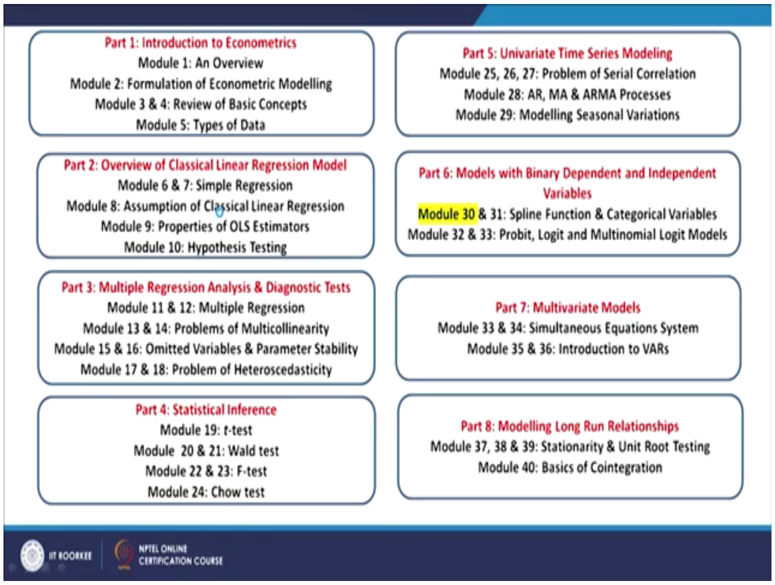
Econometric Modelling
¯Professor Sujata Kar
Department of Management Studies
Indian Institute of Technology, Roorkee
Lecture – 30
Spline Function & Categorical Variables – I

(Refer Slide Time: 0:29)



Part 1: Introduction to Econometrics
Module 1: An Overview
Module 2: Formulation of Econometric Modelling
Module 3 & 4: Review of Basic Concepts
Module 5: Types of Data

Part 2: Overview of Classical Linear Regression Model
Module 6 & 7: Simple Regression
Module 8: Assumption of Classical Linear Regression
Module 9: Properties of OLS Estimators
Module 10: Hypothesis Testing

Part 3: Multiple Regression Analysis & Diagnostic Tests
Module 11 & 12: Multiple Regression
Module 13 & 14: Problems of Multicollinearity
Module 15 & 16: Omitted Variables & Parameter Stability
Module 17 & 18: Problem of Heteroscedasticity

Part 4: Statistical Inference
Module 19: *t*-test
Module 20 & 21: Wald test
Module 22 & 23: F-test
Module 24: Chow test

Part 5: Univariate Time Series Modeling
Module 25, 26, 27: Problem of Serial Correlation
Module 28: AR, MA & ARMA Processes
Module 29: Modelling Seasonal Variations

Part 6: Models with Binary Dependent and Independent Variables
Module 30 & 31: Spline Function & Categorical Variables
Module 32 & 33: Probit, Logit and Multinomial Logit Models

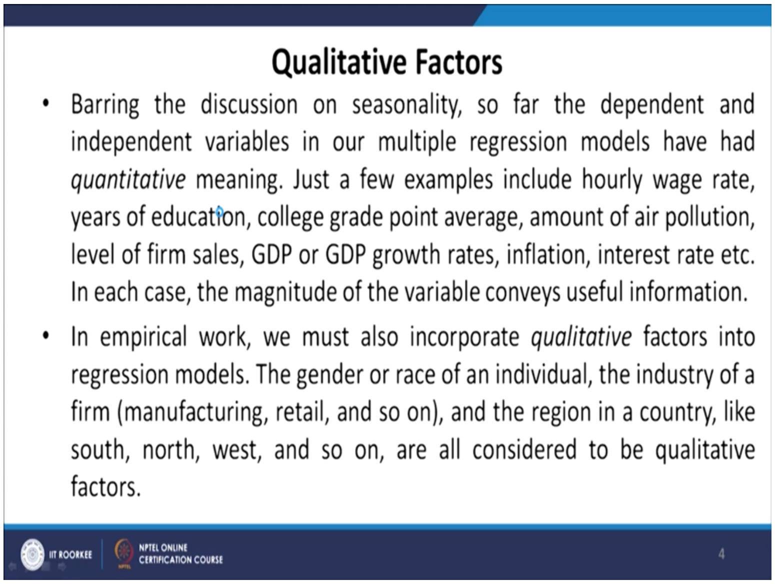
Part 7: Multivariate Models
Module 33 & 34: Simultaneous Equations System
Module 35 & 36: Introduction to VARs

Part 8: Modelling Long Run Relationships
Module 37, 38 & 39: Stationarity & Unit Root Testing
Module 40: Basics of Cointegration

Welcome back to the course on econometric modelling; this is module 30. Modules 30 and 31 are devoted to spline function estimations and categorical variables. Basically, here we are dealing with categorical variables, we would be considering categorical variables in the independent as well as dependent variables. Now, independent variables should be first considered; because at times categorical variables in the independent variable actually much easier to interpret and handle.

_

(Refer Slide Time: 00:56)



Qualitative Factors

- Barring the discussion on seasonality, so far the dependent and independent variables in our multiple regression models have had *quantitative* meaning. Just a few examples include hourly wage rate, years of education, college grade point average, amount of air pollution, level of firm sales, GDP or GDP growth rates, inflation, interest rate etc. In each case, the magnitude of the variable conveys useful information.
- In empirical work, we must also incorporate *qualitative* factors into regression models. The gender or race of an individual, the industry of a firm (manufacturing, retail, and so on), and the region in a country, like south, north, west, and so on, are all considered to be qualitative factors.

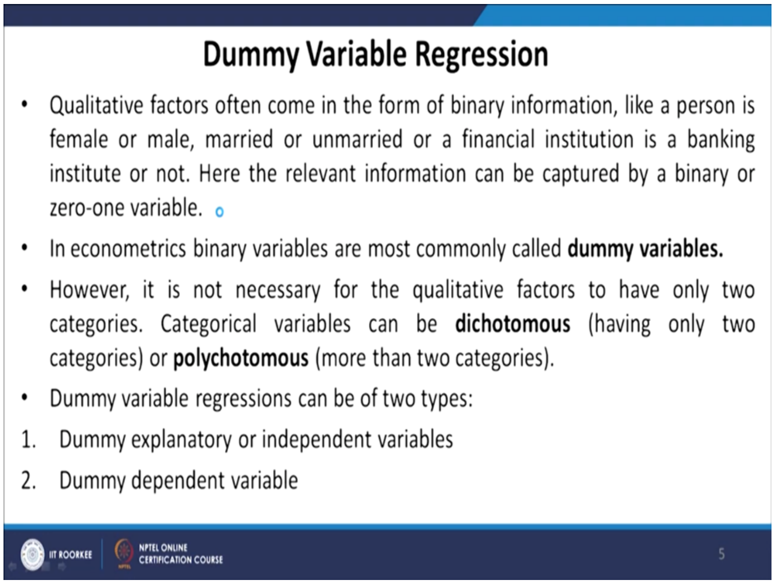
So, first, we define what are categorical variables. Barring the discussion on seasonality, so far, the dependent and independent variables in our multiple regression models have had quantitative meaning. For example, they were mostly numbers and numbers conveyed certain meaning. Just a few examples include hourly wage rate, years of education, college grade point average, amount of air pollution, level of firm sales, GDP or GDP growth rates, inflation, interest rate etcetera. In each case the magnitude of the variable conveys useful information; that is why we call them quantitative variables.

In empirical work, we must also incorporate qualitative factors into regression models. So, qualitative factors are the factors that can be assigned certain numbers. But, then the numbers themselves do not speak of anything. So, for example, gender or race of an individual, the industry of a firm like manufacturing, retail, and so on. All these are qualitative attributes or features of certain entities. So, we can assign numbers to them in order to facilitate our calculations, estimation, modelling etcetera. But the numbers themselves do not convey any meaning. So, that is why we call them qualitative factors or qualitative variables.

The other examples are the region in a country like south, north, west, and so on; they are all considered to be qualitative factors because ideally, they are qualitative features or characteristics of an entity or an individual. For example, while working with males, gender we can assign one to male and zero to female, and vice versa. So, first, it does not matter whether we are assigning one to male or one to female and zero to the other category. We certainly have a different interpretation depending on what is being coded as one or zero.

But then having coded 1 to any particular category does not impact the results; the interpretation may only differ. And simply having one zero is also not going to tell us that which attribute I am describing unless and until I specify that one here refers to the gender male. That is how the numbers here actually do not speak much; what matters is the qualitative factors.

