# Lecture 11: Assessing Studies Based on Multiple Regression

Zheng Tian

## 1 Introduction

### 1.1 Overview

The preceding lectures explain how to use multiple regression to analyze the relationship among variables. In this lecture, we step back and ask, What makes a study that uses multiple regression reliable? We answer this question by assessing regression analysis under the framework of internal and external validity.

### 1.2 Reading materials

- Chapter 9 in *Introduction to Econometrics* by Stock and Watson.

## 2 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.

### 2.1 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

By contrast, 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.

## 2.2 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.

## 2.3 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 (the actual rejection rate of the test under the null hypothesis should equal its desired significance level), and the confidence intervals should have the desired confidence level.

#### Internal validity in regression analysis

For a regression analysis of casual effects based on the OLS estimation, the requirements for internal validity are that

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

## 2.4 Threats to external validity

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

**Differences in populations**

The causal effect might not be the same in the population studied and the population of interest due to their differences in

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

**Differences in settings**

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

**How to assess the external validity of a study**

- External validity must be judged using specific knowledge of the population and settings studied and those of interest.
- We can compare two or more studies on different but related populations. Formally, this comparison can be conducted using a meta-analysis.

# 3 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.

## 3.1 Omitted variable bias

Recall that omitted variable bias arises when a variable that both determines $Y$ and is correlated with one or more of the included regressors is omitted from the regression.

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

- A trade-off between omitted variable bias and the precision of estimators
  - If you have the data on the omitted variable, or you have the data on one or more control variables for an unobserved omitted variable, we can add these additional regressors to avoid the violation of the first least squares assumption, $E(u|X) = 0$ or to let the conditional mean independence assumption hold, i.e., $E(u|X, W) = E(u|X)$, so that the coefficient on the variable of interest is unbiased and consistent.
  - 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 (i.e., its population regression coefficient is zero), or when the new variable is correlated with other regressors, resulting in imperfect multicollinearity.

    Question: Why may adding an irrelevant variable reduce the precision of other coefficients? (*Hint: What does the Gauss-Markov Theorem indicate as for the variance of the OLS estimators?*)
- Some guidelines to decide whether to include an additional variable
  1. Identify the key coefficient(s) of interest.
     - e.g., the student-teacher ratio in the test score regression.
  2. *a priori* reasoning
     - What are the most likely sources of important omitted variable?
     - Answer the question using economic theory and expert knowledge.
     - Done before analyzing data.
     - Result in a base specification and a list of additional questionable variables that might help mitigate possible omitted variable bias.
  3. Augment your base specification with the additional questionable control variables.

- If the coefficients on control variables are statistically significant or if the estimated coefficients of interest change appreciably when control variables are included, then you should consider modifying the base specification.
- If not, exclude these control variables from the regression.

4. Present an accurate summary of your results in tabular form.
   - This provides "full disclosure" to skeptical readers who can draw their conclusions.

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

Adding an omitted variable is not an option if you do not have data on that variable and if there are no adequate control variables. We introduce three ways to circumvent omitted variable bias.

- **Panel data**

  Panel data (or longitudinal data) consist of observations on the same $n$ entities at two or more time periods. If the data set contains observations on the variables $X$ and $Y$, then the data are denoted

  $$(X_{it}, Y_{it}),\ i = 1, \ldots, n \text{ and } t = 1, \ldots, T$$

  where the first subscript, $i$, refers to the entity being observed and the second subscript, $t$, refers to the date at which it is observed.

  The key of using panel data regression to circumvent omitted variable bias lies in the idea that omitted variables that represent personal characteristics do not change over time so that any changes in $Y$ over time cannot be caused by the omitted variable.

  Suppose we have $n$ entities and $T$ observations for each entity. $X_{it}$ is the observed regressor, $Y_{it}$ is the dependent variable, and $Z_i$ is the unobserved time-invariant variable representing idiosyncratic characteristics of entity $i$. We can set up a linear regression model as follows

  $$Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 Z_i + u_{it}$$

  This model is a simple representation of the **fixed effects** panel data regression model, in which $Z_i$ is usually defined as a dummy variable for entity $i$.

- **Instrumental variable**

If the omitted variable(s) cannot be measured, we can use an instrumental variables (IV) regression. Suppose that in the simple linear regression model

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

$X_i$ and $u_i$ are correlated due to unobserved omitted variables. Then we can use an instrumental variable $Z$ to account for the part in $X$ that is correlated with $u$.

For an instrumental variable $Z$ to be valid, it must satisfy two conditions:

1. **Instrument relevance**: $\text{Corr}(Z_i, X_i) \neq 0$
2. **Instrument exogeneity**: $\text{Corr}(Z_i, u_i) = 0$

The model is estimated using the Two-Stage-Least-Squares (TSLS) method which basically consists of two steps:

**Stage 1** Regress $X_i$ on $Z_i$, including an intercept, obtain the predicted values, $\hat{X}_i$.

**Stage 2** Regress $Y_i$ on $\hat{X}_i$, including an intercept; the coefficient on $\hat{X}_i$ is the TSLS estimator $\hat{\beta}_1^{TSLS}$.

- **Randomized controlled experiment**

  The third solution is to use a research design in which the effect of interest is studied using a randomized controlled experiment. Randomized controlled experiments are discussed in Chapter 12.

