Econometrics

Dummy Variables

Lec. **6**

Saeed Tajrishy

Faculty of Economics, University of Tehran

Fall 2023

Slides by **Erfān Rezaei M. N.**

Slides primarily based on *Itroductory Econometrics* by Jeffery Wooldridge (2020)

Introduction

- ✔ Qualitative data is used to:
  - ✗ Enhance regression analysis by incorporating non-numeric factors.
  - ✗ Uncover insights from categorical data in statistical models.
  - **Ex** Some examples are gender, ethnicity, educational levels, employment status, etc.
- ✔ But the conventional regression models can't handle qualitative data well!
- ✔ So there is a need for methods to integrate this type of data effectively.
- ✔ Thus, **dummy variables** (abbr. DV) are introduced!

Qualitative Information

Information describing inherent qualities or categories. Examples include gender, ethnicity, type of product, etc.

Dummy Variables

- ✔ Dummy variables are essential tools in econometrics for reflecting qualitative information into econometric models.
- ✔ Dummy variables assign binary—0 or 1—numerical values to qualitative categories.

Dummy Variables (abbr. DVs)

A.k.a. binary variables or zero-one variables, they represent categorical data, by assigning 0 or 1 to signify the absence or presence of a specific qualitative characteristics.

- ✔ Coefficients associated with dummy variables quantify the average change in the dependent variable for a specific category compared to a reference category.
- ✔ Or in simple words, these variables enable the analysis of how different groups or categories influence the dependent variable, facilitating nuanced insights.
- ✔ Also note that dummy variables can be used in various contexts, from simple two-group models to more complex scenarios involving interactions and multiple categories.

Dummy Variables - Example I

Hourly Wage Determination Model

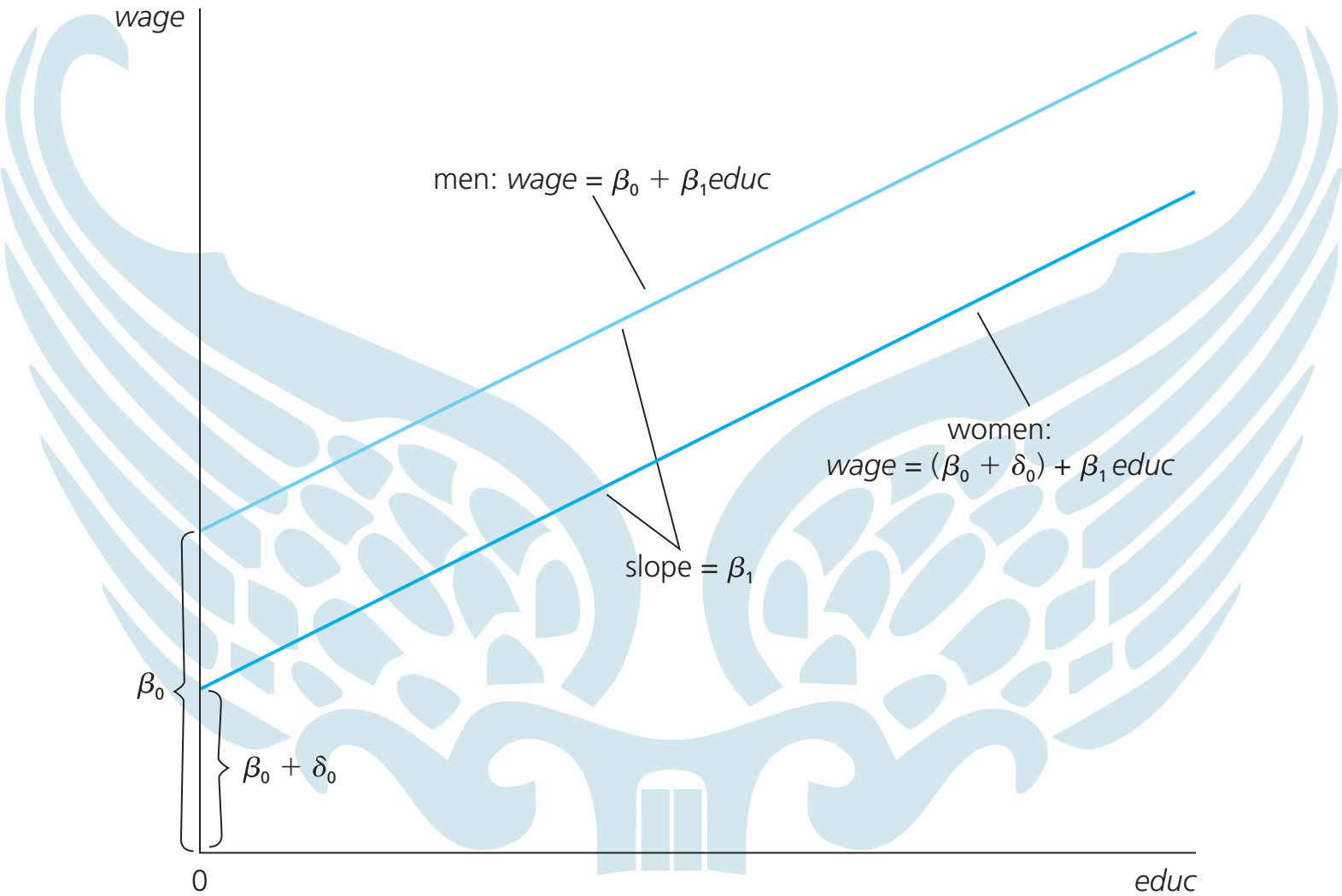
$$wage = \beta_0 + \delta_0 female + \beta_1 educ + u$$

- ✔ Here the dummy variable is *female*. When $female = 1$ the person is female and when $female = 0$ the person is male.
- ✔ So be careful that δ_0, as the coefficient of dummy variable, is the difference in hourly wage between females and males, given the same amount of education (and the same error term u).
- **Q** When wages are discriminative for women in this model?
  - **A** Short answer: when $\delta_0 < 0$.

$$\delta_0 = E(wage|\overbrace{female = 1}^{women}, educ) - E(wage|\overbrace{female = 0}^{men}, educ)$$

- ✔ The situation can be depicted graphically as an **intercept shift** between males and females. ▶ View Graph

Dummy Variables - Example I (cont'd)



Dummy Variables - Example I (cont'd)

Hourly Wage Determination Model (cont'd)

Q How to fall in **dummy variable trap** in this example?

