# Econometrics I

## Lecture 6: Endogeneity and Internal Validity

Paul T. Scott
NYU Stern

Fall 2018

# Regression Validity

Remember the broadest question: What is the effect of $X$ on $Y$?

*How do we assess whether a regression answers this question?*

# Regression Validity

Remember the broadest question: What is the effect of $X$ on $Y$?

*How do we assess whether a regression answers this question?*

- **Internal Validity:** The causal effect *in the population being studied* is *properly identified*
  - Properly identified refers to the belief that the OLS assumptions are met ($E(\varepsilon|X)$)
  - Our inference is conditional on the population under scrutiny
    *Example:* Can we identify the causal effect of education on income for American males in the 1980s?

- **External Validity:** The causal effect from the population under scrutiny can be ported to other settings
  - The causal effect of $X$ on $Y$ can depend on non-modeled conditions
    *Example:* The effect of education on income depends on the skill demands of the industries that are currently employing most workers

# Assessing External Validity I

- Consider two populations: **studied** population and a population of interest
  - Whether the effect of $X$ on $Y$ can be carried from one population to the next requires thinking hard about how similar are the environments
  - No right or wrong answer to this, judgment call

- Concrete Example: Comparing education policies across countries
  - **Studied Population:** A randomized controlled trial establishes that in the United States, 1 extra year of high school increases income by 5%
  - **Population of Interest:** Does this mean that one can increase incomes in developing countries by 20% by mandating high school completion?
    - Are the schools the same?
    - Is the demand for college educated workers the same?
    - Are the populations the same in terms of education inputs? (E.g., nutrition?)

- External validity can be understood to be not only about *populations*, but also about *data generating processes*.

# Assessing External Validity II

- External validity can be understood to be not only about *populations*, but also about *data generating processes*.
- This is often the motivation behind **structural econometric** approaches, such as dynamic models of behavior.
- Examples: Hendel and Nevo (2006) on laundry detergent demand, Scott (2013) on agricultural land use.

# Assessing Internal Validity

## What does this mean?

- "Internal validity" is a question of whether the OLS assumptions are satisfied
- External Validity and Internal Validity are not the same thing
  - A Randomized Controlled Trial (RCT) will likely be **internally valid**
  - But the population for the RCT might not be representative of the population on the whole, or of other populations so it might not be **externally valid**
- The OLS assumption in question is typically **exogeneity**. Violations of the exogeneity assumption can be referred to as **endogeneity problems**.

# What sort of exogeneity?

- The theoretical arguments we made were based on **strict exogeneity**:

$$E\left[\boldsymbol{\varepsilon}|\mathbf{X}\right] = 0$$

- Asymptotic consistency of OLS can be proved assuming that the regressors are **predetermined**:

$$E\left[\varepsilon_i \mathbf{x}_i\right] = 0$$

- The latter is a weaker assumption (the former implies the latter), and it may be more conceptually intuitive to think of endogeneity problems as violations of the latter – i.e., endogeneity means that the error term and regressor(s) are correlated.

# Assessing Internal Validity

## When is $E(\varepsilon|X) \neq 0$?

Internal validity asks when the OLS assumptions are violated.

A taxonomy of internal validity (or endogeneity) problems:

1. Omitted Variable Bias: $Z$ such that $\sigma_{X,Z}, \sigma_{\varepsilon,Z} \neq 0$

   *Solution:* Control variables, randomization (we will learn others)
2. Specification Bias: The relationship is not linear in $X$

   *Solution:* Try logs, polynomials, interactions, etc.
3. **Today:** Measurement error bias
4. Other internal validity issues:
   - Simultaneity bias: $Y$ and $X$ cause *each other*
   - Sample selection bias: Unrepresentative sample

# OVB Review

The problem: uncontrolled for variable

$$Y_i = \beta_0 + \beta_1 X_i + \underbrace{\beta_2 Z_i}_{\text{Omitted}} + \varepsilon_i$$

