TD 2: Research Design ECO 567A

Geoffrey Barrows[1]

[1]CREST, CNRS, Ecole polytechnique, Université Paris-Saclay

Jan 17, 2024

Last Time...

- Start from linear model:

$$y = X\theta + \varepsilon$$

Last Time...

- Start from linear model:

$$y = X\theta + \varepsilon$$

- Assume:

$$\varepsilon | X \sim \mathcal{N}\left(0, \sigma^2 I\right)$$

Last Time...

- Start from linear model:

$$y = X\theta + \varepsilon$$

- Assume:

$$\varepsilon | X \sim \mathcal{N}\left(0, \sigma^2 I\right)$$

  - Zero conditional expectation: On average, X is not informative about error term
  - Homoskedasticity: Error term has the same variance for all individuals

Last Time...

- Start from linear model:

$$y = X\theta + \varepsilon$$

- Assume:

$$\varepsilon | X \sim \mathcal{N}\left(0, \sigma^2 I\right)$$

  - Zero conditional expectation: On average, X is not informative about error term
  - Homoskedasticity: Error term has the same variance for all individuals

- Then, sampling distribution of the OLS estimator:

$$\hat{\theta}_{OLS} \sim \mathcal{N}\left(\theta, \sigma^2 \left(X'X\right)^{-1}\right)$$

Last Time...

- Start from linear model:

$$y = X\theta + \varepsilon$$

- Assume:

$$\varepsilon | X \sim \mathcal{N}\left(0, \sigma^2 I\right)$$

  - Zero conditional expectation: On average, X is not informative about error term
  - Homoskedasticity: Error term has the same variance for all individuals

- Then, sampling distribution of the OLS estimator:

$$\hat{\theta}_{OLS} \sim \mathcal{N}\left(\theta, \sigma^2 \left(X'X\right)^{-1}\right)$$

  - Unbiasedness: Expectation of the sampling distribution is the true value
  - Effiency: OLS has the smallest variance among unbiased linear estimators

Last Time...

- Estimate standard error:

$$se\left(\hat{\theta}_{OLS}\right) = \sqrt{\frac{e'e}{n-k}\left(X'X\right)^{-1}}$$

Last Time...

- Estimate standard error:

$$se\left(\hat{\theta}_{OLS}\right) = \sqrt{\frac{e'e}{n-k}\left(X'X\right)^{-1}}$$

  - Empirical residuals: $e = y - X\hat{\theta}_{OLS}$
  - Degrees of freedom: $k =$ Number of parameters estimated in $\hat{\theta}_{OLS}$

Last Time...

- Estimate standard error:

$$se\left(\hat{\theta}_{OLS}\right) = \sqrt{\frac{e'e}{n-k}\left(X'X\right)^{-1}}$$

  - Empirical residuals: $e = y - X\hat{\theta}_{OLS}$
  - Degrees of freedom: k = Number of parameters estimated in $\hat{\theta}_{OLS}$

- Then, sampling distribution of the t-statistic: $\frac{\hat{\theta}_{OLS}-\theta}{se(\hat{\theta}_{OLS})} \sim$ Student t

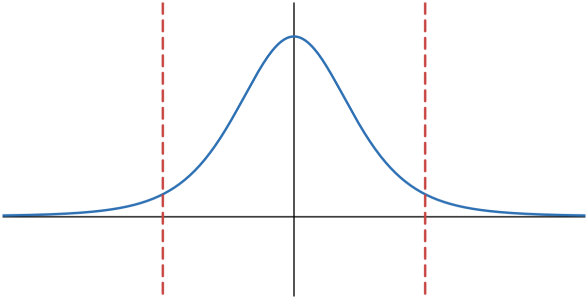
Last Time...

- Estimate standard error:

$$se\left(\hat{\theta}_{OLS}\right) = \sqrt{\frac{e'e}{n-k}\left(X'X\right)^{-1}}$$

  - Empirical residuals: $e = y - X\hat{\theta}_{OLS}$
  - Degrees of freedom: $k =$ Number of parameters estimated in $\hat{\theta}_{OLS}$

- Then, sampling distribution of the t-statistic: $\frac{\hat{\theta}_{OLS}-\theta}{se(\hat{\theta}_{OLS})} \sim$ Student t



Last Time...

- This allows hypothesis testing:

$$H0 : \theta = \theta_0 \text{ vs. } H1 : \begin{cases} \theta > \theta_0 & (i) \\ \theta < \theta_0 & (ii) \end{cases}$$

Last Time...

- This allows hypothesis testing:
$$H0 : \theta = \theta_0 \text{ vs. } H1 : \begin{cases} \theta > \theta_0 & (i) \\ \theta < \theta_0 & (ii) \end{cases}$$
  - Under H1 (i): *t-statistic is too large*; under H1 (i): *t-statistic is too small*

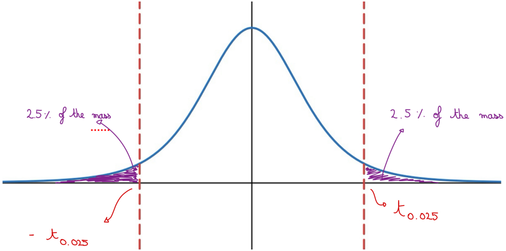
Last Time...

- This allows hypothesis testing:

$$H0 : \theta = \theta_0 \text{ vs. } H1 : \begin{cases} \theta > \theta_0 & (i) \\ \theta < \theta_0 & (ii) \end{cases}$$

  - Under H1 (i): *t-statistic is too large*; under H1 (i): *t-statistic is too small*

- Choose a confidence level, say 95%, and then:



  - E.g., if $\frac{\hat{\theta}_{OLS} - \theta_0}{se(\hat{\theta}_{OLS})} > t_{0.025}$, accept H1 (i) and reject H0

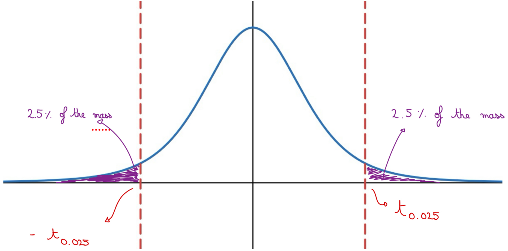
Last Time...

- This allows hypothesis testing:

$$H0 : \theta = \theta_0 \text{ vs. } H1 : \begin{cases} \theta > \theta_0 & (i) \\ \theta < \theta_0 & (ii) \end{cases}$$

  - Under H1 (i): *t-statistic is too large*; under H1 (i): *t-statistic is too small*

- Choose a confidence level, say 95%, and then:



  - E.g., if $\frac{\hat{\theta}_{OLS} - \theta_0}{se(\hat{\theta}_{OLS})} > t_{0.025}$, accept H1 (i) and reject H0

- 95% confidence interval: $(\hat{\theta}_{OLS} \pm t_{.025} * se(\hat{\theta}_{ols}))$

But...

- Can we assume by default?

$$\varepsilon | X \sim \mathcal{N}\left(0, \sigma^2 I\right)$$

- Or more generally, can we assume?

$$E\left(\varepsilon | X\right) = 0$$

- Today: (i) why we need a **research design** and (ii) how it could look like

Outline

When Do We Have $E[\varepsilon|X] \neq 0$?
- Omitted Variable Bias
- Reverse Causality / Simultaneity
- In the Potential Outcome Framework

What Happens When $E[\varepsilon|X] \neq 0$?

What To Do When $E[\varepsilon|X] \neq 0$?
- Instrumental Variables
- Randomized Control Trials (RCTs)
- Fixed Effects
- Difference-in-Differences

Review Question

When Do We Have $E\left[\varepsilon|X\right] \neq 0$?

When Do We Have $E[\varepsilon|X] \neq 0$?

Omitted Variable Bias

Definition

- First category of issues: **Omitted variable bias**

Definition

- First category of issues: **Omitted variable bias**

- Imagine we posit the following model:

$$y_i = \alpha + \beta x_i + u_i$$

Where y stands for income and x for education

Definition

- First category of issues: **Omitted variable bias**

- Imagine we posit the following model:

$$y_i = \alpha + \beta x_i + u_i$$

  Where y stands for income and x for education

- And, by doing so, we "forget" about parents' education z:

$$\begin{cases} u_i & = \gamma z_i + \epsilon_i \\ x_i & = \lambda + \mu z_i + \eta_i \end{cases}$$

Definition

- First category of issues: **Omitted variable bias**

- Imagine we posit the following model:

$$y_i = \alpha + \beta x_i + u_i$$

  Where y stands for income and x for education

- And, by doing so, we "forget" about parents' education z:

$$\begin{cases} u_i & = \gamma z_i + \epsilon_i \\ x_i & = \lambda + \mu z_i + \eta_i \end{cases}$$

  - Parents' education directly affects income: Better professional network, etc.
  - Parents' education also affects child's education: Role model, information, etc.

