Lecture 7: Assessing Stuides Based on Multiple Regression

Introduction ot Econometrics,Fall 2017

Zhaopeng Qu

Nanjing University

11/13/2017

Introduction

Definitions of internal and external validity

- **Internal validity**: the statistical inferences about causal effects are valid for the population and setting being studied.

Definitions of internal and external validity

- **Internal validity**: the statistical inferences about causal effects are valid for the population and setting being studied.
- **External validity**: the statistical inferences can be generalized from the population and setting studied to other populations and settings.

Definitions of internal and external validity

- **Internal validity**: the statistical inferences about causal effects are valid for the population and setting being studied.
- **External validity**: the statistical inferences can be generalized from the population and setting studied to other populations and settings.
- Internal and external validity distinguish between the population and setting *studied* and the population and setting to which the results are *generalized*.

Definitions of internal and external validity

- **Internal validity**: the statistical inferences about causal effects are valid for the population and setting being studied.
- **External validity**: the statistical inferences can be generalized from the population and setting studied to other populations and settings.
- Internal and external validity distinguish between the population and setting *studied* and the population and setting to which the results are *generalized*.
- Example: Class size and test score

Definitions of internal and external validity

- **Internal validity**: the statistical inferences about causal effects are valid for the population and setting being studied.
- **External validity**: the statistical inferences can be generalized from the population and setting studied to other populations and settings.
- Internal and external validity distinguish between the population and setting *studied* and the population and setting to which the results are *generalized*.
- Example: Class size and test score
    - the population studies: elementary school districts in CA

Definitions of internal and external validity

- **Internal validity**: the statistical inferences about causal effects are valid for the population and setting being studied.
- **External validity**: the statistical inferences can be generalized from the population and setting studied to other populations and settings.
- Internal and external validity distinguish between the population and setting *studied* and the population and setting to which the results are *generalized*.
- Example: Class size and test score
  - the population studies: elementary school districts in CA
  - the population of interest: high schools in CA

Definitions of internal and external validity

- **Internal validity**: the statistical inferences about causal effects are valid for the population and setting being studied.
- **External validity**: the statistical inferences can be generalized from the population and setting studied to other populations and settings.
- Internal and external validity distinguish between the population and setting *studied* and the population and setting to which the results are *generalized*.
- Example: Class size and test score
    - the population studies: elementary school districts in CA
    - the population of interest: high schools in CA
    - different populations and settings: elementary schools in MA or in China

Internal validity

Internal validity in an OLS regression model

- Suppose we are interested in the causal effect of X1 on Y and we estimate the following regression model

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + ... + \beta_k X_{k,i} + u_i, i = 1, ..., n$$

Internal validity in an OLS regression model

- Suppose we are interested in the causal effect of X1 on Y and we estimate the following regression model

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + ... + \beta_k X_{k,i} + u_i, i = 1, ..., n$$

- Internal validity has three components:

Internal validity in an OLS regression model

- Suppose we are interested in the causal effect of X1 on Y and we estimate the following regression model

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + ... + \beta_k X_{k,i} + u_i, i = 1, ..., n$$

- Internal validity has three components:
  - The OLS estimator of β_1 is *unbiased and consistent*

Internal validity in an OLS regression model

- Suppose we are interested in the causal effect of X1 on Y and we estimate the following regression model

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + ... + \beta_k X_{k,i} + u_i, i = 1, ..., n$$

- Internal validity has three components:
  - The OLS estimator of β_1 is *unbiased and consistent*
  - The value of β_1 should be *large enough* to make it sense.

Internal validity in an OLS regression model

- Suppose we are interested in the causal effect of X1 on Y and we estimate the following regression model

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + ... + \beta_k X_{k,i} + u_i, i = 1, ..., n$$

- Internal validity has three components:
  - The OLS estimator of β_1 is *unbiased and consistent*
  - The value of β_1 should be *large enough* to make it sense.
  - Hypothesis tests should have the *desired significance level* and confidence intervals should have the desired confidence level.(significant)

Threats to internal validity

- Threats to internal validity:

Threats to internal validity

- Threats to internal validity:
  - Omitted variables

Threats to internal validity

- Threats to internal validity:
    - Omitted variables
    - Function form misspecification

Threats to internal validity

- Threats to internal validity:
  - Omitted variables
  - Function form misspecification
  - Measurement error

Threats to internal validity

- Threats to internal validity:
  - Omitted variables
  - Function form misspecification
  - Measurement error
  - Simultaneous causality

Threats to internal validity

- Threats to internal validity:
  - Omitted variables
  - Function form misspecification
  - Measurement error
  - Simultaneous causality
  - Missing Data and Sample Selection

Threats to internal validity

- Threats to internal validity:
  - Omitted variables
  - Function form misspecification
  - Measurement error
  - Simultaneous causality
  - Missing Data and Sample Selection
  - Heteroskedasticity and/or correlated error terms

Threats to internal validity

- Threats to internal validity:
  - Omitted variables
  - Function form misspecification
  - Measurement error
  - Simultaneous causality
  - Missing Data and Sample Selection
  - Heteroskedasticity and/or correlated error terms
- In some way

$$Internal\ Invalidity = endogeneity\ in\ the\ estimation$$

Omitted Variable Bias(OVB): Review

- Suppose we want to estimate the causal effect of X_{1i} on Y_i,which represent STR and Test Score,respectively.

