Lecture 11: Assessing Studies Based on Multiple Regression

Zheng Tian

Outline

1. Internal and External Validity
2. Omitted Variable Bias
3. Misspecification of the Functional Form
4. Measurement Errors and Errors-in-Variable Bias
5. Missing data and sample selection
6. Simultaneous causality
7. Inconsistency of Standard Errors

1. Internal and External Validity
2. Omitted Variable Bias
3. Misspecification of the Functional Form
4. Measurement Errors and Errors-in-Variable Bias
5. Missing data and sample selection
6. Simultaneous causality
7. Inconsistency of Standard Errors

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 $E(u_i|X_{1i}, , X_{ki}) \neq 0$ 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 ⇒ 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 $\mathrm{var}(w_i) = \sigma_w^2$.
- 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 $\mathrm{cov}(w_i, \tilde{X}_i) \neq 0$, then $\mathrm{cov}(v_i, \tilde{X}_i) \neq 0$ 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 $\mathrm{cov}(w_i, \tilde{X}_i) \neq 0$, then $\mathrm{cov}(v_i, \tilde{X}_i) \neq 0$ 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.
$$\mathrm{corr}(w_i, X_i) = 0 \text{ and } \mathrm{corr}(w_i, u_i) = 0$$
- The errors are correlated with $\tilde{X}_i$, that is, $\mathrm{corr}(\tilde{X}_i, w_i) \neq 0$.
- 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 $\mathrm{var}(v_i) = \mathrm{var}(w_i) + \mathrm{var}(u_i) > \mathrm{var}(u_i)$, 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.

Missing data and sample selection

Whether missing data pose a threat to internal validity depends on why the data are missing. We consider three cases of missing data.

Missing data at random

When data are missing completely at random, unrelated with X and Y, then the effect is to reduce the sample size but not introduce any estimation bias.

Missing data based on X

When the data are missing based on the value of a regressor but unrelated with generating Y, the effect is also to reduce the sample size but not introduce bias.

Sample selection bias

When the data are missing because of a selection process that is related with 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, resulting in sample selection bias.

Sample selection bias: two examples

The sample selection problem can be cast either as a consequence of nonrandom sampling or as a missing data problem, illustrated using the following two examples.

Nonrandom sampling: Height of undergraduates

The professor of Statistics asks you to estimate the mean height of undergraduate males. You collect your data (obtain your sample) by standing outside the basketball teams locker room and recording the height of the undergraduates who enter.

Missing data: Trade volume of pairs of countries

- The amount of commodities that two countries can trade depends on GDP of two countries, industrial structures, factor abundance, etc.
- We can get the data on trade volume between pairs of countries from World Bank, Penn World Table, etc.
- Sample selection bias occurs due to the non-trading pairs.

Solutions to sample selection bias

- Collect the sample in a way that avoids sample selection.
- Heckman's two-step method.
- Randomized controlled experiment.
- Construct a model of the sample selection problem and estimate that model.

Simultaneous causality

Up to now, all we examined is how X can cause Y. What if Y causes X? If Y does cause X in some way, there is simultaneous causality problem, which lead to biased and inconsistent OLS estimator.

There are many examples of simultaneous causality in Economics. In the paper of Acemuglou et al.(2000), *The Colonial Origins of Comparative Development: An Empirical Investigation*, the authors estimate the effect of institutions on economic performance. However, the simultaneous causality (or mutual causality) comes from the fact that not only do good institutions promote economic performance, but also countries with high GDP per capita can afford good institutions and secure property rights, which in turn yield better economic performance.

Simultaneous causality bias

Simultaneous causality leads to biased estimates of the effect of X on Y, referred to as simultaneous causality bias. We can express the simultaneous causality using a simultaneous equations.

$$Y_i = \beta_0 + \beta_1 X_i + u_i \tag{3}$$

$$X_i = \gamma_0 + \gamma_1 Y_i + v_i \tag{4}$$

Intuitively, simultaneous causality comes from the following facts.

- Large u_i means large Y_i, which implies large X_i (if $\gamma_1 > 0$).
- This implies that u_i and X_i are correlated, i.e., $\mathrm{cov}(X_i, u_i) \neq 0$.
- Thus, the OLS estimator of β_1 from merely estimating Equation (3) is biased and inconsistent.