Resulting Issue

- When we estimate our wrong model, we have:

$$\mathrm{Cov}(x_i, u_i) = \mathrm{Cov}(\lambda + \mu z_i + \eta_i, \gamma z_i + \epsilon_i)$$

Resulting Issue

- When we estimate our wrong model, we have:

$$\begin{aligned}\mathrm{Cov}(x_i, u_i) =& \mathrm{Cov}(\lambda + \mu z_i + \eta_i, \gamma z_i + \epsilon_i) \\ =& \mu\gamma\mathrm{Cov}(z_i, z_i) + \mu\mathrm{Cov}(z_i, \epsilon_i) + \gamma\mathrm{Cov}(\eta_i, z_i) + \mathrm{Cov}(\eta_i, \epsilon_i)\end{aligned}$$

Resulting Issue

- When we estimate our wrong model, we have:

$$
\begin{aligned}
\mathrm{Cov}(x_i, u_i) =& \mathrm{Cov}(\lambda + \mu z_i + \eta_i, \gamma z_i + \epsilon_i) \\
=& \mu\gamma \mathrm{Cov}(z_i, z_i) + \mu \mathrm{Cov}(z_i, \epsilon_i) + \gamma \mathrm{Cov}(\eta_i, z_i) + \mathrm{Cov}(\eta_i, \epsilon_i) \\
=& \mu\gamma \mathrm{Cov}(z_i, z_i)
\end{aligned}
$$

Resulting Issue

- When we estimate our wrong model, we have:

$$\begin{aligned}\mathrm{Cov}(x_i, u_i) =& \mathrm{Cov}(\lambda + \mu z_i + \eta_i, \gamma z_i + \epsilon_i) \\ =& \mu\gamma \mathrm{Cov}(z_i, z_i) + \mu \mathrm{Cov}(z_i, \epsilon_i) + \gamma \mathrm{Cov}(\eta_i, z_i) + \mathrm{Cov}(\eta_i, \epsilon_i) \\ =& \mu\gamma \mathrm{Cov}(z_i, z_i) \\ =& \mu\gamma \mathrm{Var}(z_i, z_i) \neq 0\end{aligned}$$

Resulting Issue

- When we estimate our wrong model, we have:

$$
\begin{aligned}
\mathrm{Cov}(x_i, u_i) =& \mathrm{Cov}(\lambda + \mu z_i + \eta_i, \gamma z_i + \epsilon_i) \\
=& \mu\gamma\mathrm{Cov}(z_i, z_i) + \mu\mathrm{Cov}(z_i, \epsilon_i) + \gamma\mathrm{Cov}(\eta_i, z_i) + \mathrm{Cov}(\eta_i, \epsilon_i) \\
=& \mu\gamma\mathrm{Cov}(z_i, z_i) \\
=& \mu\gamma\mathrm{Var}(z_i, z_i) \neq 0
\end{aligned}
$$

- It follows that $E(u|x) \neq 0$ and OLS estimator may be biased

Resulting Issue

- When we estimate our wrong model, we have:

$$\begin{aligned}\mathrm{Cov}(x_i, u_i) =& \mathrm{Cov}(\lambda + \mu z_i + \eta_i, \gamma z_i + \epsilon_i) \\ =& \mu\gamma \mathrm{Cov}(z_i, z_i) + \mu \mathrm{Cov}(z_i, \epsilon_i) + \gamma \mathrm{Cov}(\eta_i, z_i) + \mathrm{Cov}(\eta_i, \epsilon_i) \\ =& \mu\gamma \mathrm{Cov}(z_i, z_i) \\ =& \mu\gamma \mathrm{Var}(z_i, z_i) \neq 0\end{aligned}$$

- It follows that $E(u|x) \neq 0$ and OLS estimator may be biased

- In the above, we see that there is no issue if:

Resulting Issue

- When we estimate our wrong model, we have:

$$\begin{aligned}\mathrm{Cov}(x_i, u_i) =& \mathrm{Cov}(\lambda + \mu z_i + \eta_i, \gamma z_i + \epsilon_i)\\ =& \mu\gamma \mathrm{Cov}(z_i, z_i) + \mu \mathrm{Cov}(z_i, \epsilon_i) + \gamma \mathrm{Cov}(\eta_i, z_i) + \mathrm{Cov}(\eta_i, \epsilon_i)\\ =& \mu\gamma \mathrm{Cov}(z_i, z_i)\\ =& \mu\gamma \mathrm{Var}(z_i, z_i) \neq 0\end{aligned}$$

- It follows that $E(u|x) \neq 0$ and OLS estimator may be biased

- In the above, we see that there is no issue if:
  - $\mu = 0$: Omitted variable is unrelated to education
  - $\gamma = 0$: Omitted variable does not affect income directly
  - $\mathrm{Var}(z_i, z_i) = 0$: Omitted variable is a constant (not super interesting)

Other Examples

- Other example: What is the **causal effect of pollution on health**?

Other Examples

- Other example: What is the **causal effect of pollution on health**?

- Places with more pollution also have more labor demand
- Which means more income
- And hence better health outcomes

Other Examples

- Other example: What is the **causal effect of pollution on health**?

- Places with more pollution also have more labor demand
- Which means more income
- And hence better health outcomes

- So, "income" is an omitted variable related to both pollution and health

Other Examples

- What is the **causal effect of government programs on poverty reduction**?

Other Examples

- What is the **causal effect of government programs on poverty reduction**?

- Governments might target areas that are growing anyways
- And that would have seen poverty decrease even absent the program

Other Examples

- What is the **causal effect of government programs on poverty reduction**?
- Governments might target areas that are growing anyways
- And that would have seen poverty decrease even absent the program
- So, "expected growth" could be a problematic omitted variable

When Do We Have $E\left[\varepsilon|X\right] \neq 0$?

Reverse Causality / Simultaneity

Definition

- Second category of issues: **Reverse causality / Simultaneity**

Definition

- Second category of issues: **Reverse causality / Simultaneity**

- Imagine we posit the following model:

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

Where y stands for crime prevalence and x for crowd density

Definition

- Second category of issues: **Reverse causality / Simultaneity**
- Imagine we posit the following model:

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

  Where y stands for crime prevalence and x for crowd density

- But we “forget” that criminality also affects people’s movements:

$$\begin{cases} y_i & = \alpha + \beta x_i + \varepsilon_i \\ x_i & = \gamma + \delta y_i + \nu_i \end{cases}$$

Definition

- Second category of issues: **Reverse causality / Simultaneity**

- Imagine we posit the following model:

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

  Where y stands for crime prevalence and x for crowd density

- But we “forget” that criminality also affects people’s movements:

$$\begin{cases} y_i & = \alpha + \beta x_i + \varepsilon_i \\ x_i & = \gamma + \delta y_i + \nu_i \end{cases}$$

  - Crowd density has a causal effect on crime: Deterrence?
  - Crime has a causal effect on density: Lower attractivity

Resulting Issue

- When we estimate our wrong model, we have:

$$\text{Cov}(x_i, u_i) = \text{Cov}(\gamma + \delta y_i + \nu_i, \varepsilon_i)$$

Resulting Issue

- When we estimate our wrong model, we have:

$$
\begin{aligned}
\mathrm{Cov}(x_i, u_i) =& \mathrm{Cov}(\gamma + \delta y_i + \nu_i, \varepsilon_i) \\
=& \delta \mathrm{Cov}(y_i, \varepsilon_i)
\end{aligned}
$$

Resulting Issue

- When we estimate our wrong model, we have:

$$\begin{aligned}\mathrm{Cov}(x_i, u_i) =& \mathrm{Cov}(\gamma + \delta y_i + \nu_i, \varepsilon_i) \\ =& \delta \mathrm{Cov}(y_i, \varepsilon_i) \\ =& \delta \mathrm{Cov}(\alpha + \beta x_i + \varepsilon_i, \varepsilon_i)\end{aligned}$$

Resulting Issue

- When we estimate our wrong model, we have:

$$
\begin{aligned}
\mathrm{Cov}(x_i, u_i) =& \mathrm{Cov}(\gamma + \delta y_i + \nu_i, \varepsilon_i) \\
=& \delta \mathrm{Cov}(y_i, \varepsilon_i) \\
=& \delta \mathrm{Cov}(\alpha + \beta x_i + \varepsilon_i, \varepsilon_i) \\
=& \beta\delta \mathrm{Cov}(x_i, \varepsilon_i) + \delta \mathrm{Var}(\varepsilon_i)
\end{aligned}
$$

