Applied Statistics and Econometrics

Giuseppe Ragusa
Luiss University

Lecture 11: Internal and External Validity of Regression Studies

Outline

- Internal and External Validity
- Threats to Internal Validity
  - Omitted variable bias
  - Functional form misspecification
  - Errors-in-variables bias
  - Missing data and sample selection bias
  - Simultaneous causality bias
- Application to Test Scores

Internal and External Validity

- Let's step back and take a broader look at regression. Is there a systematic way to assess (critique) regression studies? We know the strengths of multiple regression – but what are the pitfalls?
  - We will list the most common reasons that multiple regression estimates, based on observational data, can result in biased estimates of the causal effect of interest.
  - In the test score application, let's try to address these threats as best we can – and assess what threats remain. After all this work, what have we learned about the effect on test scores of class size reduction?

A Framework for Assessing Statistical Studies: Internal and External Validity

- **Internal validity**: the statistical inferences about causal effects are valid for the population being studied.
- **External validity**: the statistical inferences can be generalized from the population and setting studied to other populations and settings, where the "setting" refers to the legal, policy, and physical environment and related salient features.

Threats to External Validity of Multiple Regression Studies

Assessing threats to external validity requires detailed substantive knowledge and judgment on a case-by-case basis.

- How far can we generalize class size results from California?
  - Differences in populations
    - California in 2011?
    - Massachusetts in 2011?
    - Mexico in 2011?
  - Differences in settings
    - different legal requirements (e.g. special education)
    - different treatment of bilingual education
  - differences in teacher characteristics

Threats to Internal Validity of Multiple Regression Analysis

Internal validity: the statistical inferences about causal effects are valid for the population being studied.

Five threats to the internal validity of regression studies:

1. Omitted variable bias
2. Wrong functional form
3. Errors-in-variables bias
4. Sample selection bias
5. Simultaneous causality bias

All of these imply that

$$E(u_i|X_{1i},\ldots,X_{ki}) \neq 0$$

(or that conditional mean independence fails) – in which case OLS is biased and inconsistent.

Omitted variable bias

- Omitted variable bias arises if an omitted variable is both:
  1. a determinant of Y **and**
  2. correlated with at least one included regressor.
- We first discussed omitted variable bias in regression with a single X. OV bias arises in multiple regression if the omitted variable satisfies conditions 1. and 2. above.
- If the multiple regression includes control variables, then we need to ask whether there are omitted factors that are not adequately controlled for, that is, whether the error term is correlated with the variable of interest even after we have included the control variables.

Solutions to omitted variable bias

1. If the omitted causal variable can be measured, include it as an additional regressor in multiple regression;
2. If you have data on one or more controls and they are adequate (in the sense of conditional mean independence plausibly holding) then include the control variables;
3. Possibly, use panel data in which each entity (individual) is observed more than once;
4. If the omitted variable(s) cannot be measured, use instrumental variables regression;
5. Run a randomized controlled experiment. Why does this work? Remember – if X is randomly assigned, then X necessarily will be distributed independently of u; thus $E(u|X = x) = 0$.

Wrong functional form (functional form misspecification)

Arises if the functional form is incorrect – for example, an interaction term is incorrectly omitted; then inferences on causal effects will be biased.

- Solutions to functional form misspecification
  - Continuous dependent variable: use the "appropriate" nonlinear specifications in X (logarithms, interactions, etc.)
  - Discrete (example: binary) dependent variable: need an extension of multiple regression methods ("probit" or "logit" analysis for binary dependent variables).

Errors-in-variables bias

So far we have assumed that X is measured without error. In reality, economic data often have measurement error

- Data entry errors in administrative data
- Recollection errors in surveys (when did you start your current job?)
- Ambiguous questions (what was your income last year?)
- Intentionally false response problems with surveys (What is the current value of your financial assets? How often do you drink and drive?)

Errors-in-variables bias, ctd.

In general, measurement error in a regressor results in "errors-in-variables" bias.

Illustration: suppose

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

is "correct" in the sense that the three least squares assumptions hold (in particular $E(u_i|X_i) = 0$).

Let

$$X_i = \text{unmeasured true value of } X$$

$$\tilde{X}_i = \text{imprecisely measured version of } X$$

Errors-in-variables bias, ctd.