A Having dummy variables for both male and female is a simple illustration of the dummy variable trap.

$$female + male = 1$$

- ✔ The dummy variable trap occurs when there is perfect collinearity among the dummy variables, making estimation problematic.
- ✔ The choice of the **base group** or **benchmark group** is arbitrary but essential for interpretation.
- ✔ In the given example, males are chosen as the base group against which comparisons are made.

Q A loophole to avoid dummy variable trap is to omit the overall intercept of a model, but this isn't a favorable decision. Why? In this case it would be remodeled as:

$$wage = \beta_0 male + \alpha_0 female + \beta_1 educ + u$$

Dummy Variables - Example II

Effects of Training Grants on Hours of Training

$$\widehat{hrsemp} = \hat{\beta}_0 + \hat{\delta}_0 grant + \hat{\beta}_1 log(sales) + \hat{\beta}_2 log(employ)$$

Where

- $hrsemp$ = hours of training per employee at the firm level,
- $grant$ = a dummy variable equal to one if the firm received a job training grant,
- $sales$ = annual sales,
- $employ$ = number of employees.

Dummy Variables - Example III

Housing Price Regression

$$\widehat{log(price)} = \hat{\beta}_0 + \hat{\delta}_0 colonial + \hat{\beta}_1 log(lotsize) + \hat{\beta}_2 log(sqrft) + \hat{\beta}_3 bdrms$$

Where

- $price$ = house price,
- $colonial$ = a binary variable equal to one if the house is of the colonial style,
- $lotsize$ = lot size of house,
- $sqrft$ = house size in square feet,
- $bdrms$ = number of bedrooms in house.

Dummy Variables - Implications

- ✔ Example I has implications for policy analysis.
- ✔ But policy analysis often involves **program evaluation** to understand the effects of economic or social programs.
- ✔ The simplest case in program evaluation involves two groups: the **control group** and the **experimental (or treatment) group**. (Example II)
  - ✗ **Control Group**: A group of subjects or entities in an experiment that does not receive the experimental treatment or intervention.
  - ✗ **Treatment Group**: A group of subjects or entities in an experiment that receives the experimental treatment or intervention being studied.
- ✔ Also note that when independent variable is in logarithmic scale, the dummy variable coefficients have a percentage interpretation. (Example III)

Dummy Variables - Multiple Categories

✔ Several DVs could be used throughout a regression model.

A General Rule

If the regression model is to have different intercepts for, say, g groups or categories, we need to include $g - 1$ dummy variables in the model along with an intercept.

Dummy Variables - Multiple Categories (cont'd)

✔ An alternative is to include g dummy variables and exclude an overall intercept; however, this approach has practical drawbacks:

✗ First, it makes it more cumbersome to test for differences relative to a base group.

✗ Second, regression packages may use an **uncentered R-squared**, which can be misleading. The uncentered R-squared measure is different from $R^2 = 1 - {}^{SSR}/_{SST}$, since in uncentered R-squared SST is replaced with a total sum of squares that does not center y_i about its mean (SST_0). Also note that $R_0^2 \geq R^2$ is always true if $\bar{y}_i = 0$.

$$R_0^2 = 1 - \frac{SSR}{SST_0} \quad \left(\text{where } SST_0 = \sum_{i=1}^{n} y_i^2.\right)$$

Dummy Variables - Ordinal Information

- ✔ Sometimes the data is inherently **ordinal**, e.g., ratings.
- ✔ In this situation, for each level of order a DV must be defined.

The Effect of CR on MBR

Consider a study about the effect of city credit ratings (CR) on municipal bond interest rates (MBR). Suppose ratings take on integer values $\{0, 1, 2, 3, 4\}$, with 0 being the worst and 4 being the best.

$$MBR = \beta_0 + \delta_1 CR_1 + \delta_2 CR_2 + \delta_3 CR_3 + \delta_4 CR_4 + other\ factors$$

Where

- $CR_{i=\{0,1,2,3,4\}}$ = if $CR_i = 1$, then $CR_i = i$, and $CR_i = 0$ otherwise.

Q But why the following model isn't a good one (Where CR is supposed to be a self-explanatory variable)? Criticize it.

$$MBR = \beta_0 + \beta_1 CR + other\ factors$$

Dummy Variables - Interaction Term

Interaction Term

To capture the joint effect of multiple variables on the dependent variable, allowing the relationship between the variables to vary depending on the values of the interacting factors. This term is usually seen in the models as the product of interacting terms.

Effects of Training Grants on Hours of Training

$$\widehat{log(wage)} = \hat{\beta}_0 + \hat{\delta}_0 female + \hat{\delta}_1 married + \hat{\zeta}_1 female \cdot married$$

Where

- $female$ = When $female = 1$ the person is female and otherwise,
- $married$ = When $married = 1$ the person is married and otherwise,
- $female \cdot married$ = When $female = 0, married = 0$ the person is a single male and others.

Copyright Disclaimer