Resulting Issue

- When we estimate our wrong model, we have:

$$\begin{aligned}\mathrm{Cov}(x_i, u_i) =& \mathrm{Cov}(\gamma + \delta y_i + \nu_i, \varepsilon_i)\\ =& \delta \mathrm{Cov}(y_i, \varepsilon_i)\\ =& \delta \mathrm{Cov}(\alpha + \beta x_i + \varepsilon_i, \varepsilon_i)\\ =& \beta\delta \mathrm{Cov}(x_i, \varepsilon_i) + \delta \mathrm{Var}(\varepsilon_i)\end{aligned}$$

- Solving the equation: $\mathrm{Cov}(x_i, \varepsilon_i) = \frac{\delta}{1-\beta\delta}\mathrm{Var}(\varepsilon_i) \neq 0$

Resulting Issue

- When we estimate our wrong model, we have:

$$\begin{aligned}\mathrm{Cov}(x_i, u_i) =& \mathrm{Cov}(\gamma + \delta y_i + \nu_i, \varepsilon_i)\\ =& \delta \mathrm{Cov}(y_i, \varepsilon_i)\\ =& \delta \mathrm{Cov}(\alpha + \beta x_i + \varepsilon_i, \varepsilon_i)\\ =& \beta\delta \mathrm{Cov}(x_i, \varepsilon_i) + \delta \mathrm{Var}(\varepsilon_i)\end{aligned}$$

- Solving the equation: $\mathrm{Cov}(x_i, \varepsilon_i) = \frac{\delta}{1-\beta\delta}\mathrm{Var}(\varepsilon_i) \neq 0$

- It follows that $E(u|x) \neq 0$ and OLS estimator may be biased

Resulting Issue

- When we estimate our wrong model, we have:

$$\begin{aligned}\mathrm{Cov}(x_i, u_i) =& \mathrm{Cov}(\gamma + \delta y_i + \nu_i, \varepsilon_i)\\ =& \delta \mathrm{Cov}(y_i, \varepsilon_i)\\ =& \delta \mathrm{Cov}(\alpha + \beta x_i + \varepsilon_i, \varepsilon_i)\\ =& \beta\delta \mathrm{Cov}(x_i, \varepsilon_i) + \delta \mathrm{Var}(\varepsilon_i)\end{aligned}$$

- Solving the equation: $\mathrm{Cov}(x_i, \varepsilon_i) = \frac{\delta}{1-\beta\delta}\mathrm{Var}(\varepsilon_i) \neq 0$
- It follows that $E(u|x) \neq 0$ and OLS estimator may be biased
- In the above, we see that there is no issue if:

Resulting Issue

- When we estimate our wrong model, we have:

$$\begin{aligned}\mathrm{Cov}(x_i, u_i) =& \mathrm{Cov}(\gamma + \delta y_i + \nu_i, \varepsilon_i) \\ =& \delta \mathrm{Cov}(y_i, \varepsilon_i) \\ =& \delta \mathrm{Cov}(\alpha + \beta x_i + \varepsilon_i, \varepsilon_i) \\ =& \beta\delta \mathrm{Cov}(x_i, \varepsilon_i) + \delta \mathrm{Var}(\varepsilon_i)\end{aligned}$$

- Solving the equation: $\mathrm{Cov}(x_i, \varepsilon_i) = \frac{\delta}{1-\beta\delta}\mathrm{Var}(\varepsilon_i) \neq 0$
- It follows that $E(u|x) \neq 0$ and OLS estimator may be biased
- In the above, we see that there is no issue if:
  - $\delta = 0$, which essentially denies reverse causality

When Do We Have $E\left[\varepsilon|X\right] \neq 0$?

In the Potential Outcome Framework

Introducing the Potential Outcome Framework

- Imagine we study a treatment (e.g., ban on polluting cars)

$$\text{Treatment } \; X_i = \begin{cases} 1 & \text{if treated} \\ 0 & \text{if otherwise} \end{cases}$$

Introducing the Potential Outcome Framework

- Imagine we study a treatment (e.g., ban on polluting cars)

$$\text{Treatment } \; X_i = \begin{cases} 1 & \text{if treated} \\ 0 & \text{if otherwise} \end{cases}$$

- We want to assess its effect on a given outcome (e.g., pollution)

$$\text{Potential Outcome} = \begin{cases} Y_{1i} & \text{if } X_i = 1 \\ Y_{0i} & \text{if } X_i = 0 \end{cases}$$

Introducing the Potential Outcome Framework

- Imagine we study a treatment (e.g., ban on polluting cars)

$$\text{Treatment } \; X_i = \begin{cases} 1 & \text{if treated} \\ 0 & \text{if otherwise} \end{cases}$$

- We want to assess its effect on a given outcome (e.g., pollution)

$$\text{Potential Outcome} = \begin{cases} Y_{1i} & \text{if } X_i = 1 \\ Y_{0i} & \text{if } X_i = 0 \end{cases}$$

- We only observe the outcome that is actually realized

$$\text{Observed Outcome } \; Y_i \;\; = \;\; \begin{cases} Y_{1i} & \text{if } X_i = 1 \\ Y_{0i} & \text{if } X_i = 0 \end{cases} = Y_{0i} + \underbrace{(Y_{1i} - Y_{0i})}_{\text{Treatment effect}} * X_i$$

Selection Bias

- Estimate with OLS:

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

Selection Bias

- Estimate with OLS:

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

- Amounts to: $\hat{\beta} = \underbrace{\hat{E}\left(Y_i | X_i = 1\right)}_{\text{Mean of } Y \text{ among } X=1} - \underbrace{\hat{E}\left(Y_i | X_i = 0\right)}_{\text{Mean of } Y \text{ among } X=0}$

Selection Bias

- Estimate with OLS:

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

- Amounts to: $\hat{\beta} = \underbrace{\hat{E}\left(Y_i | X_i = 1\right)}_{\text{Mean of } Y \text{ among } X=1} - \underbrace{\hat{E}\left(Y_i | X_i = 0\right)}_{\text{Mean of } Y \text{ among } X=0}$

- Expectation of the sampling distribution: $E(\hat{\beta}) = E\left(Y_i | X_i = 1\right) - E\left(Y_i | X_i = 0\right)$
  - Where: $E[Y_i | X_i = 1] = \alpha + \beta + E[\varepsilon_i | X_i = 1]$
  - And: $E[Y_i | X_i = 0] = \alpha + E[\varepsilon_i | X_i = 0]$

Selection Bias

- Estimate with OLS:

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

- Amounts to: $\hat{\beta} = \underbrace{\hat{E}\left(Y_i | X_i = 1\right)}_{\text{Mean of } Y \text{ among } X=1} - \underbrace{\hat{E}\left(Y_i | X_i = 0\right)}_{\text{Mean of } Y \text{ among } X=0}$

- Expectation of the sampling distribution: $E(\hat{\beta}) = E\left(Y_i | X_i = 1\right) - E\left(Y_i | X_i = 0\right)$
  - Where: $E[Y_i | X_i = 1] = \alpha + \beta + E[\varepsilon_i | X_i = 1]$
  - And: $E[Y_i | X_i = 0] = \alpha + E[\varepsilon_i | X_i = 0]$

- We can assess unbiasedness:

$$\begin{aligned} E(\hat{\beta}) &= E[Y_i | X_i = 1] - E[Y_i | X_i = 0] \\ &= \beta + E[\varepsilon_i | X_i = 1] - E[\varepsilon_i | X_i = 0] \\ &= \beta + \underbrace{E[Y_{0i} | X_i = 1] - E[Y_{0i} | X_i = 0]}_{\text{Risk of selection bias}} \end{aligned}$$

What Happens When $E\left[\varepsilon|X\right] \neq 0$?