(Refer Slide Time: 03:36)



Dummy Variable Regression

- Qualitative factors often come in the form of binary information, like a person is female or male, married or unmarried or a financial institution is a banking institute or not. Here the relevant information can be captured by a binary or zero-one variable.
- In econometrics binary variables are most commonly called **dummy variables.**
- However, it is not necessary for the qualitative factors to have only two categories. Categorical variables can be **dichotomous** (having only two categories) or **polychotomous** (more than two categories).
- Dummy variable regressions can be of two types:

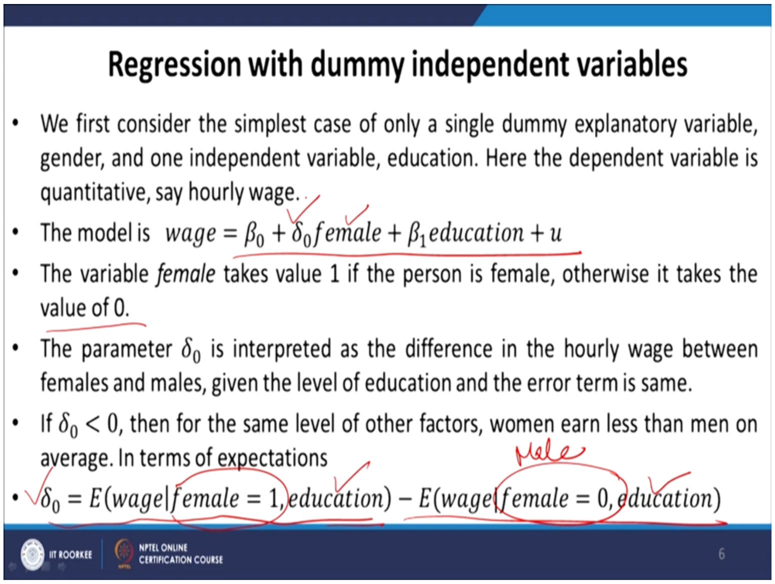
1. Dummy explanatory or independent variables
2. Dummy dependent variable

IIT ROORKEE | NPTEL ONLINE CERTIFICATION COURSE 5

So, qualitative factors often come in the form of binary information, like a person is female or male, married, or unmarried or a financial institution is a banking institute or not. So, on that basis we can assign one or zero to any of the categories; one to one category and 0 to the other category. Here the relevant information can be captured by a binary or zero-one variable. In econometrics binary variables are commonly called dummy variables, but we can also have dummy variables with more than one category. Just as we have discussed in the context of seasonal variations that seasonality is also captured using seasonal dummy variables, where we have multiple dummy variables for different seasons.

But yes, of course, one dummy variable is taking only two values that are one and zero; and that is why binary variables are most commonly known as dummy variables. However, it is not necessary for the qualitative factors to have only two categories. Categorical variables can be dichotomous that is having only two categories or polychotomous that is more than two categories. So, dummy variable regression can also be of two types, dummy explanatory or independent variables, and dummy dependent variables. So, first of all, we will be discussing dummy explanatory or independent variables.

(Refer Slide Time: 05:02)



Regression with dummy independent variables

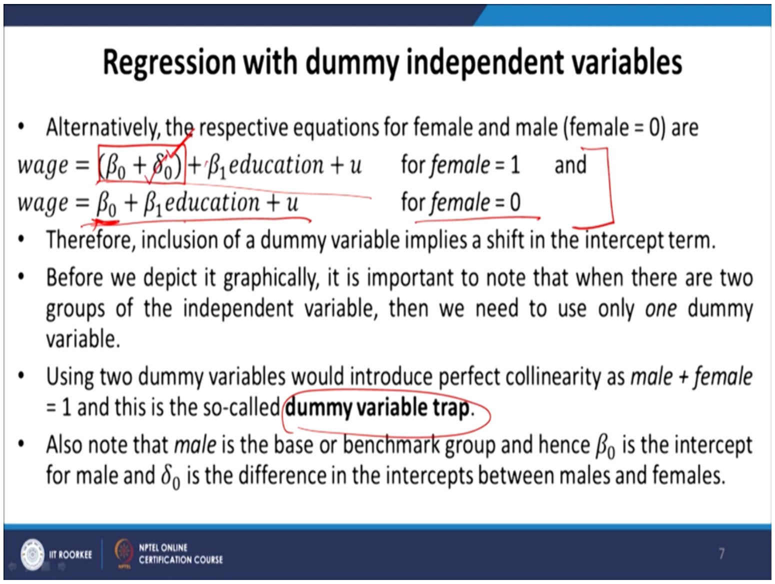
- We first consider the simplest case of only a single dummy explanatory variable, gender, and one independent variable, education. Here the dependent variable is quantitative, say hourly wage.
- The model is $wage = \beta_0 + \delta_0 female + \beta_1 education + u$
- The variable *female* takes value 1 if the person is female, otherwise it takes the value of 0.
- The parameter δ_0 is interpreted as the difference in the hourly wage between females and males, given the level of education and the error term is same.
- If $\delta_0 < 0$, then for the same level of other factors, women earn less than men on average. In terms of expectations
- $\delta_0 = E(wage|female = 1, education) - E(wage|\overset{\text{Male}}{female = 0}, education)$

IIT ROORKEE | NPTEL ONLINE CERTIFICATION COURSE 6

And we also begin with only dichotomous variables. So, we first consider the simplest case of only a single dummy explanatory variable, gender; and one independent variable, say education. Here the dependent variable is quantitative, say hourly wage. What we are trying to find out is that whether hourly wage varies between males and females; or whether it depends on the gender of an individual or not, given the education level. So, the model is (refer slide time: 5:35). Now, you can see that here we are having only females because there should be only one dummy variable when there are two categories.

And the dummy variable takes value one for females; and otherwise, it takes the value of 0. We could have also expressed the model as (refer slide time: 6:03). The parameter δ is interpreted as the difference in the hourly wage between females and males, given the level of education and the error term is the same. If $\delta_0 < 0$, then for the same level of other factors including education, women earn less than men on average; that would have been the interpretation of this parameter. In terms of expectations, δ_0 measures the expected value of wage, given that female is equal to 1 i.e. females take the value 1 in the dummy variable and also the level of education, (refer slide time: 6:52). So, education is controlled for females equal to 0, gives us the value associated with males; or expected value expected wages associated with males. And this of course gives us the expected wage associated with females. The difference between them is reflected in the parameter of this dummy variable or categorical variable.

(Refer Slide Time: 07:24)



Regression with dummy independent variables

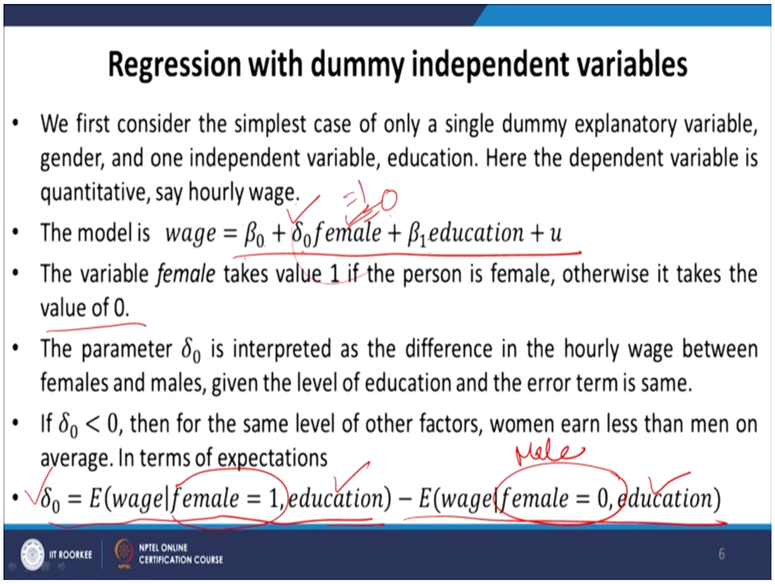
- Alternatively, the respective equations for female and male (female = 0) are

$wage = \boxed{(\beta_0 + \delta_0)} + \beta_1 education + u$ for *female* = 1 and

$wage = \beta_0 + \beta_1 education + u$ for *female* = 0

- Therefore, inclusion of a dummy variable implies a shift in the intercept term.
- Before we depict it graphically, it is important to note that when there are two groups of the independent variable, then we need to use only *one* dummy variable.
- Using two dummy variables would introduce perfect collinearity as *male* + *female* = 1 and this is the so-called **dummy variable trap**.
- Also note that *male* is the base or benchmark group and hence β_0 is the intercept for male and δ_0 is the difference in the intercepts between males and females.

