Lecture 11: Assessing Studies Based on Multiple Regression

Zheng Tian

Outline

An over view of internal and external validity

- The concepts of internal and external validity provide a general framework for assessing whether a statistical or econometric study is useful for answering a specific question of interest.
- We focus on regression analysis that have the objective of estimating the causal effect of a change in some independent variable on a dependent variable.

The population and setting studied versus the population and setting of interest

The population and setting studied

- The population studied is the population of entities-people, companies, school districts, and so forth-from which the sample is drawn.
- The setting studied refers to as the institutional, legal, social, and economic environment in which the population studied fits in and the sample is drawn.

The population and setting of interest

- The population and setting of interest is the population and setting of entities to which the causal inferences from the study are to be applied.

Definition of 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 of interest.

Threats to internal validity

Internal validity consists of two components

- The estimator of the causal effect should be unbiased and consistent.
- Hypothesis tests should have the desired significance level, and the confidence intervals should have the desired confidence level.

Internal validity in regression analysis

1. the OLS estimator is unbiased and consistent, and
2. the standard errors are computed in the correct way that makes confidence intervals have the desired confidence level.

Threats to external validity

Differences in populations

The causal effect may be different regarding different populations

- demographic and personal characteristics
- geographic and climate features
- timing

Differences in settings

- Difference in institutional environment, laws, or physical environment.

How to assess the external validity of a study

- Use specific knowledge.
- Case-by-case judgment.

Threats to Internal Validity of Multiple Regression Analysis

We introduce 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 $E(u_i|X_{1i}, \ldots, X_{ki}) \neq 0$ so as to make the OLS estimators biased and inconsistent.

Omitted variable bias

What are the two conditions for omitted variable bias?

Omitted variable bias

What are the two conditions for omitted variable bias?

1. At least one of the included regressors must be correlated with the omitted variable.
2. The omitted variable must be a determinant of the dependent variable, Y.

Solutions to omitted variable bias when the variable is observed or there are adequate control variables

- Include the omitted variables or the control variables
  - Avoid the violation of the first least squares assumption, $E(u|X) = 0$ or to let the conditional mean independence assumption hold, i.e., $E(u|X, W) = E(u|X)$

- Adding an additional independent variable may reduce the precision of the estimators of the coefficients
  - when the new variable actually does not belong to the population regression function,
  - when the new variable is correlated with other regressors.

Solutions to omitted variable bias when the variable is observed or there are adequate control variables

1. Identify the key coefficient(s) of interest.
2. *a priori* reasoning: before analyzing data, you should consider
   - What are the most likely sources of important omitted variable?
   - Answer the question using economic theory and expert knowledge.
3. Result in a base specification and a list of additional questionable variables that might help mitigate possible omitted variable bias.
4. Augment your base specification with the additional questionable control variables.
5. Present an accurate summary of your results in tabular form.

Solutions to omitted variable bias when adequate control variables are not available

- Panel data regression;
- Instrumental variables regression;
- Randomized controlled experiment.

Misspecification of the functional form of the regression function

- Functional form misspecification arises when the functional form of the estimated regression function differs from the functional form of the population regression function.
  - e.g., nonlinear vs. linear models
- Functional form misspecification bias can be considered as a type of omitted variable bias, in which the omitted variables are the terms that reflect the missing nonlinear aspects of the regression function.
  - e.g., missing the quadratic term

Solutions to functional form misspecification

- Plotting the data and the estimated regression function.
- Use a different functional form.
  - 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)

Measurement error and errors-in-variable bias

Measurement errors often happen in practice.

- respondents misstated answers to survey questions
- typographical errors when data were entered into the database
- the malfunctions of machines when recording data.

Measurement errors in

- dependent variable
- independent variable $\Rightarrow$ errors-in-variable bias.

Definition of errors-in-variable bias

- Errors-in-variables bias in the OLS estimator arises when an independent variable is measured imprecisely.
- This bias depends on the nature of the measurement error and persists even if the sample size is large.

