Internal and External Validity [1]

Jasmine(Yu) Hao

Faculty of Business and Economics
Hong Kong University

September 9, 2021

[1] This section is based on Stock and Watson (2020), Chapter 9.

Internal and External Validity I

- A framework for evaluating whether a statistical or econometric study is useful for answering a specific question of interest.
- Internal and external validity distinguish between the population and setting studied and the population and setting to which the results are generalized.
- The population studied is the population of entities. The population to which the results are generalized, or the population of interest, is the population of entities to which the causal inferences from the study are to be applied.
- By setting, we mean the institutional, legal, social, physical, and economic environment.

Threats to Internal Validity I

Internal validity has two components.

- First, the estimator of the causal effect should be unbiased and consistent. For example,
- Second, hypothesis tests should have the desired significance level and confidence intervals should have the desired confidence level.

For example, if a confidence interval is constructed as bn $STR \pm 1.96SE(\hat{\beta}STR)$, this confidence interval should contain the true population causal effect, βSTR, with 95% probability over repeated samples drawn from the population being studied.

Threats to Internal Validity II

- In regression analysis, causal effects are estimated using the estimated regression function, and hypothesis tests are performed using the estimated regression coefficients and their standard errors.
- Accordingly, in a study based on OLS regression, the requirements for internal validity are that the OLS estimator is **unbiased and consistent** and that standard errors to achieve desired confidence level.
- For various reasons, these requirements might not be met, and these reasons constitute **threats to internal validity**.
- e.g. OVB;

Threats to External Validity I

Potential threats to external validity arise from **differences between the population and setting studied and those of interest**.

1. Differences in populations.
   - ◇ e.g., laboratory studies of the toxic effects of chemicals typically use animal populations like mice (the population studied), but the results are used to write health and safety regulations for human populations (the population of interest).
   - ◇ More generally, the true causal effect might not be the same, due to different characteristics of the populations, geographical differences, or because the study is out of date.

Threats to External Validity II

2. Differences in settings.
   - ◇ e.g., on college binge drinking of an antidrinking advertising campaign might not generalize to another, identical group of college students if the legal penalties for drinking at the two colleges differ. In this case, the legal setting in which the study was conducted differs from the legal setting to which its results are applied.
   - ◇ More generally, differences in the institutional environment (public universities versus religious universities), differences in laws (differences in legal penalties), and differences in the physical environment (tailgate-party binge drinking in southern California versus Fairbanks, Alaska).

Application to test scores and the student– teacher ratio I

- The analysis on test score is based on California school districts. Suppose for the moment that these results are **internally valid**. To what other populations and settings of interest could this finding be generalized? **The closer the population and setting of the study are to those of interest, the stronger the case is for external validity**.
- e.g. College student ?
- other U.S. elementary school districts?

How to assess the external validity of a study

- External validity must be judged using specific knowledge of the populations and settings studied and those of interest.
- Important differences cast doubt on the external validity of the study.
- Two or more studies on similar populations, check the validity by comparing their results.
- For example, in Section 9.4, we analyze test score and class size data for elementary school districts in Massachusetts and compare the Massachusetts and California results.

How to design an externally valid study I

- Threats to external validity stem from a lack of comparability of populations and settings, these threats are best minimized at the early stages of a study, before the data are collected. [2]

[2] Study design is beyond the scope of this textbook, and the interested reader is referred to Shadish, Cook, and Campbell (2002).

Threats to Internal Validity of Multiple Regression Analysis

- regression analysis are internally valid if theestimated regression coefficients are **unbiased and consistent**, and standard errors yield **desired confidence level**.
- Threats: omitted variables, misspecification of the functional form of the regression function, imprecise measurement of the independent variables ("errors in variables"), sample selection, and simultaneous causality.
- All five sources of bias arise because the regressor is correlated with the error term in the population regression, violating the first least squares assumption.

Omitted Variable Bias I

- OVB arises when a variable that both determines Y and is correlated with one or more of the included regressors is omitted from the regression.
- This bias persists even in large samples.
- Solutions to omitted variable bias when the variable is observed or there are adequate control variables.
- Alternatively, if you have data on one or more control variables and if these control variables are adequate in the sense that they lead to **conditional mean independence**, then including those control variables eliminates the potential bias in the coefficient on the variable of interest.

Omitted Variable Bias II

- Balance the cost and benefit of the omitted variable bias v.s. reduces the precision of the estimators(trade-off between bias and variance).
  1. identify the key coefficient or coefficients of interest in your regression.
  2. What are the most likely sources of important omitted variable bias in this regression?
  3. Augment your base specification with the additional, questionable control variables identified in the second step.
  4. The fourth step is to present an accurate summary of your results in tabular form.

