# Day 1: Identification by Design

Peter Hull

Design-Based Regression Inference
Fall 2024

# The Design of This Course

- This is a three-day intensive in design-based causal inference
  - Far from comprehensive: will focus on core concepts with regression/IV
  - Emphasis will be on practical lessons for applied work
  - I will assume you all have a solid foundation in the basics of causal inference (e.g. Scott's intro courses)

# The Design of This Course

- This is a three-day intensive in design-based causal inference
  - Far from comprehensive: will focus on core concepts with regression/IV
  - Emphasis will be on practical lessons for applied work
  - I will assume you all have a solid foundation in the basics of causal inference (e.g. Scott's intro courses)
- 6-7 hours of lecture, two 30-minute coding demonstrations
  - Please ask questions in the Discord chat!
  - I will try to stick to the schedule but may improvise slightly
  - Code assignments will be "take home," w/solutions the following class

# The Design of This Course

- This is a three-day intensive in design-based causal inference
  - Far from comprehensive: will focus on core concepts with regression/IV
  - Emphasis will be on practical lessons for applied work
  - I will assume you all have a solid foundation in the basics of causal inference (e.g. Scott's intro courses)
- 6-7 hours of lecture, two 30-minute coding demonstrations
  - Please ask questions in the Discord chat!
  - I will try to stick to the schedule but may improvise slightly
  - Code assignments will be "take home," w/solutions the following class
- Feedback/follow-up: *peter_hull@brown.edu*

# Course Schedule

**Schedule**

| | | |
|---|---|---|
| Monday 9/9 | 6:00-7:50pm | Lecture 1: Selection-on-Observables |
| | 6:50-7:00pm | *Break* |
| | 7:00-7:50pm | Lecture 2: Design vs. Outcome Models |
| | 7:50-8:00pm | *Break* |
| | 8:00-8:50pm | Lecture 3: Design-Based IV |
| | 8:50-9:00pm | Application 1 Overview |
| Wednesday 9/11 | 6:00-6:30pm | Live-Coding Application 1 |
| | 6:30-6:40pm | *Break* |
| | 6:40-7:40pm | Lecture 4: Negative Weights |
| | 7:40-7:50pm | *Break* |
| | 7:50-8:50pm | Lecture 5: Clustering |
| | 8:50-9:00pm | Application 2 Overview |
| Friday 9/13 | 6:00-6:30pm | Live-Coding Application 2 |
| | 6:30-6:40pm | *Break* |
| | 6:40-7:40pm | Lecture 6: Recentering |
| | 7:40-7:50pm | *Break* |
| | 7:50-9:00pm | Lecture 7: Nonlinear Models |

# The Design *in* This Course

- Design-based methods use knowledge on the assignment process of as-if-randomly assigned shocks in order to estimate causal effects
  - Mimic analysis of "true" experiments, w/known randomization protocol
  - Contrasts with identification strategies that model untreated potential outcomes (e.g. parallel trends) without appealing to randomization

# The Design *in* This Course

- Design-based methods use knowledge on the assignment process of as-if-randomly assigned shocks in order to estimate causal effects
  - Mimic analysis of "true" experiments, w/known randomization protocol
  - Contrasts with identification strategies that model untreated potential outcomes (e.g. parallel trends) without appealing to randomization
- Design-based methods have several practical advantages:

# The Design *in* This Course

- Design-based methods use knowledge on the assignment process of as-if-randomly assigned shocks in order to estimate causal effects
  - Mimic analysis of "true" experiments, w/known randomization protocol
  - Contrasts with identification strategies that model untreated potential outcomes (e.g. parallel trends) without appealing to randomization

- Design-based methods have several practical advantages:
  1. A large class of robust estimators, which work without restricting how unobservables (e.g. potential outcomes) relate to observables

# The Design *in* This Course

- Design-based methods use knowledge on the assignment process of as-if-randomly assigned shocks in order to estimate causal effects
  - Mimic analysis of "true" experiments, w/known randomization protocol
  - Contrasts with identification strategies that model untreated potential outcomes (e.g. parallel trends) without appealing to randomization

- Design-based methods have several practical advantages:
  1. A large class of robust estimators, which work without restricting how unobservables (e.g. potential outcomes) relate to observables
  2. Robust regression/IV estimation, avoiding "negative weight" issues

# The Design *in* This Course

- Design-based methods use knowledge on the assignment process of as-if-randomly assigned shocks in order to estimate causal effects
  - Mimic analysis of "true" experiments, w/known randomization protocol
  - Contrasts with identification strategies that model untreated potential outcomes (e.g. parallel trends) without appealing to randomization

- Design-based methods have several practical advantages:
  1. A large class of robust estimators, which work without restricting how unobservables (e.g. potential outcomes) relate to observables
  2. Robust regression/IV estimation, avoiding "negative weight" issues
  3. Clear criteria for how to pick controls and cluster standard errors

# The Design *in* This Course

- Design-based methods use knowledge on the assignment process of as-if-randomly assigned shocks in order to estimate causal effects
  - Mimic analysis of "true" experiments, w/known randomization protocol
  - Contrasts with identification strategies that model untreated potential outcomes (e.g. parallel trends) without appealing to randomization

- Design-based methods have several practical advantages:
  1. A large class of robust estimators, which work without restricting how unobservables (e.g. potential outcomes) relate to observables
  2. Robust regression/IV estimation, avoiding "negative weight" issues
  3. Clear criteria for how to pick controls and cluster standard errors
  4. Flexibility in leveraging exogenous shocks with known "formulas"

# The Design *in* This Course

- Design-based methods use knowledge on the assignment process of as-if-randomly assigned shocks in order to estimate causal effects
  - Mimic analysis of "true" experiments, w/known randomization protocol
  - Contrasts with identification strategies that model untreated potential outcomes (e.g. parallel trends) without appealing to randomization

- Design-based methods have several practical advantages:
  1. A large class of robust estimators, which work without restricting how unobservables (e.g. potential outcomes) relate to observables
  2. Robust regression/IV estimation, avoiding "negative weight" issues
  3. Clear criteria for how to pick controls and cluster standard errors
  4. Flexibility in leveraging exogenous shocks with known "formulas"
  5. Clear(er) role of nonlinear/structural models as extrapolation devices

# The Design *in* This Course

- Design-based methods use knowledge on the assignment process of as-if-randomly assigned shocks in order to estimate causal effects
  - Mimic analysis of "true" experiments, w/known randomization protocol
  - Contrasts with identification strategies that model untreated potential outcomes (e.g. parallel trends) without appealing to randomization

- Design-based methods have several practical advantages:
  1. A large class of robust estimators, which work without restricting how unobservables (e.g. potential outcomes) relate to observables
  2. Robust regression/IV estimation, avoiding "negative weight" issues
  3. Clear criteria for how to pick controls and cluster standard errors
  4. Flexibility in leveraging exogenous shocks with known "formulas"
  5. Clear(er) role of nonlinear/structural models as extrapolation devices

- We'll get into all of this over the next few days, building up slowly...

# Outline

1. Preliminaries / Regression Recap

2. Selection on Observables

3. Design vs. Outcome Models

4. Design-Based IV

# Preliminaries: Parameters, Estimands, and Estimators

Three distinct objects, not always clearly distinguished:

# Preliminaries: Parameters, Estimands, and Estimators

Three distinct objects, not always clearly distinguished:

- **Parameters** come from economic (or other) models of the world
  - E.g. a "structural" model of supply and demand, or a potential outcome model relating schooling to earnings
  - They set the target for empirical analyses: what we want to know

# Preliminaries: Parameters, Estimands, and Estimators

Three distinct objects, not always clearly distinguished:

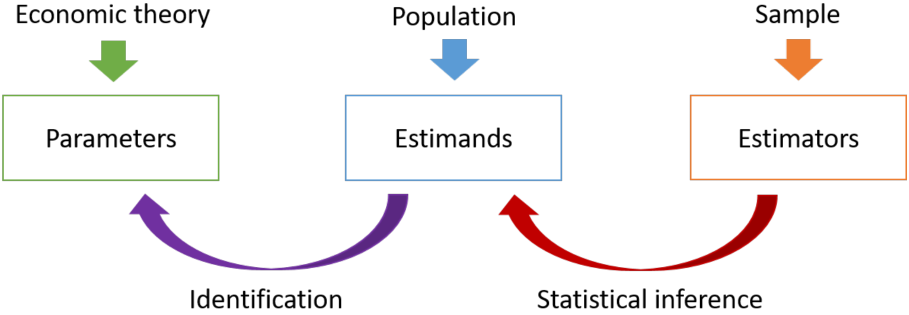
- **Parameters** come from economic (or other) models of the world
  - E.g. a "structural" model of supply and demand, or a potential outcome model relating schooling to earnings
  - They set the target for empirical analyses: what we want to know
- **Estimands** are functions of the population data distribution
  - E.g. a difference in means or ratio of population regression coef's
  - We make assumptions to link parameters & estimands (identification)

# Preliminaries: Parameters, Estimands, and Estimators

Three distinct objects, not always clearly distinguished:

- **Parameters** come from economic (or other) models of the world
  - E.g. a "structural" model of supply and demand, or a potential outcome model relating schooling to earnings
  - They set the target for empirical analyses: what we want to know
- **Estimands** are functions of the population data distribution
  - E.g. a difference in means or ratio of population regression coef's
  - We make assumptions to link parameters & estimands (identification)
- **Estimators** are functions of observed data (i.e. the "sample")
  - E.g. a difference in sample means or ratio of OLS coefficients
  - Since data are random, so are estimators. Each has a distribution
  - We use knowledge of estimator distributions to learn about estimands (inference) and thus identified parameters

# The Lay of the Land




Separating out the different kinds of tasks in identification vs. inference can help make our lives easier!

