Introduction to Econometrics: Recitation 5

Seung-hun Lee*

October 13th, 2022

1 Nonlinear regressors

Not everything in real life is correlated in a linear fashion. For instance, production function are not usually linear with respect to its inputs (Cobb-douglas production function), Wage rises, but the rate at which it rises falls over time (Mincer equation, from Mincer (1974)[1]), and the effect of classroom size can differ depending on how many students are in the class to begin with(Lazear(2001)[2]).

For correlations like this, resorting to linear regressors would not allow the regression model to have the best fit. This is where the nonlinear regressors come in. Regressions with nonlinear regressors allow us to graph relationship between our Y variable and our X variable that is not necessarily linear. When incorporating such regressors, the interpretation of each coefficient becomes trickier. In the next few subsections, I will walk through implementing and interpreting coefficients for regressors that are not linear.

1.1 Quadratic terms

Recall from the regression with single independent variable $Y = \beta_0 + \beta_1 X_1 + u$, β_1 coefficient means the marginal effect of X_1 on Y. Mathematically, there are two ways to show this. (For now, all usual assumptions hold)

*Contact me at sl4436@columbia.edu if you spot any errors or have suggestions on improving this note.

[1]Mincer, Jacob (1974) "Schooling, Experience, and Earnings", New York, National Bureau of Economic Research

[2]Lazear, Edward (2001) "Educational Production", *Quarterly Journal of Economics*, 116(3): 777-803

- **Derivatives**: Take derivatives on Y with respect to X_1, this gets us

$$\frac{\partial Y}{\partial X_1} = \frac{\partial}{\partial X_1}[\beta_0 + \beta_1 X_1 + u] = \beta_1$$

  Given that the idea of derivatives captures how much our Y variable changes with unit change in X_1, β_1 represents how much Y responds to a one unit change in X_1

- **Taking Differences**: Suppose that you raise the amount of X_1 by Δx. Now the change in the amount of Y, denoted as Δy, as a result of this change is

$$\beta_0 + \beta_1(X_1 + \Delta x) + u = Y + \Delta y$$

  By subtracting $Y = \beta_0 + \beta_1 X_1 + u$, I get

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

In this example, the marginal effect of X on Y is constant - it does not depend on X! However, there are many relationships that cannot be explained this way. For instance, it is most likely the case that the wage rises with age. However, they are not likely to be in a linear relationship. As one gets older, the rate at which wage increases with age decreases. This is where **polynomial regressors** can improve the fitting of the regression model. Let wage be W and age be X. I now regress the following model.

$$W = \beta_0 + \beta_1 X + \beta_2 X^2 + u$$

Then, the marginal impact of age on wage can be captured by

$$\frac{\partial W}{\partial X} = \beta_1 + 2\beta_2 X$$

Note now that unlike before, there is a term that depends on X. This implies that the marginal effect of age on wage now is different depending on what X variables we use (or age). When β_2 is positive (negative), then marginal impact of age on wage is higher (lower) for older people. Simply put, age rises at faster (slower) rate for older people.

1.2 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

1.3 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[3]. 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.

- **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: $E[Y|D_1 = 1] - E[Y|D_1 = 0] = \beta_1 + \beta_3 \times D_2$

[3] 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

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 external 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 an attenuation bias.
- **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.

3 Midterm checklist

Yes, it's that time of the year! Applaud yourself since you are halfway through. To make your life easier, I will organize the list of things that you need to know for the upcoming midterm exam. Ideally, you would have looked back at the key proofs and derived them yourself, run loads of regressions to be familiar with OLS formula and a lot more. If you are time-constrained, you are still okay. You can look at the items below and remind yourself of what you have learned. Moreover, the list here is intended to give you some guidance of what you should review in the last minute (hopefullly).