This provides "full disclosure" to a potential skeptic, who can then draw his or her own conclusions.

Solutions to OVB without control variables I

1. **Panel regression**: use data in which the same observational unit is observed at different points in time.
2. **IV regression**: use instrumental variables regression.
3. **RCT**: use a study design in which the effect of interest (for example, the effect of reducing class size on student achievement) is studied using a randomized controlled experiment. (Sometimes in tech industry, AB test.)

Misspecification of the Functional Form I

- This bias is a type of omitted variable bias.
  For example, if the population regression function is a quadratic polynomial, then a regression that omits the square of the independent variable would suffer from omitted variable bias.
- Solutions to functional form misspecification.
  - nonlinear regression.
  - probit/logit regression.

Measurement Error and Errors-in-Variables Bias I

Sources of measurement error. wrong answer, wrong information. Written in terms of the imprecisely measured variable Xi, the population regression equation $Y_i = \beta_0 + \beta_1 X_i + u_i$

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 \tilde{X}_i + [\beta_1 (X_i - \tilde{X}_i) + u_i] \\ &= \beta_0 + \beta_1 \tilde{X}_i + v_i, \end{aligned}$$

where $v_i = \beta_1 (X_i - \tilde{X}_i) + u_i$.

- DGP: $Y_i = \beta_0 + \beta_1 X_i + u_i$.
- only obserbe $\tilde{X}_i$,

$$\hat{\beta}_1 \to_p \frac{\sigma_X^2}{\sigma_X^2 + \sigma_w^2} \beta_1.$$

Solutions to errors-in-variables bias

1. One such method is instrumental variables regression. It relies on having another variable (the instrumental variable) that is correlated with the actual value X_i but is uncorrelated with the measurement error
2. A second method is to develop a mathematical model of the measurement error use the resulting formulas to adjust the estimates, requires specialized knowledge about the nature of the measurement error.

Missing Data and Sample Selection

- When the data are missing completely at random, effect is to reduce the sample size but not introduce bias.
- When the data are missing based on the value of a regressor, the effect also is to reduce the sample size but not to introduce bias.
- If the data are missing because of a selection process that is related to the value of the dependent variable (Y) beyond depending on the regressors (X), then this selection process can introduce correlation between the error term and the regressors.
- The sample selection problem can be cast either as a consequence of nonrandom sampling or as a missing data problem.

Simultaneous Causality I

- So far, we have assumed that causality runs from the regressors to the dependent variable (X causes Y). But what if causality also runs from the dependent variable to one or more regressors (Y causes X)?
- Suppose a government initiative subsidized hiring teachers in school districts with poor test scores.
- Simultaneous causality leads to correlation between the regressor and the error term.
- This leads to **simultaneous causality** bias and inconsistency of the OLS estimator. This correlation between the error term and the regressor can be made mathematically precise by introducing an additional equation that describes the reverse causal link.

Simultaneous Causality II

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$
$$X_i = \gamma_0 + \gamma_1 Y_i + v_i.$$

- Simultaneous causality leads to correlation between Xi and the error term ui.
- Because it can be expressed mathematically using two simultaneous equations, simultaneous causality bias is sometimes called simultaneous equations bias.

Correlation of the error term across observations I

- Population regression error can be correlated across observations, partially random sampling
  - repeated observations on the same entity over time,
  - "serial" correlation is induced in the regression error over time.
  - the error term can be correlated across observations is when sampling is based on a geographical unit.
- Does not introduce bias, but violate assumptions on the error distribution. The consequence is that the OLS standard errors are incorrect.
- Using an alternative formula for standard errors.

Internal and External Validity When the Regression Is Used for Prediction I

- When regression models are used for prediction, concerns about external validity are very important

Chapter 4 began by considering two problems.

- How much test scores will increase if reduces **class sizes**?
- Reliable prediction about test scores in a district based on data with access?

Reliable prediction using multiple regression has three requirements.

1. The data used to estimate the prediction model and the observation for which the prediction is to be made are drawn from the same distribution.
2. The list of predictors to reduce the threat of omitted variable bias.

Internal and External Validity When the Regression Is Used for Prediction II

3. The third requirement concerns the estimator itself. So far, we have focused on OLS for estimating multiple regression. In some prediction applications, however, there are very many predictors; indeed, in some applications the number of predictors can exceed the sample size. If there are very many predictors, then there are—surprisingly—some estimators that can provide more accurate out-of-sample predictions than OLS. .

References I

Stock, J. H. and Watson, M. W. (2020). *Introduction to econometrics*, volume 4. Pearson New York.