# Example: The Simplest Experimental Story

- Throughout today, we'll consider the goal of estimating **parameter** $\beta$ in the constant-effects causal model

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

$(y_i, x_i)$ are the observed outcome/treatment; $\varepsilon_i$ is the unobserved "untreated" potential outcome (i.e. the value of $y_i$ if we set $x_i$ to 0)

# Example: The Simplest Experimental Story

- Throughout today, we'll consider the goal of estimating **parameter** $\beta$ in the constant-effects causal model

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

  $(y_i, x_i)$ are the observed outcome/treatment; $\varepsilon_i$ is the unobserved "untreated" potential outcome (i.e. the value of $y_i$ if we set $x_i$ to 0)
  - On Wednesday: heterogeneous effects / multiple treatments

# Example: The Simplest Experimental Story

- Throughout today, we'll consider the goal of estimating **parameter** $\beta$ in the constant-effects causal model

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

  $(y_i, x_i)$ are the observed outcome/treatment; $\varepsilon_i$ is the unobserved "untreated" potential outcome (i.e. the value of $y_i$ if we set $x_i$ to 0)
  - On Wednesday: heterogeneous effects / multiple treatments
- Suppose $x_i$ is drawn randomly in a simple experiment: $x_i \mid \varepsilon \overset{iid}{\sim} F_x$

# Example: The Simplest Experimental Story

- Throughout today, we'll consider the goal of estimating **parameter** $\beta$ in the constant-effects causal model

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

  $(y_i, x_i)$ are the observed outcome/treatment; $\varepsilon_i$ is the unobserved "untreated" potential outcome (i.e. the value of $y_i$ if we set $x_i$ to 0)
  - On Wednesday: heterogeneous effects / multiple treatments
- Suppose $x_i$ is drawn randomly in a simple experiment: $x_i \mid \varepsilon \stackrel{iid}{\sim} F_x$
  - Treatment is random, so knowing $\varepsilon_i$ doesn't help you predict it

# Example: The Simplest Experimental Story

- Throughout today, we'll consider the goal of estimating **parameter** $\beta$ in the constant-effects causal model

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

  $(y_i, x_i)$ are the observed outcome/treatment; $\varepsilon_i$ is the unobserved "untreated" potential outcome (i.e. the value of $y_i$ if we set $x_i$ to 0)
    - On Wednesday: heterogeneous effects / multiple treatments
- Suppose $x_i$ is drawn randomly in a simple experiment: $x_i \mid \varepsilon \overset{iid}{\sim} F_x$
    - Treatment is random, so knowing $\varepsilon_i$ doesn't help you predict it
    - $F_x$ is given by the experimental protocol (e.g. $x_i \sim Bernoulli(0.5)$); note: doesn't vary with $i$

# Example: The Simplest Experimental Story

- Throughout today, we'll consider the goal of estimating **parameter** $\beta$ in the constant-effects causal model

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

  $(y_i, x_i)$ are the observed outcome/treatment; $\varepsilon_i$ is the unobserved "untreated" potential outcome (i.e. the value of $y_i$ if we set $x_i$ to 0)
    - On Wednesday: heterogeneous effects / multiple treatments
- Suppose $x_i$ is drawn randomly in a simple experiment: $x_i \mid \varepsilon \overset{iid}{\sim} F_x$
    - Treatment is random, so knowing $\varepsilon_i$ doesn't help you predict it
    - $F_x$ is given by the experimental protocol (e.g. $x_i \sim Bernoulli(0.5)$); note: doesn't vary with $i$
- Randomization makes $\beta$ coincide with the regression **estimand**

$$\beta^{OLS} = \frac{Cov(y_i, x_i)}{Var(x_i)} =$$

# Example: The Simplest Experimental Story

- Throughout today, we'll consider the goal of estimating **parameter** $\beta$ in the constant-effects causal model

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

  $(y_i, x_i)$ are the observed outcome/treatment; $\varepsilon_i$ is the unobserved "untreated" potential outcome (i.e. the value of $y_i$ if we set $x_i$ to 0)
    - On Wednesday: heterogeneous effects / multiple treatments
- Suppose $x_i$ is drawn randomly in a simple experiment: $x_i \mid \varepsilon \overset{iid}{\sim} F_x$
    - Treatment is random, so knowing $\varepsilon_i$ doesn't help you predict it
    - $F_x$ is given by the experimental protocol (e.g. $x_i \sim Bernoulli(0.5)$); note: doesn't vary with $i$
- Randomization makes $\beta$ coincide with the regression **estimand**

$$\beta^{OLS} = \frac{Cov(y_i, x_i)}{Var(x_i)} = \frac{Cov(\beta x_i + \varepsilon_i, x_i)}{Var(x_i)} =$$

# Example: The Simplest Experimental Story

- Throughout today, we'll consider the goal of estimating **parameter** $\beta$ in the constant-effects causal model

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

  $(y_i, x_i)$ are the observed outcome/treatment; $\varepsilon_i$ is the unobserved "untreated" potential outcome (i.e. the value of $y_i$ if we set $x_i$ to 0)
  - On Wednesday: heterogeneous effects / multiple treatments
- Suppose $x_i$ is drawn randomly in a simple experiment: $x_i \mid \varepsilon \overset{iid}{\sim} F_x$
  - Treatment is random, so knowing $\varepsilon_i$ doesn't help you predict it
  - $F_x$ is given by the experimental protocol (e.g. $x_i \sim Bernoulli(0.5)$); note: doesn't vary with $i$
- Randomization makes $\beta$ coincide with the regression **estimand**

$$\beta^{OLS} = \frac{Cov(y_i, x_i)}{Var(x_i)} = \frac{Cov(\beta x_i + \varepsilon_i, x_i)}{Var(x_i)} = \beta + \frac{Cov(x_i, \varepsilon_i)}{Var(x_i)} =$$

# Example: The Simplest Experimental Story

- Throughout today, we'll consider the goal of estimating **parameter** $\beta$ in the constant-effects causal model

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

  $(y_i, x_i)$ are the observed outcome/treatment; $\varepsilon_i$ is the unobserved "untreated" potential outcome (i.e. the value of $y_i$ if we set $x_i$ to 0)
  - On Wednesday: heterogeneous effects / multiple treatments
- Suppose $x_i$ is drawn randomly in a simple experiment: $x_i \mid \varepsilon \overset{iid}{\sim} F_x$
  - Treatment is random, so knowing $\varepsilon_i$ doesn't help you predict it
  - $F_x$ is given by the experimental protocol (e.g. $x_i \sim Bernoulli(0.5)$); note: doesn't vary with $i$
- Randomization makes $\beta$ coincide with the regression **estimand**

$$\beta^{OLS} = \frac{Cov(y_i, x_i)}{Var(x_i)} = \frac{Cov(\beta x_i + \varepsilon_i, x_i)}{Var(x_i)} = \beta + \frac{Cov(x_i, \varepsilon_i)}{Var(x_i)} = \beta$$

# Example: The Simplest Experimental Story

- Throughout today, we'll consider the goal of estimating **parameter** $\beta$ in the constant-effects causal model

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

  $(y_i, x_i)$ are the observed outcome/treatment; $\varepsilon_i$ is the unobserved "untreated" potential outcome (i.e. the value of $y_i$ if we set $x_i$ to 0)
  - On Wednesday: heterogeneous effects / multiple treatments
- Suppose $x_i$ is drawn randomly in a simple experiment: $x_i \mid \varepsilon \stackrel{iid}{\sim} F_x$
  - Treatment is random, so knowing $\varepsilon_i$ doesn't help you predict it
  - $F_x$ is given by the experimental protocol (e.g. $x_i \sim Bernoulli(0.5)$); note: doesn't vary with $i$
- Randomization makes $\beta$ coincide with the regression **estimand**

$$\beta^{OLS} = \frac{Cov(y_i, x_i)}{Var(x_i)} = \frac{Cov(\beta x_i + \varepsilon_i, x_i)}{Var(x_i)} = \beta + \frac{Cov(x_i, \varepsilon_i)}{Var(x_i)} = \beta$$

- We can estimate $\beta^{OLS}$ with the OLS **estimator**, $\hat{\beta}^{OLS} = \frac{\widehat{Cov}(y_i, x_i)}{\widehat{Var}(x_i)}$

# Inference vs. Identification

- Under mild conditions, the OLS estimator gets arbitrarily "close" to the regression estimand as the sample grows (i.e., $\hat{\beta}^{OLS} \xrightarrow{p} \beta^{OLS}$)
  - Moreover, the errors $\hat{\beta}^{OLS} - \beta^{OLS}$ approximately follow a known distribution (i.e. $\mathrm{N}(0, \hat{SE}^2)$, where $\hat{SE}$ is the robust standard errors)
  - We can use this to conduct inference (e.g. 95% CI) on $\beta^{OLS}$

# Inference vs. Identification

- Under mild conditions, the OLS estimator gets arbitrarily "close" to the regression estimand as the sample grows (i.e., $\hat{\beta}^{OLS} \xrightarrow{p} \beta^{OLS}$)
  - Moreover, the errors $\hat{\beta}^{OLS} - \beta^{OLS}$ approximately follow a known distribution (i.e. $\mathrm{N}(0, \hat{SE}^2)$, where $\hat{SE}$ is the robust standard errors)
  - We can use this to conduct inference (e.g. 95% CI) on $\beta^{OLS}$

- That's all Stata can tell us when we *reg y x, r*. The rest is up to us
  - Outside of true experiments, we need to ponder whether $Cov(x_i, \varepsilon_i) = 0$ in order to say whether $\beta^{OLS}$ identifies $\beta$

# Inference vs. Identification

- Under mild conditions, the OLS estimator gets arbitrarily "close" to the regression estimand as the sample grows (i.e., $\hat{\beta}^{OLS} \xrightarrow{p} \beta^{OLS}$)
  - Moreover, the errors $\hat{\beta}^{OLS} - \beta^{OLS}$ approximately follow a known distribution (i.e. $\mathrm{N}(0, \hat{SE}^2)$, where $\hat{SE}$ is the robust standard errors)
  - We can use this to conduct inference (e.g. 95% CI) on $\beta^{OLS}$

- That's all Stata can tell us when we *reg y x, r*. The rest is up to us
  - Outside of true experiments, we need to ponder whether $Cov(x_i, \varepsilon_i) = 0$ in order to say whether $\beta^{OLS}$ identifies $\beta$
  - Selection bias: units with higher untreated potential outcomes $\varepsilon_i$ tend to be more/less likely to get higher treatments $x_i$

# Inference vs. Identification

- Under mild conditions, the OLS estimator gets arbitrarily "close" to the regression estimand as the sample grows (i.e., $\hat{\beta}^{OLS} \xrightarrow{p} \beta^{OLS}$)
  - Moreover, the errors $\hat{\beta}^{OLS} - \beta^{OLS}$ approximately follow a known distribution (i.e. $\mathrm{N}(0, \hat{SE}^2)$, where $\hat{SE}$ is the robust standard errors)
  - We can use this to conduct inference (e.g. 95% CI) on $\beta^{OLS}$

- That's all Stata can tell us when we *reg y x, r*. The rest is up to us
  - Outside of true experiments, we need to ponder whether $Cov(x_i, \varepsilon_i) = 0$ in order to say whether $\beta^{OLS}$ identifies $\beta$
  - Selection bias: units with higher untreated potential outcomes $\varepsilon_i$ tend to be more/less likely to get higher treatments $x_i$
  - How can we assess / relax this strong condition?

