Econometrics II

José Fajardo

Universitat Pompeu Fabra

October 8, 2018

Outline of Todays' Lecture

- Internal and External Validity
- Instrumental Variables (IV Regresion)

Internal and External Validity

- Is there a systematic way to assess (critique) regression studies?
- We know the strengths of multiple regression
- But what are the pitfalls?

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

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.

- Omitted variable bias
- Wrong functional form
- Errors-in-variables bias
- Sample selection bias
- Simultaneous causality bias

All these imply, $E(u_i|(X_{1i}, X_{2i}, \ldots, X_{ki}) \neq 0$, that is, LSA 1 does not hold, then OLS estimators are biased.

Omitted Variables and Solutions

- If the omitted causal variable can be measured, include it as an additional regressor in multiple regression
- 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
- If the omitted variable(s) cannot be measured, use instrumental variables regression
- Run a randomized controlled experiment.

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-Variable Bias

Consider the case of two regressors: In general, measurement error in a regressor results in "errors-in-variables" bias.

A bit of math shows that errors-in-variables typically leads to correlation between the measured variable and the regression error. Consider the single-regressor model:

$$Y_i = \beta_0 + \beta_1 X_i + u_i, \ i = 1, \ldots, n$$

and suppose $E(u_i|(X_i) = 0$, and let:

X_i= unmeasured true value of X

$\tilde{X}_i$= mis-measured version of X (the observed data)

Cont'd

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

so the regression, with measurement error, you run is:

$$Y_i = \beta_0 + \beta_1 \tilde{X}_i + \tilde{u}_i, \tag{1}$$

where $\tilde{u}_i = \beta_1 (X_i - \tilde{X}_i) + u_i$, typically $\tilde{X}_i$ is correlated with $\tilde{u}_i$, so β_1 is biased:

$$Cov(\tilde{X}_i, \tilde{u}_i) = \beta_1 Cov(\tilde{X}_i, X_i - \tilde{X}_i) + Cov(\tilde{X}_i, u_i) \neq 0, \tag{2}$$

To get some intuition for the problem, consider two special cases

A) Classical measurement error

The classical measurement error model assumes that $\tilde{X}_i = X_i + v_i$, where v_i is a zero mean random noise with $corr(X_i, v_i) = 0$ and $corr(u_i, v_i) = 0$. Then:

$$Var(\tilde{X}_i) = \sigma_X^2 + \sigma_v^2$$

$$Cov(\tilde{X}_i, X_i - \tilde{X}_i) = -\sigma_v^2$$

Then, in eq. (2):

$$Cov(\tilde{X}_i, \tilde{u}_i) = \beta_1 Cov(\tilde{X}_i, X_i - \tilde{X}_i) = -\beta_1 \sigma_v^2$$

By eq. (1), we have:

$$\widehat{\beta}_1 = \frac{\widehat{Cov}(Y_i, \tilde{X}_i)}{\sigma_{\tilde{X}}} = \beta_1 + \frac{\widehat{Cov}(\tilde{u}_i, \tilde{X}_i)}{\sigma_{\tilde{X}}^2} \xrightarrow{P} \left(\frac{\sigma_X^2}{\sigma_X^2 + \sigma_v^2} \right) \beta_1$$

The classical measurement error model is special because it assumes $Corr(Xi, vi) = 0$.

B) "Best Guess" measurement error

Suppose the respondent doesn't remember X_i, but makes a best guess of the form $\tilde{X}_i = E(X_i|W_i)$, given the available inf. W, where $E(u_i|W_i) = 0$. Then,

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

- As $\tilde{X}_i$ is the Best guest it can not be correlated with the error measurement $X_i - \tilde{X}_i$, otherwise it would be useful information for predicting X_i, then $\tilde{X}_i$ would not be the Best guess for X_i.
- As $\tilde{X}_i$ is a function of the available information W, and u_i is independent of W, then $Cov(\tilde{X}_i, u_i) = 0$.
- Thus, $Cov(\tilde{X}_i, \tilde{u}_i) = 0$. So $\widehat{\beta_1}$ is unbiased.

“Best Guess” measurement error

- Under the “Best Guess” model, you still have measurement error, you don’t observe the true value of X_i , but there this measurement error doesn’t introduce bias into!
- The “best guess” model is extreme, it isn’t enough to make a good guess, you need the “best” guess $\tilde{X}_i = E(X_i|W_i)$, that is, the conditional expectation of X given W, where $E(u_i|W_i) = 0$.

Solution to Errors-in-variables bias

- Obtain better data (often easier said than done).
- 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.
- Instrumental variables regression.

Missing data and sample selection bias

Data are often missing. Sometimes missing data introduces bias, sometimes it doesn't. It is useful to consider three cases:

(1) Data are missing at random.

(2) Data are missing based on the value of one or more X's

(3) Data are missing based in part on the value of Y or u

Cases 1 and 2 don't introduce bias: the standard errors are larger than they would be if the data weren't missing, β_1 is unbiased.

Case 3 introduces "sample selection" bias.