Omitted Variable Bias(OVB): Review

- Suppose we want to estimate the causal effect of X_{1i} on Y_i,which represent STR and Test Score,respectively.
- Besides, W_i is the share of English learners which we will omit in the regression Thus

Omitted Variable Bias(OVB): Review

- Suppose we want to estimate the causal effect of X_{1i} on Y_i,which represent STR and Test Score,respectively.
- Besides, W_i is the share of English learners which we will omit in the regression Thus
- True model:

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

where $E(u_i|X_i, W_i) = 0$ and

Omitted Variable Bias(OVB): Review

- Suppose we want to estimate the causal effect of X_{1i} on Y_i,which represent STR and Test Score,respectively.
- Besides, W_i is the share of English learners which we will omit in the regression Thus
- True model:

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

  where $E(u_i|X_i, W_i) = 0$ and
- But we can't observe W_i,so we just run the following model

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

  where $v_i = \gamma W_i + u_i$

Omitted Variable Bias(OVB): Review(in Lec4)

$$
\begin{aligned}
plim\hat{\beta}_1 &= \frac{Cov(X_i, Y_i)}{VarX_i} \\
&= \frac{Cov(X_i, (\beta_0 + \beta_1 X_i + v_i))}{VarX_i} \\
&= \frac{Cov(X_i, (\beta_0 + \beta_1 X_i + \gamma W_i + u_i))}{VarX_i} \\
&= \frac{Cov(X_i, \beta_0) + \beta_1 Cov(X_i, X_i) + \gamma Cov(X_i, W_i) + Cov(X_i, u_i)}{VarX_i} \\
&= \beta_1 + \gamma \frac{Cov(X_i, W_i)}{VarX_i}
\end{aligned}
$$

Omitted Variable Bias(OVB): violation of consistency

- we have

$$plim\hat{\beta}_1 = \beta_1 + \gamma\frac{Cov(X_i, W_i)}{VarX_i}$$

Omitted Variable Bias(OVB): violation of consistency

- we have

$$plim\hat{\beta}_1 = \beta_1 + \gamma\frac{Cov(X_i, W_i)}{VarX_i}$$

- An omitted variable W_i leads to an inconsistent OLS estimate of the causal effect of X_i if both

Omitted Variable Bias(OVB): violation of consistency

- we have

$$plim\hat{\beta}_1 = \beta_1 + \gamma\frac{Cov(X_i, W_i)}{VarX_i}$$

- An omitted variable W_i leads to an inconsistent OLS estimate of the causal effect of X_i if both
  - W_i is related to X, thus $Cov(X_i, W_i) \neq 0$

Omitted Variable Bias(OVB): violation of consistency

- we have

$$plim\hat{\beta}_1 = \beta_1 + \gamma\frac{Cov(X_i, W_i)}{VarX_i}$$

- An omitted variable W_i leads to an inconsistent OLS estimate of the causal effect of X_i if both
  - W_i is related to X, thus $Cov(X_i, W_i) \neq 0$
  - W_i has some effect on Y_i, thus $\gamma \neq 0$

Omitted Variable Bias(OVB): violation of consistency

- we have

$$plim\hat{\beta}_1 = \beta_1 + \gamma\frac{Cov(X_i, W_i)}{VarX_i}$$

- An omitted variable W_i leads to an inconsistent OLS estimate of the causal effect of X_i if both
  - W_i is related to X, thus $Cov(X_i, W_i) \neq 0$
  - W_i has some effect on Y_i, thus $\gamma \neq 0$
- the OLS regression is not internally valid

Omitted Variable Bias(OVB): violation of consistency

- we have

$$plim\hat{\beta}_1 = \beta_1 + \gamma \frac{Cov(X_i, W_i)}{VarX_i}$$

- An omitted variable W_i leads to an inconsistent OLS estimate of the causal effect of X_i if both
  - W_i is related to X, thus $Cov(X_i, W_i) \neq 0$
  - W_i has some effect on Y_i, thus $\gamma \neq 0$
- the OLS regression is not internally valid
- The OLS estimator does not provide a unbiased and consistent estimate of the causal effect of X_{1i}.

Omitted Variable Bias: Should I Include More Variables in My Regression?

- Include more variables in regression, eliminate the omitted variable bias in higher possibility. But the variance can increase more.

Omitted Variable Bias: Should I Include More Variables in My Regression?

- Include more variables in regression, eliminate the omitted variable bias in higher possibility. But the variance can increase more.
- guidelines to decide whether to include an additional variable:

Omitted Variable Bias: Should I Include More Variables in My Regression?

- Include more variables in regression, eliminate the omitted variable bias in higher possibility. But the variance can increase more.
- guidelines to decide whether to include an additional variable:
  1. Be specific about the coefficient or coefficients of interest.

Omitted Variable Bias: Should I Include More Variables in My Regression?

- Include more variables in regression, eliminate the omitted variable bias in higher possibility. But the variance can increase more.
- guidelines to decide whether to include an additional variable:
  1. Be specific about the coefficient or coefficients of interest.
  2. Use a priori reasoning to identify the most important potential sources of omitted variable bias, leading to a base specification and some "questionable"variables.

Omitted Variable Bias: Should I Include More Variables in My Regression?

- Include more variables in regression, eliminate the omitted variable bias in higher possibility. But the variance can increase more.
- guidelines to decide whether to include an additional variable:
    1. Be specific about the coefficient or coefficients of interest.
    2. Use a priori reasoning to identify the most important potential sources of omitted variable bias, leading to a base specification and some "questionable"variables.
    3. Test whether additional "questionable" control variables have nonzero coefficients.

Omitted Variable Bias: Should I Include More Variables in My Regression?

- Include more variables in regression, eliminate the omitted variable bias in higher possibility. But the variance can increase more.
- guidelines to decide whether to include an additional variable:
  1. Be specific about the coefficient or coefficients of interest.
  2. Use a priori reasoning to identify the most important potential sources of omitted variable bias, leading to a base specification and some “questionable”variables.
  3. Test whether additional “questionable” control variables have nonzero coefficients.
  4. Provide “full disclosure” representative tabulations of your results so that others can see the effect of including the questionable variables on thecoefficient(s) of interest. Do your results change if you include a questionable control variable?

Solutions to OBV when adequate control variables are not available.

- use data in which the same observational unit is observed at different points in time(Panel Data).

Solutions to OBV when adequate control variables are not available.

- use data in which the same observational unit is observed at different points in time(Panel Data).
- use instrumental variables regression(IV) and other quasi-experimental methods

Solutions to OBV when adequate control variables are not available.

- use data in which the same observational unit is observed at different points in time(Panel Data).
- use instrumental variables regression(IV) and other quasi-experimental methods
- use randomized controlled experiment(RCT)

Functional form misspecification(in the last lecture)

- Functional form misspecification makes the OLS estimator biased and inconsistent.

Functional form misspecification(in the last lecture)

- Functional form misspecification makes the OLS estimator biased and inconsistent.
- It can be seen as an special case of OVB,in which the omitted variables are the terms that reflect the missing nonlinear aspects of the regression function.