# Regression Recap

- The **regression** of $y_i$ on $x_i = [x_{1i}, \ldots, x_{Ji}]'$ gives the best (MSE-minimizing) linear approximation to the CEF of $y_i \mid x_i$:

# Regression Recap

- The **regression** of $y_i$ on $x_i = [x_{1i}, \ldots, x_{Ji}]'$ gives the best (MSE-minimizing) linear approximation to the CEF of $y_i \mid x_i$:

$$\beta^{OLS} = \arg\min_b E[(y_i - x_i' b)^2]$$

# Regression Recap

- The **regression** of $y_i$ on $x_i = [x_{1i}, \ldots, x_{Ji}]'$ gives the best (MSE-minimizing) linear approximation to the CEF of $y_i \mid x_i$:

$$\beta^{OLS} = \arg\min_b E[(y_i - x_i' b)^2] = \arg\min_b E[(E[y_i \mid x_i] - x_i' b)^2]$$

# Regression Recap

- The **regression** of $y_i$ on $x_i = [x_{1i}, \dots, x_{Ji}]'$ gives the best (MSE-minimizing) linear approximation to the CEF of $y_i \mid x_i$:

$$\beta^{OLS} = \arg\min_b E[(y_i - x_i'b)^2] = \arg\min_b E[(E[y_i \mid x_i] - x_i'b)^2]$$

- Hence, regression gives the CEF when it is linear: $E[y_i \mid x_i] = x_i'\beta^{OLS}$
    - Leading example:

# Regression Recap

- The **regression** of $y_i$ on $x_i = [x_{1i}, \ldots, x_{Ji}]'$ gives the best (MSE-minimizing) linear approximation to the CEF of $y_i \mid x_i$:

$$\beta^{OLS} = \arg\min_b E[(y_i - x_i'b)^2] = \arg\min_b E[(E[y_i \mid x_i] - x_i'b)^2]$$

- Hence, regression gives the CEF when it is linear: $E[y_i \mid x_i] = x_i'\beta^{OLS}$
  - Leading example: *saturated* regression (e.g. $x_{ji}$ are group dummies)

# Regression Recap

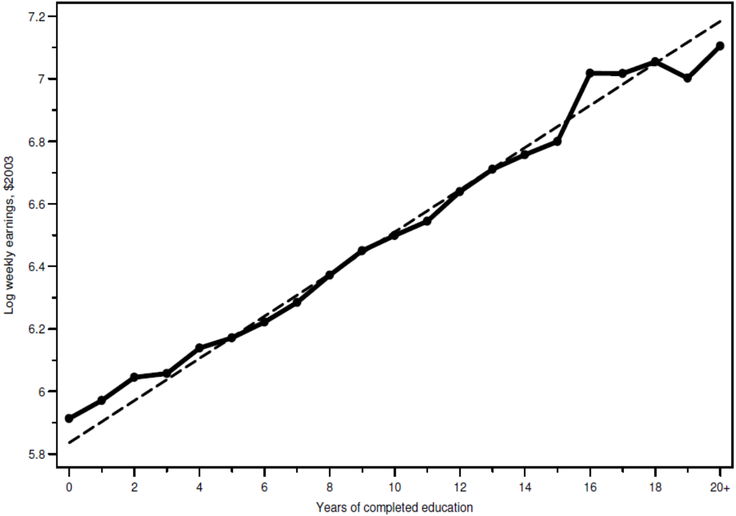
- The **regression** of $y_i$ on $x_i = [x_{1i}, \ldots, x_{Ji}]'$ gives the best (MSE-minimizing) linear approximation to the CEF of $y_i \mid x_i$:

$$\beta^{OLS} = \arg\min_b E[(y_i - x_i' b)^2] = \arg\min_b E[(E[y_i \mid x_i] - x_i' b)^2]$$

- Hence, regression gives the CEF when it is linear: $E[y_i \mid x_i] = x_i' \beta^{OLS}$
  - Leading example: *saturated* regression (e.g. $x_{ji}$ are group dummies)

- By taking FOCs: $\beta^{OLS} = E[x_i x_i']^{-1} E[x_i y_i]$ (note: non-random)
  - OLS estimator: $\hat{\beta}^{OLS} = (\sum_i x_i x_i')^{-1} \sum_i x_i y_i$ (note: random)

# Regression Linearly Approximates the CEF




Notes: CEF and linear regression of average log weekly wages given schooling for white men aged 40-49 from the 1980 IPUMS 5% sample

# Regression Anatomy

- When $x_i = [x_{1i}, 1]'$, the two elements of $E[x_i x_i']^{-1} E[x_i y_i]$ are:
  - Slope $\beta_1^{OLS} = \frac{Cov(x_{1i}, y_i)}{Var(x_{1i})}$; intercept $\beta_2^{OLS} = E[y_i] - \beta_1 E[x_{1i}]$

# Regression Anatomy

- When $x_i = [x_{1i}, 1]'$, the two elements of $E[x_i x_i']^{-1} E[x_i y_i]$ are:
  - Slope $\beta_1^{OLS} = \frac{Cov(x_{1i}, y_i)}{Var(x_{1i})}$; intercept $\beta_2^{OLS} = E[y_i] - \beta_1 E[x_{1i}]$

- The Frisch-Waugh-Lovell (FWL) theorem tells us that, more generally, the $k$-th non-constant slope coefficient is

$$\beta_k^{OLS} = \frac{Cov(\tilde{x}_{ki}, y_i)}{Var(\tilde{x}_{ki})}$$

  where $\tilde{x}_{ki}$ is the residual from regressing $x_{ki}$ on all other elements of $x_i$

# Regression Anatomy

- When $x_i = [x_{1i}, 1]'$, the two elements of $E[x_i x_i']^{-1} E[x_i y_i]$ are:
  - Slope $\beta_1^{OLS} = \frac{Cov(x_{1i}, y_i)}{Var(x_{1i})}$; intercept $\beta_2^{OLS} = E[y_i] - \beta_1 E[x_{1i}]$
- The Frisch-Waugh-Lovell (FWL) theorem tells us that, more generally, the $k$-th non-constant slope coefficient is

$$\beta_k^{OLS} = \frac{Cov(\tilde{x}_{ki}, y_i)}{Var(\tilde{x}_{ki})}$$

  where $\tilde{x}_{ki}$ is the residual from regressing $x_{ki}$ on all other elements of $x_i$
  - Also $\beta_k^{OLS} = \frac{Cov(\tilde{x}_{ki}, \tilde{y}_i)}{Var(\tilde{x}_{ki})}$ where $\tilde{y}_i$ are the analogous residuals of $y_i$

# Regression Anatomy

- When $x_i = [x_{1i}, 1]'$, the two elements of $E[x_i x_i']^{-1} E[x_i y_i]$ are:
  - Slope $\beta_1^{OLS} = \frac{Cov(x_{1i}, y_i)}{Var(x_{1i})}$; intercept $\beta_2^{OLS} = E[y_i] - \beta_1 E[x_{1i}]$

- The Frisch-Waugh-Lovell (FWL) theorem tells us that, more generally, the $k$-th non-constant slope coefficient is

$$\beta_k^{OLS} = \frac{Cov(\tilde{x}_{ki}, y_i)}{Var(\tilde{x}_{ki})}$$

  where $\tilde{x}_{ki}$ is the residual from regressing $x_{ki}$ on all other elements of $x_i$
  - Also $\beta_k^{OLS} = \frac{Cov(\tilde{x}_{ki}, \tilde{y}_i)}{Var(\tilde{x}_{ki})}$ where $\tilde{y}_i$ are the analogous residuals of $y_i$

- Notice: $\tilde{x}_{ki} = x_{ki} - E[x_{ki} \mid x_{\neg k,i}]$ when $E[x_{ki} \mid x_{\neg k,i}]$ is linear...