IIT ROORKEE | NPTEL ONLINE CERTIFICATION COURSE 7



Regression with dummy independent variables

- We first consider the simplest case of only a single dummy explanatory variable, gender, and one independent variable, education. Here the dependent variable is quantitative, say hourly wage.
- The model is $wage = \beta_0 + \delta_0 \overset{=1/0}{female} + \beta_1 education + u$
- The variable *female* takes value 1 if the person is female, otherwise it takes the value of 0.
- The parameter δ_0 is interpreted as the difference in the hourly wage between females and males, given the level of education and the error term is same.
- If $\delta_0 < 0$, then for the same level of other factors, women earn less than men on average. In terms of expectations
- $\delta_0 = E(wage|female = 1, education) - E(wage|\overset{Male}{female = 0}, education)$

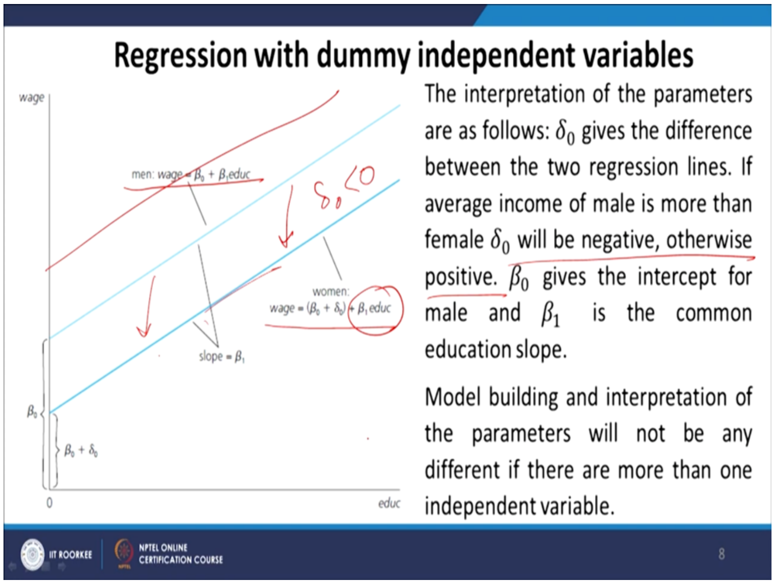
IIT ROORKEE | NPTEL ONLINE CERTIFICATION COURSE 6

Alternatively, the representative equations for female and male; that is when female equals to 0 are (refer slide time: 7:34- 8:05).

Therefore, the inclusion of a dummy variable implies a shift in the intercept term. So, whenever the dummy variable takes a value of 1, there is a shift in the intercept term. Of course, the shift can be positive or negative; if it is positive, if δ value is positive, then it shifts upward; if δ value is negative, it shifts downward. Before we depict it graphically, it is important to note that when there are two groups of the independent variable, then we need to use only one dummy variable; this has been also discussed earlier. Using two dummy variables would introduce perfect collinearity as male plus female equals 1, and this is the so-called dummy variable trap.

So, whenever we have a number of dummy variables equal to the number of categories of the independent variable; then we are actually into a dummy variable trap. Also note that male is the base or benchmark group and hence β_0 is the intercept for male, and δ_0 is the difference in the intercepts between males and females. So, this is also obvious from these two equations that β_0 is the intercept for male because here female equals to 0. And $\beta_0 + \delta_0$ gives us the intercept for females. As a result of which the difference between the two is δ_0. The δ_0 is the difference in the intercepts between males and females.

(Refer Slide Time: 09:37)



Regression with dummy independent variables

wage
men: wage = $\beta_0 + \beta_1 educ$
$\delta_0 < 0$
women:
wage = $(\beta_0 + \delta_0) + \beta_1 educ$
slope = β_1
β_0
$\beta_0 + \delta_0$
0
educ

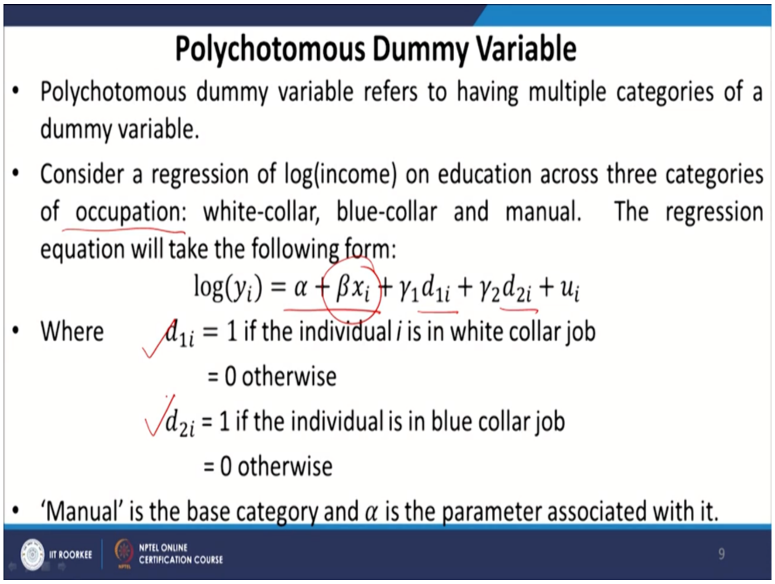
The interpretation of the parameters are as follows: δ_0 gives the difference between the two regression lines. If average income of male is more than female δ_0 will be negative, otherwise positive. β_0 gives the intercept for male and β_1 is the common education slope.

Model building and interpretation of the parameters will not be any different if there are more than one independent variable.

So, we now show it graphically. The interpretation of the parameters is as follows, δ_0 gives the difference between the two regression lines. So, this is basically the regression line for males, and this is the regression line for females. So, you can see that since the regression line for a female is down or lower than the regression line for males; so, δ_0 must be a negative number. And this slope is of course given by this which is the other independent variable that is education. If the average income of male is more than female δ_0 will be negative, otherwise positive.

So, δ_0 is positive, then this line would be the regression line for women and would be on top of the regression line of males. And this would indicate that the average income of females is more than the average income of males. β_0 gives the intercept for males and β_1 is the common education slope. Model building and interpretation of the parameters will not be any different if there are more than one independent variable. So, even if we have several independent variables, then of course we have to move on from a two-dimensional plane to an n-dimensional plane in order to show it graphically, but the interpretation remains the same.

(Refer Slide Time: 11:02)



Polychotomous Dummy Variable

- Polychotomous dummy variable refers to having multiple categories of a dummy variable.
- Consider a regression of log(income) on education across three categories of occupation: white-collar, blue-collar and manual. The regression equation will take the following form:

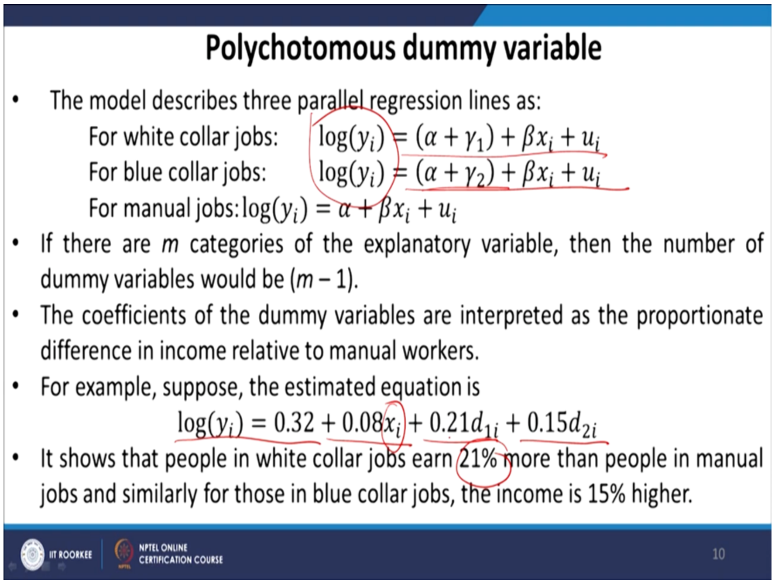
$$\log(y_i) = \alpha + \beta x_i + \gamma_1 d_{1i} + \gamma_2 d_{2i} + u_i$$

- Where d_{1i} = 1 if the individual *i* is in white collar job
  = 0 otherwise
  d_{2i} = 1 if the individual is in blue collar job
  = 0 otherwise
- 'Manual' is the base category and α is the parameter associated with it.

Now, we talk about polychotomous dummy variables. Polychotomous dummy variable refers to having multiple categories of a dummy variable. Consider a regression of log income on education across three categories of occupation: white-collar jobs, blue-collar jobs, and manual workers. The regression equation will take the following form that is (refer slide time: 11:27), and d_{1i} and d_{2i} are basically the two dummy variables corresponding to the two different categories. There are three categories of the variable measuring occupation, and we are having two dummy variables.

$d_{1i} = 1$, if the individual i is in the white-collar job; otherwise, 0. $d_{2i} = 1$, if the individual is in the blue-collar job; otherwise, 0. And as you understand that since there is no dummy for the manual workers; then manual workers or manual is the base category, and α is the parameter associated with it.

