Lecture 7: Assessing Regression Studies

Introduction to Econometrics,Spring 2023

Zhaopeng Qu

Business School,Nanjing University

April 12 2023

1 Review of previous lectures

2 Introduction

3 Internal validity

4 External validity

5 Example: Test Scores and Class Size

Review of previous lectures

Multiple OLS Regression

- The OLS regression model is

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

- The OLS estimator for β_j

$$\hat{\beta}_j = \frac{\sum_{i=1}^{n} \tilde{X}_{j,i} Y_i}{\sum_{i=1}^{n} \tilde{X}_{j,i}^2} \; for \; j = 1, 2, .., k$$

Multiple OLS Regression: Assumptions

If the four least squares assumptions in the multiple regression model hold:

- Assumption 1: The conditional distribution of u_i given $X_{1i}, \ldots, X_{ki}$ has mean zero,thus

$$E\left[u_i \middle| X_{1i}, \ldots, X_{ki}\right] = 0$$

- Assumption 2: $(Y_i, X_{1i}, \ldots, X_{ki})$ are i.i.d.
- Assumption 3: Large outliers are unlikely.
- Assumption 4: No perfect multicollinearity.

Then

- The OLS estimators $\hat{\beta}_0, \hat{\beta}_1 \ldots \hat{\beta}_k$ are *unbiased*.
- The OLS estimators $\hat{\beta}_0, \hat{\beta}_1 \ldots \hat{\beta}_k$ are *consistent*.
- The OLS estimators $\hat{\beta}_0, \hat{\beta}_1 \ldots \hat{\beta}_k$ are *normally distributed* in large samples.

Nonlinear Regression Model

1. *Nonlinear in Xs*
    - Polynomials,Logarithms and Interactions
    - The multiple regression framework can be extended to handle regression functions that are nonlinear in one or more X.
    - the difference from a standard multiple OLS regression is *how to explain estimating coefficients*.

Nonlinear Regression Model

2. **Nonlinear in β** or **Nonlinear in Y**
    - Discrete Dependent Variables or Limited Dependent Variables.
    - Linear function in Xs is not a good prediciton function or Y.
    - We need a function which parameters enter nonlinearly, such as logisitic or negative exponential functions.
    - Then the parameters can not obtained by OLS estimation any more but **Maximum Likelihood Estimation**

- Assumptions still have to be held, otherwise both OLS estimators and MLE estimators will be biased and inconsistent.
- We need a systematic way to assess the regression studies which is called **validity**.

Introduction

Definitions

- The concepts of **internal and external validity** provide a general framework for assessing whether a empirical studies answers a specific question of interest rightly and usefully.
    - **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*
    - *the population and setting to which the results are generalized.*

Differences between studied and 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(generalized).
- Example: Class size and test score
  - the population studies: elementary schools in CA
  - the population of interest: middle schools in CA
  - different populations and settings: elementary schools in MA

Warp up

- **Internal validity** is top priority in the causal inference studies.
- **External validity** is the second job only if internal validity can be secured.
- In result, we care about the internal validity over 50 times than the external validity in one studies.

Internal validity

Internal Validity in OLS Regression

- Suppose we are interested in the causal effect of X_1 on Y and we estimate the following multiple regression model

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

- Internal validity has three components:
    1. The estimators of β_1 are **unbiased and consistent**, which is the most important.
    2. Both hypothesis tests and confidence intervals should have the **desired significance level**. (at least 5% significant)
    3. The value of β_1 should be **large enough** to make it sense.

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
  - Significant coefficients or marginal effects
- In an informal way
  - Internal Invalidity = endogeneity in the estimation

Omitted Variable Bias(OVB)

OVB Review

- OLS estimator in Simple OLS

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

- Then

$$plim \hat{\beta}_1 = \frac{Cov(X_{1i}, Y_i)}{Var X_{1i}}$$

- OLS estimator in Multiple OLS

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

- Then the asymptotic estimator of β_j

$$plim \hat{\beta}_j = \frac{Cov(\tilde{X}_{ji}, Y_i)}{Var(\tilde{X}_{ji})}$$

- Where $\tilde{X}_{j,i}$ is is the fitted OLS **residual** of regress $X_{j,i}$ on other regressors,thus

$$X_{j,i} = \hat{\gamma}_0 + \hat{\gamma}_1 X_{1,i} + \hat{\gamma}_2 X_{2,i} + ... + \hat{\gamma}_{j-1} X_{j-1,i} + \hat{\gamma}_{j+1} X_{j+1,i} + \tilde{X}_{j,i}$$

OVB Review

- Suppose we want to estimate the causal effect of X_i on Y_i,which represent STR and Test Score,respectively.
- Besides, W_i is the share of English learners which is **omitted** in the regression.
- Then
  - **True model**:

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

    where $E(u_i | X_i, W_i) = 0$
  - 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 $v_i = \gamma W_i + u_i$

OVB Review

- 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 estimator does not provide a unbiased and consistent estimate of the causal effect of X_i, in other words,the OLS regression is not **internally valid**.