Mathematical illustration of errors-in-variable bias

- The population regression model is

$$Y_i = \beta_0 + \beta_1 X_i + u_i, \text{ where } E(u_i|X_i) = 0 \text{ is satisfied}$$

- Suppose a regressor X_i is imprecisely measured by $\tilde{X}_i$.
  - The measurement error is $w_i = \tilde{X}_i - X_i$.
  - Assume $E(w_i) = 0$ and $\text{var}(w_i) = \sigma_w^2$.
- Rewrite the model in terms of $\tilde{X}_i$,

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 \tilde{X}_i + [\beta_1(X_i - \tilde{X}_i) + u_i] \\ &= \beta_0 + \beta_1 \tilde{X}_i + v_i \end{aligned} \tag{1}$$

  - The new error term is $v_i = \beta_1(X_i - \tilde{X}_i) + u_i$
- If $\text{cov}(w_i, \tilde{X}_i) \neq 0$, then $\text{cov}(v_i, \tilde{X}_i) \neq 0$ and the OLS estimator $\hat{\beta}_1$ is biased since $E(v_i|\tilde{X}_i) \neq 0$.

The biased and inconsistent OLS estimator with measurement errors

- If $\text{cov}(w_i, \tilde{X}_i) \neq 0$, then $\text{cov}(v_i, \tilde{X}_i) \neq 0$ and the OLS estimator $\hat{\beta}_1$ is biased since $E(v_i|\tilde{X}_i) \neq 0$.

- The OLS estimator is inconsistent.
  - The precise size and direction of the bias in $\hat{\beta}_1$ depend on the correlation between $\tilde{X}_i$ and the measurement error w_i. This correlation depends, in turn, on the specific nature of the measurement error.

The classical measurement error model

- The classical measurement error model assumes that the errors are purely random.

$$\text{corr}(w_i, X_i) = 0 \text{ and } \text{corr}(w_i, u_i) = 0$$

- The errors are correlated with $\tilde{X}_i$, that is, $\text{corr}(\tilde{X}_i, w_i) \neq 0$.
- In the classical measurement model, the OLS estimator $\hat{\beta}_1$ is inconsistent, and its the probability limit is

$$\hat{\beta}_1 \xrightarrow{p} \frac{\sigma_X^2}{\sigma_X^2 + \sigma_w^2} \beta_1 \tag{2}$$

- Since $\frac{\sigma_X^2}{\sigma_X^2 + \sigma_w^2} < 1$, Equation (2) implies that $\hat{\beta}_1$ is biased toward 0.
  - When σ_w^2 is very large, then $\hat{\beta}_1 \xrightarrow{p} 0$;
  - When σ_w^2 is very small, then $\hat{\beta}_1 \xrightarrow{p} \beta_1$.

Measurement error in Y

The effect of measurement error in Y is different from that in X. Generally, measurement in Y that has conditional mean zero given the regressors will not induce bias in the OLS coefficients, but will lead to inefficient estimators.

- Suppose Y has the classical measurement error, that is, what we observe, $\tilde{Y}_i$, is the true value of Y_i plus a purely random error w_i. Then, the regression model is
$$\tilde{Y}_i = \beta_0 + \beta_1 X_i + v_i, \text{ where } v_i = w_i + u_i$$
- If w_i and X_i are independently distributed so that $E(w_i|X_i) = 0$, in which case $E(v_i|X_i) = 0$, so $\hat{\beta}_1$ is unbiased.
- Since $\text{var}(v_i) = \text{var}(w_i) + \text{var}(u_i) > \text{var}(u_i)$, the variance of $\hat{\beta}_1$ is larger than it would be without measurement error.

Solutions to errors-in-variable bias

- Get an accurate measure of X as possible as you can.
- Use an instrumental variable that is correlated with the actual value of X_i but is uncorrelated with the measurement error.
- Develop a mathematical model of the measurement error and use the resulting formula to adjust the estimates. This requires specific knowledge of the errors.