LECTURE 6

Introduction to Econometrics

Nonlinear specifications and dummy variables

Hieu Nguyen

Fall semester, 2024

NONLINEAR SPECIFICATION

- We will discuss different specifications nonlinear in dependent and independent variables and their interpretation

- We will define the notion of a dummy variable and we will show its different uses in linear regression models

NONLINEAR SPECIFICATION

- There is not always a linear relationship between dependent variable and explanatory variables
  - The use of OLS requires that the equation be linear in coefficients
  - However, there is a wide variety of functional forms that are linear in coefficients while being nonlinear in variables!
- We have to choose carefully the functional form of the relationship between the dependent variable and each explanatory variable
  - The choice of a functional form should be based on the underlying economic theory and/or intuition
  - Do we expect a curve instead of a straight line? Does the effect of a variable peak at some point and then start to decline?

Linear form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

- Assumes that the effect of the explanatory variable on the dependent variable is constant:

$$\frac{\partial y}{\partial x_k} = \beta_k \qquad k = 1, 2$$

- Interpretation: if x_k increases by 1 **unit** (in which x_k is measured), then y will change by β_k **units** (in which y is measured)

- Linear form is used as default functional form until strong evidence that it is inappropriate is found

LOG-LOG FORM

$$\ln y = \beta_0 + \beta_1 \ln x_1 + \beta_2 \ln x_2 + \varepsilon$$

- Assumes that the elasticity of the dependent variable with respect to the explanatory variable is constant:

$$\frac{\partial \ln y}{\partial \ln x_k} = \frac{\partial y / y}{\partial x_k / x_k} = \beta_k \qquad k = 1, 2$$

- Interpretation: if x_k increases by 1 **percent**, then y will change by β_k **percents**

- Before using a double-log model, make sure that there are no negative or zero observations in the data set

EXAMPLE

- Estimating the production function of Indian sugar industry:

$$\widehat{\ln Q} = \underset{}{2.70} + \underset{(0.14)}{0.59} \ln L + \underset{(0.17)}{0.33} \ln K$$

Q . . . output L . . . labor K . . . capital employed

Interpretation: if we increase the amount of labor by 1%, the production of sugar will increase by 0.59%, ceteris paribus.

Ceteris paribus is a Latin phrase meaning 'other things being equal'.

LOG-LINEAR FORMS

- Linear-log form:

$$y = \beta_0 + \beta_1 \ln x_1 + \beta_2 \ln x_2 + \varepsilon$$

Interpretation: if x_k increases by 1 **percent**, then y will change by $(\beta_k/100)$ **units** ($k = 1, 2$)

- Log-linear form:

$$\ln y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

Interpretation: if x_k increases by 1 **unit**, then y will change by $(\beta_k * 100)$ **percent** ($k = 1, 2$)

EXAMPLES OF LOG LINEAR FORMS

- Estimating demand for chicken meat:

$$\widehat{Y} = -6.94 - \underset{(0.19)}{0.57}\ PC + \underset{(0.11)}{0.25}\ PB + \underset{(2.81)}{12.2}\ \ln YD$$

Y . . . annual chicken consumption (kg.)
PC . . . price of chicken
PB . . . price of beef
YD . . . annual disposable income

- Interpretation: An increase in the annual disposable income by 1% increases chicken consumption by 0.12 kg per year, ceteris paribus.

EXAMPLES OF LOG LINEAR FORMS

- Estimating the influence of education and experience on wages:

$$\widehat{\ln wage} = 0.217 + \underset{(0.008)}{0.098}\ educ + \underset{(0.002)}{0.010}\ exper$$

wage	. . . annual wage (USD)
educ	. . . years of education
exper	. . . years of experience

- Interpretation: An increase in education by one year increases annual wage by 9.8%, ceteris paribus. An increase in experience by one year increases annual wage by 1%, ceteris paribus.

Polynomial form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \varepsilon$$

- To determine the effect of x_1 on y, we need to calculate the derivative:
$$\frac{\partial y}{\partial x_1} = \beta_1 + 2 \cdot \beta_2 \cdot x_1$$
- Clearly, the effect of x_1 on y is not constant, but changes with the level of x_1
- We might also have higher order polynomials, e.g.:
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_1^3 + \beta_4 x_1^4 + \varepsilon$$

Example of polynomial form

- The impact of the number of hours of studying on the grade from Introductory Econometrics:

$$\widehat{grade} = 30 + 1.4 \cdot hours - 0.009 \cdot hours^2$$

- To determine the effect of hours on grade, calculate the derivative:

$$\frac{\partial y}{\partial x} = \frac{\partial grade}{\partial hours} = 1.4 - 2 \cdot 0.009 \cdot hours = 1.4 - 0.018 \cdot hours$$