OVB Review

- OVB bias is the most possible bias when we run OLS regression using nonexperimental data.
- OVB bias means that there are some variables which **should** have been included in the regression but actually was not.
- Then the simplest way to overcome OVB:
    - Put omitted the variable into the right side of the regression, which can be denoted as **controlling** method.
- But a very important question but often overlook by many students even experienced researchers:
    - Should we control variables as many as possible to avoid OVB bias?
    - **What kind of varaibles can be as control varaibles?**

Control Variables

- **Irrelevant Variables**: the variables have a **ZERO** partial effect on the dependent variable, thus the coefficient in the population equation is zero.
- **Relevant Variables**: the variables have a **NONZERO** partial effect on the dependent variable.
  - **Non-Omitted Variables**: W is not correlated with X, thus

$$Cov(X_i, W_i) = 0$$

  - **Omitted Variables**: W is correlated with X.

$$Cov(X_i, W_i) \neq 0$$

- **Highly-correlated Variables**: Multicollinearity

Recall: the Standard Error of $\hat{\beta}$

- Our multiple OLS regression model is

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

- Under 4 *basic assumptions*,we can prove the unbiasedness of $\hat{\beta}_j$. Based on the content in multiple OLS and *partitioned regressions*,we have

$$\hat{\boldsymbol{\beta}}_j = \boldsymbol{\beta}_j + \frac{\left(\sum_{i=1}^{n} \tilde{X}_{ij} u_i\right)}{\left(\sum_{i=1}^{n} \tilde{X}_{ij}^2\right)}$$

  - Where $\tilde{X}_{ij}$ is the residual of a regression of X_j on all others X_s

- For simplicity, under the 5th assumption of multiple OLS regression: homoskedastic variance, thus

$$Var(u_i | X_{1i}, X_{2i} ..., X_{ki}) = Var(u_i | \boldsymbol{X}) = Var(u_i) = \sigma_u^2$$

  - where $\boldsymbol{X} = X_{1i} X_{2i} ..., X_{ki}$

Recall: the Standard Error of $\hat{\beta}$

- Then we have

$$
\begin{aligned}
Var(\hat{\beta}_j) &= Var\left(\boldsymbol{\beta}_j + \frac{\left(\sum_{i=1}^{n} \tilde{X}_{ij} u_i\right)}{\left(\sum_{i=1}^{n} \tilde{X}_{ij}^2\right)}\right) \\
&= \frac{\left(\sum_{i=1}^{n} \tilde{X}_{ij}^2 Var(u_i)\right)}{\left(\sum_{i=1}^{n} \tilde{X}_{i1}^2\right)^2} \\
&= \frac{\left(\sum_{i=1}^{n} \tilde{X}_{ij}^2 \sigma_u^2\right)}{\left(\sum_{i=1}^{n} \tilde{X}_{ij}^2\right)^2} \\
&= \frac{\sigma_u^2}{\left(\sum_{i=1}^{n} \tilde{X}_{ij}^2\right)}
\end{aligned}
$$

Recall: the Standard Error of $\hat{\beta}$

- **Do not forget**: The $\tilde{X}_{ij}$ is obtained from a multiple OLS regression model

$$X_{ij} = \hat{\delta}_0 + \hat{\delta}_1 X_{1i} + \hat{\delta}_2 X_{2i} + \ldots + \hat{\delta}_{j-1} X_{j-1,i} + \hat{\delta}_{j+1} X_{j+1,i} + \ldots \hat{\delta}_k X_{ki} + \tilde{X}_{ji}$$

- The **R-Squared** of this regression is

$$R_j^2 = 1 - \frac{SSR_j}{TSS_j}$$
$$\Rightarrow SSR_j = TSS_j \times (1 - R_j^2)$$
$$\Rightarrow \tilde{X}_{ij}^2 = \sum_{i=1}^{n} (X_{ji} - \bar{X}_j)^2 (1 - R_j^2)$$

  - where R_j^2 is the **R-squared** from the regression of X_j on all other Xs.

Recall: the Standard Error of $\hat{\beta}$

- Then under 4 basic assumptions and homoskedastic variance of u_i,the **variance** of the OLS estimators β_j simplify to

$$Var\left(\hat{\beta}_j\right) = \sigma^2_{\hat{\beta}_j} = \frac{\sigma^2_u}{\sum_{i=1}^{n}(X_{ij} - \bar{X})^2(1 - R^2_j)}$$

- Under 3 basic assumptions and homoskedastic variance of u_i, the **variance** of the OLS estimators β_1 simplify to

$$Var\left(\hat{\beta}_1\right) = \sigma^2_{\hat{\beta}_1} = \frac{\sigma^2_u}{\sum(X_i - \bar{X})^2}$$

Irrelevant Variables: Models

- **Irrelevant Variables**: the variables have a ZERO partial effect on the dependent variable, thus the coefficient in the population equation is zero.
- Assume that our model is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i \tag{8.1}$$

  - Where X_1 is the variable of interest or **treatment variable**.
  - X_2 is a **control variable**, which should be balanced or controlled.
  - X_3 is **irrelevant variable**, thus

$$\beta_3 = 0$$

- The model excluding irrelevant variable is

$$Y_i = \tilde{\beta}_0 + \tilde{\beta}_1 X_1 + \tilde{\beta}_2 X_{2i} + v_i \tag{8.2}$$

Irrelevant Variables: Estimate

- Then based on the OVB formula,we have

$$plim\hat{\beta}_1 = \beta_1 + \beta_3 \frac{Cov(\tilde{X}_{12,i}, X_{3i})}{Var\tilde{X}_{12,i}} = \beta_1$$

- the OLS estimator $\hat{\beta}_1$ is still **consistent**.