Functional form misspecification(in the last lecture)

- Functional form misspecification makes the OLS estimator biased and inconsistent.
- It can be seen as an special case of OVB,in which the omitted variables are the terms that reflect the missing nonlinear aspects of the regression function.
- It often can be detected by plotting the data and the estimated regression function, and it can be corrected by using a different functional form.

Measurement error

- An variable is measured imprecisely,then it might make OLS estimator biased

Measurement error

- An variable is measured imprecisely,then it might make OLS estimator biased
- **errors-in-variables bias**: This bias persists even in very large samples, so the OLS estimator is inconsistent if there is measurement error.

Measurement error

- An variable is measured imprecisely,then it might make OLS estimator biased
- **errors-in-variables bias**: This bias persists even in very large samples, so the OLS estimator is inconsistent if there is measurement error.
- for example: recall last year's earnings

Measurement error

There are different types of measurement error

1. Measurement error in the dependent variable Y

Measurement error

There are different types of measurement error

1. Measurement error in the dependent variable Y
    - Less problematic than measurement error in X

Measurement error

There are different types of measurement error

1. Measurement error in the dependent variable Y
    - Less problematic than measurement error in X
    - Usually not a violation of internal validity

Measurement error

There are different types of measurement error

1. Measurement error in the dependent variable Y
    - Less problematic than measurement error in X
    - Usually not a violation of internal validity
    - But leads to less precise estimates

Measurement error

There are different types of measurement error

1. Measurement error in the dependent variable Y
    - Less problematic than measurement error in X
    - Usually not a violation of internal validity
    - But leads to less precise estimates
2. Measurement error in the independent variable X(errors-in-variables bias)

Measurement error

There are different types of measurement error

1. Measurement error in the dependent variable Y
    - Less problematic than measurement error in X
    - Usually not a violation of internal validity
    - But leads to less precise estimates
2. Measurement error in the independent variable X(errors-in-variables bias)
    - Classical measurement error

Measurement error

There are different types of measurement error

1. Measurement error in the dependent variable Y
    - Less problematic than measurement error in X
    - Usually not a violation of internal validity
    - But leads to less precise estimates
2. Measurement error in the independent variable X(errors-in-variables bias)
    - Classical measurement error
    - Measurement error correlated with X

Measurement error

There are different types of measurement error

1. Measurement error in the dependent variable Y
   - Less problematic than measurement error in X
   - Usually not a violation of internal validity
   - But leads to less precise estimates
2. Measurement error in the independent variable X(errors-in-variables bias)
   - Classical measurement error
   - Measurement error correlated with X
   - Both types of measurement error in X are a violation of internal validity

Measurement error in X: classical measurement error

- Suppose we have the following population regression model

$$Y_i = \beta_0 + \beta_1 X_{1,i} + u_i \quad with \quad E[u_i|X_{1i}] = 0$$

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 \widetilde{X}_{1i} + \beta_1 (X_{1i} - \widetilde{X}_{1i}) + u_i \\ &= \beta_0 + \beta_1 \widetilde{X}_{1i} - \beta_1 \omega_i + u_i \end{aligned}$$

Measurement error in X: classical measurement error

- Suppose we have the following population regression model

$$Y_i = \beta_0 + \beta_1 X_{1,i} + u_i \quad with \quad E[u_i|X_{1i}] = 0$$

- Suppose that we do not observe X_{1i} but we observe $\widetilde{X}_{1i}$ a noisy measure ofX_{1i}

$$\widetilde{X}_{1i} = X_{1i} + \omega_i$$

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 \widetilde{X}_{1i} + \beta_1 (X_{1i} - \widetilde{X}_{1i}) + u_i \\ &= \beta_0 + \beta_1 \widetilde{X}_{1i} - \beta_1 \omega_i + u_i \end{aligned}$$

Measurement error in X: classical measurement error

- Suppose we have the following population regression model

$$Y_i = \beta_0 + \beta_1 X_{1,i} + u_i \quad with \quad E[u_i|X_{1i}] = 0$$

- Suppose that we do not observe X_{1i} but we observe $\widetilde{X}_{1i}$ a noisy measure ofX_{1i}

$$\widetilde{X}_{1i} = X_{1i} + \omega_i$$

- Adding and subtracting $\beta_1 \widetilde{X}_{1i}$ gives

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 \widetilde{X}_{1i} + \beta_1 (X_{1i} - \widetilde{X}_{1i}) + u_i \\ &= \beta_0 + \beta_1 \widetilde{X}_{1i} - \beta_1 \omega_i + u_i \end{aligned}$$

Measurement error in X: classical measurement error

- Suppose we have the following population regression model

$$Y_i = \beta_0 + \beta_1 X_{1,i} + u_i \quad with \quad E[u_i|X_{1i}] = 0$$

- Suppose that we do not observe X_{1i} but we observe $\widetilde{X}_{1i}$ a noisy measure ofX_{1i}

$$\widetilde{X}_{1i} = X_{1i} + \omega_i$$

- Adding and subtracting $\beta_1 \widetilde{X}_{1i}$ gives

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 \widetilde{X}_{1i} + \beta_1 (X_{1i} - \widetilde{X}_{1i}) + u_i \\ &= \beta_0 + \beta_1 \widetilde{X}_{1i} - \beta_1 \omega_i + u_i \end{aligned}$$

- Classical measurement error

Measurement error in X: classical measurement error

- Suppose we have the following population regression model

$$Y_i = \beta_0 + \beta_1 X_{1,i} + u_i \quad with \quad E[u_i|X_{1i}] = 0$$

- Suppose that we do not observe X_{1i} but we observe $\widetilde{X}_{1i}$ a noisy measure ofX_{1i}

$$\widetilde{X}_{1i} = X_{1i} + \omega_i$$

- Adding and subtracting $\beta_1 \widetilde{X}_{1i}$ gives

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 \widetilde{X}_{1i} + \beta_1 (X_{1i} - \widetilde{X}_{1i}) + u_i \\ &= \beta_0 + \beta_1 \widetilde{X}_{1i} - \beta_1 \omega_i + u_i \end{aligned}$$