# Outline

1. Regression/IV Recap✓
2. Selection on Observables
3. Design vs. Outcome Models
4. Design-Based IV

## Stratified Randomization

- Now consider a slightly more complicated experimental design: $x_i \mid w, \varepsilon \overset{iid}{\sim} F_x(w_i)$ where $w_i = \{1, 2, \ldots, K\}$ indexes some strata

# Stratified Randomization

- Now consider a slightly more complicated experimental design: $x_i \mid w, \varepsilon \overset{iid}{\sim} F_x(w_i)$ where $w_i = \{1, 2, \ldots, K\}$ indexes some strata
  - E.g. treatment is more available/rationed in across waves or cites, $k$
  - But still: knowing $\varepsilon_i$ doesn't help predict $x_i$ (as long as you know $w_i$)

# Stratified Randomization

- Now consider a slightly more complicated experimental design: $x_i \mid w, \varepsilon \overset{iid}{\sim} F_x(w_i)$ where $w_i = \{1, 2, \ldots, K\}$ indexes some strata
  - E.g. treatment is more available/rationed in across waves or cites, $k$
  - But still: knowing $\varepsilon_i$ doesn't help predict $x_i$ (as long as you know $w_i$)
- Consider regressing $y_i$ on $x_i$, controlling for strata FE, $w_{ik} = \mathbf{1}[w_i = k]$

# Stratified Randomization

- Now consider a slightly more complicated experimental design: $x_i \mid w, \varepsilon \overset{iid}{\sim} F_x(w_i)$ where $w_i = \{1, 2, \ldots, K\}$ indexes some strata
  - E.g. treatment is more available/rationed in across waves or cites, $k$
  - But still: knowing $\varepsilon_i$ doesn't help predict $x_i$ (as long as you know $w_i$)

- Consider regressing $y_i$ on $x_i$, controlling for strata FE, $w_{ik} = \mathbf{1}[w_i = k]$
  - By FWL, noting that $Cov(\tilde{x}_i, x_i) = Var(\tilde{x}_i)$,

$$\beta^{OLS} = \frac{Cov(\tilde{x}_i, y_i)}{Var(\tilde{x}_i)} = \frac{Cov(\tilde{x}_i, \beta x_i + \varepsilon_i)}{Var(\tilde{x}_i)} = \beta + \frac{Cov(\tilde{x}_i, \varepsilon_i)}{Var(\tilde{x}_i)}$$

# Stratified Randomization

- Now consider a slightly more complicated experimental design: $x_i \mid w, \varepsilon \overset{iid}{\sim} F_x(w_i)$ where $w_i = \{1, 2, \ldots, K\}$ indexes some strata
  - E.g. treatment is more available/rationed in across waves or cites, $k$
  - But still: knowing $\varepsilon_i$ doesn't help predict $x_i$ (as long as you know $w_i$)

- Consider regressing $y_i$ on $x_i$, controlling for strata FE, $w_{ik} = \mathbf{1}[w_i = k]$
  - By FWL, noting that $Cov(\tilde{x}_i, x_i) = Var(\tilde{x}_i)$,

$$\beta^{OLS} = \frac{Cov(\tilde{x}_i, y_i)}{Var(\tilde{x}_i)} = \frac{Cov(\tilde{x}_i, \beta x_i + \varepsilon_i)}{Var(\tilde{x}_i)} = \beta + \frac{Cov(\tilde{x}_i, \varepsilon_i)}{Var(\tilde{x}_i)}$$

  - Moreover, since $E[x_i \mid w_i]$ is linear, $\tilde{x}_i = x_i - E[x_i \mid w_i]$. And by the LIE:

$$Cov(\tilde{x}_i, \varepsilon_i) = E[(x_i - E[x_i \mid w_i])\varepsilon_i] = E[(E[x_i \mid w, \varepsilon] - E[x_i \mid w_i])\varepsilon_i$$

# Stratified Randomization

- Now consider a slightly more complicated experimental design: $x_i \mid w, \varepsilon \stackrel{iid}{\sim} F_x(w_i)$ where $w_i = \{1, 2, \ldots, K\}$ indexes some strata
  - E.g. treatment is more available/rationed in across waves or cites, $k$
  - But still: knowing $\varepsilon_i$ doesn't help predict $x_i$ (as long as you know $w_i$)
- Consider regressing $y_i$ on $x_i$, controlling for strata FE, $w_{ik} = \mathbf{1}[w_i = k]$
  - By FWL, noting that $Cov(\tilde{x}_i, x_i) = Var(\tilde{x}_i)$,

$$\beta^{OLS} = \frac{Cov(\tilde{x}_i, y_i)}{Var(\tilde{x}_i)} = \frac{Cov(\tilde{x}_i, \beta x_i + \varepsilon_i)}{Var(\tilde{x}_i)} = \beta + \frac{Cov(\tilde{x}_i, \varepsilon_i)}{Var(\tilde{x}_i)}$$

  - Moreover, since $E[x_i \mid w_i]$ is linear, $\tilde{x}_i = x_i - E[x_i \mid w_i]$. And by the LIE:

$$Cov(\tilde{x}_i, \varepsilon_i) = E[(x_i - E[x_i \mid w_i])\varepsilon_i] = E[(E[x_i \mid w, \varepsilon] - E[x_i \mid w_i])\varepsilon_i$$

  - Finally, by conditional random assignment, $E[x_i \mid w, \varepsilon] - E[x_i \mid w] = 0$

# Stratified Randomization

- Now consider a slightly more complicated experimental design: $x_i \mid w, \varepsilon \overset{iid}{\sim} F_x(w_i)$ where $w_i = \{1, 2, \ldots, K\}$ indexes some strata
  - E.g. treatment is more available/rationed in across waves or cites, $k$
  - But still: knowing $\varepsilon_i$ doesn't help predict $x_i$ (as long as you know $w_i$)

- Consider regressing $y_i$ on $x_i$, controlling for strata FE, $w_{ik} = \mathbf{1}[w_i = k]$
  - By FWL, noting that $Cov(\tilde{x}_i, x_i) = Var(\tilde{x}_i)$,

$$\beta^{OLS} = \frac{Cov(\tilde{x}_i, y_i)}{Var(\tilde{x}_i)} = \frac{Cov(\tilde{x}_i, \beta x_i + \varepsilon_i)}{Var(\tilde{x}_i)} = \beta + \frac{Cov(\tilde{x}_i, \varepsilon_i)}{Var(\tilde{x}_i)}$$

  - Moreover, since $E[x_i \mid w_i]$ is linear, $\tilde{x}_i = x_i - E[x_i \mid w_i]$. And by the LIE:

$$Cov(\tilde{x}_i, \varepsilon_i) = E[(x_i - E[x_i \mid w_i])\varepsilon_i] = E[(E[x_i \mid w, \varepsilon] - E[x_i \mid w_i])\varepsilon_i$$

  - Finally, by conditional random assignment, $E[x_i \mid w, \varepsilon] - E[x_i \mid w] = 0$
  - Thus, $Cov(\tilde{x}_i, \varepsilon_i) = 0$ and we have identification: $\beta^{OLS} = \beta$

# Selection on Observables

- Design-based regressions in observational data appeal to such experimental ideals:
  1. Claim $x_i$ is as-good-as-randomly assigned conditional on some $w_i$: formally, that $x_i \mid w, \varepsilon \sim F_x(w_i)$

# Selection on Observables

- Design-based regressions in observational data appeal to such experimental ideals:
  1. Claim $x_i$ is as-good-as-randomly assigned conditional on some $w_i$: formally, that $x_i \mid w, \varepsilon \sim F_x(w_i)$
  2. Control flexibly for $w_i$, such that the auxiliary regression of $x_i$ on this estimates $E[x_i \mid w_i]$ ($\implies$ the controlled reg uses $\tilde{x}_i = x_i - E[x_i \mid w_i]$)

# Selection on Observables

- Design-based regressions in observational data appeal to such experimental ideals:
  1. Claim $x_i$ is as-good-as-randomly assigned conditional on some $w_i$: formally, that $x_i \mid w, \varepsilon \sim F_x(w_i)$
  2. Control flexibly for $w_i$, such that the auxiliary regression of $x_i$ on this estimates $E[x_i \mid w_i]$ ($\implies$ the controlled reg uses $\tilde{x}_i = x_i - E[x_i \mid w_i]$)
- Two steps to make design claims convincing:

# Selection on Observables

- Design-based regressions in observational data appeal to such experimental ideals:
  1. Claim $x_i$ is as-good-as-randomly assigned conditional on some $w_i$: formally, that $x_i \mid w, \varepsilon \sim F_x(w_i)$
  2. Control flexibly for $w_i$, such that the auxiliary regression of $x_i$ on this estimates $E[x_i \mid w_i]$ ($\implies$ the controlled reg uses $\tilde{x}_i = x_i - E[x_i \mid w_i]$)

- Two steps to make design claims convincing:
  1. Tell a clear *ex ante* story about where the $x_i \mid w_i$ variation comes from and why it is unlikely to be correlated with $\varepsilon_i$

# Selection on Observables

- Design-based regressions in observational data appeal to such experimental ideals:
  1. Claim $x_i$ is as-good-as-randomly assigned conditional on some $w_i$: formally, that $x_i \mid w, \varepsilon \sim F_x(w_i)$
  2. Control flexibly for $w_i$, such that the auxiliary regression of $x_i$ on this estimates $E[x_i \mid w_i]$ ($\implies$ the controlled reg uses $\tilde{x}_i = x_i - E[x_i \mid w_i]$)

- Two steps to make design claims convincing:
  1. Tell a clear *ex ante* story about where the $x_i \mid w_i$ variation comes from and why it is unlikely to be correlated with $\varepsilon_i$
  2. Use *ex post* balance tests to check that $x_i$ is not correlated, conditional on $w_i$, with other observables that may proxy for $\varepsilon_i$

# Selection on Observables

- Design-based regressions in observational data appeal to such experimental ideals:
  1. Claim $x_i$ is as-good-as-randomly assigned conditional on some $w_i$: formally, that $x_i \mid w, \varepsilon \sim F_x(w_i)$
  2. Control flexibly for $w_i$, such that the auxiliary regression of $x_i$ on this estimates $E[x_i \mid w_i]$ ($\implies$ the controlled reg uses $\tilde{x}_i = x_i - E[x_i \mid w_i]$)
- Two steps to make design claims convincing:
  1. Tell a clear *ex ante* story about where the $x_i \mid w_i$ variation comes from and why it is unlikely to be correlated with $\varepsilon_i$
  2. Use *ex post* balance tests to check that $x_i$ is not correlated, conditional on $w_i$, with other observables that may proxy for $\varepsilon_i$
- Best to use group-dummy $w_i$, such that linear $E[x_i \mid w_i]$ is trivial
  - Otherwise, good to check sensitivity to more flexible control specs (e.g. add interactions or higher-order polynomials)

# Example: Dale and Krueger (2002)

- D&K estimate effects of attending a more selective college (e.g., a private school) on adult earnings
  - They have data on schooling and earnings, as well as information on which colleges individuals applied to and got into