Monte Carlo Experiment

- Assume a true model

$$
\begin{aligned}
y_i &= \alpha + \beta * x_i + a_i + \varepsilon_i \\
\alpha &= 10 \\
\beta &= 2
\end{aligned}
$$

Monte Carlo Experiment

- Assume a true model

$$
\begin{aligned}
y_i &= \alpha + \beta * x_i + a_i + \varepsilon_i \\
\alpha &= 10 \\
\beta &= 2
\end{aligned}
$$

- Examine three cases:
  - Case 1: Exogenous a
    - Variable is omitted but it is unrelated to the regressor of interest
  - Case 2: Endogenous a, and we don't know it
    - Variable is omitted, and it is related to the regressor of interest
  - Case 3: Endogenous a, but we know it
    - Variable is related to the regressor of interest, but it is not omitted

Case 1: Exogenous Ability — Theoretically

- True model:

$$\begin{aligned} y_i &= \alpha + \beta x_i + a_i + \varepsilon_i \\ E(a + \varepsilon | x) &= E(a|x) + E(\varepsilon|x) = E(a|x) = 0 \end{aligned}$$

Case 1: Exogenous Ability — Theoretically

- True model:

$$
\begin{aligned}
y_i &= \alpha + \beta x_i + a_i + \varepsilon_i \\
E(a + \varepsilon | x) &= E(a|x) + E(\varepsilon|x) = E(a|x) = 0
\end{aligned}
$$

- Estimate:

$$
y_i = \alpha + \beta x_i + \underbrace{u_i}_{=a_i + \varepsilon_i}
$$

Case 1: Exogenous Ability — Theoretically

- True model:

$$\begin{aligned} y_i &= \alpha + \beta x_i + a_i + \varepsilon_i \\ E(a + \varepsilon | x) &= E(a|x) + E(\varepsilon|x) = E(a|x) = 0 \end{aligned}$$

- Estimate:

$$y_i = \alpha + \beta x_i + \underbrace{u_i}_{=a_i + \varepsilon_i}$$

- Prediction:

Case 1: Exogenous Ability — Theoretically

- True model:

$$\begin{aligned} y_i &= \alpha + \beta x_i + a_i + \varepsilon_i \\ E(a + \varepsilon | x) &= E(a|x) + E(\varepsilon|x) = E(a|x) = 0 \end{aligned}$$

- Estimate:

$$y_i = \alpha + \beta x_i + \underbrace{u_i}_{= a_i + \varepsilon_i}$$

- Prediction: OLS estimator should be **unbiased**

Case 1: Exogenous ability — Experiment

- Monte Carlo experiment:

$$
\begin{aligned}
x_i &\sim \mathcal{U}(0,20) \\
a_i &\sim \mathcal{N}(0,5) \\
\varepsilon_i &\sim \mathcal{N}(0,5) \\
y_i &= \alpha + \beta x_i + a_i + \varepsilon_i
\end{aligned}
$$

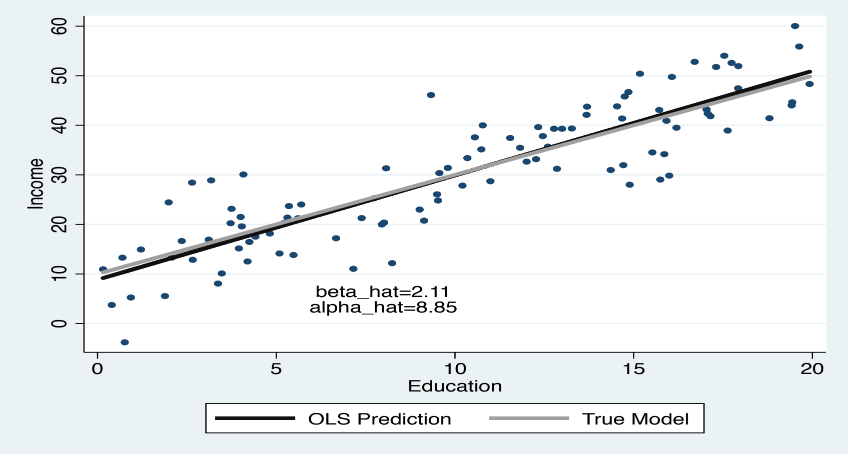
- Data:

Table: Summary statistics

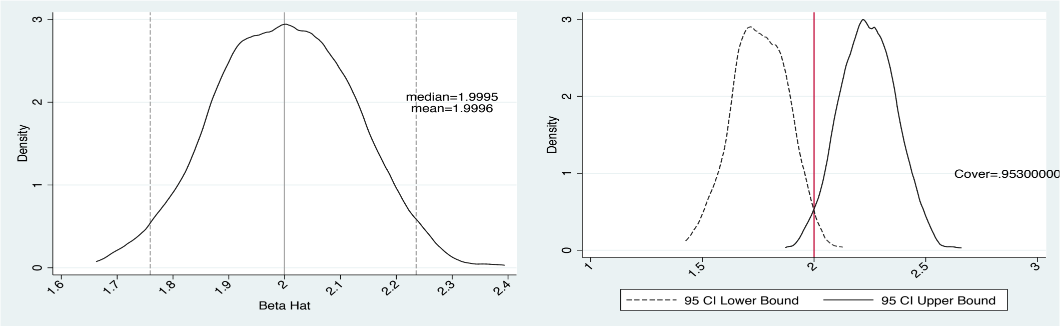
Variable	**Mean**	**Std. Dev.**	**Min.**	**Max.**	**N**
y	30.268	13.734	-3.795	60.018	100
x	10.169	5.758	0.156	19.923	100
a	0.012	4.899	-13.5	12.748	100
eps	-0.083	5.251	-11.975	11.012	100

Case 1: Exogenous ability — Results

Estimation equation: $y_i = \alpha + \beta x_i + \underbrace{u_i}_{=a_i + \varepsilon_i}$



Case 1: Exogenous ability — Results



Case 2: Endogenous a, and we don't know it — Theoretically

- True model:

$$\begin{aligned} y_i &= \alpha_1 + \beta x_i + a_i + \varepsilon_i \\ x_i &= \alpha_2 + \delta a_i + \eta_i \\ \delta &= 0.5 \end{aligned}$$

Case 2: Endogenous a, and we don't know it — Theoretically

- True model:

$$\begin{aligned} y_i &= \alpha_1 + \beta x_i + a_i + \varepsilon_i \\ x_i &= \alpha_2 + \delta a_i + \eta_i \\ \delta &= 0.5 \end{aligned}$$

- Estimate:

$$y_i = \alpha_1 + \beta x_i + \underbrace{u_i}_{=a_i+\varepsilon_i}$$

Case 2: Endogenous a, and we don't know it — Theoretically

- True model:

$$\begin{aligned} y_i &= \alpha_1 + \beta x_i + a_i + \varepsilon_i \\ x_i &= \alpha_2 + \delta a_i + \eta_i \\ \delta &= 0.5 \end{aligned}$$

- Estimate:

$$y_i = \alpha_1 + \beta x_i + \underbrace{u_i}_{=a_i+\varepsilon_i}$$

- Prediction:

Case 2: Endogenous a, and we don't know it — Theoretically

- True model:

$$\begin{aligned} y_i &= \alpha_1 + \beta x_i + a_i + \varepsilon_i \\ x_i &= \alpha_2 + \delta a_i + \eta_i \\ \delta &= 0.5 \end{aligned}$$

- Estimate:

$$y_i = \alpha_1 + \beta x_i + \underbrace{u_i}_{=a_i + \varepsilon_i}$$

- Prediction: OLS estimator is **likely biased**

Case 2: Endogenous a, and we don't know it — Experiment

- Monte Carlo experiment:

$$\begin{aligned} \eta &\sim N(0,1)\ ;\quad a_i \sim N(0,5)\ ;\quad \varepsilon_i \sim N(0,5) \\ x_i &= \alpha_2 + \delta a_i + \eta_i \\ y_i &= \alpha_1 + \beta x_i + a_i + \varepsilon_i \end{aligned}$$

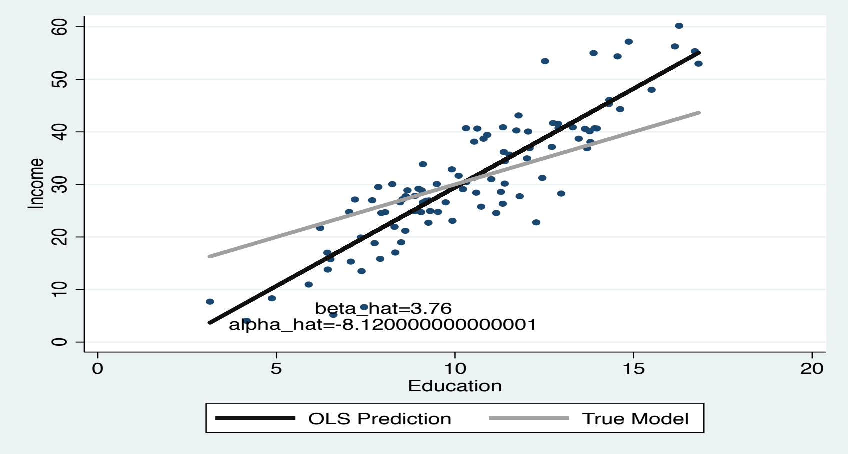
- Data:

Table: Summary statistics

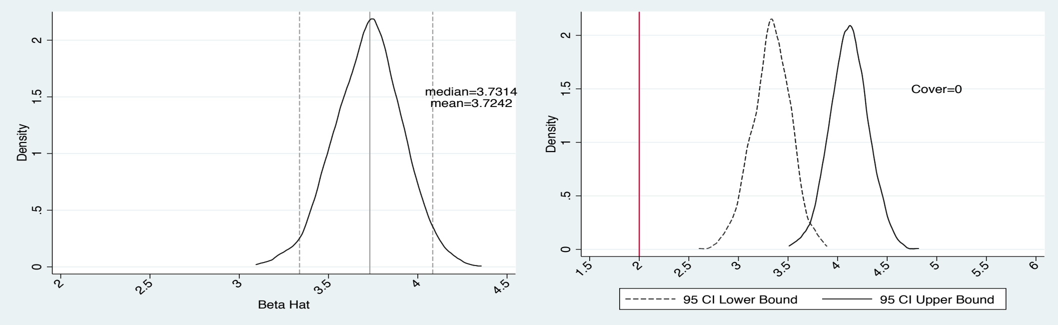
Variable	**Mean**	**Std. Dev.**	**Min.**	**Max.**	**N**
y	31.212	12.035	4.037	60.175	100
x	10.472	2.852	3.142	16.819	100
a	0.940	5.418	-11.25	14.065	100
eps	-0.672	4.947	-13.5	10.32	100
eta	0.002	1.02	-2.068	2.252	100

Case 2: Endogenous a, and we don't know it — Results

Estimation Equation: $y_i = \alpha_1 + \beta * x_i + \underbrace{u_i}_{=a_i+\varepsilon_i}$



Case 2: Endogenous a, and we don't know it — Results



Case 3: Endogenous a, but we know it — Theoretically

- True Model

$$\begin{aligned} y_i &= \alpha_1 + \beta x_i + a_i + \varepsilon_i \\ x_i &= \alpha_2 + \delta a_i + \eta_i \\ \delta &= 0.5 \end{aligned}$$

Case 3: Endogenous a, but we know it — Theoretically

- True Model

$$\begin{aligned} y_i &= \alpha_1 + \beta x_i + a_i + \varepsilon_i \\ x_i &= \alpha_2 + \delta a_i + \eta_i \\ \delta &= 0.5 \end{aligned}$$

- Estimate:

$$y_i = \alpha_1 + \beta * x_i + a_i + \varepsilon_i$$

Case 3: Endogenous a, but we know it — Theoretically

- True Model

$$\begin{aligned} y_i &= \alpha_1 + \beta x_i + a_i + \varepsilon_i \\ x_i &= \alpha_2 + \delta a_i + \eta_i \\ \delta &= 0.5 \end{aligned}$$

- Estimate:

$$y_i = \alpha_1 + \beta * x_i + a_i + \varepsilon_i$$

- Prediction:

Case 3: Endogenous a, but we know it — Theoretically

- True Model

$$\begin{aligned} y_i &= \alpha_1 + \beta x_i + a_i + \varepsilon_i \\ x_i &= \alpha_2 + \delta a_i + \eta_i \\ \delta &= 0.5 \end{aligned}$$

- Estimate:

$$y_i = \alpha_1 + \beta * x_i + a_i + \varepsilon_i$$

- Prediction: OLS estimator should be **unbiased**

Case 3: Endogenous a, but we know it — Experiment

- Monte Carlo experiment:

$$\begin{aligned} \eta &\sim N(0,1)\ ;\quad a_i \sim N(0,5)\ ;\quad \varepsilon_i \sim N(0,5) \\ x_i &= \alpha_2 + \delta a_i + \eta_i \\ y_i &= \alpha_1 + \beta x_i + a_i + \varepsilon_i \end{aligned}$$

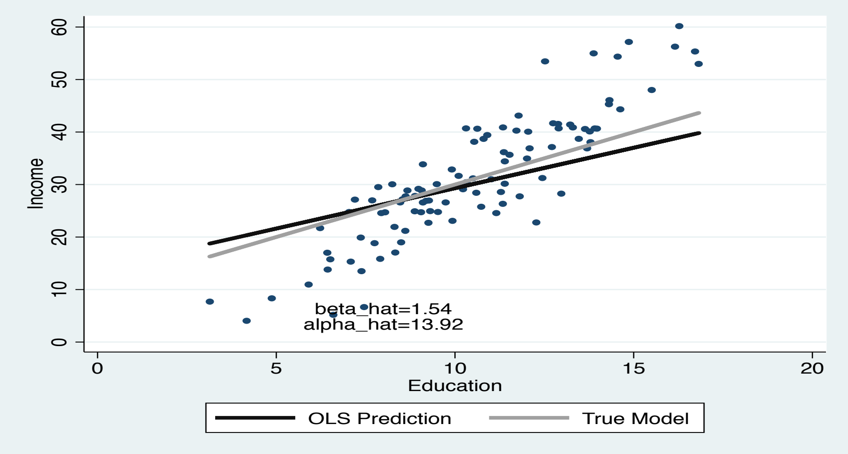
- Data:

Table: Summary statistics

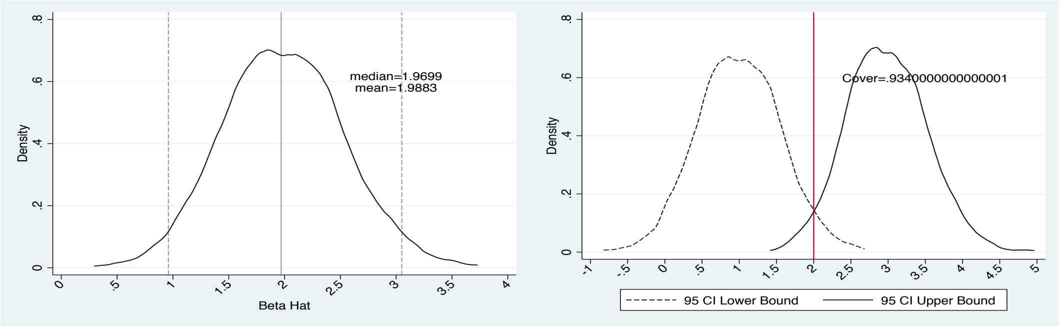
Variable	**Mean**	**Std. Dev.**	**Min.**	**Max.**	**N**
y	31.212	12.035	4.037	60.175	100
x	10.472	2.852	3.142	16.819	100
a	0.940	5.418	-11.25	14.065	100
eps	-0.672	4.947	-13.5	10.32	100
eta	0.002	1.02	-2.068	2.252	100

Case 3: Endogenous a, but we know it — Results

Estimation equation: $y_i = \alpha_1 + \beta x_i + a_i + \varepsilon_i$



Case 3: Endogenous *a*, but we know it — Results



What To Do When $E\left[\varepsilon | X\right] \neq 0$?

What To Do When $E[\varepsilon|X] \neq 0$?

Instrumental Variables

Context

- Suppose $E[\varepsilon|X] \neq 0$. E.g., x correlated with a, which also determines y:

$$x_i = \alpha_2 + \delta a_i + z_i + \eta_i$$

Context

- Suppose $E\left[\varepsilon|X\right] \neq 0$. E.g., x correlated with a, which also determines y:

$$x_i = \alpha_2 + \delta a_i + z_i + \eta_i$$

- If you omit a and simply estimate:

$$y_i = \alpha_1 + \beta x_i + \underbrace{u_i}_{=a_i + \varepsilon_i}$$

Then, we fall into omitted variable bias, and OLS is biased

Context

- Suppose $E[\varepsilon|X] \neq 0$. E.g., x correlated with a, which also determines y:

$$x_i = \alpha_2 + \delta a_i + z_i + \eta_i$$

- If you omit a and simply estimate:

$$y_i = \alpha_1 + \beta x_i + \underbrace{u_i}_{=a_i+\varepsilon_i}$$

  Then, we fall into omitted variable bias, and OLS is biased

- Instrumental variable approach:
  - Find another z variable which correlates with x...
  - ...but not with a...
  - ...and use it to recover β!