- Classical measurement error

Measurement error in X: classical measurement error

- Suppose we estimate the following regression model

$$Y_i = \beta_0 + \beta_1 \widetilde{X}_{1i} + e_i \quad with e_i = -\beta_1 w_i + u_i$$

$$plim(\hat{\beta}_1) = \beta_1 + \frac{Cov(\widetilde{X}_{1i}, e_i)}{Var(\widetilde{X}_{1i})}$$

$$plim(\hat{\beta}_1) = \beta_1 + \frac{Cov(X_{1i} + w_i, -\beta_1 w_i + u_i)}{Var(X_{1i} + w_i)}$$

Measurement error in X: classical measurement error

- Suppose we estimate the following regression model

$$Y_i = \beta_0 + \beta_1 \widetilde{X}_{1i} + e_i \quad with e_i = -\beta_1 w_i + u_i$$

- With classical measurement error the OLS estimate of β_i is inconsistent

$$plim(\hat{\beta}_1) = \beta_1 + \frac{Cov(\widetilde{X}_{1i}, e_i)}{Var(\widetilde{X}_{1i})}$$

$$plim(\hat{\beta}_1) = \beta_1 + \frac{Cov(X_{1i} + w_i, -\beta_1 w_i + u_i)}{Var(X_{1i} + w_i)}$$

Measurement error in X: classical measurement error

- Suppose we estimate the following regression model

$$Y_i = \beta_0 + \beta_1 \widetilde{X}_{1i} + e_i \quad with e_i = -\beta_1 w_i + u_i$$

- With classical measurement error the OLS estimate of β_i is inconsistent

$$plim(\hat{\beta}_1) = \beta_1 + \frac{Cov(\widetilde{X}_{1i}, e_i)}{Var(\widetilde{X}_{1i})}$$

- Substituting $\widetilde{X}_{1i} = X_{1i} + \omega_i$ and $e_i = -\beta_1 w_i + u_i$ gives

$$plim(\hat{\beta}_1) = \beta_1 + \frac{Cov(X_{1i} + w_i, -\beta_1 w_i + u_i)}{Var(X_{1i} + w_i)}$$

Measurement error in X: classical measurement error

- we have

$$plim(\hat{\beta}_1) = \beta_1 + \frac{Cov(X_{1i} + w_i, -\beta_1 w_i + u_i)}{Var(X_{1i} + w_i)}$$

Measurement error in X: classical measurement error

- we have

$$plim(\hat{\beta}_1) = \beta_1 + \frac{Cov(X_{1i} + w_i, -\beta_1 w_i + u_i)}{Var(X_{1i} + w_i)}$$

- Because

$$Cov(X_{1i}, w_i) = Cov(\omega_i, u_i) = Cov(X_{1i}, u_i) = 0$$

Measurement error in X: classical measurement error

- we have

$$plim(\hat{\beta}_1) = \beta_1 + \frac{Cov(X_{1i} + w_i, -\beta_1 w_i + u_i)}{Var(X_{1i} + w_i)}$$

- Because

$$Cov(X_{1i}, w_i) = Cov(\omega_i, u_i) = Cov(X_{1i}, u_i) = 0$$

- With classical measurement error β_1 is biased towards 0, which is called **attenuation bias**

$$\begin{aligned} plim(\hat{\beta}_1) &= \beta_1 - \frac{Cov(w_i, w_i)}{Var(X_{1i}) + Var(w_i)} \\ &= \beta_1 \left(1 - \frac{Var(w_i)}{Var(X_{1i}) + Var(w_i)}\right) \\ &= \beta_1 \left(\frac{Var(X_{1i})}{Var(X_{1i}) + Var(w_i)}\right) = \beta_1 \frac{\sigma^2_{X_{1i}}}{\sigma^2_{X_{1i}} + \sigma^2_w} \end{aligned}$$

Measurement error in the dependent variable Y

- Measurement error in Y is generally less problematic than measurement error in X

$$\widetilde{Y}_i = \beta_0 + \beta_1 X_i + e_i$$

where $e_i = u_i + \omega_i$

Measurement error in the dependent variable Y

- Measurement error in Y is generally less problematic than measurement error in X
- Suppose Y is measured with classical error: $\widetilde{Y}_i = Y_i + \omega_i$

$$\widetilde{Y}_i = \beta_0 + \beta_1 X_i + e_i$$

where $e_i = u_i + \omega_i$

Measurement error in the dependent variable Y

- Measurement error in Y is generally less problematic than measurement error in X
- Suppose Y is measured with classical error: $\widetilde{Y}_i = Y_i + \omega_i$
- and we estimate

$$\widetilde{Y}_i = \beta_0 + \beta_1 X_i + e_i$$

where $e_i = u_i + \omega_i$

Measurement error in the dependent variable Y

- Measurement error in Y is generally less problematic than measurement error in X
- Suppose Y is measured with classical error: $\widetilde{Y}_i = Y_i + \omega_i$
- and we estimate

$$\widetilde{Y}_i = \beta_0 + \beta_1 X_i + e_i$$

where $e_i = u_i + \omega_i$

- The OLS estimate $\hat{\beta}_1$ will be unbiased and consistent because $E[e_i|X_i] = 0$

Measurement error in the dependent variable Y

- Measurement error in Y is generally less problematic than measurement error in X
- Suppose Y is measured with classical error: $\widetilde{Y}_i = Y_i + \omega_i$
- and we estimate

$$\widetilde{Y}_i = \beta_0 + \beta_1 X_i + e_i$$

where $e_i = u_i + \omega_i$

- The OLS estimate $\hat{\beta}_1$ will be unbiased and consistent because $E[e_i|X_i] = 0$
- Nevertheless,the OLS estimate will be less precise because

$$Var(e_i) > Var(u_i)$$

Solutions to errors-in-variables bias

-

Solutions to errors-in-variables bias

-
- The best way to solve the errors-in-variables problem is to get an accurate measure of X.(Say nothing useful)

Solutions to errors-in-variables bias

-
- The best way to solve the errors-in-variables problem is to get an accurate measure of X.(Say nothing useful)
- instrumental variables regression

Solutions to errors-in-variables bias

-
- The best way to solve the errors-in-variables problem is to get an accurate measure of X.(Say nothing useful)
- instrumental variables regression
  - It relies on having another variable (the “instrumental” variable) that is correlated with the actual value Xi but is uncorrelated with the measurement error.