Irrelevant Variables: Variance

- The variance of $\hat{\beta}_1$ in 8.1 is

$$Var(\hat{\beta}_1) = \frac{\sigma_u^2}{\sum_{i=1}^{n}(X_{1i} - \bar{X})^2(1 - R_{23}^2)} \tag{8.3}$$

Where R_{23}^2 is the R-Squared of the regression of X_1 on X_2 and X_3

- The variance of $\hat{\hat{\beta}}_1$ in 8.2 is

$$Var(\hat{\hat{\beta}}_1) = \frac{\sigma_v^2}{\sum_{i=1}^{n}(X_{1i} - \bar{X})^2(1 - R_2^2)} \tag{8.4}$$

Where R_2^2 is the R-Squared of the regression of X_1 on X_2

Irrelevant Variables: Variance

- Based on 8.1 and 8.2, we have

$$u_i = Y - \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i}$$
$$v_i = Y - \tilde{\beta}_0 + \tilde{\beta}_1 X_{1i} + \tilde{\beta}_2 X_{2i}$$

- Because $\beta_3 = 0$ then $Var(u_i) = Var(v_i) \Rightarrow \sigma_u^2 = \sigma_v^2$
- Because $R_2^2 \leq R_{23}^2$ then we have

$$Var(\hat{\beta}_1) \geq Var(\hat{\tilde{\beta}}_1)$$

- It means controlling an irrelevant variable will only enlarge the variance of the estimator, in other words, make our estimate less precise.

Irrelevant Variables: Wrap up

- The OLS estimator is still unbiased and consistent.
- It increase the variance of estimator, in other words,it will make the estimate **less precise**.
- **Conclusion**: *we should avoid to put irrelevant variables into our regression.*

Relevant Variables: Non-Omitted

- What about *Relevant* but *Non-Omitted* variables? Our regression models is still 8.1 and 8.2, but X_3 now is not an irrelevant variable but a **Non-omitted variable**, thus

$$Cov(X_{1i}, X_{3i}) = 0$$
$$Cov(X_{2i}, X_{3i}) = 0$$

- Then based on the OVB formula,we have

$$plim\hat{\beta}_1 = \beta_1 + \beta_3 \frac{Cov(\tilde{X}_{12,i}, X_{3i})}{Var\tilde{X}_{12,i}} = \beta_1$$

- the OLS estimator $\hat{\beta}_1$ is still **consistent**.

Relevant Variables: Non-Omitted

- Because $Cov(X_{1i}, X_{3i}) = 0$ and $Cov(X_{2i}, X_{3i}) = 0$, then we also have

$$R_2^2 = R_{23}^2$$

- Then the variance of $\hat{\beta}_1$ and $\hat{\hat{\beta}}_1$ are following respectively.

$$Var(\hat{\beta}_1) = \frac{\sigma_u^2}{\sum_{i=1}^{n}(X_{1i} - \bar{X})^2(1 - R_2^2)}$$

$$Var(\hat{\hat{\beta}}_1) = \frac{\sigma_v^2}{\sum_{i=1}^{n}(X_{1i} - \bar{X})^2(1 - R_2^2)}$$

Revevant Variables: Non-Omitted

- Because $\beta_3 \neq 0$ and $Cov(X_{1i}, X_{3i}) = 0$ and $Cov(X_{2i}, X_{3i}) = 0$,then

$$Var(u_i) \leq Var(v_i) \Rightarrow \sigma_u^2 \leq \sigma_v^2$$

- then we have

$$Var(\hat{\beta}_1) \leq Var(\hat{\tilde{\beta}}_1)$$

- It decrease the variance of estimator, in other words,it will make the estimate **more precise**.
- **Conclusion**: *we should always put Relevant but Non-Omitted Variables into our regression.*

Bad Controls v.s Omitted Variable Bias

- It seems that controlling for more covariates always increases the likelihood that regression estimates have a causal interpretation.
  - often true, but not always.
- eg. Some researchers regressing earnings(Y_i) on schooling(S_i) (and experience) include controls for occupation(O_i). Thus our regression model is

$$Y_i = \beta_0 + \beta_1 S_i + \gamma O_i + u_i$$

  where β_1 is the most of interest coefficient.
- Clearly we can also think of schooling(S_i) affecting the access to higher level occupations(O_i),
  - e.g. you need a Ph.D. to become a university professor. thus

$$O_i = \lambda_0 + \lambda_1 S_i + e_i$$

Bad Controls v.s Omitted Variable Bias

- Assume that the true relation is a two equation system: a simultaneous equations system

$$
\begin{aligned}
Y_i &= \beta_0 + \beta_1 S_i + \gamma O_i + e_i \\
O_i &= \lambda_0 + \lambda_1 S_i + u_i
\end{aligned}
$$

- In the case, Occupation O_i is an *endogenous variable*.
- As a result, you could not necessarily estimate the first equation by OLS, which means that the estimation of β_1 is not *unbiased* and *consistent*, because of controlling Occupation(O_i).

Bad Controls: Occupation

- Let us come back to the wage premium of college graduation: the conditional expectation.But now we have additional control variable-*occupations*: *white-color* and *blue-olor*
- Two reasonable assumptions:
    1. white-collar jobs, on average, pay more than blue-collar jobs.
    2. graduating college increases the likelihood of a white-collar job.
- **Question**: Is occupation an omitted variable in the regression of college degree on wage?
- However, should we control for occupation type when considering the effect of college graduation on wages?

Bad Controls: Occupation

- Assume that college degrees are randomly assigned, then we just need to compare the wage difference between workers with college degrees and those without degrees.
- Now we **control** the occupation, which means when we do as follows conditional on occupation:
  - compare degree-earners who chose blue-collar jobs to non-degree-earners who chose blue-collar jobs.
  - or compare degree-earners who chose white-collar jobs to non-degree-earners who chose white-collar jobs.
- Note: the assumption of random degrees says nothing about random job selection.