Ex. 1: Data are Missed at Random

Suppose you took a simple random sample of 100 workers and recorded the answers on paper, but your dog ate 20 of the response sheets (selected at random) before you could enter them into the computer. This is equivalent to your having taken a simple random sample of 80 workers (think about it), so your dog didn't introduce any bias.

Ex. 2: Data are missing based on the value of one of the X's

In the test score/class size application, suppose you restrict your analysis to the subset of school districts with $STR < 20$. By only considering districts with small class sizes you won't be able to say anything about districts with large class sizes, but focusing on just the small-class districts doesn't introduce bias. This is equivalent to having missing data, where the data are missing if $STR > 20$. More generally, if data are missing based only on values of $X's$, the fact that data are missing doesn't bias the OLS estimator.

Ex. 3: Data are missing based in part on the value of Y or u

Sample selection bias arises when a selection process:

(i) influences the availability of data and

(ii) is related to the dependent variable.

Research Question: Do actively managed mutual funds outperform “hold-the-market” funds?

Ex. 3: 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 SP500 ($\widehat{excreturn}_i$)
- Is there sample selection bias? (Equivalently, are data missing based in part on the value of Y or u?)

Ex. 3: Sample Selection Bias

- Regression Model:

$$excreturn_i = \beta_0 + \beta_1 managedfund_i + u_i$$

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

Solutions to Sample Bias Selection

- 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)
- Randomized controlled experiment.
- Construct a model of the sample selection problem and estimate that model (we won't do this).

Simultaneous Causality Bias

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

- Causal effect on Y of X:

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

- Causal effect on X of Y:

$$X_i = \gamma_0 + \gamma_1 Y_i + v_i$$

Large u_i, means large Y_i, which affects X_i.Thus, $corr(X_i, u_i) \neq 0$.Then β_1 is unbiased and inconsistent.

Solutions to Simultaneous Causality

- Run a 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).

Internal and External Validity When the Regression is Used for Forecasting

Forecasting and estimation of causal effects are quite different objectives. For forecasting,

- $\overline{R}^2$ matters (a lot!)
- Omitted variable bias isn't a problem!
- Interpreting coefficients in forecasting models is not important, the important thing is a good fit and a model you can "trust" to work in your application
- External validity is paramount: the model estimated using historical data must hold into the (near) future
- More on forecasting when we take up time series data

Instrumental Variables

Outline

- IV Regression: Why and What; Two Stage Least Squares
- The General IV Regression Model
- Checking Instrument Validity
  - Weak and strong instruments
  - Instrument exogeneity
- Examples: Where Do Instruments Come From?

The IV Estimator with a Single Regressor and a Single Instrument

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

- IV regression breaks X into two parts: a part that might be correlated with u, and a part that is not. By isolating the part that is not correlated with u, it is possible to estimate β_1.
- This is done using an instrumental variable, Z_i, which is correlated with X_i but uncorrelated with u_i.

Exogeneity and Endogeneity

An endogenous variable is one that is correlated with u.

An exogenous variable is one that is uncorrelated with u.

In IV regression, we focus on the case that X is endogenous and there is an instrument, Z, which is exogenous.

Digression on terminology: "Endogenous" literally means "determined within the system." If X is jointly determined with Y, then a regression of Y on X is subject to simultaneous causality bias. But this definition of endogeneity is too narrow because IV regression can be used to address OV bias and errors-in-variable bias. Thus we use the broader definition of endogeneity above.

Two Conditions fo a Valid Instrument

For an instrumental variable (an "instrument") Z to be valid, it must satisfy two conditions:

1. Instrument relevance: $corr(Z_i, X_i) \neq 0$
2. Instrument exogeneity: $corr(Z_i, u_i) = 0$

Suppose for now that you have such a Z_i (we'll discuss how to find instrumental variables later). How can you use Z_i to estimate β_1?

The IV estimator with one X and one Z

Explanation 1: Two Stage Least Squares (TSLS)

As it sounds, TSLS has two stages, two regressions:

(1) Isolate the part of X that is uncorrelated with u by regressing X on Z using OLS:

$$X_i = \pi_0 + \pi_1 Z_i + v_i .....(1)$$

- Because Z_i is uncorrelated with u_i, $\pi_0 + \pi_1 Z_i$ is uncorrelated with u_i. We don't know π_0 or π_1 but we have estimated them, so
- Compute the predicted values of X_i, given by

$$\hat{X}_i = \hat{\pi}_0 + \hat{\pi}_1 Z_i, \ \ i = 1, \ldots, n.$$

The IV estimator with one X and one Z

(2) Replace X_i by $\hat{X}_i$ in the regression of interest, regress Y on $\hat{X}$ using OLS:

$$Y_i = \beta_0 + \beta_1 \hat{X}_i + u_i .....(2)$$

- Because $\hat{X}_i$ is uncorrelated with u_i, the first least squares assumption holds for regression (2). This requires n to be large so that π_0 and π_1 are precisely estimated.
- Thus, in large samples β_1 can be estimated by OLS using regression (2)
- The resulting estimator is called the Two Stage Least Squares (TSLS) estimator, β_1^{TSLS}.