Simultaneous Causality

- So far we assumed that X affects Y, but what if Y also affects X?

Simultaneous Causality

- So far we assumed that X affects Y, but what if Y also affects X?
  - thus we have $Y_i = \beta_0 + \beta_1 X_1 + u_i$

Simultaneous Causality

- So far we assumed that X affects Y, but what if Y also affects X?
  - thus we have $Y_i = \beta_0 + \beta_1 X_1 + u_i$
  - we also have $X_i = \gamma_0 + \gamma_1 Y_1 + u_i$

Simultaneous Causality

- So far we assumed that X affects Y, but what if Y also affects X?
  - thus we have $Y_i = \beta_0 + \beta_1 X_1 + u_i$
  - we also have $X_i = \gamma_0 + \gamma_1 Y_1 + u_i$
- we assume that $Cov(v_i, u_i) = 0$, then

$$
\begin{aligned}
Cov(X_i, u_i) &= Cov(\gamma_0 + \gamma_1 Y_1 + u_i, u_i) \\
&= Cov(\gamma_1 Y_i, u_i) \\
&= Cov(\gamma_1(\beta_0 + \beta_1 X_1 + u_i), u_i) \\
&= \gamma_1 \beta_1 Cov(X_i, u_i) + \gamma_1 Var(u_i)
\end{aligned}
$$

Simultaneous Causality

- So far we assumed that X affects Y, but what if Y also affects X?
  - thus we have $Y_i = \beta_0 + \beta_1 X_1 + u_i$
  - we also have $X_i = \gamma_0 + \gamma_1 Y_1 + u_i$
- we assume that $Cov(v_i, u_i) = 0$, then

$$\begin{aligned} Cov(X_i, u_i) &= Cov(\gamma_0 + \gamma_1 Y_1 + u_i, u_i) \\ &= Cov(\gamma_1 Y_i, u_i) \\ &= Cov(\gamma_1(\beta_0 + \beta_1 X_1 + u_i), u_i) \\ &= \gamma_1 \beta_1 Cov(X_i, u_i) + \gamma_1 Var(u_i) \end{aligned}$$

- Simultaneous causality leads to biased & inconsistent OLS estimate.

$$Cov(X_i, u_i) = \frac{\gamma_1}{1 - \gamma_1 \beta_1} Var(u_i)$$

Simultaneous causality bias

- Substituting $Cov(X_i, u_i)$ in the formula for the $\hat{\beta}_1$

$$
\begin{aligned}
plim\hat{\beta}_1 &= \beta_1 + \frac{Cov(X_i, u_i)}{Var(X_{1i})} \\
&= \beta_1 + \frac{\gamma_1 Var(u_i)}{1 - \gamma_1 \beta_1 Var(X_{1i})} \neq \beta_1
\end{aligned}
$$

Simultaneous causality bias

- Substituting $Cov(X_i, u_i)$ in the formula for the $\hat{\beta}_1$

$$
\begin{aligned}
plim\hat{\beta}_1 &= \beta_1 + \frac{Cov(X_i, u_i)}{Var(X_{1i})} \\
&= \beta_1 + \frac{\gamma_1 Var(u_i)}{1 - \gamma_1 \beta_1 Var(X_{1i})} \neq \beta_1
\end{aligned}
$$

- Class size and test score: Simultaneous causality is more likely a threat to internal validity

Solutions to simultaneous causality bias

- instrumental variables regression

Solutions to simultaneous causality bias

- instrumental variables regression
- and other experimental designs

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.

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 3 types of missing data

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 3 types of missing data

1. Data are missing at random: this will not impose a threat to internal validity.

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 3 types of missing data

1. Data are missing at random: this will not impose a threat to internal validity.
    - the effect is to reduce the sample size but not introduce bias.

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 3 types of missing data

1. Data are missing at random: this will not impose a threat to internal validity.
    - the effect is to reduce the sample size but not introduce bias.
2. Data are missing based on X: This will not impose a threat to internal validity.

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 3 types of missing data

1. Data are missing at random: this will not impose a threat to internal validity.
    - the effect is to reduce the sample size but not introduce bias.
2. Data are missing based on X: This will not impose a threat to internal validity.
    - suppose that we used only the districts in which the student–teacher ratio exceeds 20. Although we are not able to draw conclusions about what happens when $STR \leq 20$, this would not introduce bias into our analysis of the class size effect for districts with $STR \geq 20$

Missing Data and Sample Selection

3. Data are missing because of a selection process that is related to 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: **Sample Selection Bias**

Missing Data and Sample Selection

3. Data are missing because of a selection process that is related to 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: **Sample Selection Bias**

- the sample selection method (randomly selecting phone numbers of automobile owners) was related to the dependent variable (who the individual supported for president in 1936), because in 1936 car owners with phones were more likely to be Republicans.

Missing Data and Sample Selection

3. Data are missing because of a selection process that is related to 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: **Sample Selection Bias**

- the sample selection method (randomly selecting phone numbers of automobile owners) was related to the dependent variable (who the individual supported for president in 1936), because in 1936 car owners with phones were more likely to be Republicans.
- The mechanism by which the data are missing is related to the dependent variable, leading to sample selection bias.

Missing Data and Sample Selection

3. Data are missing because of a selection process that is related to 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: **Sample Selection Bias**

- the sample selection method (randomly selecting phone numbers of automobile owners) was related to the dependent variable (who the individual supported for president in 1936), because in 1936 car owners with phones were more likely to be Republicans.
- The mechanism by which the data are missing is related to the dependent variable, leading to sample selection bias.
- Solutions to selection bias.

