Dummy Variables

Class 10

Wonmun Shin

(wonmun.shin@sejong.ac.kr)

Department of Economics, Sejong University

* This lecture note is written based on Professor Chang Sik Kim's lecture notes.

Introduction to Dummy Variables

Dummy Variables

- Qualitative variables may be also important in explaining a dependent variable.

$$\begin{cases} \text{Quantitative:} & \text{Income, Cost, Prices, Wages, } \cdots \\ \text{Qualitative:} & \text{Sex, Race, Religion, Geographic Region, } \cdots \end{cases}$$

  - How can we quantify these attributes?
  - Use **dummy variables** (or **indicator variables**)
    - Example:

$$D_i = \begin{cases} 1 & \text{if the } i\text{-th person is female} \\ 0 & \text{if the } i\text{-th person is male} \end{cases}$$

One Dummy Variable

$$Y_i = \beta_1 + \beta_2 D_i + e_i$$

- Y_i: annual salary
- D_i takes a value of 1 for college graduates and 0 otherwise (non college graduates):

$$D_i = \begin{cases} 1 & \text{if college graduate} \\ 0 & \text{if non college graduate} \end{cases}$$

- What is the interpretation for dummy coefficient β_2?
  - $E[Y_i \mid D_i = 1] = \beta_1 + \beta_2$: Average salary of college graduates
  - $E[Y_i \mid D_i = 0] = \beta_1$: Average salary of non college graduates
  - Therefore, β_2 is <u>**difference in average salaries between two groups.**</u>

One Dummy Variable [cont'd]

$$Y_i = \beta_1 + \beta_2 D_i + e_i$$

- Note that we have two categories: college graduates and non college graduates
  - D_i takes a value 1 for college graduates $\rightarrow$ Dummy variable is assigned for the category of college graduates.
  - The category for which no dummy variable is assigned is known as the **benchmark category** $\rightarrow$ Non college graduate is the benchmark category.
  - The intercept value (β_1) represents the mean value of the benchmark category.
  - The coefficients attached to the dummy variable (β_2) is known as **differential intercept coefficient** 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.
- Regression models containing only dummy variables are called *ANOVA (Analysis of Variance)* models (because it analyzes variation between groups).

One Dummy Variable [cont'd]

$$Y_i = \beta_1 + \beta_2 D_i + e_i$$

- Estimation result:

$$\widehat{\text{salary}} = \underset{(57.74)}{18.0} + \underset{(7.44)}{3.28}\, D_i$$

(t-ratios in parentheses)

  - Unit of Y : \$1,000
  - Sample mean of salary of college graduates : $\hat{\beta}_1 + \hat{\beta}_2 = 18.0 + 3.28 = 21.28$
  - Sample mean of salary of non college graduates : $\hat{\beta}_1 = 18.0$

- Estimated difference $= 3.28$
  - Consider $H_0 : \beta_2 = 0$ (v.s. $H_1 : \beta_2 > 0$) $\Rightarrow$ We reject H_0 at 5% significance level.
  - College graduates receive ***significantly more*** salary on average than non college graduates!

Dummy Variable Regression Models

Multiple Dummies: Example 1

- Let's continue to consider ANOVA models *(salary and region)*.
- Suppose

$$\text{Region} \begin{cases} \text{Seoul} \\ \text{Incheon} \\ \text{Gyeonggi} \end{cases}$$

  - The qualitative variable "Region" has three categories.
  - Choose the benchmark category: "Gyeonggi"
  - We will introduce two dummies: D_{1i}, D_{2i}

Multiple Dummies: Example 1 [cont'd]

$$Y_i = \beta_1 + \beta_2 D_{1i} + \beta_3 D_{2i} + e_i$$

where

$$D_{1i} = \begin{cases} 1 & \text{if Seoul} \\ 0 & \text{otherwise} \end{cases}$$

$$D_{2i} = \begin{cases} 1 & \text{if Incheon} \\ 0 & \text{otherwise} \end{cases}$$

- β_1: Average salary of persons living in Gyeonggi
- $\beta_1 + \beta_2$: Average salary of persons living in Seoul
- $\beta_1 + \beta_3$: Average salary of persons living in Incheon
  - How about $\beta_1 + \beta_2 + \beta_3$? $\Rightarrow$ Not happens, because it is impossible to live in two regions at the same time.
- We can test $\beta_2 = 0$, $\beta_3 = 0$ to see if there are significant differences in salary across regions.