Bad Controls: Occupation

More formally,

- Y_i denotes *i*'s earnings
- W_i is also a dummy for whether individual *i* has a white-collar job
- D_i a dummy variable, refers to *i*'s college-graduation status which is randomly assigned, which indicates

$$\left\{Y_1, Y_0 \perp D\right\} \ and \ \left\{W_1, W_0 \perp D\right\}$$

- Then

$$\mathrm{Y}_i = \mathrm{D}_i \mathrm{Y}_{1i} + \left(1 - \mathrm{D}_i\right) \mathrm{Y}_{0i}$$

$$\mathrm{W}_i = \mathrm{D}_i \mathrm{W}_{1i} + \left(1 - \mathrm{D}_i\right) \mathrm{W}_{0i}$$

Bad Controls: Occupation

- Because we've assumed D_i is randomly assigned, differences in means yield causal estimates, *i.e.*

$$E\left[Y_i \mid D_i = 1\right] - E\left[Y_i \mid D_i = 0\right] = E\left[Y_{1i} - Y_{0i}\right]$$
$$E\left[W_i \mid D_i = 1\right] - E\left[W_i \mid D_i = 0\right] = E\left[W_{1i} - W_{0i}\right]$$

Bad Controls: Occupation

- What happens when we estimate the wage-effect of college graduation for white-collar jobs by controlling occupations?

$$
\begin{aligned}
&E\left[Y_i \mid W_i = 1,\ D_i = 1\right] - E\left[Y_i \mid W_i = 1,\ D_i = 0\right] \\
&= E\left[Y_{1i} \mid W_{1i} = 1,\ D_i = 1\right] - E\left[Y_{0i} \mid W_{0i} = 1,\ D_i = 0\right] \\
&= E\left[Y_{1i} \mid W_{1i} = 1\right] - E\left[Y_{0i} \mid W_{0i} = 1\right] \\
&= E\left[Y_{1i} \mid W_{1i} = 1\right] - E\left[Y_{0i} \mid W_{1i} = 1\right] + E\left[Y_{0i} \mid W_{1i} = 1\right] - E\left[Y_{0i} \mid W_{0i} = 1\right] \\
&= \underbrace{E\left[Y_{1i} - Y_{0i} \mid W_{1i} = 1\right]}_{\text{ATT on white-collar workers}} + \underbrace{E\left[Y_{0i} \mid W_{1i} = 1\right] - E\left[Y_{0i} \mid W_{0i} = 1\right]}_{\text{Selection bias}}
\end{aligned}
$$

- By introducing a *bad control*, we introduced **selection bias** into a setting that did not have selection bias without controls.

Bad Controls: Occupation

- Specifically,

$$\underbrace{E\left[Y_{1i} - Y_{0i} \mid W_{1i} = 1\right]}_{\text{ATT on white-collar workers}} + \underbrace{E\left[Y_{0i} \mid W_{1i} = 1\right] - E\left[Y_{0i} \mid W_{0i} = 1\right]}_{\text{Selection bias}}$$

- The First term: Expected potential non-college earnings, given that potential white collar status associated with college education is equal to 1.
- If the occupational choice between white-collor and blue-collor is randomly assigned, then

$$E\left[Y_{0i} \mid W_{1i} = 1\right] = E\left[Y_{0i} \mid W_{0i} = 1\right]$$

- It describes how college graduation changes the composition of the pool of white-collar workers, which in turn change the wage premium between college and high school graduates.
- Even if the true wage causal effect is zero, this selection bias need not be zero.

Bad Controls v.s Omitted Variable Bias

- Putting a bunch of "control" variables might actually be a really **bad idea**: when these variables are themselves **outcomes** of the X variable of interest(another Y).
- But if you don't control more variables,you may suffer **Omitted Variable Bias**, which also lead a unbiased and inconsistent estimate.
- How to deal with bad control and omitted variable bias, "one of the hard questions in the social sciences" King(2010).
- Bad controls are variables that are themselves **outcomes** of the treatment variable.
  - In general control variables should be fixed characteristics or pre-determined by the time of treatment.

Wrap up?

- Which variables belong on the right hand side of a regression equation?
  - Relevant and Omitted Variables : variables determining the treatment and correlated with the outcome.
    - in general these variables will be fixed characteristics or pre-determined by the time of treatment.(Not bad controls)
  - Relevant but Non-omitted Variables: Variables uncorrelated with the treatment but correlated with the outcome.
    - these variables may help reducing standard errors.
- Which variables should NOT be included in the right hand side of the equation?
  - Variables which are outcomes of the treatment itself. These are bad controls.
  - Variables are irrelevant.
  - Variables are highly correlated.

Functional form misspecification

Functional form misspecification

- Functional form misspecification also 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 functions, and it can be corrected by using different functional forms.
- It can also use nonparametric or semi-parametric methods to make a robust estimate.
    - Matching and Propensity Scores Matching

Measurement error

Introduction

- When a variable is **measured imprecisely**,then it might make OLS estimator biased.
- 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

Types of Measurement errors

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 the dependent variable Y

- Suppose the true population regression model(Simple OLS) is

$$Y_i = \beta_0 + \beta_1 X_i + u_i \quad with \quad E[u_i|X_i] = 0$$

- Suppose because Y is measured with errors, thus we can not observe Y_i but observe $\tilde{Y}_i$, which is a noisy measure of Y_i,thus

$$\tilde{Y}_i = Y_i + \omega_i$$

- The noisy part of $\tilde{Y}_i$, ω_i, satisfies

$$E[\omega_i|Y_i] = 0$$

- It means that $Cov(\omega_i, Y_i) = 0$ and $Cov(\omega_i, u_i) = 0$,which is a key hypothesis and is called **classical measurement error**
- For example: measurement error due to someone making random mistakes when imputing data in a database.