Missing Data and Sample Selection

3. Data are missing because of a selection process that is related to 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: **Sample Selection Bias**

- the sample selection method (randomly selecting phone numbers of automobile owners) was related to the dependent variable (who the individual supported for president in 1936), because in 1936 car owners with phones were more likely to be Republicans.
- The mechanism by which the data are missing is related to the dependent variable, leading to sample selection bias.
- Solutions to selection bias.
    - Heckman Selection Model

Missing Data and Sample Selection

3. Data are missing because of a selection process that is related to 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: **Sample Selection Bias**

- the sample selection method (randomly selecting phone numbers of automobile owners) was related to the dependent variable (who the individual supported for president in 1936), because in 1936 car owners with phones were more likely to be Republicans.
- The mechanism by which the data are missing is related to the dependent variable, leading to sample selection bias.
- Solutions to selection bias.
  - Heckman Selection Model
  - Control function model

Sources of Inconsistency of OLS Standard Errors

- a different threat to internal validity. Even if the OLS estimator is consistent and the sample is large, inconsistent standard errors will let you make a bad judgement about the effect of the interest.

Sources of Inconsistency of OLS Standard Errors

- a different threat to internal validity. Even if the OLS estimator is consistent and the sample is large, inconsistent standard errors will let you make a bad judgement about the effect of the interest.
- There are two main reasons for inconsistent standard errors:

Sources of Inconsistency of OLS Standard Errors

- a different threat to internal validity. Even if the OLS estimator is consistent and the sample is large, inconsistent standard errors will let you make a bad judgement about the effect of the interest.
- There are two main reasons for inconsistent standard errors:

1. Heteroskedasticity: The solution to this problem is to use *heteroskedasticity-robust standard errors* and to construct F-statistics using a heteroskedasticity-robust variance estimator.

Sources of Inconsistency of OLS Standard Errors

- a different threat to internal validity. Even if the OLS estimator is consistent and the sample is large, inconsistent standard errors will let you make a bad judgement about the effect of the interest.
- There are two main reasons for inconsistent standard errors:

1. Heteroskedasticity: The solution to this problem is to use *heteroskedasticity-robust standard errors* and to construct F-statistics using a heteroskedasticity-robust variance estimator.
2. Correlation of the error term across observations.

Sources of Inconsistency of OLS Standard Errors

- a different threat to internal validity. Even if the OLS estimator is consistent and the sample is large, inconsistent standard errors will let you make a bad judgement about the effect of the interest.
- There are two main reasons for inconsistent standard errors:

1. Heteroskedasticity: The solution to this problem is to use *heteroskedasticity-robust standard errors* and to construct F-statistics using a heteroskedasticity-robust variance estimator.
2. Correlation of the error term across observations.
    - This will not happen if the data are obtained by sampling at random from the population.

Sources of Inconsistency of OLS Standard Errors

- a different threat to internal validity. Even if the OLS estimator is consistent and the sample is large, inconsistent standard errors will let you make a bad judgement about the effect of the interest.
- There are two main reasons for inconsistent standard errors:

1. Heteroskedasticity: The solution to this problem is to use *heteroskedasticity-robust standard errors* and to construct F-statistics using a heteroskedasticity-robust variance estimator.
2. Correlation of the error term across observations.
    - This will not happen if the data are obtained by sampling at random from the population.
    - Sometimes, however, sampling is only partially random.

Sources of Inconsistency of OLS Standard Errors

- a different threat to internal validity. Even if the OLS estimator is consistent and the sample is large, inconsistent standard errors will let you make a bad judgement about the effect of the interest.
- There are two main reasons for inconsistent standard errors:

1. Heteroskedasticity: The solution to this problem is to use *heteroskedasticity-robust standard errors* and to construct F-statistics using a heteroskedasticity-robust variance estimator.
2. Correlation of the error term across observations.
    - This will not happen if the data are obtained by sampling at random from the population.
    - Sometimes, however, sampling is only partially random.
    - when the data are repeated observations on the same entity over time, the omitted variables that constitute the regression error may be persistent (like district demographics), then "serial" correlation is induced in the regression error over time.

Sources of Inconsistency of OLS Standard Errors

- a different threat to internal validity. Even if the OLS estimator is consistent and the sample is large, inconsistent standard errors will let you make a bad judgement about the effect of the interest.
- There are two main reasons for inconsistent standard errors:

1. Heteroskedasticity: The solution to this problem is to use *heteroskedasticity-robust standard errors* and to construct F-statistics using a heteroskedasticity-robust variance estimator.
2. Correlation of the error term across observations.
   - This will not happen if the data are obtained by sampling at random from the population.
   - Sometimes, however, sampling is only partially random.
   - when the data are repeated observations on the same entity over time, the omitted variables that constitute the regression error may be persistent (like district demographics), then "serial" correlation is induced in the regression error over time.
   - Another situation in which the error term can be correlated across

Wrap Up

- There are five primary threats to the internal validity of a multiple regression study:

Wrap Up

- There are five primary threats to the internal validity of a multiple regression study:

1. Omitted variables

Wrap Up

- There are five primary threats to the internal validity of a multiple regression study:

1. Omitted variables
2. Functional form misspecification

Wrap Up

- There are five primary threats to the internal validity of a multiple regression study:

1. Omitted variables
2. Functional form misspecification
3. Errors in variables (measurement error in the regressors)

Wrap Up

- There are five primary threats to the internal validity of a multiple regression study:

1. Omitted variables
2. Functional form misspecification
3. Errors in variables (measurement error in the regressors)
4. Sample selection

Wrap Up

- There are five primary threats to the internal validity of a multiple regression study:

1. Omitted variables
2. Functional form misspecification
3. Errors in variables (measurement error in the regressors)
4. Sample selection
5. Simultaneous causality

Wrap Up

- There are five primary threats to the internal validity of a multiple regression study:

1. Omitted variables
2. Functional form misspecification
3. Errors in variables (measurement error in the regressors)
4. Sample selection
5. Simultaneous causality

- Each of these, if present, results in failure of the first least squares assumption,which in turn means that the OLS estimator is biased and inconsistent.

Wrap Up

- There are five primary threats to the internal validity of a multiple regression study:

1. Omitted variables
2. Functional form misspecification
3. Errors in variables (measurement error in the regressors)
4. Sample selection
5. Simultaneous causality

- Each of these, if present, results in failure of the first least squares assumption,which in turn means that the OLS estimator is biased and inconsistent.
- Incorrect calculation of the standard errors also poses a threat to internal validity.