Decreasing returns to hours of studying: more hours implies higher grade, but the positive effect of additional hour of studying decreases with more hours

Choice of correct functional form

- The functional form has to be correctly specified in order to avoid biased and inconsistent estimates

  Remember that one of the OLS assumptions is that the model is correctly specified

- Ideally: the specification is given by underlying theory of the equation
- In reality: underlying theory does not give precise functional form
- In most cases, either linear form is adequate, or common sense will point out an easy choice from among the alternatives

Choice of correct functional form

- Nonlinearity of independent variables
  - often approximated by polynomial form
  - missing higher powers of a variable can be detected as omitted variables (see next lecture)

- Nonlinearity of dependent variable
  - harder to detect based on statistical fit of the regression
  - R^2 is incomparable across models where the y is transformed
  - dependent variables are often transformed to log-form in order to make their distribution closer to the normal distribution

DUMMY VARIABLES

- Dummy variable - takes on the values of 0 or 1, depending on a qualitative attribute

- Examples of dummy variables:

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

$$Weekend = \begin{cases} 1 & \text{if the day is on weekend} \\ 0 & \text{if the day is a work day} \end{cases}$$

$$NewStadium = \begin{cases} 1 & \text{if the team plays on new stadium} \\ 0 & \text{if the team plays on old stadium} \end{cases}$$

INTERCEPT DUMMY

- Dummy variable included in a regression alone (not interacted with other variables) is an intercept dummy
- It changes the intercept for the subset of data defined by a dummy variable condition:

$$y_i = \beta_0 + \beta_1 D_i + \beta_2 x_i + \varepsilon_i$$

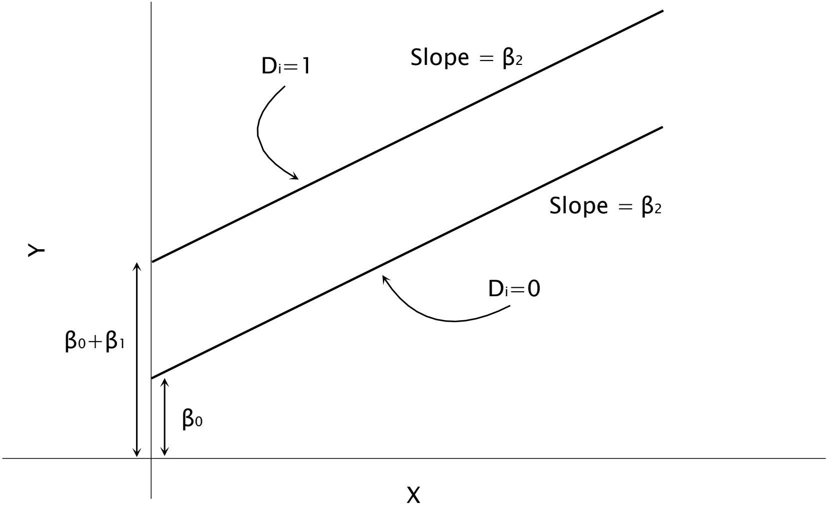
where

$$D_i = \begin{cases} 1 & \text{if the } i\text{-th observation meets a particular condition} \\ 0 & \text{otherwise} \end{cases}$$

- We have

$$\begin{aligned} y_i &= (\beta_0 + \beta_1) + \beta_2 x_i + \varepsilon_i \quad \text{if } D_i = 1 \\ y_i &= \beta_0 + \beta_2 x_i + \varepsilon_i \quad \text{if } D_i = 0 \end{aligned}$$

INTERCEPT DUMMY



EXAMPLE

Estimating the determinants of wages:

$$\widehat{wage}_i = -3.890 + \underset{(0.270)}{2.156}\ M_i + \underset{(0.051)}{0.603}\ educ_i + \underset{(0.064)}{0.010}\ exper_i$$

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

$wage$ ... average hourly wage in USD

Interpretation of the dummy variable M: men earn on average \$2.156 per hour more than women, ceteris paribus

Slope dummy

- If a dummy variable is interacted with another variable (x), it is a slope dummy.
- It changes the relationship between x and y for a subset of data defined by a dummy variable condition:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 (x_i \cdot D_i) + \varepsilon_i$$