Measurement error in the dependent variable Y

- And we can only 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 estimate will be less precise because

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

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

Measurement error in X: classical measurement error

- The true model is

$$Y_i = \beta_0 + \beta_1 X_{1i} + u_i$$

  with $E[u_i|X_i] = 0$
- Due to the **classical measurement error**,we only have X_{1i}^* thus $X_{1i}^* = X_{1i} + w_i$,we have to estimate the model is

$$Y_i = \beta_0 + \beta_1 X_{1i}^* + e_i$$

- where $e_i = -\beta_1 w_i + u_i$

Measurement error in X: classical measurement error

- Similar to OVB bias in simple OLS model

$$
\begin{aligned}
plim(\hat{\beta}_1) &= \frac{Cov(Y_i, X^*_{1i})}{Var(X^*_{1i})} \\
&= \frac{Cov\left[\beta_0 + \beta_1 X_{1i} + u_i, (X_{1i} + w_i)\right]}{Var(X_{1i} + w_i)} \\
&= \frac{\beta_1 Cov(X_{1i}, X_{1i})}{Var(X_{1i} + w_i)} \\
&= \beta_1 \left(\frac{Var(X_{1i})}{Var(X_{1i}) + Var(w_i)} \right) \\
&= \beta_1 \frac{\sigma^2_{X_{1i}}}{\sigma^2_{X_{1i}} + \sigma^2_w}
\end{aligned}
$$

Measurement error in X: classical measurement error

- Because

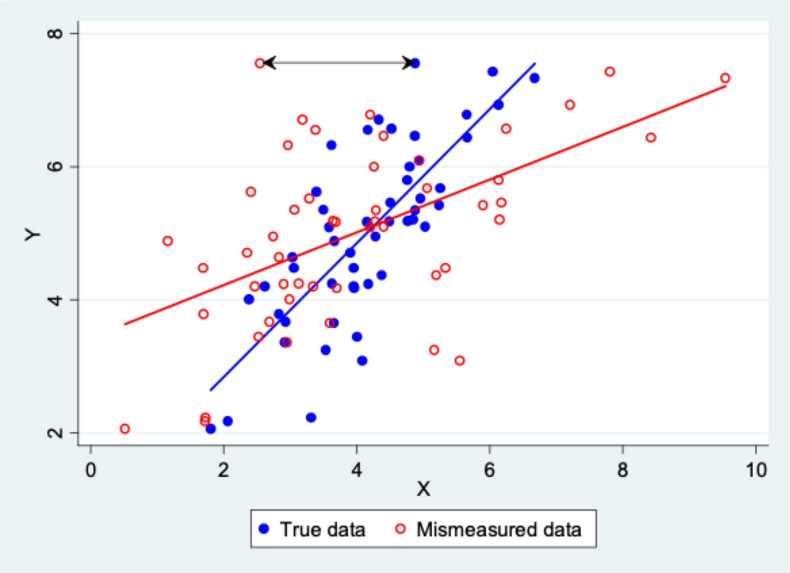
$$0 \leq \frac{\sigma_{X_{1i}}^2}{\sigma_{X_{1i}}^2 + \sigma_w^2} \leq 1$$

- we have

$$plim(\hat{\beta}_1) = \beta_1 \frac{\sigma_{X_{1i}}^2}{\sigma_{X_{1i}}^2 + \sigma_w^2} \leq \beta_1$$

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

Measurement error in X: classical measurement error



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**
  - It relies on having another variable (the “instrumental” variable) that is correlated with the actual value X_i but is uncorrelated with the measurement error. We will discuss it later on.

Simultaneous Causality

Introduction

- 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 $X_i = \gamma_0 + \gamma_1 Y_1 + v_i$
- Assume that $Cov(v_i, u_i) = 0$, then

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

- 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 $\hat{\beta}_1$

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

- OLS estimate is **inconsistent** if simultaneous causality bias exits.

Solutions to simultaneous causality bias

- Instrumental Variables
- and other experimental designs

Missing Data and Sample Selection

Introduction

- 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 $STR \geq 20$

Introduction

3. Data are missing because of a *selection process* that is related to the value of the dependent variable (Y),then this selection process can introduce correlation between the error term and the regressors: **Sample Selection Bias**
    - Eg.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.

- Solutions to sample selection bias:
    - Instrumental Variables or other quasi-experimental methods
    - **Heckman Selection Model(or Heckit Model)**

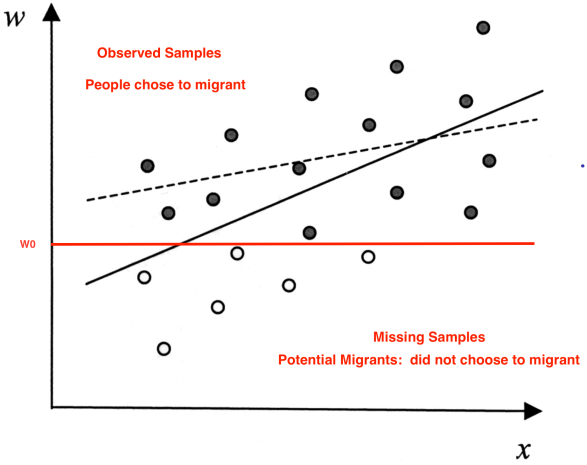
Wage determination of working women

- A Classical Example: wage determination for migrants

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

  - Y_i is logwage
  - X_i is schooling years
- The sample selection problem arises in that the sample consists only of migrants who chose to migrant from other places.
  - If the selection to migration is random,then OK.
  - But in reality, people choose to migrant probably they are *smarter,more ambitious* and more risk-preferent which normally can not observed or measured in the data.