Wrap Up

- There are five primary threats to the internal validity of a multiple regression study:

1. Omitted variables
2. Functional form misspecification
3. Errors in variables (measurement error in the regressors)
4. Sample selection
5. Simultaneous causality

- Each of these, if present, results in failure of the first least squares assumption,which in turn means that the OLS estimator is biased and inconsistent.
- Incorrect calculation of the standard errors also poses a threat to internal validity.
- Applying this list of threats to a multiple regression study provides a systematic way to assess the internal validity of that study.

External validity

Defination

- Suppose we estimate a regression model that is internally valid.

Defination

- Suppose we estimate a regression model that is internally valid.
- Can the statistical inferences be generalized from the population and setting studied to other populations and settings?

Threats to external validity

1. Differences in populations

Threats to external validity

1. Differences in populations
    - The population from which the sample is drawn might differ from the population of interest

Threats to external validity

1. Differences in populations
   - The population from which the sample is drawn might differ from the population of interest
   - For example, if you estimate the returns to education for *men*, these results might not be informative if you want to know the returns to education for *women*.

Threats to external validity

1. Differences in populations
   - The population from which the sample is drawn might differ from the population of interest
   - For example, if you estimate the returns to education for *men*, these results might not be informative if you want to know the returns to education for *women*.
2. Differences in settings

Threats to external validity

1. Differences in populations
   - The population from which the sample is drawn might differ from the population of interest
   - For example, if you estimate the returns to education for *men*, these results might not be informative if you want to know the returns to education for *women*.
2. Differences in settings
   - The setting studied might differ from the setting of interest due to differences in laws, institutional environment and physical environment.

Threats to external validity

1. Differences in populations
   - The population from which the sample is drawn might differ from the population of interest
   - For example, if you estimate the returns to education for *men*, these results might not be informative if you want to know the returns to education for *women*.
2. Differences in settings
   - The setting studied might differ from the setting of interest due to differences in laws, institutional environment and physical environment.
   - For example, the estimated returns to education using data from the U.S might not be informative for China.

Threats to external validity

1. Differences in populations
   - The population from which the sample is drawn might differ from the population of interest
   - For example, if you estimate the returns to education for *men*, these results might not be informative if you want to know the returns to education for *women*.
2. Differences in settings
   - The setting studied might differ from the setting of interest due to differences in laws, institutional environment and physical environment.
   - For example, the estimated returns to education using data from the U.S might not be informative for China.
   - the educational system is different and different institutions of the labor market.

Application to the case of class size and test score

- This analysis was based on test results for California school districts.

Application to the case of class size and test score

- This analysis was based on test results for California school districts.
- Suppose for the moment that these results are internally valid. To what other populations and settings of interest could this finding be generalized?

Application to the case of class size and test score

- This analysis was based on test results for California school districts.
- Suppose for the moment that these results are internally valid. To what other populations and settings of interest could this finding be generalized?
    - generalize to colleges: it is implausible

Application to the case of class size and test score

- This analysis was based on test results for California school districts.
- Suppose for the moment that these results are internally valid. To what other populations and settings of interest could this finding be generalized?
  - generalize to colleges: it is implausible
  - generalize to other U.S. elementary school districts: it is plausible

Wrap up

- It is not easy to make your studies valid internally.

Wrap up

- It is not easy to make your studies valid internally.
- Even harder when you consider generalize your findings.

Wrap up

- It is not easy to make your studies valid internally.
- Even harder when you consider generalize your findings.
- Then common way to generalize the findings actually is to repeat to make the studies internal valid.

Wrap up

- It is not easy to make your studies valid internally.
- Even harder when you consider generalize your findings.
- Then common way to generalize the findings actually is to repeat to make the studies internal valid.
- Then we make a generalizing conclusions based on a bunch of internal valid studies.

Example: Test Scores and Class Size

External Validity

- Whether the California analysis can be generalized—that is, whether it is externally valid—depends on the population and setting to which the generalization is made.

External Validity

- Whether the California analysis can be generalized—that is, whether it is externally valid—depends on the population and setting to which the generalization is made.
- we consider whether the results can be generalized to other elementary public school districts in the United States.

External Validity

- Whether the California analysis can be generalized—that is, whether it is externally valid—depends on the population and setting to which the generalization is made.
- we consider whether the results can be generalized to other elementary public school districts in the United States.
  - more specifically, 220 public school districts in *Massachusetts* in 1998.

External Validity

- Whether the California analysis can be generalized—that is, whether it is externally valid—depends on the population and setting to which the generalization is made.
- we consider whether the results can be generalized to other elementary public school districts in the United States.
  - more specifically, 220 public school districts in *Massachusetts* in 1998.
  - if we find similar results in the California and Massachusetts, it would be evidence of external validity of the findings in California.

External Validity

- Whether the California analysis can be generalized—that is, whether it is externally valid—depends on the population and setting to which the generalization is made.
- we consider whether the results can be generalized to other elementary public school districts in the United States.
  - more specifically, 220 public school districts in *Massachusetts* in 1998.
  - if we find similar results in the California and Massachusetts, it would be evidence of external validity of the findings in California.
  - Conversely, finding different results in the two states would raise questions about the internal or external validity of at least one of the studies.

Comparison of the California and Massachusetts data.

TABLE 9.1 Summary Statistics for California and Massachusetts Test Score Data Sets

	California		Massachusetts	
	Average	Standard Deviation	Average	Standard Deviation
Test scores	654.1	19.1	709.8	15.1
Student–teacher ratio	19.6	1.9	17.3	2.3
% English learners	15.8%	18.3%	1.1%	2.9%
% Receiving lunch subsidy	44.7%	27.1%	15.3%	15.1%
Average district income ($)	$15,317	$7226	$18,747	$5808
Number of observations	420		220	
Year	1999		1998	