- Intimately related to Selection Bias
- OVB $\Rightarrow$ $X$ and $\varepsilon$ correlated
- Solutions:
  - Randomization
  - Control directly for $Z$
  - Panel data and instruments

# Selection Bias Review

The problem: selection into a group is non-random

$$Y_i = \beta_0 + \beta_1 T_i + (Y_{i0} - \beta_0)$$

- Examples:
  - Experiments: People select into treatment versus control
  - Education: People who get a college degree are not random
- Non-random selection $\Rightarrow$ baseline outcomes $Y_{i0}$ are correlated with $T_i$
- Solutions:
  - Randomization via an RCT
  - Control variables
  - More careful sample selection
  - Model selection bias explicitly

- **Sample Selection Bias** occurs when the sample under consideration is not selected randomly from the population under consideration
- Education and Income example:
  - Only working people have an income
  - So estimated effect of education on income is only the effect on *already employed* persons
  - If education $\Rightarrow$ a higher probability of working, then *total effect* should take into account this probability
- If selection makes $E(\varepsilon|X) \neq 0$ then OLS is biased

- **Sample Selection Bias** occurs when the sample under consideration is not selected randomly from the population under consideration
- This is *not* the same as **selection bias**
  - Selection is about who is assigned treatment versus control
  - Sample selection is about whether we do not *observe* data for some groups
- **Solutions:**
  - Typically hard to deal with
  - Either get more data from the full population...
  - ... or model the selection issue

# Functional Form Misspecification

The problem: the relationship isn't linear:

$$Y_i = \beta_0 + \beta_1 X_i + \underbrace{\beta_2 X_i^2 + \beta_3 X_i^3 + ...}_{\text{Ought to be included}} + \varepsilon_i$$

- Technically the omitted non-linear terms are like OVB
- Could also be about log versus linear or interaction terms
- Solutions:
  - Include non-linear terms (polynomials or logarithms)
  - Include interaction terms (if the issue is that $\beta$ varies)
  - Do some model selection to avoid over-fitting

# Measurement Error (Errors-in-Variables) Bias

The problem: the $X$ variable is measured with noise

- The idea mathematically:

$$\textbf{TRUTH:} Y = \beta_0 + \beta_1 X + \varepsilon$$
$$\textbf{DATA:} Y = \beta_0 + \beta_1 X^* + \varepsilon^*$$

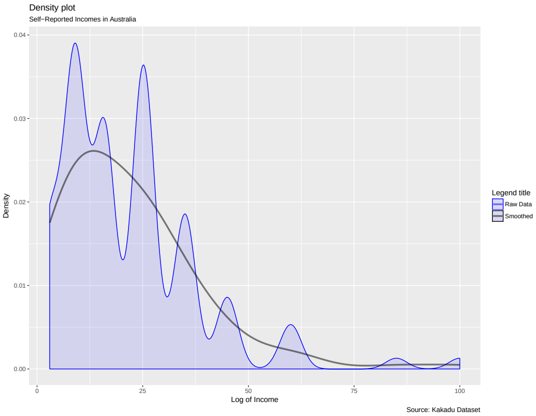
  where $X^*$ is a noisy measure of $X$
- Examples:
  - Recording errors in data entry
  - Recollection errors in survey data (these are frequent)
  - Rounding errors
  - Hard to measure variables (e.g., a firm's capital stock)
- Note: we can also have measurement error in $Y$. Turns out these behave differently

# Visualizing Measurement Error: Self-Reported Australian Incomes

Survey question: What is your income in thousands?

# Visualizing Measurement Error: Self-Reported Australian Incomes

Survey question: What is your income in thousands?