Wage determination of working women



- When y is regressed on X and W: OLS throws away the red area and just uses blue to estimate β.

Heckman Sample Selection Model

- A **two-equation behavioral model**

1. *selection equation*

$$Z_i^* = W_i'\gamma + e_i$$

where Z_i is a latent variable which indicates the propensity of working for a married woman

- and the error term e_i satisfies

$$E[e_i|W_i] = 0$$

- Then Z_i is a dummy variable to represent whether a woman to work or not actually,thus

$$Z_i = \begin{cases} 1 \ if \ Z^* > 0 \\ 0 \ if \ Z^* \leq 0 \end{cases}$$

Heckman Sample Selection Model

2. *outcome equation*

$$Y_i^* = X_i'\beta + u_i$$

- where the outcome(Y_i) can be observed only when Z_i=1 or $Z_i^* > 0$

$$Y_i^* = \begin{cases} Y_i \; if \; Z_i = 1 \\ 0 \; or \; missing \; if \; Z_i = 0 \end{cases}$$

```
- The error term $u_i$ satisfies $E[u_i|X_i]=0$
```

Heckman Sample Selection Model

- The conditional expectation of wages on X_i is

$$E[Y_i^*|X_i] = X_i'\beta$$

- The conditional expectation of wages on X_i is *only for women who work*$(Z^* > 0)$

$$\begin{aligned} E[Y_i^*|X_i, Z_i^* > 0] &= E[Y_i|X_i, Z_i^* > 0] \\ &= E[X_i'\beta + u_i|X_i, Z_i^* > 0] \\ &= X_i'\beta + E[u_i|Z_i^* > 0] \\ &= X_i'\beta + E[u_i|e_i > -W_i'\gamma] \end{aligned}$$

Heckman Sample Selection Model

- If u_i and e_i is independent, then $E[u_i|e_i > -W_i'\gamma] = 0$, then

$$E[Y_i^*|X_i, Z_i^* > 0] = E[Y_i^*|X_i] = X_i'\beta$$

  - It means that only using sample-selected data does not make the estimation of β biased.
- But in reality, unobservables in the two equations, thus u_i and e_i, are likely to be **correlated**
  - eg. innate ability,ambitions,...
- Instead assume that u_i and e_i are **jointly normal distributed**, which means that

$$\begin{pmatrix} u_i \\ e_i \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_u \\ \mu_e \end{pmatrix}, \begin{pmatrix} \sigma_u^2 & \sigma_{eu} \\ \sigma_{ue} & \sigma_e^2 \end{pmatrix}\right) = \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_u^2 & \rho\sigma_u \\ \rho\sigma_u & 1 \end{pmatrix}\right)$$

Math Review: Two Normal Distributed R.V.s

Two Normal Distributed R.V.s

For any two normal variables (n_0, n_1) with zero mean, we can write $n_1 = \alpha_0 n_0 + \eta$, where $\eta \sim N(0, \sigma_\eta)$ and $E(\eta | n_0) = 0$.Then we have

$$\alpha_0 = \frac{Cov(n_0, n_1)}{Var(n_0)}$$

or

$$E(n_1 \mid n_0) = \frac{Cov(n_0, n_1)}{Var(n_0)} n_0$$

Then

$$n_1 = E(n_1 \mid n_0) + \eta = \frac{Cov(n_0, n_1)}{Var(n_0)} n_0 + \eta$$

Heckman Sample Selection Model

- For two normal variables u_i and e_i with zero means, we have

$$\alpha_0 = \frac{Cov(u_i, e_i)}{Var(e_i)} = \frac{\sigma_{ue}}{\sigma_e^2}$$

- Then

$$u_i = \alpha_0 e_i + \eta = \frac{\sigma_{ue}}{\sigma_e^2} e_i + \eta$$

where $\eta \sim N\left(0, \sigma_\eta\right)$ and $E\left(\eta | e_i\right) = 0$

Heckman Sample Selection Model

- Then the conditional expectation of u_i

$$
\begin{aligned}
E[u_i|e_i > -W_i'\gamma] &= E[\frac{\sigma_{ue}}{\sigma_e^2} e_i + \eta | e_i > -W_i'\gamma] \\
&= \frac{\sigma_{ue}}{\sigma_e^2} E[e_i|e_i > -W_i'\gamma] + E[\eta|e_i > -W_i'\gamma] \\
&= \frac{\sigma_{ue}}{\sigma_e^2} E[e_i|e_i > -W_i'\gamma]
\end{aligned}
$$

Math Review: Truncated Density Function

Truncated Density Function

If a continuous random variable X has p.d.f. $f(x)$ and c.d.f. $F(x)$ and a is a constant, then the conditional density function

$$f(x|x > a) = \begin{cases} \frac{f(x)}{1-F(a)} & \textit{if } x > a \\ 0 & \textit{if } x \leq a \end{cases}$$