The most important things, i.e. things you should know by heart by now, are listed with $\star\star\star$. The more stars there are, the more crucial they are to understand what we have learned so far. However, it does not mean that the items with fewer stars can easily be ignored. I am giving some priority weights

3.1 Concepts you should know

$\star\star\star$ **Randomized control test**: This is the core design that any econometric regressions should strive for. The key idea to remember is that experimental designs should have a clear control vs treatment group structure and that the groups are otherwise identical except for treatment status. If you have this setup, an OLS regression is going to give you an accurate estimate of the (average) treatment effect. This setup is not always feasible, for which we have concepts that will be covered after the midterm exam.

$\star\star\star$ **Hypothesis test**: Some people say that this is in the core of econometrics. By now you should know how to set up a null/alternative hypothesis, what type of test distribution to use (F, t, or normal), know your test statistic, and come to the conclusion using either the test statistics, p-value, or confidence interval. Also be familiar with R^2 and $\bar{R}^2$

$\star\star\star$ **Interpreting coefficients**: The most important thing when you get your regression output is understanding what it means. Especially in multivariate case, you should know that the $\hat{\beta}$ coefficient in front of your independent variable is the marginal effect of changing X by 1 unit on Y, *holding other variables constant*. Interpretation is very crucial for nonlinear regressors. Be familiar with the regressions that involve logs and interaction terms.

$\star\star\star$ **Omitted variable bias**: This is what motivates the use of multivariate regressions. Do remember what the two conditions are and how you can mathematically show them.

⋆⋆⋆ **OLS assumptions**: Not that you need to memorize them. You do need to remind yourself how you proved unbiasedness, how you showed OVB, etc. with which assumptions.

⋆⋆ **Estimating the OLS**: Star docked off since you may be given the access to the formula during the exam. However, it is essential to know where the OLS estimators originated from.

⋆⋆ **Statistical properties of OLS**: Specifically, you need to be familiar with how to prove OLS estimator is unbiased. It does not mean you should memorize every line of the proof. It helps, however, to remind yourself which assumptions were used in the proof.

⋆⋆ **Heteroskedasticity**: Remember how this affects t-statistics. Specifically, that it alters the standard errors but not the coefficients.

⋆⋆ **Multicollinearity**: Know two types of multicollinearity and what causes them. In particular, if you are given dummy variables, understand what to do with it. Specifically, *If you have K dummy variables for K categories, include* $K - 1$ *variables.*

⋆⋆ **Binary (independent) variables**: Know how to set them up. Also, know that the coefficient in front of it implies the mean difference with the "benchmark" group.

⋆⋆ **Internal/External Validity**: When the question asks you to assess the regression model and results, you can use these ideas to critique the model.

⋆ **Reviewing Statistics**: This should have been taken care of as you moved onto the later chapters. If you are still not comfortable, look at this again as soon as possible.

3.2 Things to do

⋆⋆⋆ **Do decent amount of studying, but definitely not too much**: Do get some enough sleep. It is never a good idea to cramp your econometrics knowledge overnight before the day of the exam. Always remember that slow and steady wins the race (especially for econometrics).

⋆⋆⋆ **Use office hours**: I am talking about OH for both TA's and the professors. Our job is to help you get the most out of this course. We are ready to help you in any way we can. Just visit remaining OH as often as possible and ask questions. It may just help to see what others are asking too.

⋆⋆⋆ **Review problem sets**: Not all of them. However, the "pen(cil) and paper" questions are a must. Since we assume that you are familiar with the problem set, it is important to walk back through the questions you solved before.

⋆⋆⋆ **Review extra questions the professor revealed**: The questions there reflect the ideal questions that the professor wants you to know before the exam. This will also be covered on Tuesday.

⋆⋆ **Ungraded Problems in the problem set**: Some of them are empirical exercises. Nevertheless, they can also be a valuable source of new questions for those who need them.

⋆ **End of the Chapter questions**: Since we did not cover some of the topics in the book, not all of the questions are relevant. Also note that the book approaches some matter differently from the class. Some of the questions that match what we learned, however, can be valuable for you. This is true if you need new questions to work on.