

# Introduction to Econometrics: Recitation 5

Seung-hun Lee

October 25th, 2021

## 1 Assessing the model: Internal/external 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  $\beta$  coefficient will be biased.

## 2 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 (hopefully).

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

### 2.1 Concepts you should know

- $\star\star\star$  **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.
- $\star\star\star$  **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.
- $\star\star$  **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.

## 2.2 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 cram 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.
- ★★ **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.