Simultaneous causality bias (cont'd)

Formally, we can prove that $\mathrm{cov}(X_i, u_i) \neq 0$, resulting in the bias in the OLS estimator of β_1.

Proof.

$$\begin{aligned}
\mathrm{cov}(X_i, u_i) &= \mathrm{cov}(\gamma_0 + \gamma_1 Y_i + v_i, u_i) \\
&= \gamma_1 \mathrm{cov}(Y_i, u_i) + \mathrm{cov}(v_i, u_i)(\text{ Assuming } \mathrm{cov}(v_i, u_i) = 0) \\
&= \gamma_1 \mathrm{cov}(\beta_0 + \beta_1 X_i + u_i, u_i) \\
&= \gamma_1 \mathrm{cov}(X_i, u_i) + \gamma_1 \sigma_u^2
\end{aligned}$$

Solving for $\mathrm{cov}(X_i, u_i)$ yields the result $\mathrm{cov}(X_i, u_i) = \gamma_1 \sigma_u^2 / (1 - \gamma_1 \beta_1)$, which is not equal to zero unless $\gamma_1 = 0$, i.e., the simultaneous causality does exist. □

Solutions to simultaneous causality bias

- 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)

Sources of inconsistency of OLS standard errors

Inconsistent standard errors pose a different threat to internal validity. Even if the OLS estimator is consistent and the sample is large, inconsistent standard errors will produce hypothesis tests with size that differs from the desired significance level and "95%" confidence intervals that fail to include the true value in 95% of repeated samples.

There are two main reasons for inconsistent standard errors: improperly handled heteroskedasticity and correlation of the error term across observations.

Heteroskedasticity

If the errors are heteroskedastic and you mistakenly use the homoskedasticity-only standard errors that are reported by some software by default, then the t-test and the F-test based on the wrong standard errors do not have the desired size.

The solution to this problem is to use heteroskedasticity-robust standard errors of the OLS estimators and to construct t- and F-statistics using a heteroskedasticity-robust variance estimator, which is provided as an option in modern software packages.

The Breusch-Pagan test for heteroskedasticity

We can test whether heteroskedasticity exists in a regression model using the Breusch-Pagan test. The test consist of the following steps:

1. Estimate a regression model, $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + u$, and obtain the squared OLS residuals, $\hat{u}^2$.
2. Run a regression of $\hat{u}^2 = \delta_0 + \delta_1 X_1 + \cdots + \delta_k X_k + v$, and obtain the R^2 of this regression, denoted as $R^2_{\hat{u}^2}$.
3. Test the null hypothesis, $H_0 : E(u^2|X_1, \ldots, X_k) = \sigma^2$, i.e., homoskedasticity, against the alternative hypothesis for heteroskedasticity. The test statistics can be the overall F statistics for the regression in the second step, which is

   $$F = \frac{R^2_{\hat{u}^2}/k}{(1 - R^2_{\hat{u}^2})/(n-k-1)} \sim F(k, n-k-1)$$

   Or we can compute an LM test statistics, which is

   $$LM = nR^2_{\hat{u}^2} \sim \chi^2(k)$$

   where n is the number of observations.
4. Based on the F-statistic or the LM statistic, compute the p-value. If the p-value is smaller than the significance level, we can reject the null hypothesis of homoskedasticity.

Correlation of the error term across observations

In the lease squares assumptions, we assume that (X_i, Y_i) for $i = 1, \dots, n$ are i.i.d., which implies that u_i are uncorrelated across observations. However, in some setting, the population regression error can be correlated across observations. There are mainly two types of correlation in consideration: serial correlation and spatial correlation.

- Serial correlation arises from the repeated observations over the same entity over time. It is a prevalent problem in time series data.
- Spatial correlation arises from the influence of contiguous (neighboring) observations over geographic units.
- The OLS estimator with serial correlation or spatial correlation is still unbiased and consistent, but inference based on no correlation assumption is not valid.

Solutions to correlation of the error term across observations

- Use the heteroskedasticity-and-auto-correlation-consistent standard errors (HAC). We will learn how to handle serial correlation in time series data in the next two semesters.
- Model the spatial correlation specifically. Spatial econometrics is a branch of econometrics that deals with spatial correlation.