- The lumps occur at 5s and 0s
- E.g., people making 41k might say “40” or “45” instead of 41

# Measurement Error in $X$

## The Math of Measurement Error

- Consider adding and subtracting noise to the truth:

$$
\begin{aligned}
Y_i =& \beta_0 + \beta_1 X_i + \varepsilon_i \\
=& \beta_0 + \beta_1 X_i^* + \underbrace{\beta_1 (X_i - X_i^*) + \varepsilon_i}_{\text{“New” Error Term:} V_i} \\
=& \beta_0 + \beta_1 X_i^* + V_i
\end{aligned}
$$

- $X_i^*$ and $V_i$ likely to be correlated because $X^*$ is part of $V$
- Extra assumptions:
  - **Classical Errors-in-Variables:**

$$X_i = X_i^* + u_i$$

    with $u$ being independent of $X$ and $\varepsilon$ (just noise)
  - If $u$ and $X$ are correlated this is very complicated

Under classical measurement error:

$$Y_i = \beta_0 + \beta_1 X_i^* + \underbrace{\beta_1 u_i + \varepsilon_i}_{V_i}$$

- Using the population definition of $\hat{\beta}^{OLS}$:

$$\begin{aligned}
\hat{\beta}^{OLS} &\rightarrow \frac{Cov(X^*, Y)}{Var(X^*)} \\
&= \frac{Cov(X + u, \beta_0 + \beta_1 X + \varepsilon)}{Var(X + u)} \\
&= \frac{Cov(X, \beta_0 + \beta_1 X + \varepsilon) + Cov(u, \beta_0 + \beta_1 X + \varepsilon)}{Var(X + u)} \\
&= \frac{\beta_1 Cov(X, X) + 0}{Var(X + u)}
\end{aligned}$$

# Measurement Error in $X$

## Classical Measurement Error, Cont'd

- Continuing from above:

$$\begin{aligned}
\hat{\beta}^{OLS} \to & \frac{\beta_1 Cov(X, X) + 0}{Var(X + u)} \\
= & \beta_1 \times \frac{Var(X)}{Var(X + u)} \\
= & \beta_1 \times \underbrace{\frac{\sigma_X^2}{\sigma_X^2 + \sigma_u^2}}_{\text{Attenuation}}
\end{aligned}$$

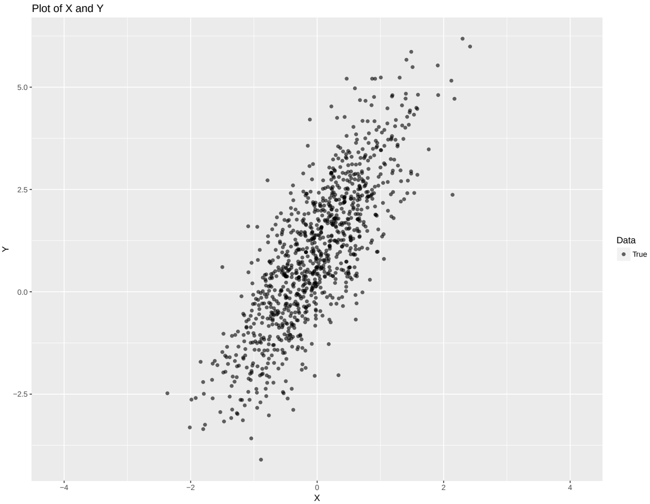
- Estimated coefficient converges to the truth times an attenuation term
  - Attenuation term is less than 1 $\Rightarrow$ pushes coefficient towards zero
    - **NB:** It does *not* make the term more negative—it dampens the coefficient and preserves the sign
  - This "attenuates" the effect of $X$ on $Y$. Hence the name: **attenuation bias**

- Why does measurement error attenuate the estimated effect of $X$ on $Y$?
  - Noise in measured $X$ makes it more likely to see high $X$ and low $X$ with some $Y$ because of randomness
  - Extreme example: fix $X$ and keep adding noise—eventually $X^*$ will look like noise itself
- Notice the following rearranging of the attenuation term:

$$\frac{\sigma_X^2}{\sigma_X^2 + \sigma_u^2} = \frac{1}{1 + (\sigma_u^2/\sigma_X^2)}$$