Math Review: Truncated Density Function

Truncated Density Function

The proof follows from the definition of a conditional probability is

$$Pr(A|B) = \frac{Pr(AB)}{Pr(B)}$$

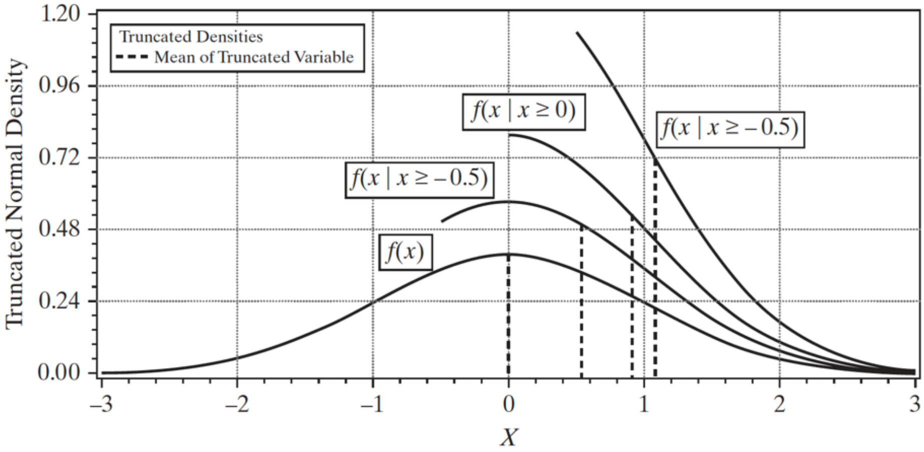
then,

$$\begin{aligned} F(x|X > c) = \frac{Pr(X < x, X > c)}{Pr(X > c)} &= \frac{Pr(c < X < x)}{1 - F(c)} \\ &= \frac{F(x) - F(c)}{1 - F(c)} \end{aligned}$$

then,

$$f(x|x > c) = \frac{d}{dx} F(x|X > c) = \frac{\frac{d}{dx}[F(x)] - 0}{1 - F(c)} = \frac{f(x)}{1 - F(c)}$$

Math Review: Truncated Density Function



- It amounts merely to **scaling** the density so that it integrates to one over the range above a.

Standard Normal Truncated Density Function

- If X is distributed as standard normal, thus $X \sim N\left(0,1\right)$, then the p.d.f and c.d.f are as follow

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$
$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt$$

- And c is a scalar, then we can get the Truncated Density Function of an R.V. distributed in Standard Normal

$$f\left(x \mid x > c\right) = \frac{\phi\left(x\right)}{1 - \Phi\left(c\right)}$$

- The Expectation of in a standard normal truncated p.d.f

$$E(x|x > c) = \frac{f(c)}{1 - \Phi\left(c\right)} \equiv \lambda\left(c\right)$$

where $\lambda\left(c\right)$ is called by Inverse Mills Ratio.

The Expectation in a Standard Normal Truncated

Proof

$$
\begin{aligned}
E(x|x>c) &= \int_c^{+\infty} xf(x|x>c)dx = \int_c^{+\infty} x\frac{\phi(x)}{1-\Phi(c)}dx \\
&= \frac{1}{1-\Phi(c)}\int_c^{+\infty} x\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx \\
&= \frac{1}{1-\Phi(c)}\int_c^{+\infty} \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}d(\frac{x^2}{2}) \\
&= \frac{1}{1-\Phi(c)}\int_{\frac{c^2}{2}}^{+\infty} \frac{1}{\sqrt{2\pi}}e^{-t}d(t) \\
&= \frac{1}{1-\Phi(c)}\times\frac{1}{\sqrt{2\pi}} - e^{-t}\ |_{\frac{c^2}{2}}^{+\infty} \\
&= \frac{1}{1-\Phi(c)}\times\frac{1}{\sqrt{2\pi}}e^{-\frac{c^2}{2}} = \frac{f(c)}{1-\Phi(c)}
\end{aligned}
$$

Heckman Sample Selection Model

- Then the conditional expectation of u_i

$$
\begin{aligned}
E[u_i|e_i > -W_i'\gamma] &= \frac{\sigma_{ue}}{\sigma_e^2}E[e_i|e_i > -W_i'\gamma] \\
&= \frac{\sigma_{ue}}{\sigma_e}E[\frac{e_i}{\sigma_e}|\frac{e_i}{\sigma_e} > \frac{-W_i'\gamma}{\sigma_e}] \\
&= \frac{\sigma_{ue}}{\sigma_e}\frac{\phi(-W_i'\gamma/\sigma_e)}{1-\Phi(-W_i'\gamma/\sigma_e)} \\
&= \frac{\sigma_{ue}}{\sigma_e}\frac{\phi(W_i'\gamma/\sigma_e)}{\Phi(W_i'\gamma/\sigma_e)} \\
&= \sigma_\lambda\lambda(W_i'\gamma)
\end{aligned}
$$

Heckman Sample Selection Model

- Then the conditional expectation of wages on X_i is only for women who work($Z^* > 0$)

$$E[Y_i^* | X_i, Z_i^* > 0] = E[Y_i | X_i, Z_i = 1] = X_i' \beta + \sigma_\lambda \lambda(W_i' \gamma)$$

- It means that if we could include $\lambda(W_i'\gamma)$ as **an additional regressor** into the outcome equation, thus we run

$$Y_i = X_i' \beta + \sigma_\lambda \lambda(W_i' \gamma) + u_i$$

  then we can obtain the **unbiased** and **consistent** estimate β using a self-selected sample.
- The coefficient before $\lambda(\cdot)$ can be testing significance to indicate whether the term should be included in the regression, in other words, *whether the selection should be corrected*.

Heckit Model Estimation: a two-step method

1. Estimate selection equation using **all observations**,thus

$$Z_i = W_i'\gamma + e_I$$

- obtain estimates of parameters $\hat{\gamma}$
- computer the Inverse Mills Ratio(IMR) $\dfrac{\phi(W_i'\hat{\gamma})}{\Phi(W_i'\hat{\gamma})} = \hat{\lambda}(W_i'\hat{\gamma})$

2. Estimate the outcome equation using **only the selected observations**.

$$Y_i = X_i'\beta + \sigma_\lambda \hat{\lambda}(W_i'\hat{\gamma}) + u_i$$

- **Note**: standard error is not right, have to be adjusted because we use $\hat{\lambda}(W_i'\hat{\gamma})$ instead of $\lambda(W_i'\gamma)$ in the estimation.