(Refer Slide Time: 12:13)



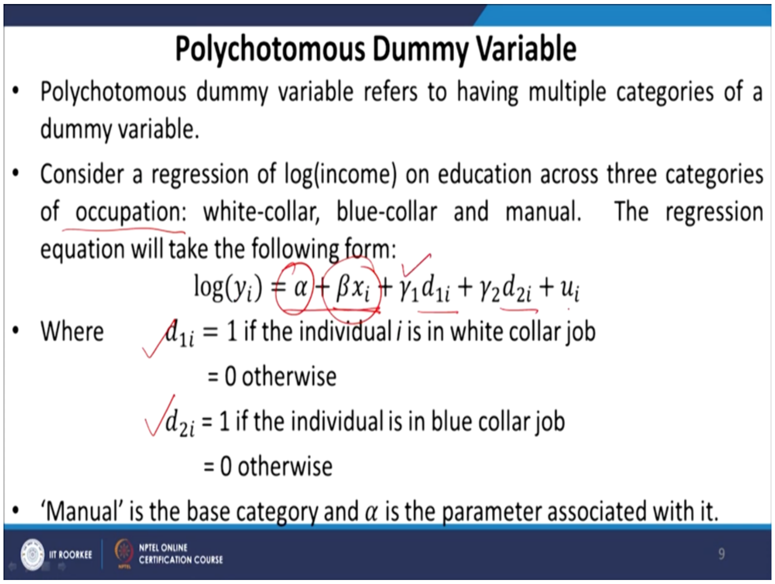
Polychotomous dummy variable

- The model describes three parallel regression lines as:
  For white collar jobs: $\log(y_i) = (\alpha + \gamma_1) + \beta x_i + u_i$
  For blue collar jobs: $\log(y_i) = (\alpha + \gamma_2) + \beta x_i + u_i$
  For manual jobs: $\log(y_i) = \alpha + \beta x_i + u_i$
- If there are *m* categories of the explanatory variable, then the number of dummy variables would be ($m - 1$).
- The coefficients of the dummy variables are interpreted as the proportionate difference in income relative to manual workers.
- For example, suppose, the estimated equation is

$$\log(y_i) = 0.32 + 0.08x_i + 0.21d_{1i} + 0.15d_{2i}$$

- It shows that people in white collar jobs earn 21% more than people in manual jobs and similarly for those in blue collar jobs, the income is 15% higher.

IIT ROORKEE | NPTEL ONLINE CERTIFICATION COURSE 10



Polychotomous Dummy Variable

- Polychotomous dummy variable refers to having multiple categories of a dummy variable.
- Consider a regression of log(income) on education across three categories of occupation: white-collar, blue-collar and manual. The regression equation will take the following form:

$$\log(y_i) = \alpha + \beta x_i + \gamma_1 d_{1i} + \gamma_2 d_{2i} + u_i$$

- Where d_{1i} = 1 if the individual *i* is in white collar job
  = 0 otherwise
  d_{2i} = 1 if the individual is in blue collar job
  = 0 otherwise
- 'Manual' is the base category and α is the parameter associated with it.

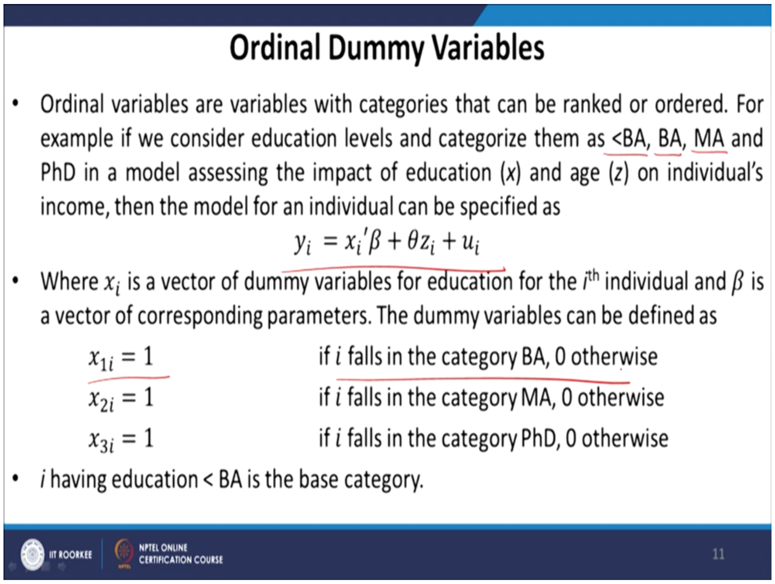
IIT ROORKEE | NPTEL ONLINE CERTIFICATION COURSE 9

The model describes three parallel regression lines here. For white-collar jobs we have (refer slide time: 12:20). The coefficients of the dummy variables are interpreted as the proportionate difference in the income relative to manual workers. Now, here we are talking about the proportionate difference in income because we are having a logarithm of income as the dependent variable.

Earlier in the previous example with the dichotomous variable, we had only wage. So, if we consider wage, then this simply measures the difference in the wages. If I am considering log wage or log income; and then we are basically the interpretation would be that the proportionate difference in the log in the income of people across different occupations. The coefficients of the dummy variables are interpreted as the proportionate difference in income relative to manual workers. For example, suppose, the estimated equation is (refer slide time: 13:43).

It shows that people in white-collar jobs earn 21 percent more than people in manual jobs; and similarly, for those in blue-collar jobs, the income is 15 percent higher.

(Refer Slide Time: 14:07)



Ordinal Dummy Variables

- Ordinal variables are variables with categories that can be ranked or ordered. For example if we consider education levels and categorize them as <BA, BA, MA and PhD in a model assessing the impact of education (x) and age (z) on individual's income, then the model for an individual can be specified as

$$y_i = x_i'\beta + \theta z_i + u_i$$

- Where x_i is a vector of dummy variables for education for the i^{th} individual and β is a vector of corresponding parameters. The dummy variables can be defined as

$x_{1i} = 1$ if i falls in the category BA, 0 otherwise

$x_{2i} = 1$ if i falls in the category MA, 0 otherwise

$x_{3i} = 1$ if i falls in the category PhD, 0 otherwise

- i having education < BA is the base category.

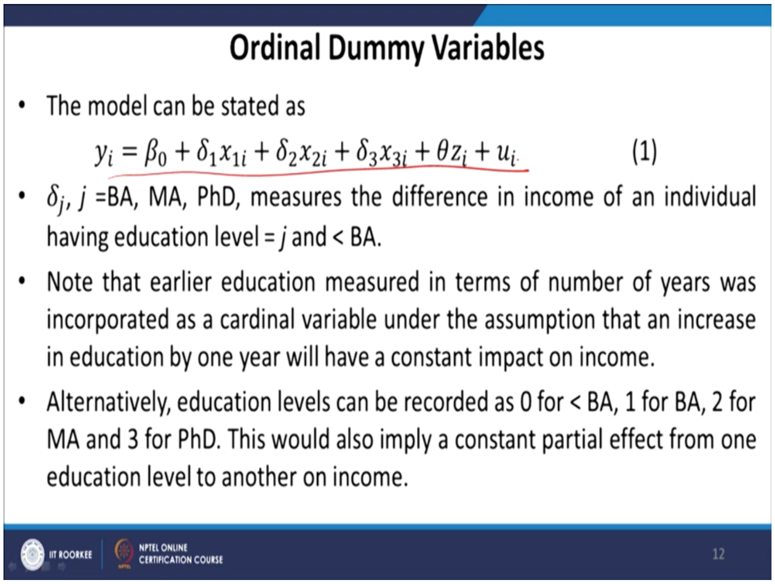
Now, we talk about ordinal variables. When we were discussing previously the categories, we could see that there was no ordering of the variables. So, male and female, there is no ordering in the sense that if we assign number 1 to females; that does not imply that females are better than male. Or if we assign number 1 to male, then males are better than females. Similarly, when it comes to occupation also, we are not getting into any ordering of the jobs; like whether white-collar jobs are better than blue-collar jobs or manual jobs are better etcetera.

But, then in certain cases, very clear-cut orderings are possible. For example, whenever there are some rankings, then we can go for orderings. So, ordinal variables are variables with categories that can be ranked or ordered. For example, if we consider education levels and categorize them as less than bachelor's degree, BA. So, we are just considering, for example, bachelor's in arts;

one can consider any other discipline like bachelor's in science, B-tech and things like that. But, just, for example, less than Bachelor of Arts is one category; then a bachelor of arts, then master of arts, and finally Ph.D.

So, these things can certainly be ranked because one degree is higher than the other degree. So, if we can categorize them like this, the education levels in a model assessing the impact of education and age on an individual's income. Then the model for an individual can be specified like this, where we are having ordinal dummy variables. So, (refer slide time: 16:01). So, we are again having three dummy variables for four categories.

