Econometrics: Regression Analysis With Qualitative Information

Burcu Eke

UC3M

Introduction

- In the regression model, there are often variables of interest that are qualitative and can not be measured as a quantitative variable.
- These variables, called "dummy", or "binary" variables, measure some qualitative characteristics such as:
  - Gender male or female;
  - Immigration status: immigrant or not;
  - Marital status: married or not;
  - Residence status reside in a particular city or not;
  - Sector of a company: manufacturing or service sector;
  - Size of a company: big or small;
  - Month of the year, and so on.

Dummy Variables

- Using dummy variables, we can measure the effect of the qualitative factor on our dependent variable

- Typically, the dummy variables take value 1 in a category and value 0 "otherwise". "Otherwise" can represent one or more other categories. For example:

$$\text{Female} = \begin{cases} 1 & \text{if the individual is female} \\ 0 & \text{if the individual is male} \end{cases}$$

$$\text{Male} = \begin{cases} 1 & \text{if the individual is male} \\ 0 & \text{if the individual is female} \end{cases}$$

Dummy Variables

$$\text{Small} = \begin{cases} 1 & \text{if the firm is small} \\ 0 & \text{otherwise} \end{cases}$$

$$\text{Medium} = \begin{cases} 1 & \text{if the firm is medium size} \\ 0 & \text{otherwise} \end{cases}$$

$$\text{Big} = \begin{cases} 1 & \text{if the firm is big} \\ 0 & \text{otherwise} \end{cases}$$

Dummy Variables

- Dummy variables help us with two different aspects
  - **Additive** dummy variables measure differences in groups with respect to the intercept term
  - **Interaction** dummy variables measure differences in groups with respect to the slope term
- **Dummy variable trap**: Suppose you have a set of multiple dummy variables for multiple categories and every observation falls in one and only one category. Then, if you include all these dummy variables and a constant term (β_0), you will have perfect multicollinearity. Also known as dummy variable trap.

Additive Dummy Variables

- Additive dummy variables result in different intercepts for different populations.

- Consider the following model
  $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i \; i = 1, \ldots, n$, where
  - Y_i is the wage rate of individual i,
  - X_{1i} is the years of schooling for individual i, and
  - $X_{2i} = \begin{cases} 1 & \text{if the individual is female} \\ 0 & \text{if the individual is male} \end{cases}$

- So, we have $E[Y|X_{1i}, X_{2i}] = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}$. This implies
  - For females: $E[Y|X_{1i}, X_{2i} = 1] = (\beta_0 + \beta_2) + \beta_1 X_{1i}$
  - For males: $E[Y|X_{1i}, X_{2i} = 0] = \beta_0 + \beta_1 X_{1i}$

Additive Dummy Variables

- $\beta_2 = E[Y|X_{1i},\ \text{female}] - E[Y|X_{1i},\ \text{male}]$ is the average difference between a women and a man for a given level of education.
- Assuming that $\beta_2 < 0$, graphically we have:

Additive Dummy Variables

- There are two alternative formulations for this model:
  1. $Y_i = \alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{3i} + \varepsilon \; i = 1, \ldots, n$, where:
  $$X_{3i} = \begin{cases} 1 & \text{if the individual is male} \\ 0 & \text{if the individual is female} \end{cases}$$
  2. $Y_i = \delta_1 X_{1i} + \delta_2 X_{2i} + \delta_3 X_{3i} + \varepsilon \; i = 1, \ldots, n$

Additive Dummy Variables: Alternative Model (1)

- $Y_i = \alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{3i} + \varepsilon_i \; i = 1, \ldots, n$. Now we have:
- $E[Y|X_{1i}, X_3] = \alpha_0 + \alpha_1 X_1 + \alpha_3 X_3$, hence
  - $E[Y|X_{1i},\ \text{female}] = E[Y|X_{1i}, X_{3i} = 0] = \alpha_0 + \alpha_1 X_1$,
  - $E[Y|X_{1i},\ \text{male}] = E[Y|X_{1i}, X_{3i} = 1] = (\alpha_0 + \alpha_2) + \alpha_1 X_1$,
  - $\alpha_2 = E[Y|X_{1i},\ \text{male}] - E[Y|X_{1i},\ \text{female}]$ is the average difference between a women and a man for a given level of education.
  - Therefore our model should satisfy:

$$\begin{aligned} \alpha_1 &= \beta_1 \\ \alpha_0 &= \beta_0 + \beta_2 \\ \alpha_0 + \alpha_2 &= \beta_0 \end{aligned}$$