## 3.2 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)

## 3.3 Measurement error and errors-in-variable bias

Measurement errors often happen in practice. They may come from respondents misstated answers to survey questions, from typographical errors when data were entered into the database for the first time, and from the malfunctions of machines when recording data.

Measurement errors can occur in independent variables as well as the dependent variable, of which their effects on the estimated coefficients depend on the nature of the errors. Let's first focus on errors in independent variable, which cause biased estimated coefficients, referred to as **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

Suppose a regressor $X_i$ is imprecisely measured by $\tilde{X}_i$. That means that we observe $\tilde{X}_i$ and use it in estimation.

Then consider a simple regression model

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

in which $E(u_i|X_i) = 0$ is satisfied.

Since we use $\tilde{X}_i$ other than $X_i$ in estimation, we rewrite the model in terms of $\tilde{X}_i$, that is,

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

where $v_i = \beta_1(X_i - \tilde{X}_i) + u_i$ in which we define the measurement error as $w_i = \tilde{X}_i - X_i$, and assume $E(w_i) = 0$ and $\text{Var}(w_i) = \sigma_w^2$.

If the measurement errors $w_i$ is correlated with $\tilde{X}_i$, then the regressor $\tilde{X}_i$ is correlated with the new error term $v_i$ and $\hat{\beta}_i$ will be biased and inconsistent The OLS estimator $\hat{\beta}_1$ is biased since $E(v_i|\tilde{X}_i) \neq 0$.

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 so that we assume $\text{Corr}(w_i, X_i) = 0$ and $\text{Corr}(w_i, u_i) = 0$, but the errors are correlated with $\tilde{X}_i$, that is, $\text{Corr}(\tilde{X}_i, w_i) \neq 0$. Then, we can prove that in this model, the OLS estimator $\hat{\beta}_1$ of Equation (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 $\sigma_w^2$ is very large, then $\hat{\beta}_1 \xrightarrow{p} 0$;
- When $\sigma_w^2$ is very small, then $\hat{\beta}_1 \xrightarrow{p} \beta_1$.

*Proof.* Since $\tilde{X}_i = X_i + w_i$, we have $\text{Var}(\tilde{X}_i) = \sigma_X^2 + \sigma_w^2$.

According to Equation (2) and $\text{Cov}(X_i, u_i) = 0$, we have

$$\begin{aligned} v_i &= \beta_1(X_i - \tilde{X}_i) + u_i = -\beta_1 w_i + u_i \\ \text{Cov}(\tilde{X}_i, w_i) &= \text{Cov}(X_i + w_i, w_i) = \sigma_w^2 \\ \text{Cov}(\tilde{X}_i, v_i) &= -\beta_1 \text{Cov}(\tilde{X}_i, w_i) + \text{Cov}(\tilde{X}_i, u_i) = -\beta_1 \sigma_w^2 \end{aligned}$$

Recall that in Chapter 6 for a simple regression model, when the error term is correlated with the regressor, like $\text{Cov}(\tilde{X}_i, v_i) \neq 0$, then $\hat{\beta}_1$ has the probability limit

$$\hat{\beta}_1 \xrightarrow{p} \beta_1 + \frac{\text{Cov}(\tilde{X}_i, v_i)}{\text{Var}(\tilde{X}_i)}$$

for which the probability limit is just

$$\beta_1 - \beta_1 \frac{\sigma_w^2}{\sigma_{\tilde{X}_i}^2} = \frac{\sigma_X^2}{\sigma_X^2 + \sigma_w^2}\beta_1$$

□

### 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.

- 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

$$\tilde{Y}_i = \beta_0 + \beta_1 + v_i, \text{ where } v_i = w_i + u_i$$

- 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.

## 3.4 Missing data and sample selection

Missing data are a common feature of economic data sets. 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. For example, we repeat an experiment examining the influence of X on Y on several days and save the results at different time. Suppose that time is a regressor, and we miss the all data from 1 pm to 2 pm. If the missing data do not affect the process of doing the experiment, then the estimate of the causal effect of X on Y will still be unbiased.

**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**.

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 team's locker room and recording the height of the undergraduates who enter.

  - Is this a good research design – will it yield an unbiased estimate of undergraduate height?
  - You have sampled individuals in a way that was related to the outcome Y (height), resulting in bias.

- 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.
  - Using the data of observed trade volume between pairs of countries can lead to sample selection bias because the sample selection process omit the pairs of countries that do not trade with each other. But the fact that two countries do not trade may also bear some economic meaning that can influence the causal effect of the variables of interest on trade volume.

- Solutions to sample selection bias
  - Collect the sample in a way that avoids sample
  - Randomized controlled experiment.
  - Construct a model of the sample selection problem and estimate that model.

### 3.5 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 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., $\text{Cov}(X_i, u_i) \neq 0$.
- Thus, the OLS estimator of $\beta_1$ from merely estimating Equation (3) is biased and inconsistent.

Formally, we can prove that $\text{Cov}(X_i, u_i) \neq 0$, resulting in the bias in the OLS estimator of $\beta_1$.

*Proof.*

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

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

**Solutions to simultaneous causality bias**

1. 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).
2. 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.
3. Use instrumental variables regression to estimate the causal effect of interest (effect of X on Y, ignoring effect of Y on X)

## 3.6 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 $i = 1, \ldots, n$ 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.
- Solution:

- 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.