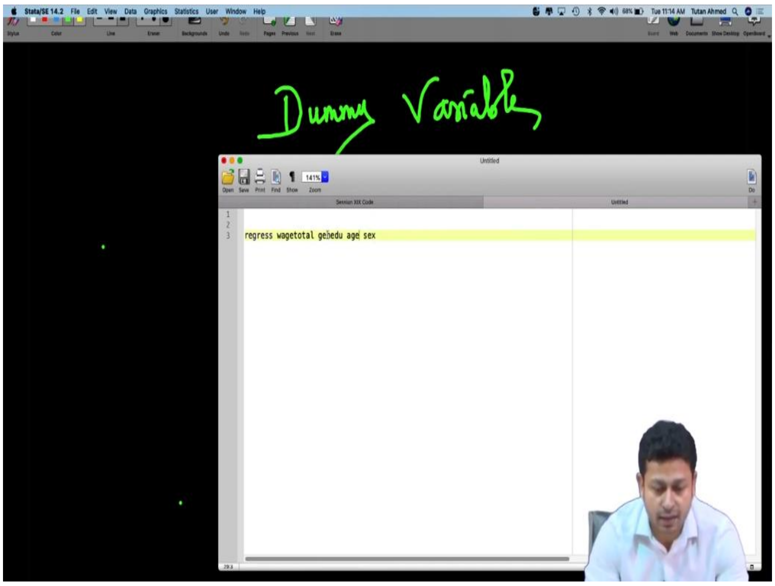
So, they all will come under the dummy variable category. And that is why it is so important; when we want to use all these different variables in our regression, it is really very important that we understand the concept of dummy variable, and how to use the dummy variable. So, let us say we want to run a regression on wage, and how we really do that. But before that, let me actually show you how we actually represent the dummy variables.

So, here, you see the data set. In data set, we have the values; because, values like 1, 2, 3, 4, 5 and so forth, but you actually want to; so, they have not really considered how you are going to use the variables; they simply have assigned some numerical value to represent different categories; but you have to use it. So, let us talk about the first category and how to use it. So, let us talk about the male, female category.

So, the dummy variable with binary categories. Let us say we will deal with the gender and let us say rural, urban, which is in our data set is sector. So, let us see how we are going to use these 2 dummy variables in our regression equation. So, let me go back. Let me actually run a regression. Let us see what are the different variables we have. We can see the pane on the right side. That actually talks about all the different columns.

And let us say I want to run a regression on wage total. So, that is the wage we have, total wage. So, that is my dependent variable. And I will include some of the independent variables. I have variables like; I will include gender, sex and age I will include; age or experience. And let us say I will include the levels of education. That is your general education. And let us say I will include sector; so, rural, urban. So, let us first run a very simple equation.

(Refer Slide Time: 14:50)

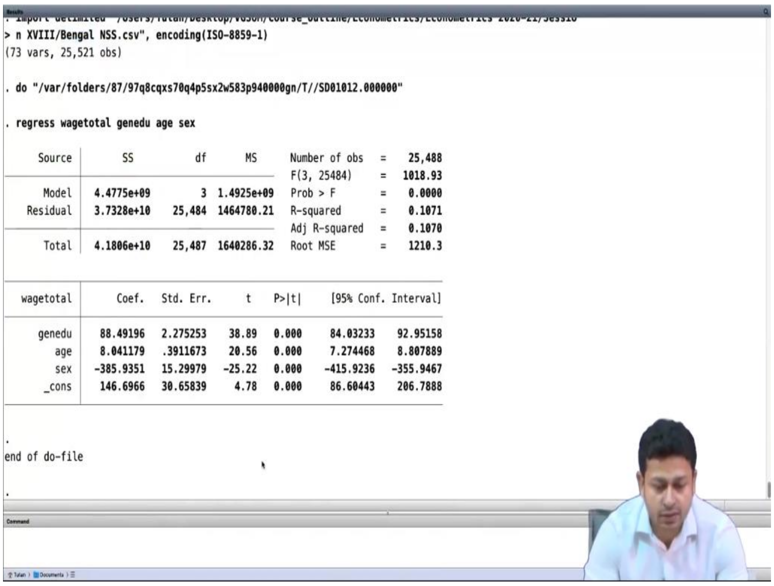


Dummy Variable

```
regress wagetotal genedu age sex
```

I can use a do file to run the code or I can type simply there, regress, wage total, then general education, then let us say I put age. So, for example, I could have used experience, but I am not; I have to create experience using other variables like age and years of education. We can do it later on. This is just for the illustration of the dummy variable. Now, we have sector, and we have sex. Or let us first take only one variable sex, male and female.

(Refer Slide Time: 15:34)



```
. import delimited "/Users/Tutan/Desktop/VGSOM/Course_Outline/Econometrics/Econometrics 2020-21/Sessio
> n XVIII/Bengal NSS.csv", encoding(ISO-8859-1)
(73 vars, 25,521 obs)

. do "/var/folders/87/97q8cqxs70q4p5sx2w583p940000gn/T//SD01012.000000"

. regress wagetotal genedu age sex

      Source |       SS           df       MS      Number of obs   =    25,488
-------------+----------------------------------   F(3, 25484)     =   1018.93
       Model |  4.4775e+09         3  1.4925e+09   Prob > F        =    0.0000
    Residual |  3.7328e+10    25,484  1464780.21   R-squared       =    0.1071
-------------+----------------------------------   Adj R-squared   =    0.1070
       Total |  4.1806e+10    25,487  1640286.32   Root MSE        =    1210.3

------------------------------------------------------------------------------
   wagetotal |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      genedu |   88.49196   2.275253    38.89   0.000     84.03233    92.95158
         age |   8.041179   .3911673    20.56   0.000     7.274468    8.807889
         sex |  -385.9351   15.29979   -25.22   0.000    -415.9236   -355.9467
       _cons |   146.6966   30.65839     4.78   0.000     86.60443    206.7888
------------------------------------------------------------------------------

.
end of do-file

.
```

And if I run it, I will get the regression here. And my regression shows that; you get some value for sex. So, it is -385.93 and it is significant; P value is 0. So, it has an important contribution to the regression equation. R square value is not very high. And other variables we have included, all comes out to be significant. So, essentially, how do we explain this regression equation is that, you explain it; so, you have taken the values of the dummy variable as 1 and 2.

So, here Stata we will choose one reference category; you have not defined it that way. So, let us say, in this case, Stata has chosen 1 as a reference category; and 1 is male and 2 is female. So, the moment you move from 1 to 2, the moment you increase the value, it actually shows that your income is declining, and it is declining 385.9 rupees. And let us say this income data is per week.

So, it is per week; if you are a female, then your income is going to be less than male by rupees 385. So, that is basically what it is saying. So, with this, we will end the lecture here. In the next lecture, I am going to detail, basically, I am going to explain the regression equation and like how to incorporate the dummy variable in the regression equation in detail. Thank you.