Additive Dummy Variables: Alternative Model (2)

- $Y_i = \delta_1 X_{1i} + \delta_2 X_{2i} + \delta_3 X_{3i} + \varepsilon_i \; i = 1, \ldots, n$. Now we have:
- $E[Y|X_{1i}, X_2, X_3] = \delta_1 X_{1i} + \delta_2 X_{2i} + \delta_3 X_{3i}$, hence
  - $E[Y|X_{1i}, \text{ female}] = E[Y|X_{1i}, X_{2i} = 1, X_{3i} = 0] = \delta_2 + \delta_1 X_{1i}$,
  - $E[Y|X_{1i}, \text{ male}] = E[Y|X_{1i}, X_{2i} = 0, X_{3i} = 1] = \delta_3 + \delta_1 X_{1i}$,
  - $\delta_3 - \delta_2 = E[Y|X_{1i}, \text{ male}] - E[Y|X_{1i}, \text{ female}]$ is the average difference between a women and a man for a given level of education.
  - Therefore our model should satisfy:

$$
\begin{aligned}
\delta_1 &= \alpha_1 = \beta_1 \\
\delta_2 &= \alpha_0 = \beta_0 + \beta_2 \\
\delta_3 &= \alpha_0 + \alpha_2 = \beta_0
\end{aligned}
$$

Additive Dummy Variables

- However, notice that a model like
  $Y_i = \delta_0 + \delta_1 X_{1i} + \delta_2 X_{2i} + \delta_3 X_{3i} + \varepsilon_i \; i = 1, \ldots, n$
  Would **not** be valid due to multicollinearity (Recall problem 2 of set 3)

Additive Dummy Variables

- How would we test if there are significant differences between the two groups: male and female?
  - For model $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i \;\Rightarrow\; H_0 : \beta_2 = 0$ vs. $H_1 : \beta_2 \neq 0$
  - For model $Y_i = \alpha_0 + \alpha_1 X_{1i} + \alpha_3 X_{3i} + \varepsilon_i \;\Rightarrow\; H_0 : \alpha_3 = 0$ vs. $H_1 : \alpha_3 \neq 0$
  - For model $Y_i = \delta_1 X_{1i} + \delta_2 X_{2i} + \delta_3 X_{3i} + \varepsilon_i \;\Rightarrow\; H_0 : \delta_2 = \delta_3$ vs. $H_1 : \delta_2 \neq \delta_3$

Interaction Dummy Variables

- We use interaction dummy variables to account for the changes due to the dummy categories, in the effect of the independent variables, i.e., X_1: education,? on Y
- Consider an example with additive and interaction effects: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{4i} + \varepsilon_i \; i = 1, \ldots, n$, where $X_{4i} = X_{1i} \times X_{2i}$.
- In this case, $X_{4i} = \begin{cases} X_{1i} & \text{if the individual is female} \\ 0 & \text{if the individual is male} \end{cases}$
- So, we have $E[Y|X_{1i}, X_{2i}, X_{4i}] = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{4i}$. This implies
  - For females: $E[Y|X_{1i},\ \text{female}] = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_{1i}$
  - For males: $E[Y|X_{1i}, X_{2i} = 0] = \beta_0 + \beta_1 X_{1i}$

Interaction Dummy Variables

- β_2 measures the difference in the intercept term between men and women. That is, it is the difference on the mean income of men and women
- β_3 measures the difference in the slope term between men and women. That is, if education (X_1) increases by 1 year, the on average, the hourly wage increases by:
  - $\beta_1 + \beta_3$ units for women, and
  - β_1 units for men.
  - Thus, measures the differences in the average effect of education on wages due to different genders

Interaction Dummy Variables

- How to test if there are significant differences between genders for the effect of education on the wage rate
  - $\Rightarrow\; H_0 : \beta_3 = 0$ vs. $H_1 : \beta_3 \neq 0$
- How to test if there are significant differences between genders, on average
  - $\Rightarrow\; H_0 : \beta_2 = 0$ vs. $H_1 : \beta_2 \neq 0$
- How to test if there are any significant difference between men and women
  - $\Rightarrow\; H_0 : \beta_2 = \beta_3 = 0$ vs. $H_1 : \beta_2 \neq 0$ and/or $\beta_2 \neq 0$