(Refer Slide Time: 16:37)



Ordinal Dummy Variables

- The model can be stated as

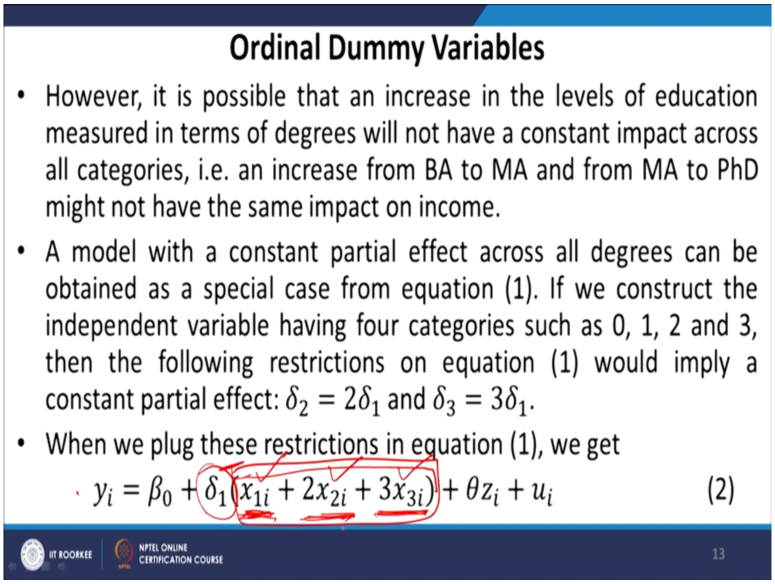
$$y_i = \beta_0 + \delta_1 x_{1i} + \delta_2 x_{2i} + \delta_3 x_{3i} + \theta z_i + u_i \qquad (1)$$

- δ_j, j =BA, MA, PhD, measures the difference in income of an individual having education level = j and < BA.
- Note that earlier education measured in terms of number of years was incorporated as a cardinal variable under the assumption that an increase in education by one year will have a constant impact on income.
- Alternatively, education levels can be recorded as 0 for < BA, 1 for BA, 2 for MA and 3 for PhD. This would also imply a constant partial effect from one education level to another on income.

The model can be stated as (refer slide time: 16:42- 17:05). Note that earlier education measured in terms of the number of years was incorporated as a cardinal variable or quantitative variable. Under the assumption that an increase in education by one year will have a constant impact on income. So, whenever we go for a regression line, and we are having education as the independent variable; then the interpretation is that if education increases by one percent or one unit or by one year. Then of course, when we are measuring education in terms of the number of years spent on education, then one unit increases in education and then how it is going to impact the income. So, how much, by what percent income is increasing, or how income is changing. So, that is how we interpret the coefficients of a quantitative variable. Alternatively, education levels can be recorded as, 0 for less than BA, 1 for BA, 2 for MA, and 3 for Ph.D. This would also imply a constant partial effect from one education level to another on income.

So, instead of having education level measured in terms of ten years in on education, eleven, twelve, thirteen, and so on. By focusing on degree, we can also have only one variable; and the variable will have four alternative values 0 for BA, less than BA, 1 for BA, 2 for MA, and 3 for Ph.D. So, this is a single variable, this is not a dummy variable; a single variable having only four alternative values.

(Refer Slide Time: 18:50)

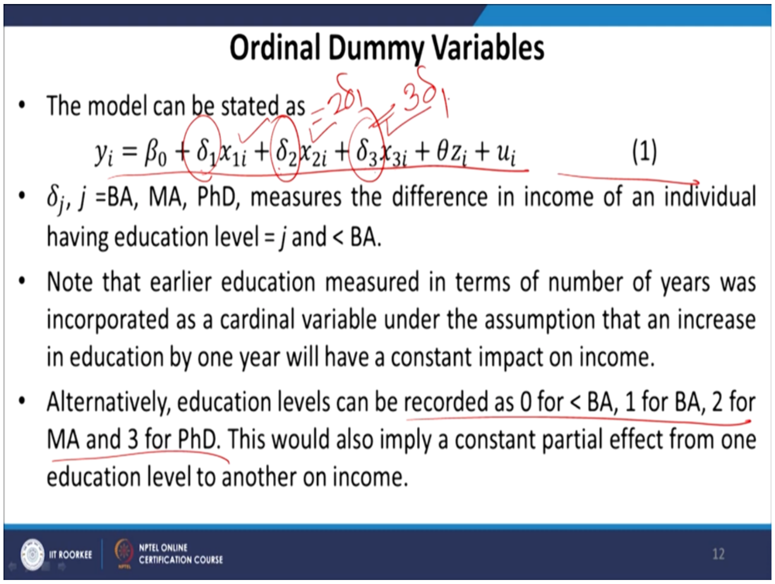


Ordinal Dummy Variables

- However, it is possible that an increase in the levels of education measured in terms of degrees will not have a constant impact across all categories, i.e. an increase from BA to MA and from MA to PhD might not have the same impact on income.
- A model with a constant partial effect across all degrees can be obtained as a special case from equation (1). If we construct the independent variable having four categories such as 0, 1, 2 and 3, then the following restrictions on equation (1) would imply a constant partial effect: $\delta_2 = 2\delta_1$ and $\delta_3 = 3\delta_1$.
- When we plug these restrictions in equation (1), we get

$$y_i = \beta_0 + \boxed{\delta_1}\boxed{(\underline{x_{1i}} + 2\underline{x_{2i}} + 3\underline{x_{3i}})} + \theta z_i + u_i \qquad (2)$$

IIT ROORKEE | NPTEL ONLINE CERTIFICATION COURSE 13



Ordinal Dummy Variables

- The model can be stated as $= 2\delta_1 \quad = 3\delta_1$

$$y_i = \beta_0 + \delta_1 x_{1i} + \delta_2 x_{2i} + \delta_3 x_{3i} + \theta z_i + u_i \qquad (1)$$

- δ_j, j =BA, MA, PhD, measures the difference in income of an individual having education level = j and < BA.
- Note that earlier education measured in terms of number of years was incorporated as a cardinal variable under the assumption that an increase in education by one year will have a constant impact on income.
- Alternatively, education levels can be recorded as 0 for < BA, 1 for BA, 2 for MA and 3 for PhD. This would also imply a constant partial effect from one education level to another on income.

IIT ROORKEE | NPTEL ONLINE CERTIFICATION COURSE 12

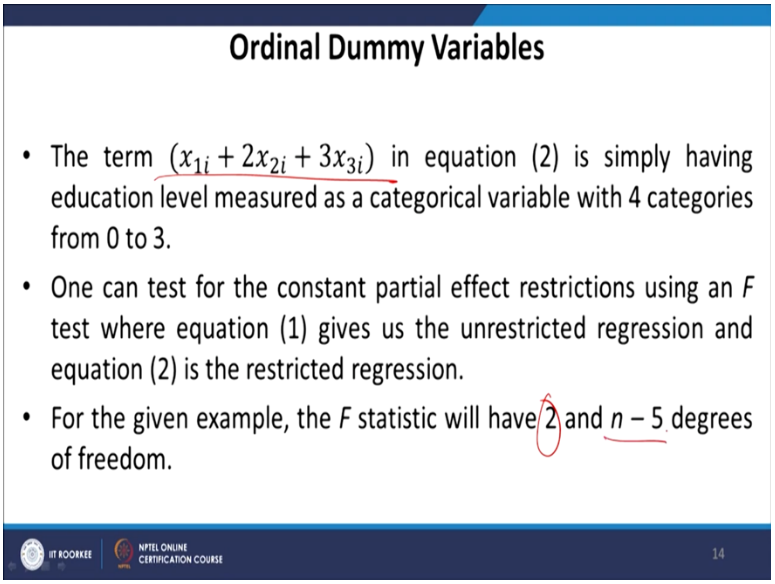
So, this is also a variable that can capture the impact of degrees on income. But the assumption is that whenever we move from one level to another level of education; the impact of it on income remains constant. That is whenever we are moving from BA to MA, the impact income is increasing by certain percentages. And the income increases in that percent only, whenever a person moves from MA to Ph.D. So, the partial effect on the income of different degrees or change in different degrees remains constant. However, it is possible that an increase in the levels of education measured in terms of degrees will not have a constant impact across all categories.

That is an increase from BA to MA and from MA to Ph.D. might not have this same impact on income. So, in that case, that the partial impact of education or different degrees will not have the same impact on income. So, a model with a constant partial effect across all degrees can be obtained as a special case from equation 1. So, equation-1 is not an equation since it takes into consideration different dummy variables for different degrees. And it allows for variation in the coefficient associated with different degrees on income. We are not assuming a common constant partial effect.

A common constant partial effect emerges when we have a single variable recording the degrees like this. So, our model inclusion of dummy variables allows for changes in the partial effect. But a model with a constant partial effect across all degrees can be obtained as a special case of equation-1 that is of the dummy variable case. If we construct the independent variable having four categories, such as 0, 1, 2, and 3, then the following restrictions on equation-1 would imply a constant partial effect. That is what the restrictions are (refer slide time: 21:14- 23:28).