Multiple Dummies: Example 2

- Imagine we are interested in the effect of more than one qualitative variables on salary.
- Suppose now we want to see the effects of ***(a) region*** *(Seoul or not)*, and ***(b) sex*** *(male or female)*

$$Seoul_i = \begin{cases} 1 & \text{if person } i \text{ lives in Seoul} \\ 0 & \text{otherwise} \end{cases}$$

$$Male_i = \begin{cases} 1 & \text{if person } i \text{ is male} \\ 0 & \text{otherwise} \end{cases}$$

Multiple Dummies: Example 2 [cont'd]

$$Y_i = \beta_1 + \beta_2 Seoul_i + \beta_3 Male_i + e_i$$

1. β_1 (benchmark): Average salary of females living outside Seoul
2. $\beta_1 + \beta_2$: Average salary of females living in Seoul
3. $\beta_1 + \beta_3$: Average salary of males living outside Seoul
4. $\beta_1 + \beta_2 + \beta_3$: Average salary of males living in Seoul

- Regional difference: ② − ① = ④ − ③ = β_2
- Sexual difference: ③ − ① = ④ − ② = β_3

One Dummy + One Quantitative Variable

$$Y_i = \beta_1 + \beta_2 D_i + \beta_3 X_i + e_i$$

- Y_i: annual salary of a teacher, X_i: years of experience
- D_i takes a value of 1 for male teachers and 0 for female teachers:

$$D_i = \begin{cases} 1 & \text{if male} \\ 0 & \text{if female} \end{cases}$$

- What is the interpretation for dummy coefficient β_2?
  - $E\left[Y_i \mid D_i = 1\right] = \beta_1 + \beta_2 + \beta_3 X_i$: Average salary of male teachers
  - $E\left[Y_i \mid D_i = 0\right] = \beta_1 + \beta_3 X_i$: Average salary of female teachers
  - β_2 is **difference in average salaries between two groups, *controlling for experience years*.**
- Regression models containing qualitative as well as quantitative variables are called *ANCOVA (Analysis of Covariance)* models.

One Dummy + One Quantitative Variable [cont'd]

- Estimation result:

$$\widehat{\text{salary}} = \underset{(93.61)}{17.969} + \underset{(38.45)}{3.3336}\, D_i + \underset{(21.46)}{1.3707}\, X_i$$

(t-ratios in parentheses)

  - Unit of Y : \$1,000
  - $\hat{\beta}_3 = 1.3707$: On average, one additional year of experience increases teacher's salary by about 1.37 thousand dollars.
  - $\hat{\beta}_2$ is the estimated differential intercept coefficient, controlling for X_i.
- Estimated difference $= 3.3336$
  - Is there sexual discrimination? *i.e.* Given the same experience, do male teachers earn more than female teachers on average?
  - Consider $H_0 : \beta_2 = 0$ (v.s. $H_1 : \beta_2 > 0$) $\Rightarrow$ Reject H_0.
  - $\therefore$ Yes, there is sexual discrimination!

Dummy Variable Trap

Dummy Variable Trap

$$Y_i = \beta_1 + \beta_2 D_{1i} + \beta_3 D_{2i} + e_i$$

where

$$D_{1i} = \begin{cases} 1 & \text{if male} \\ 0 & \text{if female} \end{cases}$$

$$D_{2i} = \begin{cases} 1 & \text{if female} \\ 0 & \text{if male} \end{cases}$$

- In order to figure out the existence of sexual discrimination, you might want to test

$$H_0 : \beta_2 = \beta_3 \quad v.s. \quad H_1 : \beta_2 > \beta_3$$

Dummy Variable Trap [cont'd]

- **HOWEVER**, the above model can't be estimated due to the problem of **perfect multicollinearity**!
  - We have $D_{1i} + D_{2i} = 1$, and 1 is the constant regressor.
- **General rule**: If a qualitative variable has m categories, introduce only $(m - 1)$ dummy variables.
  - For each qualitative regressor, the number of dummy variables introduced must be one less than the categories of that variable.
- If you do not follow this rule, you will fall into what is called the **dummy variable trap**.
- One way to circumvent this trap: If we do omit the intercept in the model, we can introduce as many dummy variables as the number of categories.
  - Caution: Make sure that when you run this regression, you use the no-intercept option in your regression package.