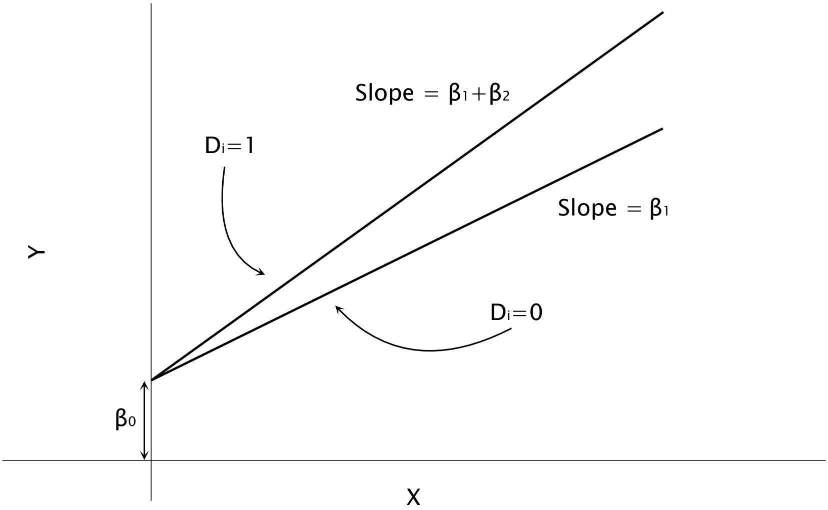
where

$$D_i = \begin{cases} 1 & \text{if the } i\text{-th observation meets a particular condition} \\ 0 & \text{otherwise} \end{cases}$$

- We have

$$\begin{aligned} y_i &= \beta_0 + (\beta_1 + \beta_2) x_i + \varepsilon_i \quad \text{if } D_i = 1 \\ y_i &= \beta_0 + \beta_1 x_i + \varepsilon_i \quad \text{if } D_i = 0 \end{aligned}$$

SLOPE DUMMY



EXAMPLE

- Estimating the determinants of wages:

$$\widehat{wage}_i = -2.620 + \underset{(0.054)}{0.450}\ educ_i + \underset{(0.021)}{0.170}\ M_i \cdot educ_i + \underset{(0.065)}{0.010}\ exper_i$$

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

$wage$ . . . average hourly wage in USD

- Interpretation: men gain on average 17 cents per hour more than women for each additional year of education, ceteris paribus

Slope and intercept dummies

- Allow both for different slope and intercept for two subsets of data distinguished by a qualitative condition:

$$y_i = \beta_0 + \beta_1 D_i + \beta_2 x_i + \beta_3 (x_i \cdot D_i) + \varepsilon_i$$

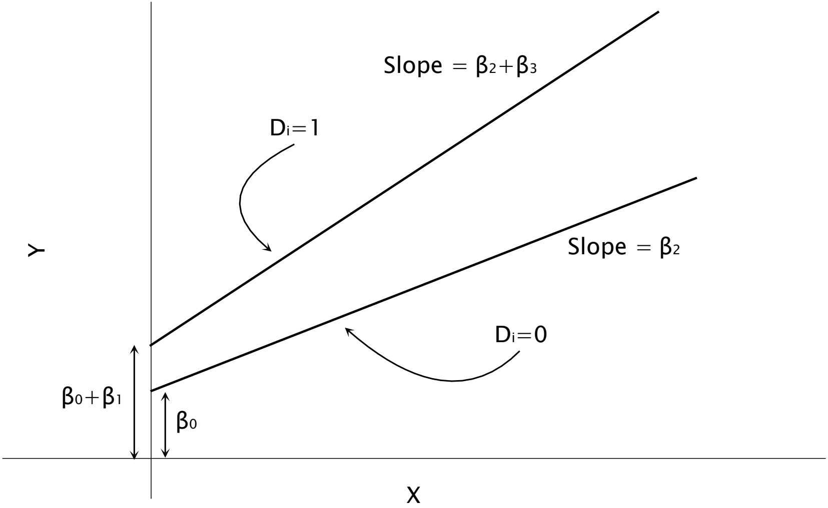
where

$$D_i = \begin{cases} 1 & \text{if the } i\text{-th observation meets a particular condition} \\ 0 & \text{otherwise} \end{cases}$$

- We have

$$\begin{aligned} y_i &= (\beta_0 + \beta_1) + (\beta_2 + \beta_3) x_i + \varepsilon_i \quad \text{if } D_i = 1 \\ y_i &= \beta_0 + \beta_2 x_i + \varepsilon_i \quad \text{if } D_i = 0 \end{aligned}$$

SLOPE AND INTERCEPT DUMMIES



Dummy variables - multiple categories

- What if a variable defines three or more qualitative attributes?
- Example: level of education - elementary school, high school, and college
- Define and use a set of dummy variables:

$$H = \begin{cases} 1 & \text{if high school} \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad C = \begin{cases} 1 & \text{if college} \\ 0 & \text{otherwise} \end{cases}$$

- Should we include also a third dummy in the regression, which is equal to 1 for people with elementary education?

  No, unless we exclude the intercept!

  Using full set of dummies leads to perfect multicollinearity (dummy variable trap)

SUMMARY

- We discussed different nonlinear specifications of a regression equation and their interpretation

- We defined the concept of a dummy variable and we showed its use

- Further readings:

  Studenmund, Chapter 7

  Wooldridge, Chapters 6 & 7