Introduction to Econometrics: Recitation 6

Seung-hun Lee

1 Nonlinear regressors

1.1 ln (natural logarithm) terms

We might be interested in how the changes in the independent variable X in terms of *percentages*, not levels, affect changes in the dependent variable Y. This is where **log regressors** can be used. Log regressors captures the idea of percentage changes. To see why, 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.

There are three different types of regression to go over. For now, I will use log to refer to the logarithm with base e, so equivalent with ln,

- **lin-**log: Consider the model $Y = \beta_0 + \beta_1 \log X + u$. I now back out β_1 to see what this coefficient implies. 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, the following relationship holds

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

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 $\beta_1 \times 0.01$. Put differently,

$$\Delta y = \beta_1 \times \frac{\Delta x}{X}$$

Also, change in X by 1% implies that $\frac{\Delta x}{X}$ changes by 0.01. This is why we multiply 0.01 to β_1.

- 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 the following result:

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

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 $100 \times \beta_1$%. To see this,

note that the above equation implies

$$\frac{\Delta y}{Y} = \beta_1 \Delta x$$

By multiplying 100, the percentage change can be obtained.

- 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

1.2 Interaction terms

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. For instance, one might guess that schools in districts where there are more funding (X_2) are more likely to enjoy the benefits of small school classroom[1]. In mathematical terms, 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 $X_1 \times X_2$

There are three types of interaction terms we can use: interaction between continuous variables, interaction between binary variables, and interaction with one binary and one continuous variable.

[1] This is inspired by the following work:
Lafortune, Julien, Jesse Rothstein, Diane W. Schanzenbach (2018) "School Finance Reform and the Distribution of Student Achievement", *American Economic Journal: Applied Economics*, 10(2): 1-26

- **Binary $\times$ Binary:** Suppose that there are two binary variables, D_1, D_2. For example, let D_1 be union membership, and D_2 be workers working under fixed contract (hereafter regular-contract worker). The dependent variable of interest is wage. 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 all workers with union membership and β_2 captures the same for all regular-contract workers. What do we want to do if we want to know the group difference in average for regular-contract workers with union membership?

  Note that $D_1 \times D_2$ becomes 1 only if the individual in question is a regular-contract worker with a union membership. This group average is captured by β_3 coefficient. To sum up:

  - Union members: $E[Y|D_1 = 1] = \beta_0 + \beta_1 + \beta_2 \times D_2 + \beta_3 \times D_2$
  - Non-members: $E[Y|D_1 = 0] = \beta_0 + \beta_2 \times D_2$
  - Effect of union membership: $E[Y|D_1 = 1] - E[Y|D_1 = 0] = \beta_1 + \beta_3 \times D_2$

  So the effect of union membership differs depending on the worker being a regular-contract worker or not. We can see that this setup allows us to analyze the difference in effect of binary variable depending on another binary factor.

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

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

  Let's go back to the example in the beginning of this subsection. In that 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. If the district receives the funding, the effect of classroom size would be $\beta_2 + \beta_3$. If the district does not get funding, the effect of classroom size would be β_2. By using this interaction term, we can analyze the effect of the continuous variable X_1 differently across different groups (who have different values of D_i).

- **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. Similar can be said of marginal impact of X_2 on Y. An example would be letting Y be wage, X_1 be age, and X_2 be years of experience.

2 Assessing the model: Internal/exteral validity

Not all regression models are perfect. The result may only apply to limited context. There may be factors within the regression model that prevents the model from delivering accurate verdict.

External validity is concerned with the applicability of the regression model to other contexts. This requires the deep understanding of the case being studied by the model and how they are similar/different from other cases. 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. Specifically, this approaches questions the validity of the statistical inference within the given dataset. There are five threats to internal validity.

- **Omitted variable bias:** Whenever the regression model contains control variables, we should be concerned about whether the researcher left out factors that could affect Y and are correlated with X. If so, the error term and X is correlated and the estimates are biased.
- **Wrong functional form:** This occurs when the researcher does not include X properly. Perhaps it is a mistake to contain X alone. One can suggest including X in a quadratic form or with some interaction term.
- **Errors in variables bias:** Also known as measurement error. Perhaps our variable X does not have a clear cut measure. If so, some part of X which is related to Y and the observed version of X is not included, resulting in similar problems with the case of OVB.
- **Sample selection bias:** Consider a survey. An ideal survey would represent various groups in the population. Now, let's say people don't respond anymore. If certain groups are more likely to not respond, then the survey fails to represent population and the estimates based on this data is biased.
- **Simultaneous causality bias:** We assumed that X causes Y. However, there are cases where Y causes X too. We assumed that class size, our X causes some result in test scores Y. However, it is possible that after observing smaller size class performs well, schools start reducing class size. In this case, our β coefficient will be biased.