Those are who are less than BA. So, that refers to that one variable that has only four categories: 0 for less than BA, 1 for BA, 2 for MA, and 3 for Ph.D. And we obtained a single parameter estimate δ_1. So, this is the case of constant partial effect which can be obtained as a special case of the dummy variable thing.

(Refer Slide Time: 23:54)

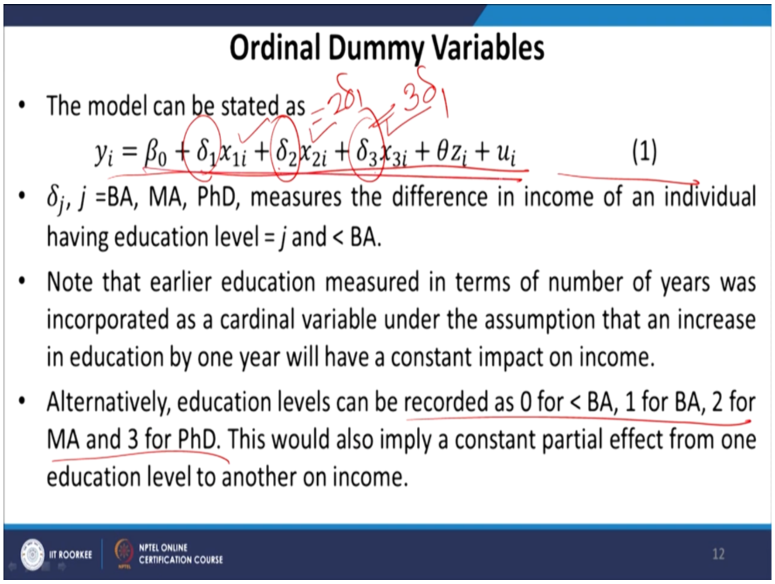


Ordinal Dummy Variables

- The term $(x_{1i} + 2x_{2i} + 3x_{3i})$ in equation (2) is simply having education level measured as a categorical variable with 4 categories from 0 to 3.
- One can test for the constant partial effect restrictions using an F test where equation (1) gives us the unrestricted regression and equation (2) is the restricted regression.
- For the given example, the F statistic will have 2 and $n-5$ degrees of freedom.

IIT ROORKEE | NPTEL ONLINE CERTIFICATION COURSE

14



Ordinal Dummy Variables

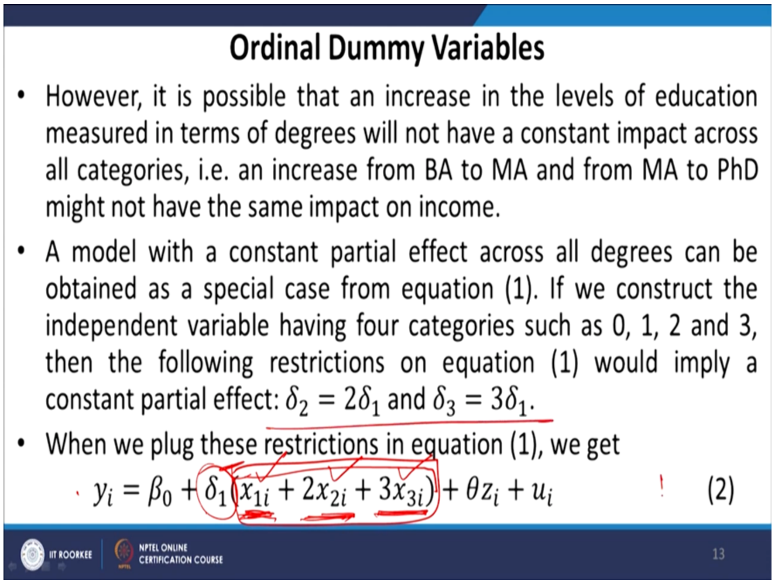
- The model can be stated as

$$y_i = \beta_0 + \delta_1 x_{1i} + \overset{=2\delta_1}{\delta_2} x_{2i} + \overset{=3\delta_1}{\delta_3} x_{3i} + \theta z_i + u_i \qquad (1)$$

- δ_j, j =BA, MA, PhD, measures the difference in income of an individual having education level = j and < BA.
- Note that earlier education measured in terms of number of years was incorporated as a cardinal variable under the assumption that an increase in education by one year will have a constant impact on income.
- Alternatively, education levels can be recorded as 0 for < BA, 1 for BA, 2 for MA and 3 for PhD. This would also imply a constant partial effect from one education level to another on income.

IIT ROORKEE | NPTEL ONLINE CERTIFICATION COURSE

12



Ordinal Dummy Variables

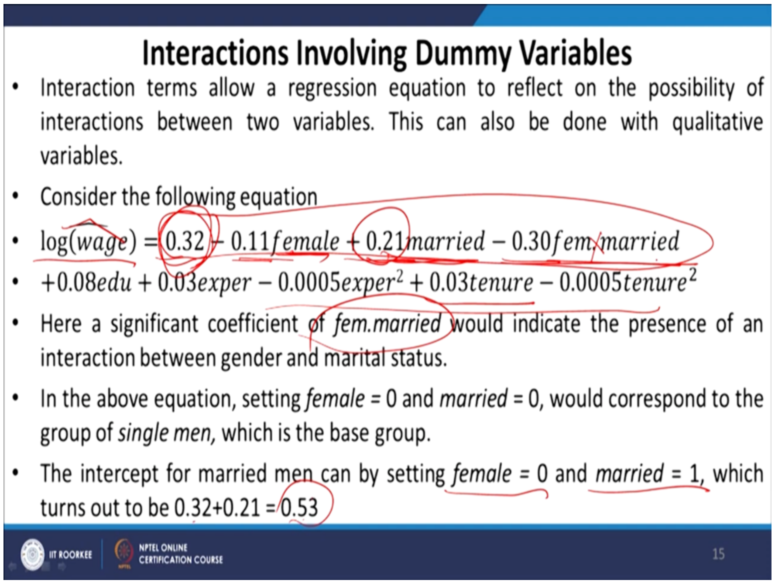
- However, it is possible that an increase in the levels of education measured in terms of degrees will not have a constant impact across all categories, i.e. an increase from BA to MA and from MA to PhD might not have the same impact on income.
- A model with a constant partial effect across all degrees can be obtained as a special case from equation (1). If we construct the independent variable having four categories such as 0, 1, 2 and 3, then the following restrictions on equation (1) would imply a constant partial effect: $\delta_2 = 2\delta_1$ and $\delta_3 = 3\delta_1$.
- When we plug these restrictions in equation (1), we get

$$y_i = \beta_0 + \delta_1(x_{1i} + 2x_{2i} + 3x_{3i}) + \theta z_i + u_i \qquad (2)$$

IIT ROORKEE | NPTEL ONLINE CERTIFICATION COURSE 13

The term this in equation-2 is simply having education level measured as a categorical variable with 4 categories from 0 to 3. One can test for the constant partial effect restrictions using an F-test, where equation-1 gives us the unrestricted regression. This is our equation-1, this is the unrestricted regression, and after incorporating these restrictions equation-2 is the restricted regression. For the given example, the F-statistics will have 2 and $n - 5$ degrees of freedom. 2 because there are only two restrictions; and there have been originally 5 independent variables. And that is how $n - 5$ degrees of freedom for the denominator.

(Refer Slide Time: 24:44)



Interactions Involving Dummy Variables

- Interaction terms allow a regression equation to reflect on the possibility of interactions between two variables. This can also be done with qualitative variables.
- Consider the following equation
- $\log(wage) = 0.32 - 0.11 female + 0.21 married - 0.30 fem.married$
- $+0.08 edu + 0.03 exper - 0.0005 exper^2 + 0.03 tenure - 0.0005 tenure^2$
- Here a significant coefficient of *fem.married* would indicate the presence of an interaction between gender and marital status.
- In the above equation, setting *female* = 0 and *married* = 0, would correspond to the group of *single men,* which is the base group.
- The intercept for married men can by setting *female* = 0 and *married* = 1, which turns out to be 0.32+0.21 = 0.53