# Example: Dale and Krueger (2002)

- D&K estimate effects of attending a more selective college (e.g., a private school) on adult earnings
  - They have data on schooling and earnings, as well as information on which colleges individuals applied to and got into

- *Ex ante* selection-on-observables story:
  - Conditional on the colleges $i$ applied to / was admitted to, the decision to go to a more elite school is unrelated to latent earnings potential

# Example: Dale and Krueger (2002)

- D&K estimate effects of attending a more selective college (e.g., a private school) on adult earnings
  - They have data on schooling and earnings, as well as information on which colleges individuals applied to and got into

- *Ex ante* selection-on-observables story:
  - Conditional on the colleges $i$ applied to / was admitted to, the decision to go to a more elite school is unrelated to latent earnings potential
  - Formally, private school attendance $x_i$ is independent of potential outcomes $\varepsilon_i$ given a vector of application/admission dummies $w_i$

# Example: Dale and Krueger (2002)

- D&K estimate effects of attending a more selective college (e.g., a private school) on adult earnings
  - They have data on schooling and earnings, as well as information on which colleges individuals applied to and got into

- *Ex ante* selection-on-observables story:
  - Conditional on the colleges $i$ applied to / was admitted to, the decision to go to a more elite school is unrelated to latent earnings potential
  - Formally, private school attendance $x_i$ is independent of potential outcomes $\varepsilon_i$ given a vector of application/admission dummies $w_i$
  - Group dummy controls, so the auxiliary regression estimates $E[x_i \mid w_i]$

# Example: Dale and Krueger (2002)

- D&K estimate effects of attending a more selective college (e.g., a private school) on adult earnings
  - They have data on schooling and earnings, as well as information on which colleges individuals applied to and got into

- *Ex ante* selection-on-observables story:
  - Conditional on the colleges $i$ applied to / was admitted to, the decision to go to a more elite school is unrelated to latent earnings potential
  - Formally, private school attendance $x_i$ is independent of potential outcomes $\varepsilon_i$ given a vector of application/admission dummies $w_i$
  - Group dummy controls, so the auxiliary regression estimates $E[x_i \mid w_i]$

- *Ex post* empirical validation:
  - Conditional on the selection controls, $x_i$ appears uncorrelated with other baseline observables (demographics, etc)

# Dale and Krueger Estimates (from MHE)

| | No Selection Controls | | | Selection Controls | | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Private School | 0.135 | 0.095 | 0.086 | 0.007 | 0.003 | 0.013 |
| | (0.055) | (0.052) | (0.034) | (0.038) | (0.039) | (0.025) |
| Own SAT score/100 | | 0.048 | 0.016 | | 0.033 | 0.001 |
| | | (0.009) | (0.007) | | (0.007) | (0.007) |
| Predicted log(Parental Income) | | | 0.219 | | | 0.190 |
| | | | (0.022) | | | (0.023) |
| Female | | | -0.403 | | | -0.395 |
| | | | (0.018) | | | (0.021) |
| Black | | | 0.005 | | | -0.040 |
| | | | (0.041) | | | (0.042) |
| Hispanic | | | 0.062 | | | 0.032 |
| | | | (0.072) | | | (0.070) |
| Asian | | | 0.170 | | | 0.145 |
| | | | (0.074) | | | (0.068) |
| Other/Missing Race | | | -0.074 | | | -0.079 |
| | | | (0.157) | | | (0.156) |
| High School Top 10 Percent | | | 0.095 | | | 0.082 |
| | | | (0.027) | | | (0.028) |
| High School Rank Missing | | | 0.019 | | | 0.015 |
| | | | (0.033) | | | (0.037) |
| Athlete | | | 0.123 | | | 0.115 |
| | | | (0.025) | | | (0.027) |
| Selection Controls | N | N | N | Y | Y | Y |

Notes: Columns (1)-(3) include no selection controls. Columns (4)-(6) include a dummy for each group formed by matching students according to schools at which they were accepted or rejected. Each model is estimated using only observations with Barron's matches for which different students attended both private and public schools. The sample size is 5,583. Standard errors are shown in parentheses.

# Aside: The Link to Propensity Scores

- Notice we haven't assumed that the treatment $x_i$ is binary
  - In our constant-effects model, $y_i = \beta x_i + \varepsilon_i$, things don't really get more complicated with multivalued/continuous $x_i$

# Aside: The Link to Propensity Scores

- Notice we haven't assumed that the treatment $x_i$ is binary
  - In our constant-effects model, $y_i = \beta x_i + \varepsilon_i$, things don't really get more complicated with multivalued/continuous $x_i$
- If $x_i \in \{0,1\}$, then $E[x_i \mid w_i] = Pr(x_i = 1 \mid w_i)$ is the *propensity score*
  - Usually we're used to using these for matching/weighting estimators
  - Now we've seen another use: *controlling* for the propensity score

# Aside: The Link to Propensity Scores

- Notice we haven't assumed that the treatment $x_i$ is binary
  - In our constant-effects model, $y_i = \beta x_i + \varepsilon_i$, things don't really get more complicated with multivalued/continuous $x_i$
- If $x_i \in \{0,1\}$, then $E[x_i \mid w_i] = Pr(x_i = 1 \mid w_i)$ is the *propensity score*
  - Usually we're used to using these for matching/weighting estimators
  - Now we've seen another use: *controlling* for the propensity score
- If we know/estimate the propensity score in a first step, we could alternatively use the *recentered* $\tilde{x}_i = x_i - Pr(x_i = 1 \mid w_i)$ directly
  - We'll come back to this idea...

# Outline

# Why are Multi-Way FE Different?

- Consider a two-way fixed effect (FE) regression estimated in a panel of individuals $i$ observed over time periods $t$:

$$y_{it} = \beta x_{it} + \alpha_i + \tau_t + \nu_{it}$$

# Why are Multi-Way FE Different?

- Consider a two-way fixed effect (FE) regression estimated in a panel of individuals $i$ observed over time periods $t$:

$$y_{it} = \beta x_{it} + \alpha_i + \tau_t + \nu_{it}$$

Q: Can we justify this specification by selection-on-observables?

# Why are Multi-Way FE Different?

- Consider a two-way fixed effect (FE) regression estimated in a panel of individuals $i$ observed over time periods $t$:

$$y_{it} = \beta x_{it} + \alpha_i + \tau_t + \nu_{it}$$

  Q: Can we justify this specification by selection-on-observables?

- Note that the unit & time FE controls uniquely identify observations

# Why are Multi-Way FE Different?

- Consider a two-way fixed effect (FE) regression estimated in a panel of individuals $i$ observed over time periods $t$:

$$y_{it} = \beta x_{it} + \alpha_i + \tau_t + \nu_{it}$$

  Q: Can we justify this specification by selection-on-observables?

- Note that the unit & time FE controls uniquely identify observations
  - What would it mean for $x_{it}$ to be as-if-randomly-assigned given $(i, t)$?

# Why are Multi-Way FE Different?

- Consider a two-way fixed effect (FE) regression estimated in a panel of individuals $i$ observed over time periods $t$:

$$y_{it} = \beta x_{it} + \alpha_i + \tau_t + \nu_{it}$$

  Q: Can we justify this specification by selection-on-observables?

- Note that the unit & time FE controls uniquely identify observations
  - What would it mean for $x_{it}$ to be as-if-randomly-assigned given $(i, t)$?

- The auxiliary regression is of $x_{it}$ on two-way FEs (no interactions)
  - Is additivity, i.e. $E[x_{it} \mid (i,t)] = \mu_i + \gamma_t$, realistic to impose?

# Why are Multi-Way FE Different?

- Consider a two-way fixed effect (FE) regression estimated in a panel of individuals $i$ observed over time periods $t$:

$$y_{it} = \beta x_{it} + \alpha_i + \tau_t + \nu_{it}$$

  Q: Can we justify this specification by selection-on-observables?

- Note that the unit & time FE controls uniquely identify observations
  - What would it mean for $x_{it}$ to be as-if-randomly-assigned given $(i, t)$?

- The auxiliary regression is of $x_{it}$ on two-way FEs (no interactions)
  - Is additivity, i.e. $E[x_{it} \mid (i,t)] = \mu_i + \gamma_t$, realistic to impose?
  - Clearly can't make this specification more flexible without "dummying out" observations (there's no observed variation in $x_{it}$ given $(i, t)$)

# Why are Multi-Way FE Different?

- Consider a two-way fixed effect (FE) regression estimated in a panel of individuals $i$ observed over time periods $t$:

$$y_{it} = \beta x_{it} + \alpha_i + \tau_t + \nu_{it}$$

  Q: Can we justify this specification by selection-on-observables?

- Note that the unit & time FE controls uniquely identify observations
  - What would it mean for $x_{it}$ to be as-if-randomly-assigned given $(i,t)$?

- The auxiliary regression is of $x_{it}$ on two-way FEs (no interactions)
  - Is additivity, i.e. $E[x_{it} \mid (i,t)] = \mu_i + \gamma_t$, realistic to impose?
  - Clearly can't make this specification more flexible without "dummying out" observations (there's no observed variation in $x_{it}$ given $(i,t)$)

- We need a different justification for this sort of regression...

