Recitation 5: Extended specifications and assessing the models

Seung-hun Lee

Columbia University
Undergraduate Introduction to Econometrics Recitation

October 13th, 2022

Extended models: Quadratic, log, and interacted regressors

Not everything in world is linearly related

- The effect of X may grow larger/smaller as X increases (think of wage and age)
- For correlations like this, nonlinear regressors are necessary
- When incorporating such regressors, the interpretation of each coefficient becomes trickier.
- Quadratic relations: Think about wage and age - wages increase with age, but (usually) at a decreasing pace

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

- The marginal effect of X on W can be written as

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

Back to linear models: So how is the marginal effect different?

- In a linear regressor only format, the marginal effect is

$$W = \beta_0 + \beta_1 X + u$$

$$\frac{\partial W}{\partial X} = \beta_1$$

- The difference is that with quadratic terms, we can express cases where *marginal* changes to W with respect to X is not a constant, but depends on some value of X
  - In the above case, if $\beta_2 > 0$, marginal increase in W increases with X (and vice versa)

We are interested in percent changes (because measurement units)...

- 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 function 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.

How much *level* changes in Y due to *percentage* changes in X?

- **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 $\beta_1 \times 0.01$.

How much *percentage* changes in Y due to *level* changes in X?

- 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 $(100 \times \beta_1)\%$.

How much *percentage* changes in Y due to *percentage* changes in X?

- 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

IRL: Marginal effect of X_1 on Y maybe a function of some other variables!

- 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 $X_1 \times X_2$

Binary × 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.

Binary × 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

Continuous × Continuous

Interaction terms

- 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 model

Critiquing the regressions

- **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 external 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.

Is the model itself set properly?

- **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 attenuation bias
- **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)

Midterm checklistl

Midterm Checklist

- 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 (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

Concepts you should know

⋆⋆⋆ **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.

⋆⋆⋆ **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 $\bar{R}^2$

Concepts you should know

- ⋆ ⋆ ⋆ **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.
- ⋆ ⋆ ⋆ **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.

Concepts you should know

- ⋆⋆ **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.

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.

Things to do

- ⋆⋆ **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.