Application: Interaction Dummy

Interaction Dummy

- Recall, Example 2 of salary with region and sex dummies.

$$Y_i = \beta_1 + \beta_2 Seoul_i + \beta_3 Male_i + e_i$$

1. β_1 (benchmark): Average salary of females living outside Seoul
2. $\beta_1 + \beta_2$: Average salary of females living in Seoul
3. $\beta_1 + \beta_3$: Average salary of males living outside Seoul
4. $\beta_1 + \beta_2 + \beta_3$: Average salary of males living in Seoul

- Implicit in this model is the assumption that the differential effect of the region dummy $Seoul_i$ is **constant** across the two categories of sex, and the differential effect of the sex dummy $Female_i$ is also **constant** across the regions.
  - If the mean salary is higher for Seoul residents, this is so whether they are male or not.
  - If the mean salary is higher for males than for females, this is so whether they live in Seoul or not.

Interaction Dummy [cont'd]

- In many applications, such an assumption may be untenable.
  - A male Seoul resident may earn more wages than a male non-Seoul resident.
  - In other words, there may be interaction between the two qualitative variables $Seoul_i$ and $Male_i$.

$$Y_i = \beta_1 + \beta_2 Seoul_i + \beta_3 Male_i + \beta_4 \left(Seoul_i \cdot Male_i \right) + e_i$$

1. β_1 (benchmark): Average salary of females living outside Seoul
2. $\beta_1 + \beta_2$: Average salary of females living in Seoul
3. $\beta_1 + \beta_3$: Average salary of males living outside Seoul
4. $\beta_1 + \beta_2 + \beta_3 + \beta_4$: Average salary of males living in Seoul

Interaction Dummy [cont'd]

- Now, the regional differences in two groups are not same!
  - Male wage difference between Seoul and non-Seoul $= ④ - ③ = \beta_2 + \beta_4$
  - Female wage difference between Seoul and non-Seoul $= ② - ① = \beta_2$
- Is there more regional difference (Seoul vs. non-Seoul) among males than among females?
  - Test $H_0 : \beta_4 = 0 \quad vs. \quad H_1 : \beta_4 > 0$

Slope Dummy

- One can interact dummy variables with other quantitative regressors.
- Consider the following model:

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 (D_i X_i) + e_i$$

  - D_i is a dummy variable and X_i is a quantitative variable.
  - $D_i X_i$ is the product of a dummy variable and a (quantitative) regressor, called a **slope dummy** because it allows for a change in the slope of relationship.

$$E(Y_i) = \begin{cases} \beta_1 + (\beta_2 + \beta_3) X_i & \text{when } D_i = 1 \\ \beta_1 + \beta_2 X_i & \text{when } D_i = 0 \end{cases}$$

$$\implies \frac{\partial E(Y_i)}{\partial X_i} = \begin{cases} \beta_2 + \beta_3 & \text{when } D_i = 1 \\ \beta_2 & \text{when } D_i = 0 \end{cases}$$

Slope Dummy [cont'd]

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 (D_i X_i) + e_i$$

- **Example:** $Y_i =$ house price, $X_i =$ size of house (in square feet), $D_i = 1$ if the house is in the desirable neighborhoods, $D_i = 0$ if the house is in other neighborhoods.
- $D_i X_i$: Indicates the interaction effect of location and size on house price
- Interpretation
  - In the desirable neighborhoods, as the house size goes up by 1 square feet, the house price goes up by $\beta_2 + \beta_3$.
  - In other neighborhoods, the effect of house size on prices is β_2.