Interaction Dummy Variables: Additional Comments

- As in additive dummy variable models, there are alternative specifications for the interaction dummy variable models.
  - For example:
    $Y_i = \alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{3i} + \alpha_3 X_{5i} + \varepsilon_i \; i = 1, \ldots, n$, where $X_{5i} = X_{1i} \times X_{3i}$
    - In this case, $X_{3i} = \begin{cases} 1 & \text{if the individual is male} \\ 0 & \text{if the individual is falemale} \end{cases}$
    - In this case, $X_{5i} = \begin{cases} X_{1i} & \text{if the individual is male} \\ 0 & \text{if the individual is falemale} \end{cases}$
  - Alternatively:
    $Y_i = \delta_1 X_{2i} + \delta_2 X_{3i} + \delta_3 X_{4i} + \delta_4 X_{5i} + \varepsilon_i \; i = 1, \ldots, n$

Interaction Dummy Variables

- However, a model like the following will **not** be valid: $Y_i = \gamma_1 X_{1i} + \gamma_2 X_{2i} + \gamma_3 X_{3i} + \gamma_4 X_{4i} + \gamma_5 X_{5i} + \varepsilon_i \; i = 1, \ldots, n$ since it violates **A4** (no perfect multicollinearity) because $X_{4i} + X_{5i} = X_{1i} \; \forall i \in 1, \ldots, n$

Interaction Dummy Variables

- We may have more than two categories for our dummy variable. For example, assume that firms are divided into three sectors, i.e., services, manufacturing, and agriculture

- $V_i = \alpha_0 + \alpha_1 S_{1i} + \alpha_2 S_{2i} + \alpha_3 P_i + \alpha_4 (P_i \times S_{1i}) + \alpha_5 (P_i \times S_{2i}) + \varepsilon_i \; i = 1, \ldots, n$, where
  - V_i = Sales of the company i
  - P_i = Advertising expenditures of the company i
  - $S_{1i} = \begin{cases} 1 & \text{if the company } i \text{ belongs to sector 1} \\ 0 & \text{otherwise} \end{cases}$
  - $S_{2i} = \begin{cases} 1 & \text{if the company } i \text{ belongs to sector 2} \\ 0 & \text{otherwise} \end{cases}$

Interaction Dummy Variables

- Then:
  - $E[V_i|P_i,\text{ sector 1}] = (\alpha_0 + \alpha_1) + (\alpha_3 + \alpha_4)P_i$
  - $E[V_i|P_i,\text{ sector 2}] = (\alpha_0 + \alpha_2) + (\alpha_3 + \alpha_5)P_i$
  - $E[V_i|P_i,\text{ sector 3}] = \alpha_0 + \alpha_3 P_i$

Interaction Dummy Variables

- In this particular representation of the model, in order to include both the constant term and the variable P_i , we exclude the additive and interaction effects corresponding to sector 3, and only included those of sector 1 ans 2
  - α_0 corresponds to the additive dummy for sector 3 (the constant term for sector 3)
  - α_3 corresponds to the interaction dummy for sector 3 (the effect of advertising on sector 3 sales)which we ignore (Sector 3)
  - The intercept for the other sectors, namely, 1 and 2 are $(\alpha_0 + \alpha_1)$ and $(\alpha_0 + \alpha_2)$, respectively
  - The slopes for the other sectors, namely, 1 and 2 are $(\alpha_3 + \alpha_4)$ and $(\alpha_3 + \alpha_5)$, respectively

Interaction Dummy Variables

- There are many alternative representations for this model. One possible way is: $V_i = \delta_1 S_{1i} + \delta_2 S_{2i} + \alpha_3 S_{3i} + \delta_4 \left(P_i \times S_{1i}\right) + \delta_5 \left(P_i \times S_{2i}\right) + \delta_6 \left(P_i \times S_{3i}\right) + \varepsilon_i \; i = 1, \ldots, n$

- Comparing both representation, what are the relationships between α_j's and δ_j's?

- How would you test for the effects of sector on sales?