An Example: Wage Equation for Married Women

TABLE 17.7 Wage Offer Equation for Married Women

Dependent Variable: log(*wage*)		
Independent Variables	**OLS**	**Heckit**
educ	.108 (.014)	.109 (.016)
exper	.042 (.012)	.044 (.016)
$exper^2$	−.00081 (.00039)	−.00086 (.00044)
constant	−.522 (.199)	−.578 (.307)
$\hat{\lambda}$	—	.032 (.134)
Sample size *R*-squared	428 .157	428 .157

Sources of Inconsistency of OLS Standard Errors

Introduction

- 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 judgment 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

2. Correlation of the error term across observations.
    - This will not happen if the data are obtained by sampling at random from the population.(i.i.d)
    - Sometimes, however, sampling is only partially random.
        - When the data are repeated observations on the same entity over time
        - Another situation in which the error term can be correlated across observations is when sampling is based on a geographical unit.(**cluster**)

- Both situation means that the assumptions

$$Cov(u_i, u_j) \neq 0$$

,the second key assumption in OLS is partially violated.

- the OLS estimator is unbiased and consistent, but inconsistent standard errors is not right.

Clustering Standard Error

- Suppose we focus on the topic of class size and student performance, but now the data are collecting on students rather than school district.
- Our regression model is

$$TestScore_{ig} = \beta_0 + \beta_1 ClassSize_g + u_{ig}$$

- $TestScore_{ig}$ is the dependent variable for student i in class g, with G groups.
- $ClassSize_g$ the independent variable, **varies only at the group level**.
- Intuitively,the test score of students in the same class(g) tend to be correlated. Thus

$$Cov[u_{ig}, u_{jg}] = \rho\sigma_u^2$$

where ρ is the intraclass correlation coefficient.

Clustering Standard Error

- Stata: use option `vce(cluster clustvar)`. Where `clustvar` is a variable that identifies the groups in which on observables are allowed to correlate.
- R: the `vcovHC()` function from `plm` package

Magnitude of β_1

Introduction

- The value of β_1 should be **large enough** to make it sense.
    - Question: How large is **large enough**?
- Recall: the explanation of β_1 is **the effect of one unit** X **change on** Y
- However, the scale on which these tests are scored is often arbitrary and not easy to interpret.
- If we are interested in how a particular individual's score compares with the population.
- Thus, instead of asking about the effect on hourly wage if, say, a test score is 10 points higher, it makes more sense to ask what happens when the test score is **one or two standard deviation** higher.

Standardized Variables

- Assume Xs and Y are all continuous variables, then we run a multiple regression model

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + \cdots + \hat{\beta}_k X_{ik} + \hat{u}_i$$

- Because $\Sigma \hat{u}_i = 0$ and $\bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{X}_1 + \cdots + \hat{\beta}_k \bar{X}_k$,then

$$Y_i - \bar{Y} = \hat{\beta}_1 (X_{i1} - \bar{X}_1) + \hat{\beta}_2 (X_{i2} - \bar{X}_2) + \cdots + \hat{\beta}_k (X_{ik} - \bar{X}_k) + \hat{u}_i$$

- Then, we obtain following expressions

$$\begin{aligned} \frac{Y_i - \bar{Y}}{\sigma_y} = & \hat{\beta}_1 \frac{\sigma_{x_1}}{\sigma_y} \frac{(X_{i1} - \bar{X}_1)}{\sigma_{x_1}} + \hat{\beta}_2 \frac{\sigma_{x_1}}{\sigma_y} \frac{(X_{i2} - \bar{X}_2)}{\sigma_{x_2}} + \cdots + \\ & \hat{\beta}_k \frac{\sigma_{x_1}}{\sigma_y} \frac{(X_{i2} - \bar{X}_k)}{\sigma_{x_k}} + \frac{\hat{u}_i}{\sigma_y} \end{aligned}$$

Standardized Variables

- Then we have a standardized regression model

$$Z_y = \hat{\phi}_1 Z_1 + \hat{\phi}_2 Z_2 + \cdots + \hat{\phi}_k Z_k + v_i$$

  where Z_y denotes the Z-score of Y, Z_1 denotes the **Z-score** of X_1,and so on.

- The estimate coefficients

$$\hat{\phi}_j = (\hat{\sigma}_j / \hat{\sigma}_y)\,\hat{\beta}_j \text{ for } j = 1, \ldots, k$$

- $\hat{\phi}_j$ are traditionally called **standardized coefficients** or **beta coefficients** , which can be explained as if X_j increases by **1 standard deviation**, then Y changes by ϕ standard deviations.

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
- Besides, the data structure may violate the 2th OLS regression assumption, thus random sampling.
  1. Times series
  2. Cluster data
  3. Spatial data
- Last but not least, the magnitude of β_1 matters.

Wrap Up

- 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

Definition

- 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
   - 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.
   - Because 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.
- 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.
- 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.
- 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	

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.

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.

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 or migrating to a better district.
- **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.
- **Simultaneous causality**: it would arise if the performance on tests affected the student–teacher ratio.
- **Heteroskedasticity** and **correlation of the error term** across observations.
  - It does not threaten internal validity.
  - Correlation of the error term across observations, however, could threaten the consistency of the standard errors because the assumption of simple random sampling is violated.