Then

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

or

$$Y_i = \beta_0 + \beta_1 \tilde{X}_i + \tilde{u}_i, \quad \tilde{u}_i = \beta_1 (X_i - \tilde{X}_i) + u_i$$

If $\tilde{X}_i$ is correlated with $\tilde{u}_i$ then $\hat{\beta}_1$ will be biased

$$\begin{aligned} cov(\tilde{X}_i, \tilde{u}_i) &= cov(\tilde{X}_i, \beta_1 (X_i - \tilde{X}_i) + u_i) \\ &= \beta_1 cov(\tilde{X}_i, (X_i - \tilde{X}_i)) + cov(\tilde{X}_i, u_i) \\ &= \beta_1 [cov(\tilde{X}_i, X_i) - var(\tilde{X}_i)] \end{aligned}$$

Errors-in-variables bias, ctd.

$$Y_i = \beta_0 + \beta_1 \tilde{X}_i + \tilde{u}_i, \quad \tilde{u}_i = \beta_1 (X_i - \tilde{X}_i) + u_i$$

- If X_i is measured with error, $\tilde{X}_i$ is in general correlated with X_i, so $\hat{\beta}_1$ is biased and inconsistent.
- It is possible to derive formulas for this bias, but they require making specific mathematical assumptions about the measurement error process (for example, that $\tilde{u}_i$ and X_i are uncorrelated). Those formulas are special and particular, but the observation that measurement error in X results in bias is general.

Potential solutions to errors-in-variables bias

- Obtain better data.
- Develop a specific model of the measurement error process.
  - This is only possible if a lot is known about the nature of the measurement error – for example a subsample of the data are cross-checked using administrative records and the discrepancies are analyzed and modeled. (Very specialized; we won't pursue this here.)
- Instrumental variables regression.

Sample selection bias

So far we have assumed simple random sampling of the population. In some cases, simple random sampling is thwarted because the sample, in effect, "**selects itself**."

- Sample selection bias arises when a selection process
  1. influences the availability of data and
  2. that process is related to the dependent variable.

Example #1: Mutual funds

- Do actively managed mutual funds outperform "hold-the-market" funds?
- **Empirical strategy:**
  - Sampling scheme: simple random sampling of mutual funds available to the public on a given date.
  - Data: returns for the preceding 10 years.
  - Estimator: average ten-year return of the sample mutual funds, minus ten-year return on S&P500
  - **Is there sample selection bias?**

Example #1: Mutual funds

Sample selection bias induces correlation between a regressor and the error term.

Mutual fund example:

$$return = \beta_0 + \beta_1 managed\_fund_i + u_i$$

Being a managed fund in the sample ($managed\_fund_i = 1$) means that your return was better than failed managed funds, which are not in the sample – so $corr(manage\_fund_i, u_i) \neq 0$

Example #2: returns to education

- What is the return to an additional year of education?
- Empirical strategy:
  - Sampling scheme: simple random sampling of workers
  - Data: earnings and years of education
  - Estimator: regress ln(*earnings*) on *years_education*
  - Ignore issues of omitted variable bias and measurement error – is there sample selection bias?

Potential solutions to sample selection bias

- Collect the sample in a way that avoids sample selection.
  - Mutual funds example: change the sample population from those available at the end of the ten-year period, to those available at the beginning of the period (include failed funds)
  - Returns to education example: sample college graduates, not workers (include the unemployed)
  - Randomized controlled experiment. Construct a model of the sample selection problem and estimate that model.

Simultaneous causality bias

So far we have assumed that X causes Y. What if Y causes X, too?

- Example: Class size effect
  - Low *str* results in better test scores
  - But suppose districts with low test scores are given extra resources: as a result of a political process they also have low *str*
  - What does this mean for a regression of *testscore* on *str*?

Simultaneous causality bias in equations

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$
$$X_i = \gamma_0 + \gamma_1 Y_i + \nu_i$$

- Large u_i means large Y_i, which implies large X_i (if $\beta_1 > 0$)
- Thus $corr(X_i, u_i) \neq 0$
- Thus is $\hat{\beta}_1$ is biased and inconsistent.
- e.g.: A district with particularly bad test scores given the *str* (negative u_i) receives extra resources, thereby lowering its *str*; so *str*i and u_i are correlated

Simultaneous causality, ctd.

- Randomized controlled experiment. Because X_i is chosen at random by the experimenter, there is no feedback from the outcome variable to Y_i (assuming perfect compliance).
- Develop and estimate a complete model of both directions of causality. This is the idea behind many large macro models (e.g. Federal Reserve Bank-US). This is extremely difficult in practice.
- Use instrumental variables regression to estimate the causal effect of interest (effect of X on Y, ignoring effect of Y on X).

Applying this Framework: Test Scores and Class Size

Objective: Assess the threats to the internal and external validity of the empirical analysis of the California test score data.

- External validitiy:
  - Compare results for California and Massachusetts
  - Think hard...
- Internal validity
  - Go through the list of five potential threats and internal validity think hard...

Check of external validity

Compare the California study to one using Massachusetts data

- The Massachusetts data set
  - 220 elementary school districts
  - Test: 1998 MCAS test – fourth grade total (Math + English + Science)
  - **Variables:** *str*, *testscore*,*english*, *lunch*, Income