# Two-Way FE as Outcome Modeling

- Continue to assume a constant-effects causal model: $y_{it} = \beta x_{it} + \varepsilon_{it}$
  - Assume $x_{it}$ is *deterministic* in the set of unit and time indicators, $w_{it}$: once I know the unit and period, I know the treatment status
  - No scope for selection-on-observables

# Two-Way FE as Outcome Modeling

- Continue to assume a constant-effects causal model: $y_{it} = \beta x_{it} + \varepsilon_{it}$
  - Assume $x_{it}$ is *deterministic* in the set of unit and time indicators, $w_{it}$: once I know the unit and period, I know the treatment status
  - No scope for selection-on-observables

- Add in a model for untreated potential outcomes: $E[\varepsilon_{it} \mid w_{it}] = \alpha_i + \tau_t$
  - "Parallel trends": units with different treatment paths have different outcome levels but common outcome changes

# Two-Way FE as Outcome Modeling

- Continue to assume a constant-effects causal model: $y_{it} = \beta x_{it} + \varepsilon_{it}$
  - Assume $x_{it}$ is *deterministic* in the set of unit and time indicators, $w_{it}$: once I know the unit and period, I know the treatment status
  - No scope for selection-on-observables
- Add in a model for untreated potential outcomes: $E[\varepsilon_{it} \mid w_{it}] = \alpha_i + \tau_t$
  - "Parallel trends": units with different treatment paths have different outcome levels but common outcome changes
- Putting both models together, we have:

$$E[y_{it} \mid x_{it}, w_{it}] =$$

# Two-Way FE as Outcome Modeling

- Continue to assume a constant-effects causal model: $y_{it} = \beta x_{it} + \varepsilon_{it}$
  - Assume $x_{it}$ is *deterministic* in the set of unit and time indicators, $w_{it}$: once I know the unit and period, I know the treatment status
  - No scope for selection-on-observables
- Add in a model for untreated potential outcomes: $E[\varepsilon_{it} \mid w_{it}] = \alpha_i + \tau_t$
  - "Parallel trends": units with different treatment paths have different outcome levels but common outcome changes
- Putting both models together, we have:

$$E[y_{it} \mid x_{it}, w_{it}] = \beta x_{it} + E[\varepsilon_{it} \mid w_{it}] =$$

# Two-Way FE as Outcome Modeling

- Continue to assume a constant-effects causal model: $y_{it} = \beta x_{it} + \varepsilon_{it}$
  - Assume $x_{it}$ is *deterministic* in the set of unit and time indicators, $w_{it}$: once I know the unit and period, I know the treatment status
  - No scope for selection-on-observables

- Add in a model for untreated potential outcomes: $E[\varepsilon_{it} \mid w_{it}] = \alpha_i + \tau_t$
  - "Parallel trends": units with different treatment paths have different outcome levels but common outcome changes

- Putting both models together, we have:

$$E[y_{it} \mid x_{it}, w_{it}] = \beta x_{it} + E[\varepsilon_{it} \mid w_{it}] = \beta x_{it} + \alpha_i + \tau_t$$

# Two-Way FE as Outcome Modeling

- Continue to assume a constant-effects causal model: $y_{it} = \beta x_{it} + \varepsilon_{it}$
  - Assume $x_{it}$ is *deterministic* in the set of unit and time indicators, $w_{it}$: once I know the unit and period, I know the treatment status
  - No scope for selection-on-observables

- Add in a model for untreated potential outcomes: $E[\varepsilon_{it} \mid w_{it}] = \alpha_i + \tau_t$
  - "Parallel trends": units with different treatment paths have different outcome levels but common outcome changes

- Putting both models together, we have:

$$E[y_{it} \mid x_{it}, w_{it}] = \beta x_{it} + E[\varepsilon_{it} \mid w_{it}] = \beta x_{it} + \alpha_i + \tau_t$$

i.e. the CEF of $y_{it} \mid x_{it}, w_{it}$ is linear, with a causal coefficient of $\beta$
$\implies$ regression identifies it

# Two-Way FE as Outcome Modeling

- Continue to assume a constant-effects causal model: $y_{it} = \beta x_{it} + \varepsilon_{it}$
  - Assume $x_{it}$ is *deterministic* in the set of unit and time indicators, $w_{it}$: once I know the unit and period, I know the treatment status
  - No scope for selection-on-observables

- Add in a model for untreated potential outcomes: $E[\varepsilon_{it} \mid w_{it}] = \alpha_i + \tau_t$
  - "Parallel trends": units with different treatment paths have different outcome levels but common outcome changes

- Putting both models together, we have:

$$E[y_{it} \mid x_{it}, w_{it}] = \beta x_{it} + E[\varepsilon_{it} \mid w_{it}] = \beta x_{it} + \alpha_i + \tau_t$$

  i.e. the CEF of $y_{it} \mid x_{it}, w_{it}$ is linear, with a causal coefficient of $\beta$
  $\implies$ regression identifies it
  - Logic clearly extends to more than two FEs, time-varying controls, unit-specific trends, or any other model for $E[\varepsilon \mid w]$

# Example: Finkelstein (2007)

- Boiling multi-way FE regression specs down to simpler "diff-in-diff' comparisons can make the content of the outcome model clearer
  - Useful fact: in two periods, $\beta$ in $y_{it} = \beta x_{it} + \alpha_i + \tau_t + \nu_{it}$ is given by the regression of $y_{i,Post} - y_{i,Pre}$ on $x_{i,Post} - x_{i,Pre}$ (and a constant)

# Example: Finkelstein (2007)

- Boiling multi-way FE regression specs down to simpler "diff-in-diff' comparisons can make the content of the outcome model clearer
  - Useful fact: in two periods, $\beta$ in $y_{it} = \beta x_{it} + \alpha_i + \tau_t + \nu_{it}$ is given by the regression of $y_{i,Post} - y_{i,Pre}$ on $x_{i,Post} - x_{i,Pre}$ (and a constant)
- Finkelstein is interested in estimating market-level effects of health insurance coverage, using the 1965 introduction of Medicare

# Example: Finkelstein (2007)

- Boiling multi-way FE regression specs down to simpler "diff-in-diff' comparisons can make the content of the outcome model clearer
  - Useful fact: in two periods, $\beta$ in $y_{it} = \beta x_{it} + \alpha_i + \tau_t + \nu_{it}$ is given by the regression of $y_{i,Post} - y_{i,Pre}$ on $x_{i,Post} - x_{i,Pre}$ (and a constant)

- Finkelstein is interested in estimating market-level effects of health insurance coverage, using the 1965 introduction of Medicare
  - Effectively estimates $y_{it} = \beta x_{it} + \alpha_i + \tau_t + \nu_{it}$ where $x_{it}$ is the share of elderly in market $i$ and year $t$ with health insurance

# Example: Finkelstein (2007)

- Boiling multi-way FE regression specs down to simpler "diff-in-diff' comparisons can make the content of the outcome model clearer
    - Useful fact: in two periods, $\beta$ in $y_{it} = \beta x_{it} + \alpha_i + \tau_t + \nu_{it}$ is given by the regression of $y_{i,Post} - y_{i,Pre}$ on $x_{i,Post} - x_{i,Pre}$ (and a constant)
- Finkelstein is interested in estimating market-level effects of health insurance coverage, using the 1965 introduction of Medicare
    - Effectively estimates $y_{it} = \beta x_{it} + \alpha_i + \tau_t + \nu_{it}$ where $x_{it}$ is the share of elderly in market $i$ and year $t$ with health insurance
    - Post 1965, $x_{it} = 1$ for all markets; previously far from random

# Example: Finkelstein (2007)

- Boiling multi-way FE regression specs down to simpler “diff-in-diff’ comparisons can make the content of the outcome model clearer
  - Useful fact: in two periods, $\beta$ in $y_{it} = \beta x_{it} + \alpha_i + \tau_t + \nu_{it}$ is given by the regression of $y_{i,Post} - y_{i,Pre}$ on $x_{i,Post} - x_{i,Pre}$ (and a constant)
- Finkelstein is interested in estimating market-level effects of health insurance coverage, using the 1965 introduction of Medicare
  - Effectively estimates $y_{it} = \beta x_{it} + \alpha_i + \tau_t + \nu_{it}$ where $x_{it}$ is the share of elderly in market $i$ and year $t$ with health insurance
  - Post 1965, $x_{it} = 1$ for all markets; previously far from random
- Consider two-period version: equivalent to regressing $y_{i,Post} - y_{i,Pre}$ on $x_{i,Post} - x_{i,Pre} = 1 - x_{i,Pre}$: the pre-Medicare uninsured share in $i$

# Example: Finkelstein (2007)

- Boiling multi-way FE regression specs down to simpler "diff-in-diff' comparisons can make the content of the outcome model clearer
  - Useful fact: in two periods, $\beta$ in $y_{it} = \beta x_{it} + \alpha_i + \tau_t + \nu_{it}$ is given by the regression of $y_{i,Post} - y_{i,Pre}$ on $x_{i,Post} - x_{i,Pre}$ (and a constant)
- Finkelstein is interested in estimating market-level effects of health insurance coverage, using the 1965 introduction of Medicare
  - Effectively estimates $y_{it} = \beta x_{it} + \alpha_i + \tau_t + \nu_{it}$ where $x_{it}$ is the share of elderly in market $i$ and year $t$ with health insurance
  - Post 1965, $x_{it} = 1$ for all markets; previously far from random
- Consider two-period version: equivalent to regressing $y_{i,Post} - y_{i,Pre}$ on $x_{i,Post} - x_{i,Pre} = 1 - x_{i,Pre}$: the pre-Medicare uninsured share in $i$
  - Outcome model: if not for the introduction of Medicare, markets with higher/lower uninsured shares would have been on parallel trends

# Example: Finkelstein (2007)

- Boiling multi-way FE regression specs down to simpler "diff-in-diff' comparisons can make the content of the outcome model clearer
  - Useful fact: in two periods, $\beta$ in $y_{it} = \beta x_{it} + \alpha_i + \tau_t + \nu_{it}$ is given by the regression of $y_{i,Post} - y_{i,Pre}$ on $x_{i,Post} - x_{i,Pre}$ (and a constant)
- Finkelstein is interested in estimating market-level effects of health insurance coverage, using the 1965 introduction of Medicare
  - Effectively estimates $y_{it} = \beta x_{it} + \alpha_i + \tau_t + \nu_{it}$ where $x_{it}$ is the share of elderly in market $i$ and year $t$ with health insurance
  - Post 1965, $x_{it} = 1$ for all markets; previously far from random
- Consider two-period version: equivalent to regressing $y_{i,Post} - y_{i,Pre}$ on $x_{i,Post} - x_{i,Pre} = 1 - x_{i,Pre}$: the pre-Medicare uninsured share in $i$
  - Outcome model: if not for the introduction of Medicare, markets with higher/lower uninsured shares would have been on parallel trends
  - Event study version: $y_{it} = \alpha_i + \tau_t + \sum_s \beta_s (1 - x_{i,Pre}) \mathbf{1}[t = s] + \nu_{it}$; expect flat pre/post trends if the model is right...