Example

- Back to the effect of crowd density (x) on crime prevalence (y)

Example

- Back to the effect of crowd density (x) on crime prevalence (y)
- Our instrumental variable (z) must satisfy:
  - Strong first stage: z must correlate with x
  - Exclusion restriction: z must not correlate with the error term

Example

- Back to the effect of crowd density (x) on crime prevalence (y)
- Our instrumental variable (z) must satisfy:
  - Strong first stage: z must correlate with x
  - Exclusion restriction: z must not correlate with the error term
- Alternative framing: z should be related with y solely through its effect on x

Example

- Back to the effect of crowd density (x) on crime prevalence (y)
- Our instrumental variable (z) must satisfy:
  - Strong first stage: z must correlate with x
  - Exclusion restriction: z must not correlate with the error term
- Alternative framing: z should be related with y solely through its effect on x
- E.g., subway station maintenance operations
  - Strong first stage: Clear negative correlation with crowd density
  - Exclusion restriction: Arguably random with respect to crime prevalence

Example

- Back to the effect of crowd density (x) on crime prevalence (y)
- Our instrumental variable (z) must satisfy:
  - Strong first stage: z must correlate with x
  - Exclusion restriction: z must not correlate with the error term
- Alternative framing: z should be related with y solely through its effect on x
- E.g., subway station maintenance operations
  - Strong first stage: Clear negative correlation with crowd density
  - Exclusion restriction: Arguably random with respect to crime prevalence
- Strong first stage can be tested, not the exclusion restriction in general

In Practice: Two-stage Least Squares (2SLS)

1. First, use z to predict x:

$$
\begin{aligned}
x_i &= \pi_1 + \pi_2 z_i + \underbrace{\gamma_i}_{=\delta a_i + \eta_i} \\
\hat{x}_i &= \hat{\pi}_1 + \hat{\pi}_2 * z_i
\end{aligned}
$$

In Practice: Two-stage Least Squares (2SLS)

1. First, use z to predict x:

$$
\begin{aligned}
x_i &= \pi_1 + \pi_2 z_i + \underbrace{\gamma_i}_{=\delta a_i + \eta_i} \\
\hat{x}_i &= \hat{\pi}_1 + \hat{\pi}_2 * z_i
\end{aligned}
$$

2. Second, run OLS on $\hat{x}$, not x:

$$
y_i = \alpha + \beta \hat{x}_i + \underbrace{u_i}_{=a_i + \varepsilon_i}
$$

In Practice: Two-stage Least Squares (2SLS)

1. First, use z to predict x:

$$
\begin{aligned}
x_i &= \pi_1 + \pi_2 z_i + \underbrace{\gamma_i}_{=\delta a_i + \eta_i} \\
\hat{x}_i &= \hat{\pi}_1 + \hat{\pi}_2 * z_i
\end{aligned}
$$

2. Second, run OLS on $\hat{x}$, not x:

$$
y_i = \alpha + \beta \hat{x}_i + \underbrace{u_i}_{=a_i + \varepsilon_i}
$$

- Then: $E(u_i|\hat{x}_i) = E(a_i + \varepsilon_i|\hat{\pi}_1 + \hat{\pi}_2 * z_i) = E(a_i|\hat{\pi}_1 + \hat{\pi}_2 * z_i) + E(\varepsilon_i|\hat{\pi}_1 + \hat{\pi}_2 * z_i) = 0$
- $\hat{x}$ uncorrelated with a: Extracted the random component in x, and **OLS is unbiased**

Monte Carlo Experiment — Theoretically

- True model:

$$
\begin{aligned}
y_i &= \alpha_1 + \beta x_i + a_i + \varepsilon_i \\
x_i &= \alpha_2 + \delta a_i + z_i + \eta_i \\
\delta &= 0.5
\end{aligned}
$$

Monte Carlo Experiment — Theoretically

- True model:

$$
\begin{aligned}
y_i &= \alpha_1 + \beta x_i + a_i + \varepsilon_i \\
x_i &= \alpha_2 + \delta a_i + z_i + \eta_i \\
\delta &= 0.5
\end{aligned}
$$

- Estimate:

$$
\begin{aligned}
x_i &= \pi_1 + \pi_2 * z_i + \underbrace{\gamma_i}_{=\delta a_i + \eta_i} \implies \hat{x}_i = \hat{\pi}_1 + \hat{\pi}_2 * z_i \\
y_i &= \alpha_1 + \beta * \hat{x}_i + \underbrace{u_i}_{=a_i + \varepsilon_i}
\end{aligned}
$$

Monte Carlo Experiment — Theoretically

- True model:

$$\begin{aligned} y_i &= \alpha_1 + \beta x_i + a_i + \varepsilon_i \\ x_i &= \alpha_2 + \delta a_i + z_i + \eta_i \\ \delta &= 0.5 \end{aligned}$$

- Estimate:

$$\begin{aligned} x_i &= \pi_1 + \pi_2 * z_i + \underbrace{\gamma_i}_{=\delta a_i + \eta_i} \implies \hat{x}_i = \hat{\pi}_1 + \hat{\pi}_2 * z_i \\ y_i &= \alpha_1 + \beta * \hat{x}_i + \underbrace{u_i}_{=a_i + \varepsilon_i} \end{aligned}$$

- Prediction:

Monte Carlo Experiment — Theoretically

- True model:

$$\begin{aligned} y_i &= \alpha_1 + \beta x_i + a_i + \varepsilon_i \\ x_i &= \alpha_2 + \delta a_i + z_i + \eta_i \\ \delta &= 0.5 \end{aligned}$$

- Estimate:

$$\begin{aligned} x_i &= \pi_1 + \pi_2 * z_i + \underbrace{\gamma_i}_{=\delta a_i + \eta_i} \implies \hat{x}_i = \hat{\pi}_1 + \hat{\pi}_2 * z_i \\ y_i &= \alpha_1 + \beta * \hat{x}_i + \underbrace{u_i}_{=a_i + \varepsilon_i} \end{aligned}$$

- Prediction: **2SLS estimator should be unbiased**

Monte Carlo Experiment — Specification

$$\begin{aligned} \eta &\sim N(0,1) \\ \varepsilon_i &\sim N(0,5) \\ a_i &\sim N(0,5) \\ z_i &\sim N(0,2) \end{aligned}$$

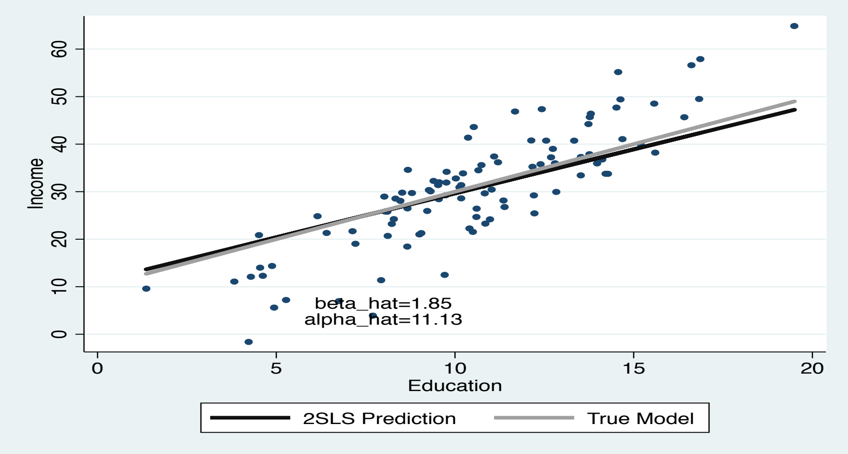
$$\begin{aligned} x_i &= \alpha_2 + \delta * a_i + z_i + \eta_i \\ y_i &= \alpha_1 + \beta * x_i + a_i + \varepsilon_i \end{aligned}$$

Table: Summary statistics

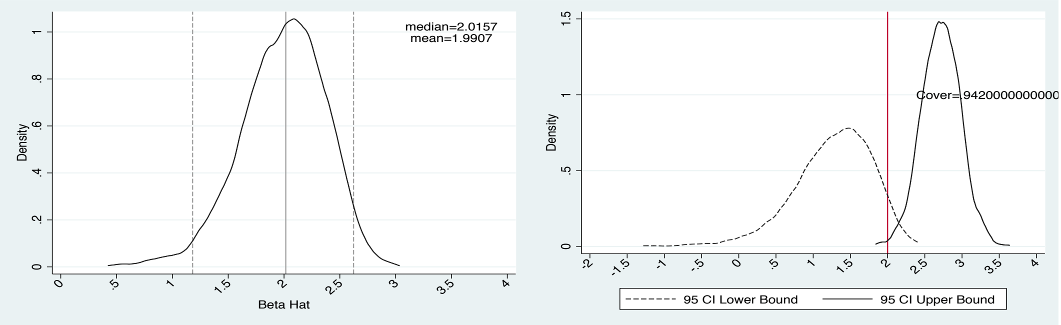
Variable	**Mean**	**Std. Dev.**	**Min.**	**Max.**	**N**
y	30.581	12.304	-1.62	64.849	100
x	10.502	3.345	1.364	19.494	100
a	0.248	5.039	-13.628	10.604	100
eps	-0.672	4.947	-13.5	10.32	100
eta	0.002	1.02	-2.068	2.252	100
z	0.376	2.167	-4.5	5.626	100

Monte Carlo Experiment — Results

Estimation equation: $y_i = \alpha + \beta \hat{x}_i + \underbrace{u_i}_{=a_i + \varepsilon_i}$



Monte Carlo Experiment — Results



Monte Carlo Experiment — Results

	First Stage	Second Stage		
	x	y	y	y
z	0.88^{***}			
	(0.13)			
x		3.16^{***}	1.97^{***}	
		(0.19)	(0.21)	
a			1.11^{***}	
			(0.14)	
$\hat{x}$				1.85^{***}
				(0.62)
Constant	10.17^{***}	-2.61	9.59^{***}	11.13^{*}
	(0.28)	(2.09)	(2.26)	(6.62)
# Observations	100	100	100	100
R squared	0.329	0.738	0.840	0.083
Mean Dep. Var	10.502	30.581	30.581	30.581

What To Do When $E[\varepsilon|X] \neq 0$?

