Recitation 6

Seung-hun Lee

Columbia University

Nonlinear regressors

Natural logs

- We might be interested in how the changes in the X in terms of *percentages* affect changes in the dependent variable Y.
- (Natural) Log regressors captures the idea of percentage changes.
- Recall from calculus the first order approximation: For any differentiable funtion f, and a very small change in x, the following relationship holds.

$$f(x + \Delta x) \simeq f(x) + f'(x)[(x + \Delta x) - x]$$

- Define $y = f(x) = \ln x$. Then $f(x + \Delta x) = y + \Delta y$ and $f'(x) = \frac{1}{x}$. We then get

$$\Delta y = \frac{\Delta x}{x} \implies \ln(x + \Delta x) - \ln x \simeq \frac{\Delta x}{x} \implies \ln\left(1 + \frac{\Delta x}{x}\right) \simeq \frac{\Delta x}{x}$$

- Therefore, the changes in log values allows us to capture changes in percentages, at least for very small amount of change.

Nonlinear regressors

Types of log regressions

- **lin-**log: Consider the model $Y = \beta_0 + \beta_1 \log X + u$.
- I take a before and after approach by changing X by Δx.
- Then the total amount of Y would be $Y + \Delta y$. Formally,

$$Y + \Delta y = \beta_0 + \beta_1 \log(X + \Delta x) + u$$

Subtract $Y = \beta_0 + \beta_1 \log X + u$ from this equation to get

$$\Delta y = \beta_1 \log(X + \Delta x) - \beta_1 \log X = \beta_1 \log\left(1 + \frac{\Delta x}{X}\right) = \beta_1 \frac{\Delta x}{X}$$

- Therefore,

$$\beta_1 = \frac{\Delta y}{(\Delta x / X)}$$

Note that the percentage change in X is $\frac{\Delta x}{X} \times 100$. In words, change in X by 1 percent, raises Y by $\beta_1 \times 0.01$.

Nonlinear regressors

Types of log regressions

- log-**lin**: Consider the model $\log Y = \beta_0 + \beta_1 X + u$.
- Conduct a similar before and after analysis as we did before to get:

$$\log(Y + \Delta y) = \beta_0 + \beta_1(X + \Delta x) + u$$

- Then, subtract $\log Y = \beta_0 + \beta_1 X + u$. This gets us

$$\log\left(1 + \frac{\Delta y}{Y}\right) \simeq \frac{\Delta y}{Y} = \beta_1 \Delta x$$

- Then, β_1 can be backed out as

$$\beta_1 = \frac{(\Delta y / Y)}{\Delta x}$$

Again, using the fact that percentage change in Y can be represented as $\frac{\Delta y}{Y} \times 100$. This implies that a 1 unit change in X raises Y by $(100 \times \beta_1)\%$.

Nonlinear regressors

Types of log regressions

- log-log: Consider the equation $\log Y = \beta_0 + \beta_1 \log X + u$.
- Similar approach allows us to write

$$\log(Y + \Delta y) = \beta_0 + \beta_1 \log(X + \Delta x) + u$$

- Subtract the original equation to obtain

$$\log\left(1 + \frac{\Delta y}{Y}\right) = \beta_1 \log\left(1 + \frac{\Delta x}{X}\right) \implies \frac{\Delta y}{Y} = \beta_1 \frac{\Delta x}{X}$$

- which implies

$$\beta_1 = \frac{(\Delta y / Y)}{(\Delta x / X)}$$

This implies that 1% change in X leads to β_1% change in Y. This is the *elasticity* interpretation

Nonlinear regressors

Interaction terms: Motivation

- Suppose that we are interested in the relationship between test scores (Y) and class size(X_1).
- However, one might guess that the effect of class size may differ depending on some other variables.
  - e.g. schools in districts where there are more funding (X_2) are more likely to enjoy the benefits of small school classroom
- In math, the marginal effect of X_1 on Y may depend on X_2.
- To capture this idea in a model, we can incorporate an **interaction term** involving X_1 and X_2, which can be written as $X_1 \times X_2$

Nonlinear regressors

Interaction terms

- **Binary $\times$ Binary:** Suppose that there are two binary variables, D_1, D_2.
- Notice the regression equation below

$$Y = \beta_0 + \beta_1 D_1 + \beta_2 D_2 + \beta_3 (D_1 \times D_2) + u$$

- By now, you know that β_1 captures the group difference in average for those in group D_1 and β_2 captures the same for those in group D_2.
- Note that $D_1 \times D_2$ becomes 1 only individual is in both groups. The average for these individuals are captured by β_3 coefficient.
- To sum up:
  - $E[Y|D_1 = 1] = \beta_0 + \beta_1 + \beta_2 \times D_2 + \beta_3 \times D_2$
  - $E[Y|D_1 = 0] = \beta_0 + \beta_2 \times D_2$
  - $E[Y|D_1 = 1] - E[Y|D_1 = 0] = \beta_1 + \beta_3 \times D_2$
- So the effect of D_1 differs depending on D_2 as well.
- This setup allows us to analyze the difference in effect of binary variable depending on another binary factor.

Nonlinear regressors

Interaction terms

- **Binary $\times$ Continuous:** Instead of a second binary variable, we include a continuous variable X_1.
- Now we write

$$Y = \beta_0 + \beta_1 D_1 + \beta_2 X_1 + \beta_3 (D_1 \times X_1) + u$$

- In earlier example, classroom size would be a continuous variable, so that will be our X_1.
- Now, define $D_i = 1$ if a school district receives funding and 0 if otherwise.
- The effect of classroom size would now be

$$\frac{\partial Y}{\partial X_1} = \beta_2 + \beta_3 D_1$$

- Now the effect of X_1 depends on D_1 - whether the district receives funding or not

Nonlinear regressors

Interaction terms

- **Continuous $\times$ Continuous:** Consider this regression, where both X_1 and X_2 are continuous variables.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 \times X_2) + u$$

- In this equation, the effect of X_1 on Y, and that for X_2 on Y are

$$\frac{\partial Y}{\partial X_1} = \beta_1 + \beta_3 X_2 \quad \frac{\partial Y}{\partial X_2} = \beta_2 + \beta_3 X_1$$

- Now you see that the marginal impact of X_1 on Y is dependent of X_2.

Assessing the regression

Internal and external validity

- **External validity** is concerned with the applicability of the regression model to other contexts.
  - For instance, you may wonder whether regression model from schools in California could explain what happens in the classrooms in South Korea
  - Any critiques to the model questioning the replicability of the result to other dataset is assessing the externali validity.
- **Internal validity** assesses the model from the point of view of the population being studied.
  - This approaches questions the validity of the statistical inference within the given dataset. There are five threats to internal validity.

Assessing the regression

5 Internal validity threats

- **Omitted variable bias:** Did the researcher left out factors that could affect Y and are correlated with X?
- **Wrong functional form:** Did the researcher incorporate X in a proper functional form? (quadratic, logs?)
- **Errors in variables bias:** Measurement error. If X does not have a clear cut measure, part of X correlated to Y and the observed version of X is left out, resulting in similar problems with the case of OVB.
- **Sample selection bias:** If certain groups are more likely to not respond, then the data fails to represent population and the estimates based on this data is biased.
- **Simultaneous causality bias:** There are cases where Y causes X too, also leading to the bias (consider supply and demand)