# Finkelstein Event Study

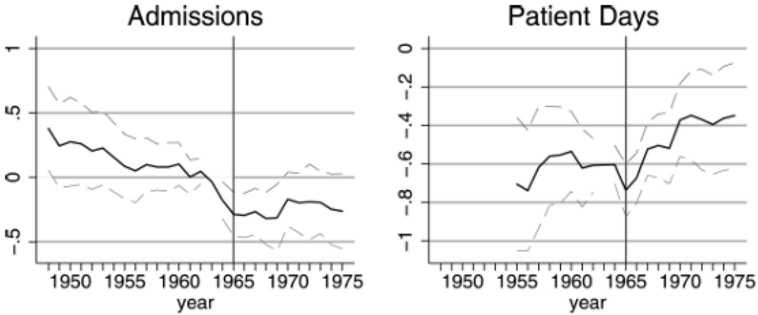



Figure II graphs the pattern of the $\lambda_t$ coefficients from estimating (1) for the log of the dependent variable given above each graph. The scale of the graph is normalized so that in the reference year (1963) it is the average difference in the dependent variable between the south and west (where Medicare had a larger impact) relative to the north and northeast (where Medicare had a smaller impact). The dashed lines show the 95 percent confidence interval on each coefficient relative to the reference year (1963). Time varying state-level controls ($X_{st}$) in all analyses consist of eight indicator variables for the number of years before (or since) the implementation of Medicaid in state $s$ (see text for more details).

# Different Roles of Controls

- In a design-based specification, the controls can be understood as specifying *which treated/control observations* are valid to compare
  - E.g. private vs. non-private students w/same applications+admissions

# Different Roles of Controls

- In a design-based specification, the controls can be understood as specifying *which treated/control observations* are valid to compare
  - E.g. private vs. non-private students w/same applications+admissions
  - Think about how the auxiliary regression isolates variation in $x_{it}$

# Different Roles of Controls

- In a design-based specification, the controls can be understood as specifying *which treated/control observations* are valid to compare
  - E.g. private vs. non-private students w/same applications+admissions
  - Think about how the auxiliary regression isolates variation in $x_{it}$

- In an outcome-model-based specification, the controls can be seen as specifying *what transformations of the outcomes* are valid to compare
  - E.g. TWFE regressions compare trends in the outcome, allowing the outcome levels to be confounded

# Different Roles of Controls

- In a design-based specification, the controls can be understood as specifying *which treated/control observations* are valid to compare
  - E.g. private vs. non-private students w/same applications+admissions
  - Think about how the auxiliary regression isolates variation in $x_{it}$
- In an outcome-model-based specification, the controls can be seen as specifying *what transformations of the outcomes* are valid to compare
  - E.g. TWFE regressions compare trends in the outcome, allowing the outcome levels to be confounded
  - Think about what "diff-in-diff" comparisons are underlying the spec

# Different Roles of Controls

- In a design-based specification, the controls can be understood as specifying *which treated/control observations* are valid to compare
  - E.g. private vs. non-private students w/same applications+admissions
  - Think about how the auxiliary regression isolates variation in $x_{it}$
- In an outcome-model-based specification, the controls can be seen as specifying *what transformations of the outcomes* are valid to compare
  - E.g. TWFE regressions compare trends in the outcome, allowing the outcome levels to be confounded
  - Think about what "diff-in-diff" comparisons are underlying the spec
- Both strategies have *ex post* validations (balance tests / pre-trend checks), but the *ex ante* case for design is arguably easier to make
  - What $\varepsilon_{it}$ model is best? E.g. does parallel trends hold in levels or logs?

# Outline

1. Regression/IV Recap✓
2. Selection on Observables✓
3. Design vs. Outcome Models✓
4. Design-Based IV

# The Simplest IV Story

- Again start w/constant fx model $y_i = \beta x_i + \varepsilon_i$, now $Cov(x_i, \varepsilon_i) \neq 0$
  - E.g. $x_i$ is enrollment in this class and $y_i$ is later wages/happiness
  - "Endogeneity": students in this class have systematically higher $\varepsilon_i$

# The Simplest IV Story

- Again start w/constant fx model $y_i = \beta x_i + \varepsilon_i$, now $Cov(x_i, \varepsilon_i) \neq 0$
    - E.g. $x_i$ is enrollment in this class and $y_i$ is later wages/happiness
    - "Endogeneity": students in this class have systematically higher $\varepsilon_i$

- Imagine the course was "oversubscribed"; I chose students by lottery
    - $z_i \in \{0,1\}$ indicates randomized admission to the course
    - Randomness + no direct effects of $z_i$ on $y_i$ implies $Cov(z_i, \varepsilon_i) = 0$

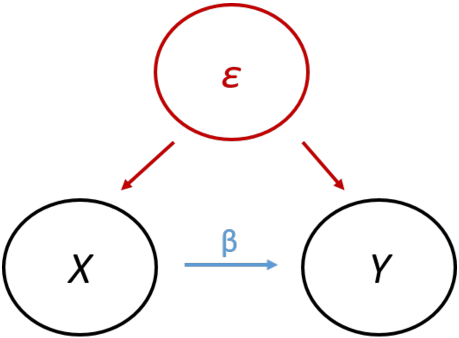
# The Simplest IV Story

- Again start w/constant fx model $y_i = \beta x_i + \varepsilon_i$, now $Cov(x_i, \varepsilon_i) \neq 0$
  - E.g. $x_i$ is enrollment in this class and $y_i$ is later wages/happiness
  - "Endogeneity": students in this class have systematically higher $\varepsilon_i$
- Imagine the course was "oversubscribed"; I chose students by lottery
  - $z_i \in \{0, 1\}$ indicates randomized admission to the course
  - Randomness + no direct effects of $z_i$ on $y_i$ implies $Cov(z_i, \varepsilon_i) = 0$
- Plugging in the model for $\varepsilon_i = y_i - \beta x_i$, we have IV identification:

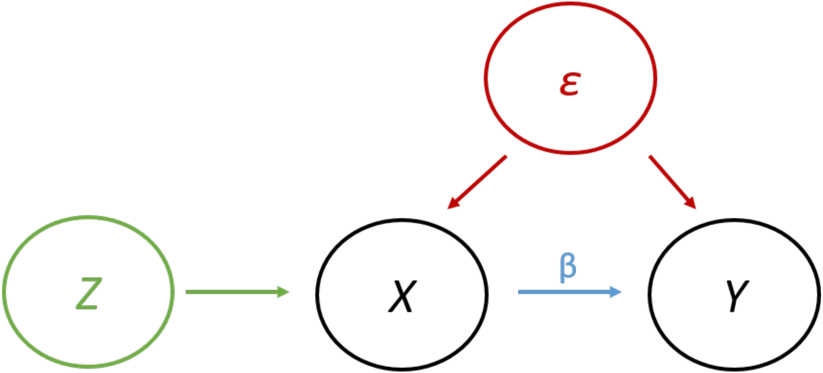
$$Cov(z_i, y_i - \beta x_i) = 0 \implies \frac{Cov(z_i, y_i)}{Cov(z_i, x_i)} = \beta$$

so long as $Cov(z_i, x_i) \neq 0$ ("relevance")

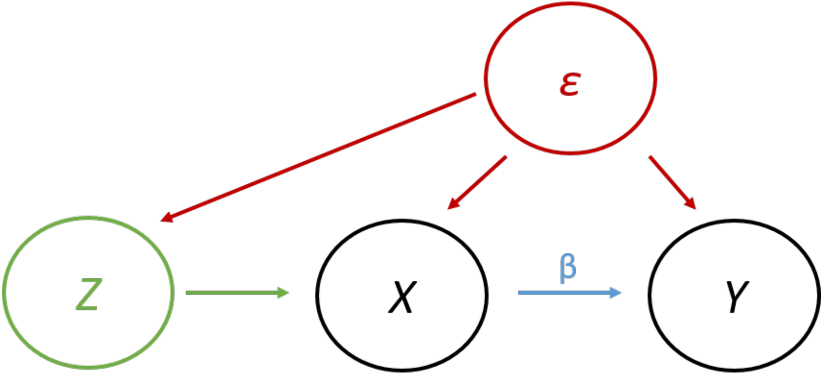
# Regression “Endogeneity”



# Instrument "Exogeneity" / "Validity"

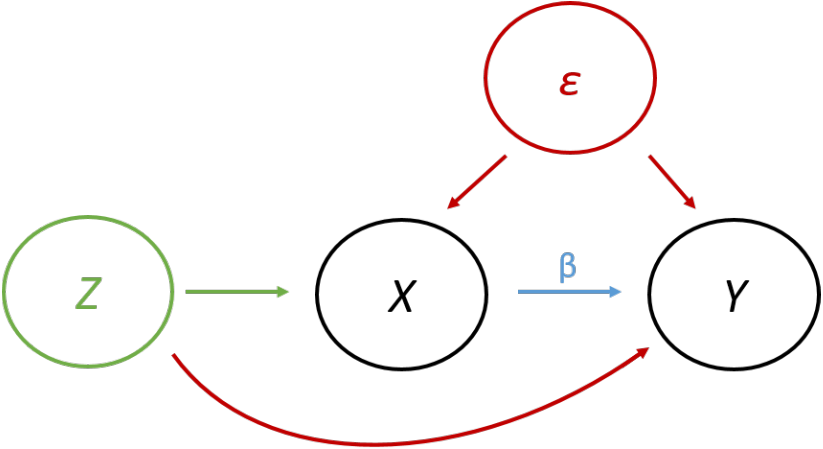


# Threats to Validity: Instrument Assignment




We will later formalize this as a failure of instrument "independence"

# Threats to Validity: Direct Effects




We will later formalize this as a failure of instrument "exclusion"