Randomized Control Trials (RCTs)

Principle

- In the previous section:
  - We introduced the principle of instrumental variables
  - We discussed cases where a possibly relevant z arises in the data

Principle

- In the previous section:
  - We introduced the principle of instrumental variables
  - We discussed cases where a possibly relevant z arises in the data

- But **researchers can also create their own** z!

Principle

- In the previous section:
  - We introduced the principle of instrumental variables
  - We discussed cases where a possibly relevant z arises in the data
- But **researchers can also create their own** z!
- E.g., effect of disposable income on consumption expenditures[1]
  - Many possible confounders: Type of job, education, household composition, etc.
  - Solution?

Principle

- In the previous section:
  - We introduced the principle of instrumental variables
  - We discussed cases where a possibly relevant z arises in the data

- But **researchers can also create their own** z!

- E.g., effect of disposable income on consumption expenditures[1]
  - Many possible confounders: Type of job, education, household composition, etc.
  - Solution? Randomize (part of) disposable income thanks to direct transfers

Principle

- In the previous section:
  - We introduced the principle of instrumental variables
  - We discussed cases where a possibly relevant z arises in the data

- But **researchers can also create their own** z!

- E.g., effect of disposable income on consumption expenditures[1]
  - Many possible confounders: Type of job, education, household composition, etc.
  - Solution? Randomize (part of) disposable income thanks to direct transfers

- Sounds crazy?
  - Boehm, Fize, and Jaravel (2025) spent around €300,000
  - Some customers of a bank receive €300, some don't; allocation is random
  - Use the transfer as an instrument for income and recover the desired parameter

In Practice

- What we call **Randomized Control Trials (RCTs)**

In Practice

- What we call **Randomized Control Trials (RCTs)**
- Often used in Development Economics, Labor Economics, etc.
- 2019 Nobel Prize to Esther Duflo, Abhijit Banerjee, and Michael Kremer

In Practice

- What we call **Randomized Control Trials (RCTs)**

- Often used in Development Economics, Labor Economics, etc.
- 2019 Nobel Prize to Esther Duflo, Abhijit Banerjee, and Michael Kremer

- In terms of identification, the best we can do

In Practice

- What we call **Randomized Control Trials (RCTs)**

- Often used in Development Economics, Labor Economics, etc.
- 2019 Nobel Prize to Esther Duflo, Abhijit Banerjee, and Michael Kremer

- In terms of identification, the best we can do

- Potential limitations:
  - Feasibility / Cost
  - Ethical concerns: € transfers are OK, but could we randomize education choices?
  - **External validity**

What To Do When $E[\varepsilon|X] \neq 0$?

Fixed Effects

Context

- We now move **from cross-sectional to panel data**
  - Instead of solely comparing units i (individuals, countries, cities, etc.) once in time
  - We observe different units i over time t

Context

- We now move **from cross-sectional to panel data**
  - Instead of solely comparing units i (individuals, countries, cities, etc.) once in time
  - We observe different units i over time t

- Suppose the true model is

$$y_{it} = \alpha + \beta x_{it} + \underbrace{a_i + \varepsilon_{it}}_{u_{it}} \text{ with } E\left[\varepsilon_{it} | x_{it}\right] = 0 \text{ and } E\left[u_{it} | x_{it}\right] \neq 0$$

Some **unobserved, time-invariant heterogeneity** a_i correlates with x_{it}

Context

- We now move **from cross-sectional to panel data**
  - Instead of solely comparing units i (individuals, countries, cities, etc.) once in time
  - We observe different units i over time t
- Suppose the true model is

$$y_{it} = \alpha + \beta x_{it} + \underbrace{a_i + \varepsilon_{it}}_{u_{it}} \text{ with } E\left[\varepsilon_{it}|x_{it}\right] = 0 \text{ and } E\left[u_{it}|x_{it}\right] \neq 0$$

  Some **unobserved, time-invariant heterogeneity** a_i correlates with x_{it}

- OLS is biased when estimated on:

$$y_{it} \quad = \quad \alpha + \beta * x_{it} + u_{it}$$

Two Main Approaches

- **Within estimator**:
  - Let: $\bar{y}_{it} = \frac{1}{T}\sum_{t=1}^{T}\left(\alpha + \beta x_{it} + \underbrace{a_i + \varepsilon_{it}}_{u_{it}}\right) = \alpha + \beta \bar{x}_{it} + a_i + \bar{\varepsilon}_{it}$

Two Main Approaches

- **Within estimator**:
  - Let: $\bar{y}_{it} = \frac{1}{T}\sum_{t=1}^{T}\left(\alpha + \beta x_{it} + \underbrace{a_i + \varepsilon_{it}}_{u_{it}}\right) = \alpha + \beta \bar{x}_{it} + a_i + \bar{\varepsilon}_{it}$
  - Then: $y_{it} - \bar{y}_{it} = \beta * (x_{it} - \bar{x}_{it}) + (\varepsilon_{it} - \bar{\varepsilon}_{it})$

Two Main Approaches

- **Within estimator**:
  - Let: $\bar{y}_{it} = \frac{1}{T}\sum_{t=1}^{T}\left(\alpha + \beta x_{it} + \underbrace{a_i + \varepsilon_{it}}_{u_{it}}\right) = \alpha + \beta \bar{x}_{it} + a_i + \bar{\varepsilon}_{it}$
  - Then: $y_{it} - \bar{y}_{it} = \beta * (x_{it} - \bar{x}_{it}) + (\varepsilon_{it} - \bar{\varepsilon}_{it})$
  - And exogeneity condition is satisfied: $E(\varepsilon_{it} - \bar{\varepsilon}_{it} | x_{it} - \bar{x}_{it}) = 0$

Two Main Approaches

- **Within estimator**:
  - Let: $\bar{y}_{it} = \frac{1}{T}\sum_{t=1}^{T}\left(\alpha + \beta x_{it} + \underbrace{a_i + \varepsilon_{it}}_{u_{it}}\right) = \alpha + \beta\bar{x}_{it} + a_i + \bar{\varepsilon}_{it}$
  - Then: $y_{it} - \bar{y}_{it} = \beta * (x_{it} - \bar{x}_{it}) + (\varepsilon_{it} - \bar{\varepsilon}_{it})$
  - And exogeneity condition is satisfied: $E(\varepsilon_{it} - \bar{\varepsilon}_{it} | x_{it} - \bar{x}_{it}) = 0$

- **First-difference estimator**:
  - Consider instead:

$$\Delta y_{it} = y_{it} - y_{it-1} = \beta\Delta x_{it} + \Delta\varepsilon_{it}$$

Two Main Approaches

- **Within estimator**:
  - Let: $\bar{y}_{it} = \frac{1}{T}\sum_{t=1}^{T}\left(\alpha + \beta x_{it} + \underbrace{a_i + \varepsilon_{it}}_{u_{it}}\right) = \alpha + \beta \bar{x}_{it} + a_i + \bar{\varepsilon}_{it}$
  - Then: $y_{it} - \bar{y}_{it} = \beta * (x_{it} - \bar{x}_{it}) + (\varepsilon_{it} - \bar{\varepsilon}_{it})$
  - And exogeneity condition is satisfied: $E(\varepsilon_{it} - \bar{\varepsilon}_{it} | x_{it} - \bar{x}_{it}) = 0$

- **First-difference estimator**:
  - Consider instead:

$$\Delta y_{it} = y_{it} - y_{it-1} = \beta \Delta x_{it} + \Delta \varepsilon_{it}$$

  - Again, exogeneity condition is satisfied: $E\left(\Delta \varepsilon_{it} | \Delta x_{it}\right) = 0$

Fixed Effects

- First-difference estimator

$$\Delta y_{it} \quad = \quad \beta * \Delta x_{it} + \Delta \varepsilon_{it}$$

- Is functionally equivalent to estimating **fixed effect model**:

$$y_{it} \quad = \quad \alpha + \beta x_{it} + \sum_{j=1}^{N} \alpha_j \mathbb{1}\{j = i\} + \varepsilon_{it}$$

  I.e., controlling for a set of dummy variables for all units $\{\mathbb{1}\{j = i\}\}_{j \in [1;N]}$

- **NB**: To lighten notations, we often write that we estimate the model:

$$y_{it} \quad = \quad \alpha + \beta x_{it} + \alpha_i + \varepsilon_{it}$$

What To Do When $E[\varepsilon|X] \neq 0$?