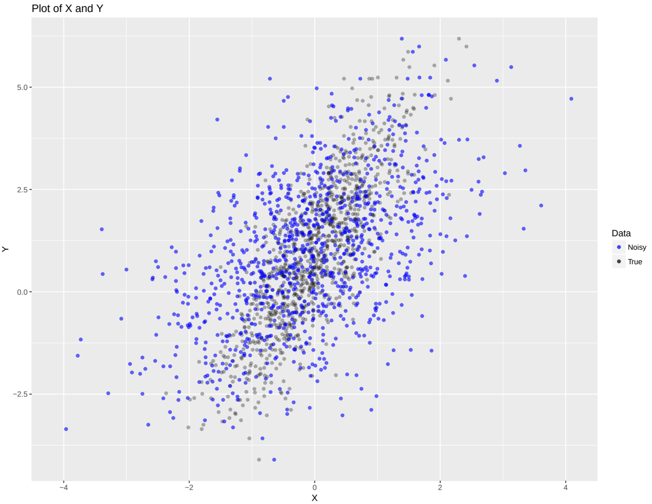
  - $\sigma_X^2/\sigma_u^2$ is called the **signal-to-noise ratio**
  - Larger signal to noise ratio $\Rightarrow$ smaller bias
  - What matter is the *relative* variance of $X$ to $u$
  - If $\sigma_u^2$ is small *relative* to $\sigma_X^2$ then attenuation bias isn't too large

# Visualizing Measurement Error



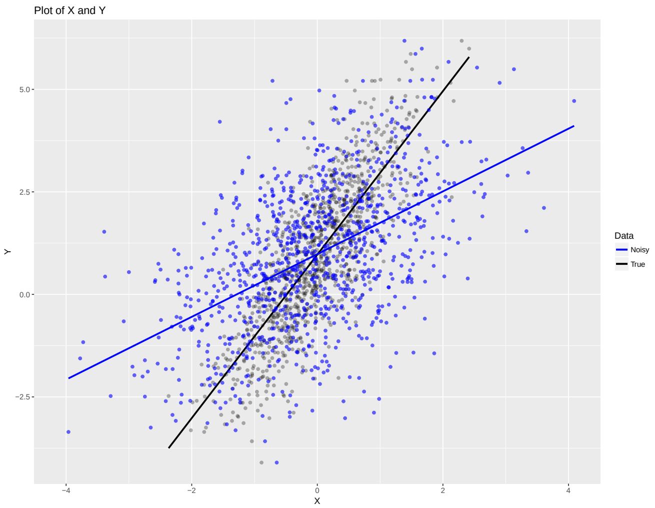

- Suppose this is the true data. But...

# Visualizing Measurement Error




- ... you only observe the noisy data...

# Visualizing Measurement Error




- … then what happens to the estimated slope?

# Measurement Error in $Y$

## No problemo

- Suppose the econometrician observes $Y^* = Y + u$, where $Y$ is the true value.

$$
\begin{aligned}
Y_i &= \beta_0 + \beta_1 X_i + \varepsilon_i \\
Y_i + Y_i^* &= \beta_0 + \beta_1 X_i + \varepsilon_i + Y_i^* \\
Y_i^* &= \beta_0 + \beta_1 X_i + \varepsilon_i + \left(Y_i^* - Y_i\right)
\end{aligned}
$$

- Again, we can see this as changing the error term:

$$Y_i^* = \beta_0 + \beta_1 X_i + V_i$$

  where $V_i = \varepsilon_i + \left(Y_i^* - Y_i\right)$.
- As long as this is classical measurmenet error, i.e. $E\left[u|X\right] = 0$, then as long as exogeneity is satisfied with the original error term $E\left[\varepsilon|X\right] = 0$, it will still be satisfied for the modified error term $E\left[V|X\right] = 0$.