Figure 1: pic

Test scores and class size in MA

Regressor	(1)	(2)	(3)	(4)	(5)	(6)
Student–teacher ratio (STR)	−1.72** (0.50)	−0.69* (0.27)	−0.64* (0.27)	12.4 (14.0)	−1.02** (0.37)	−0.67* (0.27)
STR^2				−0.680 (0.737)		
STR^3				0.011 (0.013)		
% English learners		−0.411 (0.306)	−0.437 (0.303)	−0.434 (0.300)		
% English learners > median? (Binary, $HiEL$)					−12.6 (9.8)	
$HiEL \times STR$					0.80 (0.56)	
% Eligible for free lunch		−0.521** (0.077)	−0.582** (0.097)	−0.587** (0.104)	−0.709** (0.091)	−0.653** (0.72)
District income (logarithm)		16.53** (3.15)				
District income			−3.07 (2.35)	−3.38 (2.49)	−3.87* (2.49)	−3.22 (2.31)
District income^2			0.164 (0.085)	0.174 (0.089)	0.184* (0.090)	0.165 (0.085)
District income^3			−0.0022* (0.0010)	−0.0023* (0.0010)	−0.0023* (0.0010)	−0.0022* (0.0010)
Intercept	739.6** (8.6)	682.4** (11.5)	744.0** (21.3)	665.5** (81.3)	759.9** (23.2)	747.4** (20.3)

Test scores and class size in MA

F-Statistics and p-Values Testing Exclusion of Groups of Variables

	(1)	(2)	(3)	(4)	(5)	(6)
All *STR* variables and interactions $= 0$				2.86 (0.038)	4.01 (0.020)	
$STR^2, STR^3 = 0$				0.45 (0.641)		
$Income^2, Income^3$			7.74 (< 0.001)	7.75 (< 0.001)	5.85 (0.003)	6.55 (0.002)
$HiEL, HiEL \times STR$					1.58 (0.208)	
SER	14.64	8.69	8.61	8.63	8.62	8.64
$\overline{R}^2$	0.063	0.670	0.676	0.675	0.675	0.674

These regressions were estimated using the data on Massachusetts elementary school districts described in Appendix 9.1. Standard errors are given in parentheses under the coefficients, and p-values are given in parentheses under the F-statistics. Individual coefficients are statistically significant at the *5% level or **1% level.

Figure 3: pic

Test scores and average district income in MA & CA

Test scores and class size in MA

TABLE 9.3 Student–Teacher Ratios and Test Scores: Comparing the Estimates from California and Massachusetts

	OLS Estimate $\hat{\beta}_{STR}$	Standard Deviation of Test Scores Across Districts	Estimated Effect of Two Fewer Students per Teacher, In Units of: Points on the Test	Standard Deviations
California				
Linear: Table 9.3(2)	−0.73 (0.26)	19.1	1.46 (0.52)	0.076 (0.027)
Cubic: Table 9.3(7) *Reduce STR from 20 to 18*	—	19.1	2.93 (0.70)	0.153 (0.037)
Cubic: Table 9.3(7) *Reduce STR from 22 to 20*	—	19.1	1.90 (0.69)	0.099 (0.036)
Massachusetts				
Linear: Table 9.2(3)	−0.64 (0.27)	15.1	1.28 (0.54)	0.085 (0.036)

Standard errors are given in parentheses.

Figure 4: pic

Internal Validity

- The similarity of the results for California and Massachusetts does not ensure their internal validity.

Internal Validity

- The similarity of the results for California and Massachusetts does not ensure their internal validity.
- **Omitted variables**: teacher quality or a low student– teacher ratio might have families that are more committed to enhancing their children's learning at home.

Internal Validity

- The similarity of the results for California and Massachusetts does not ensure their internal validity.
- **Omitted variables**: teacher quality or a low student– teacher ratio might have families that are more committed to enhancing their children's learning at home.
- **Functional form**: Although further functional form analysis could be carried out, this suggests that the main findings of these studies are unlikely to be sensitive to using different nonlinear regression specifications

Internal Validity

- The similarity of the results for California and Massachusetts does not ensure their internal validity.
- **Omitted variables**: teacher quality or a low student– teacher ratio might have families that are more committed to enhancing their children's learning at home.
- **Functional form**: Although further functional form analysis could be carried out, this suggests that the main findings of these studies are unlikely to be sensitive to using different nonlinear regression specifications
- **Errors in variables**: The average student–teacher ratio in the district is a broad and potentially inaccurate measure of class size.

Internal Validity

- The similarity of the results for California and Massachusetts does not ensure their internal validity.
- **Omitted variables**: teacher quality or a low student– teacher ratio might have families that are more committed to enhancing their children's learning at home.
- **Functional form**: Although further functional form analysis could be carried out, this suggests that the main findings of these studies are unlikely to be sensitive to using different nonlinear regression specifications
- **Errors in variables**: The average student–teacher ratio in the district is a broad and potentially inaccurate measure of class size.
- because students' mobility, the STR might not accurately represent the actual class sizes, which in turn could lead to the estimated class size effect being biased toward zero.

Internal Validity

- **Selection**: data cover all the public elementary school districts in the state that satisfy minimum size restrictions, so there is no reason to believe that sample selection is a problem here.

Internal Validity

- **Selection**: data cover all the public elementary school districts in the state that satisfy minimum size restrictions, so there is no reason to believe that sample selection is a problem here.
- **Simultaneous causality**: it would arise if the performance on tests affected the student–teacher ratio.

Internal Validity

- **Selection**: data cover all the public elementary school districts in the state that satisfy minimum size restrictions, so there is no reason to believe that sample selection is a problem here.
- **Simultaneous causality**: it would arise if the performance on tests affected the student–teacher ratio.
- **Heteroskedasticity** and correlation of the error term across observations

Internal Validity

- **Selection**: data cover all the public elementary school districts in the state that satisfy minimum size restrictions, so there is no reason to believe that sample selection is a problem here.
- **Simultaneous causality**: it would arise if the performance on tests affected the student–teacher ratio.
- **Heteroskedasticity** and correlation of the error term across observations
  - so heteroskedasticity does not threaten internal validity.

Internal Validity

- **Selection**: data cover all the public elementary school districts in the state that satisfy minimum size restrictions, so there is no reason to believe that sample selection is a problem here.
- **Simultaneous causality**: it would arise if the performance on tests affected the student–teacher ratio.
- **Heteroskedasticity** and correlation of the error term across observations
  - so heteroskedasticity does not threaten internal validity.
  - Correlation of the error term across observations, however, could threaten the consistency of the standard errors because simple random sampling was not used.