Difference-in-Differences

Context

- Consider again panel data; suppose the true model is:

$$y_{it} = \alpha + \beta * x_{it} + \underbrace{a_i + \gamma_t + \varepsilon_{it}}_{u_{it}} \text{ with } E\left[\varepsilon_{it}|x_{it}\right] = 0 \text{ and } E\left[u_{it}|x_{it}\right] \neq 0$$

Compared with the above, **unobserved trend** γ_t may correlate with x_{it}

Context

- Consider again panel data; suppose the true model is:

$$y_{it} = \alpha + \beta * x_{it} + \underbrace{a_i + \gamma_t + \varepsilon_{it}}_{u_{it}} \text{ with } E\left[\varepsilon_{it}|x_{it}\right] = 0 \text{ and } E\left[u_{it}|x_{it}\right] \neq 0$$

  Compared with the above, **unobserved trend** γ_t may correlate with x_{it}

- E.g., think about health (y_{it}) and pollution (x_{it})
  - Some regions are most developed industrially, which drives pollution and health (a_i)
  - COVID (part of γ_t) affects all regions with an impact on x_{it} and y_{it}

Context

- Consider again panel data; suppose the true model is:

$$y_{it} = \alpha + \beta * x_{it} + \underbrace{a_i + \gamma_t + \varepsilon_{it}}_{u_{it}} \text{ with } E\left[\varepsilon_{it} | x_{it}\right] = 0 \text{ and } E\left[u_{it} | x_{it}\right] \neq 0$$

  Compared with the above, **unobserved trend** γ_t may correlate with x_{it}

- E.g., think about health (y_{it}) and pollution (x_{it})
  - Some regions are most developed industrially, which drives pollution and health (a_i)
  - COVID (part of γ_t) affects all regions with an impact on x_{it} and y_{it}

- Again, OLS is biased when estimated on: $y_{it} = \alpha + \beta * x_{it} + u_{it}$

Spirit

- Take the first difference along the time dimension:

$$\Delta y_{it} = \beta \Delta x_{it} + \Delta \gamma_t + \Delta \varepsilon_{it}$$

- Take another difference between units:

$$\Delta y_{it} - \Delta y_{jt} = \beta \left(\Delta x_{it} - \Delta x_{jt} \right) + \left(\Delta \varepsilon_{it} - \Delta \varepsilon_{jt} \right)$$

- Applying OLS to this model yields an unbiased estimator of β if:

$$E \left(\Delta \varepsilon_{it} - \Delta \varepsilon_{jt} \middle| \Delta x_{it} - \Delta x_{jt} \right) = 0$$

In Practice

- Often used when x_{it} is a binary treatment (e.g., getting a subsidy or not)
  - Treatment group (T) receives the treatment past a given point in time
  - Control group (C) never receives the treatment

In Practice

- Often used when x_{it} is a binary treatment (e.g., getting a subsidy or not)
  - Treatment group (T) receives the treatment past a given point in time
  - Control group (C) never receives the treatment

- Method = Compare the T-C difference after the treatment with that before

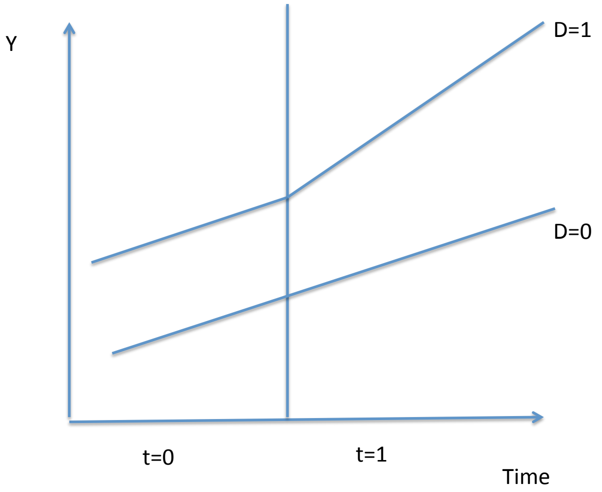
In Practice

- Often used when x_{it} is a binary treatment (e.g., getting a subsidy or not)
  - Treatment group (T) receives the treatment past a given point in time
  - Control group (C) never receives the treatment
- Method = Compare the T-C difference after the treatment with that before
- **Assumption**: Absent the treatment, both groups evolve similarly in post-period
  - Fundamental **"parallel trends"** assumption that researchers must defend

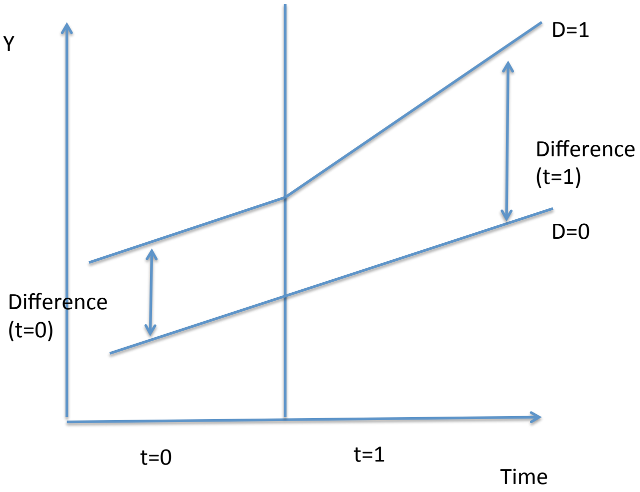
In Practice

- Often used when x_{it} is a binary treatment (e.g., getting a subsidy or not)
  - Treatment group (T) receives the treatment past a given point in time
  - Control group (C) never receives the treatment
- Method = Compare the T-C difference after the treatment with that before
- **Assumption**: Absent the treatment, both groups evolve similarly in post-period
  - Fundamental **"parallel trends"** assumption that researchers must defend
- Can we test this assumption?
  - We can check that both groups evolve similarly before the treatment
  - But we cannot test this after the treatment is deployed!

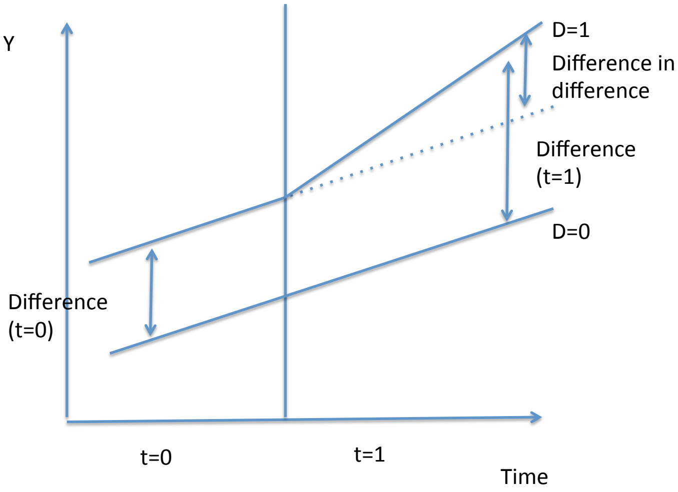
Graphical Illustration



Graphical Illustration



Graphical Illustration



Review Question

Review Question

- You want to learn about the relationship between temperature and GDP. You have a cross section for a single year with the GDP and the average annual temperature for many different countries.

Review Question

- You want to learn about the relationship between temperature and GDP. You have a cross section for a single year with the GDP and the average annual temperature for many different countries.
  - Under what condition is OLS unbiased?

Review Question

- You want to learn about the relationship between temperature and GDP. You have a cross section for a single year with the GDP and the average annual temperature for many different countries.
  - Under what condition is OLS unbiased?
  - Is this condition likely to be met?

Review Question

- You want to learn about the relationship between temperature and GDP. You have a cross section for a single year with the GDP and the average annual temperature for many different countries.
  - Under what condition is OLS unbiased?
  - Is this condition likely to be met?
  - What can you do to recover the causal impact of temperature on GDP?

References

- Joshua, D. Angrist. MOSTLY HARMLESS ECONOMETRICS. 2009.