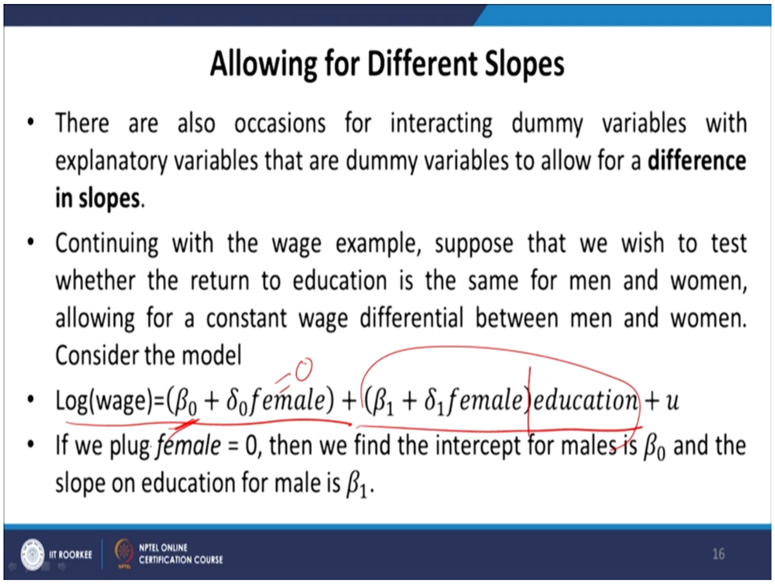
IIT ROORKEE | NPTEL ONLINE CERTIFICATION COURSE 15

Now, we will talk about interactions involving dummy variables. Interaction terms allow a regression equation to reflect on the possibility of interactions between two variables. So, how these two variables would interact with each other, other than impacting the independent variable solely or on their own complete. This can also be done with qualitative variables. Consider the following equation, where we have log $log\ \overset{\wedge}{(}wage)$. Hat here implies that this is an estimated equation because we are not including any residual term here. And we are also mentioning the estimates, (refer slide time: 25:24).

So, we are interested in the dummy variable components and their interactions; and not in these independent variables as such. Here is a significant coefficient of females multiplied by married would indicate the presence of an interaction between gender and marital status. And which is separate from the impact of gender and the impact of marital status. So, in the above equation setting female equals to 0 and married equals to 0 would correspond to the group of single men which is the base group. So, the base group is female equals 0 and also married equals 0; so, the base group is 0.32. So, 0.32 is referred to the coefficient associated with single men. The intercept for a married man or married man can be obtained by setting female equal to 0, and married equals to 1. So, female equal to 0, married equals to 1 gives us 0.21; and female equals to 0 gives us 0.32.

So, the intercept for married men can turn out to be $0.32 + 0.21 = 0.53$, so this is how we can interpret the different components of dummy variables as well as their interactions. Now, the interaction here -0.30; 0.30 implies that female married women are having 30 percent lower wages compared to the base category of single men.

(Refer Slide Time: 27:40)

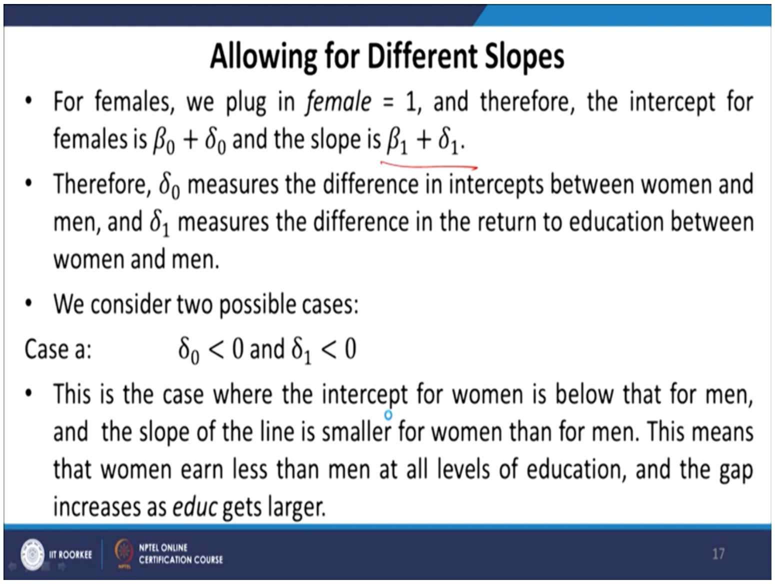


Allowing for Different Slopes

- There are also occasions for interacting dummy variables with explanatory variables that are dummy variables to allow for a **difference in slopes**.
- Continuing with the wage example, suppose that we wish to test whether the return to education is the same for men and women, allowing for a constant wage differential between men and women. Consider the model
- $\text{Log(wage)}=(\beta_0 + \delta_0 \overset{0}{female}) + (\beta_1 + \delta_1 female)education + u$
- If we plug $female$ = 0, then we find the intercept for males is β_0 and the slope on education for male is β_1.

There are also occasions for interacting dummy variables with explanatory variables that are dummy variables to allow for differences in slopes. So, so far, we have been considering only indifferences in the intercept; now, considering differences in slopes. So, continuing with the wage example, suppose that we wish to test whether the return to education is the same for men and women, allowing for a constant wage differential between men and women. So, considering the model, (refer slide time: 28:12- 28:44).

(Refer Slide Time: 28:43)



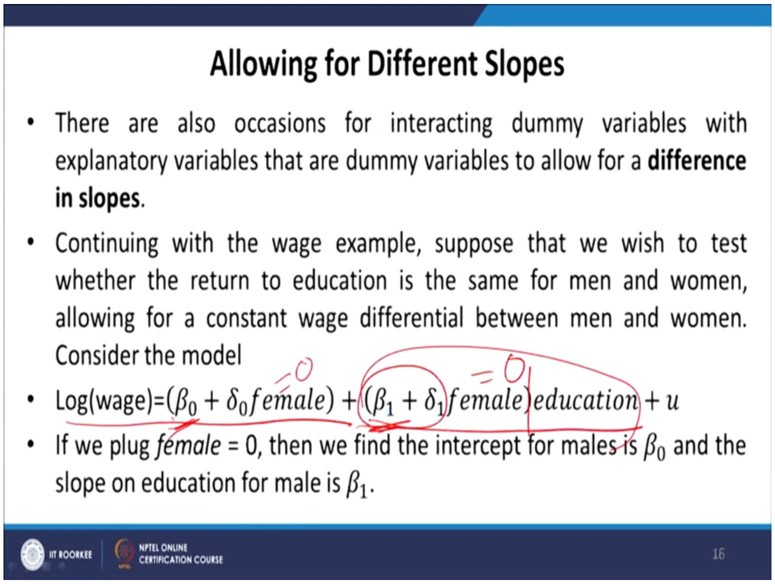
Allowing for Different Slopes

- For females, we plug in *female* = 1, and therefore, the intercept for females is $\beta_0 + \delta_0$ and the slope is $\beta_1 + \delta_1$.
- Therefore, δ_0 measures the difference in intercepts between women and men, and δ_1 measures the difference in the return to education between women and men.
- We consider two possible cases:

Case a: $\delta_0 < 0$ and $\delta_1 < 0$

- This is the case where the intercept for women is below that for men, and the slope of the line is smaller for women than for men. This means that women earn less than men at all levels of education, and the gap increases as *educ* gets larger.

IIT ROORKEE NPTEL ONLINE CERTIFICATION COURSE 17



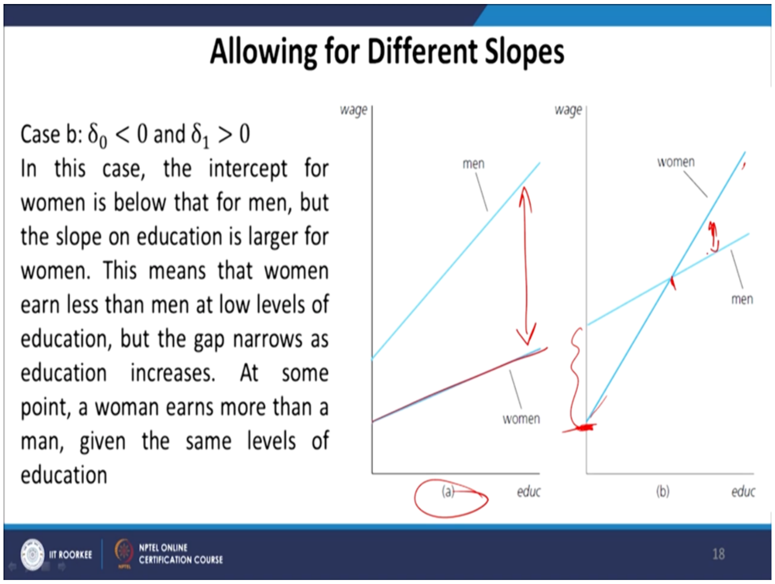
Allowing for Different Slopes

- There are also occasions for interacting dummy variables with explanatory variables that are dummy variables to allow for a **difference in slopes**.
- Continuing with the wage example, suppose that we wish to test whether the return to education is the same for men and women, allowing for a constant wage differential between men and women. Consider the model
- Log(wage)=$(\beta_0 + \delta_0 \overset{=0}{female}) + (\beta_1 + \delta_1 \overset{=0}{female})education + u$
- If we plug *female* = 0, then we find the intercept for males is β_0 and the slope on education for male is β_1.

IIT ROORKEE NPTEL ONLINE CERTIFICATION COURSE 16

(Refer slide time: 28:45- 29:10). That is the slope of the regression line; we consider two possible cases. So, the first case is $\delta_0 < 0$, and $\delta_1 < 0$. This is the case where the intercept for women is below that for men, and the slope of the line is smaller for women than for men. This means that women earn less than men at all levels of education, and the gap increases as education get larger.

(Refer Slide Time: 29:36)



Allowing for Different Slopes

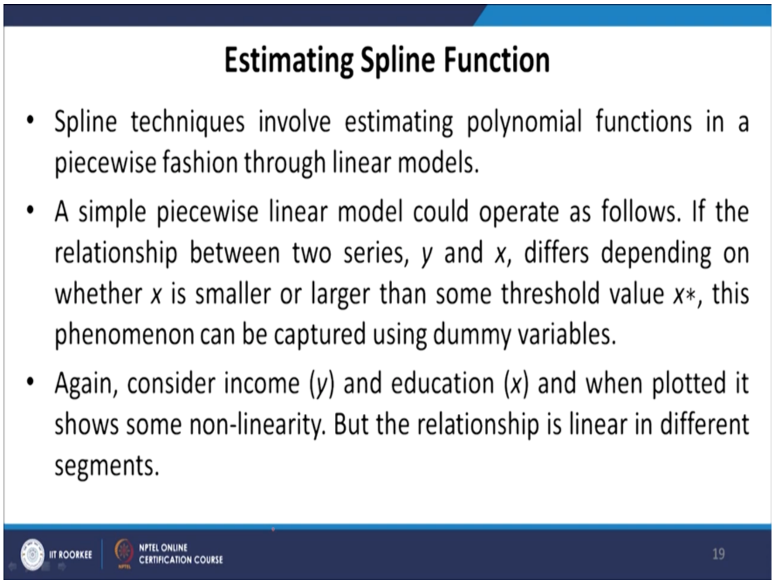
Case b: $\delta_0 < 0$ and $\delta_1 > 0$

In this case, the intercept for women is below that for men, but the slope on education is larger for women. This means that women earn less than men at low levels of education, but the gap narrows as education increases. At some point, a woman earns more than a man, given the same levels of education

So, this is the case of a. You can see that women have a lower slope, and the gap increases with an increase in education. Then, the other cases $\delta_0 < 0$, and $\delta_1 > 0$. In this case, the intercept for women is below that for men; so, to begin with, women of course obtain lower wages. And that is why it has a lower intercept, but the slope on education is larger for women. This means that women earn less than men at low levels of education, but the gap narrows as education increases. At some point, a woman earns more than a man, given the same levels of education; so, there is a crossover.

So, after some point, the income of women increases or the income of women is proportionately more than that of men, given the same levels of education. So, this is how we allow differences in slopes.

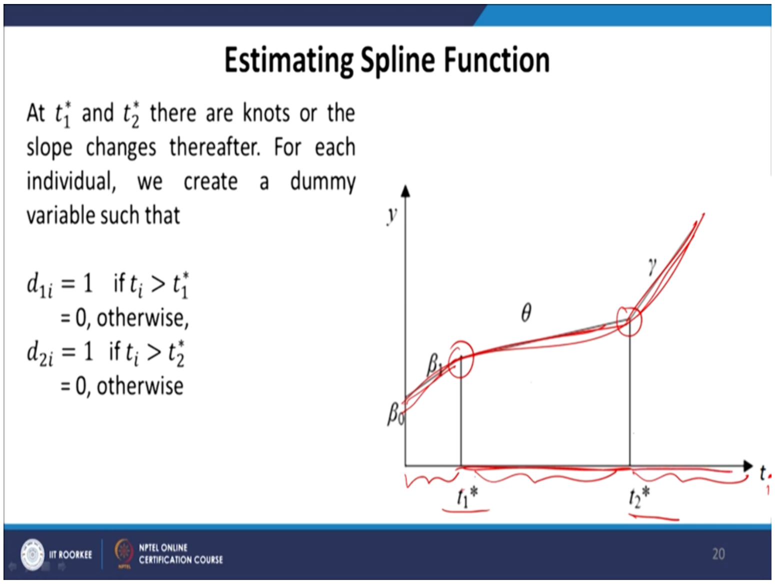
(Refer Slide Time: 30:42)



Estimating Spline Function

- Spline techniques involve estimating polynomial functions in a piecewise fashion through linear models.
- A simple piecewise linear model could operate as follows. If the relationship between two series, y and x, differs depending on whether x is smaller or larger than some threshold value $x*$, this phenomenon can be captured using dummy variables.
- Again, consider income (y) and education (x) and when plotted it shows some non-linearity. But the relationship is linear in different segments.

IIT ROORKEE | NPTEL ONLINE CERTIFICATION COURSE 19



Estimating Spline Function

At t_1^* and t_2^* there are knots or the slope changes thereafter. For each individual, we create a dummy variable such that

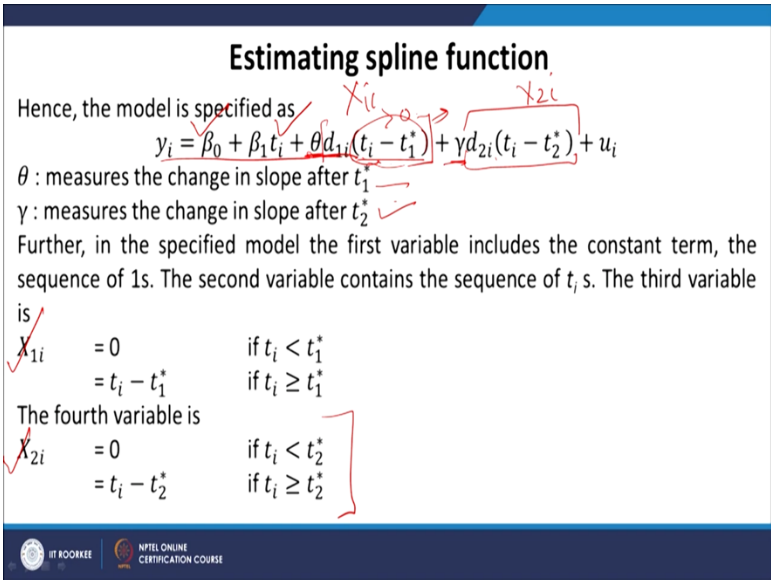
$$d_{1i} = 1 \quad \text{if } t_i > t_1^*$$
$$= 0, \text{otherwise,}$$
$$d_{2i} = 1 \quad \text{if } t_i > t_2^*$$
$$= 0, \text{otherwise}$$

IIT ROORKEE | NPTEL ONLINE CERTIFICATION COURSE 20

And finally, I talk about estimating the spline function. Spline techniques involve estimating polynomial functions in a piecewise fashion through linear models. A simple piecewise linear model could operate as follows. If the relationship between two series, y , and x, differs depending on whether x is smaller or larger than some threshold value x^*; this phenomenon can be captured using dummy variables. Again, consider income (y) and education (x), and when plotted it shows some non-linearity. But the relationship is linear in different segments; so, this is how it may look like.

So, you can see that overall, the relationship is non-linear; but it is actually linear in the segment. So, we can go for piecewise linear estimation, and this is what we called spline technique or spline estimation. So, how do we do it? Very simply using dummy variables. So, for different segments, the regression is also done piecewise. So, first of all, at t_1^* *and* t_2^* there are knots or the slope changes; the slopes change after t_1^* *and* t_2^*. For each individual, we create a dummy variable such (refer slide time: 32:07).

(Refer Slide Time: 32:26)



Estimating spline function

Hence, the model is specified as

$$y_i = \beta_0 + \beta_1 t_i + \theta d_{1i}(t_i - t_1^*) + \gamma d_{2i}(t_i - t_2^*) + u_i$$

θ : measures the change in slope after t_1^*

γ : measures the change in slope after t_2^*

Further, in the specified model the first variable includes the constant term, the sequence of 1s. The second variable contains the sequence of t_i s. The third variable is

X_{1i} $= 0$ if $t_i < t_1^*$
$= t_i - t_1^*$ if $t_i \geq t_1^*$

The fourth variable is

X_{2i} $= 0$ if $t_i < t_2^*$
$= t_i - t_2^*$ if $t_i \geq t_2^*$

And the model is specified as (refer slide time: 32:29- 34:07). Now, the interpretation of γ and θ are very similar to the dummy variable interpretation of the coefficients of dummy variables; there is nothing specific about it. The only thing is that it shows how we can estimate and non-linear function by going for piecewise linearity, and that is done by including dummy variables.

That brings me to the end of the discussion on the models that could be estimated with dummy independent variables. In the next module, I will be discussing dummy dependent or regression models with dummy dependent variables. Thank you.