# Adding Controls and Instruments

- Basic IV is $\frac{Cov(z_i,y_i)}{Cov(z_i,x_i)} = \frac{Cov(z_i,y_i)/Var(z_i)}{Cov(z_i,x_i)/Var(z_i)} = \rho/\pi$ from the regressions:

$$y_i = \kappa + \rho z_i + \nu_i \quad \text{, the "reduced form"}$$

$$x_i = \mu + \pi z_i + \eta_i \quad \text{, the "first stage"}$$

# Adding Controls and Instruments

- Basic IV is $\frac{Cov(z_i,y_i)}{Cov(z_i,x_i)} = \frac{Cov(z_i,y_i)/Var(z_i)}{Cov(z_i,x_i)/Var(z_i)} = \rho/\pi$ from the regressions:

$$y_i = \kappa + \rho z_i + \nu_i \quad \text{, the ``reduced form''}$$
$$x_i = \mu + \pi z_i + \eta_i \quad \text{, the ``first stage''}$$

- IV with controls works similarly: $\rho/\pi$ from the controlled regressions:

$$y_i = \kappa + \rho z_i + w_i'\gamma + \nu_i$$
$$x_i = \mu + \pi z_i + w_i'\gamma + \eta_i$$

# Adding Controls and Instruments

- Basic IV is $\frac{Cov(z_i,y_i)}{Cov(z_i,x_i)} = \frac{Cov(z_i,y_i)/Var(z_i)}{Cov(z_i,x_i)/Var(z_i)} = \rho/\pi$ from the regressions:

$$y_i = \kappa + \rho z_i + \nu_i \quad \text{, the ``reduced form''}$$
$$x_i = \mu + \pi z_i + \eta_i \quad \text{, the ``first stage''}$$

- IV with controls works similarly: $\rho/\pi$ from the controlled regressions:

$$y_i = \kappa + \rho z_i + w_i'\gamma + \nu_i$$
$$x_i = \mu + \pi z_i + w_i'\gamma + \eta_i$$

- Can also have multiple instruments: $(\pi' w \pi)^{-1}\pi' w \rho$ for some $w$
  - $w$ governs how different RF/FS's are weighted together (e.g. 2SLS)

# Adding Controls and Instruments

- Basic IV is $\frac{Cov(z_i,y_i)}{Cov(z_i,x_i)} = \frac{Cov(z_i,y_i)/Var(z_i)}{Cov(z_i,x_i)/Var(z_i)} = \rho/\pi$ from the regressions:

$$y_i = \kappa + \rho z_i + \nu_i \quad \text{, the "reduced form"}$$
$$x_i = \mu + \pi z_i + \eta_i \quad \text{, the "first stage"}$$

- IV with controls works similarly: $\rho/\pi$ from the controlled regressions:

$$y_i = \kappa + \rho z_i + w_i'\gamma + \nu_i$$
$$x_i = \mu + \pi z_i + w_i'\gamma + \eta_i$$

- Can also have multiple instruments: $(\pi' w \pi)^{-1}\pi' w \rho$ for some $w$
  - $w$ governs how different RF/FS's are weighted together (e.g. 2SLS)

- RF&FS are the nuclei of IV; the design-based approach starts w/them

# Bridging the Gap

- Design-based IV applies the earlier selection-on-observables logic to $z_i$:
    1. Claim $z_i$ is as-good-as-randomly assigned conditional on some $w_i$

# Bridging the Gap

- Design-based IV applies the earlier selection-on-observables logic to $z_i$:
  1. Claim $z_i$ is as-good-as-randomly assigned conditional on some $w_i$
  2. Control flexibly for $w_i$ such that regressing $z_i$ on it estimates $E[z_i \mid w_i]$

# Bridging the Gap

- Design-based IV applies the earlier selection-on-observables logic to $z_i$:
  1. Claim $z_i$ is as-good-as-randomly assigned conditional on some $w_i$
  2. Control flexibly for $w_i$ such that regressing $z_i$ on it estimates $E[z_i \mid w_i]$
- This makes both reduced form and first stage regressions causal

# Bridging the Gap

- Design-based IV applies the earlier selection-on-observables logic to $z_i$:
  1. Claim $z_i$ is as-good-as-randomly assigned conditional on some $w_i$
  2. Control flexibly for $w_i$ such that regressing $z_i$ on it estimates $E[z_i \mid w_i]$
- This makes both reduced form and first stage regressions causal
  - As before, best to both tell a clear *ex ante* story about where the $z_i \mid w_i$ variation comes from and validate as-if-random assignment *ex post*

# Bridging the Gap

- Design-based IV applies the earlier selection-on-observables logic to $z_i$:
    1. Claim $z_i$ is as-good-as-randomly assigned conditional on some $w_i$
    2. Control flexibly for $w_i$ such that regressing $z_i$ on it estimates $E[z_i \mid w_i]$
- This makes both reduced form and first stage regressions causal
    - As before, best to both tell a clear *ex ante* story about where the $z_i \mid w_i$ variation comes from and validate as-if-random assignment *ex post*
- New twist: have to also argue exclusion in order to interpret RF/FS
    - Can both argue *ex ante* and sometimes test *ex post*: e.g. by looking at effects of $z_i$ on other plausible treatment channels

# Example: Abdulkadiroglu et al. (2016)

- AAHP are interested in the effect of “takeover” charter schools: ones which convert a low-performing traditional public school (TPS)
  - Lots of evidence of charter effectiveness from admission lotteries, but external validity is an open question

# Example: Abdulkadiroglu et al. (2016)

- AAHP are interested in the effect of “takeover” charter schools: ones which convert a low-performing traditional public school (TPS)
  - Lots of evidence of charter effectiveness from admission lotteries, but external validity is an open question

- Reduced-form selection-on-observables strategy: compare students in the TPS pre-takeover to others in similarly low-performing TPS
  - Specifically, match each takeover school to a TPS using baseline test score performance and control for match cell fixed effects

# Example: Abdulkadiroglu et al. (2016)

- AAHP are interested in the effect of "takeover" charter schools: ones which convert a low-performing traditional public school (TPS)
  - Lots of evidence of charter effectiveness from admission lotteries, but external validity is an open question
- Reduced-form selection-on-observables strategy: compare students in the TPS pre-takeover to others in similarly low-performing TPS
  - Specifically, match each takeover school to a TPS using baseline test score performance and control for match cell fixed effects
- Exclusion: takeovers only affect later test scores via charter enrollment
  - Check whether there are takeover effects in the transition (pre-charter) year 0; develop a strategy to use these effects to relax exclusion

# Abdulkadiroglu et al. Results

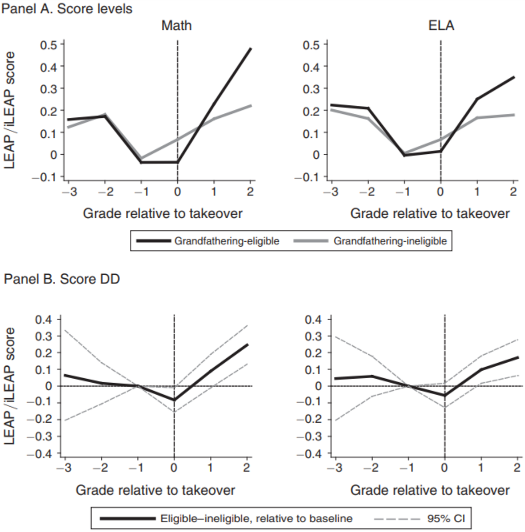



FIGURE 2. TEST SCORES IN THE RSD GRANDFATHERING SAMPLE

*Notes:* Panel A plots average LEAP/iLEAP math and ELA scores of students in the RSD legacy middle school matched sample. Panel B plots achievement growth relative to the baseline (−1) grade. Estimates in both panels control for matching cell fixed effects. Scores are standardized to those of students at direct-run schools in New Orleans RSD, by grade and year. Grade 0 is the last grade of legacy school enrollment.

# Abdulkadiroglu et al. Results (Cont.)

| | | Comparison group mean (1) | OLS (2) | 2SLS First stage (3) | 2SLS Enrollment effect (4) |
|---|---|---|---|---|---|
| *Panel A. All grades* | | | | | |
| (Fifth through eighth) | Math (N = 5,625) | −0.089 | 0.123<br>(0.020) | 1.073<br>(0.052) | 0.212<br>(0.038) |
| | ELA (N = 5,621) | −0.092 | 0.082<br>(0.018) | 1.075<br>(0.052) | 0.143<br>(0.039) |
| *Panel B. By grade* | | | | | |
| Fifth and sixth grades | Math (N = 2,579) | −0.091 | 0.099<br>(0.035) | 0.738<br>(0.041) | 0.165<br>(0.068) |
| | ELA (N = 2,579) | −0.116 | 0.023<br>(0.033) | 0.745<br>(0.042) | 0.101<br>(0.070) |
| Seventh and eighth grades | Math (N = 3,046) | −0.086 | 0.133<br>(0.020) | 1.355<br>(0.070) | 0.231<br>(0.037) |
| | ELA (N = 3,042) | −0.071 | 0.104<br>(0.019) | 1.352<br>(0.070) | 0.171<br>(0.036) |

# Abdulkadiroglu et al.: Comparison to Lottery IV

| | | | | 2SLS | | |
|---|---|---|---|---|---|---|
| | | | | First stage | | |
| | | Comparison group mean (1) | OLS (2) | Immediate offer (3) | Waitlist offer (4) | Enrollment effect (5) |
| *Panel A. All grades* | | | | | | |
| (Sixth through eighth) | Math (N = 2,202) | 0.059 | 0.301 | 0.760 | 0.562 | 0.270 |
| | | | (0.022) | (0.063) | (0.067) | (0.056) |
| | ELA (N = 2,205) | 0.103 | 0.148 | 0.759 | 0.562 | 0.118 |
| | | | (0.020) | (0.063) | (0.067) | (0.051) |
| *Panel B. By potential exposure* | | | | | | |
| First exposure year (sixth and seventh grades) | Math (N = 881) | 0.056 | 0.347 | 0.519 | 0.397 | 0.365 |
| | | | (0.044) | (0.034) | (0.038) | (0.086) |
| | ELA (N = 882) | 0.058 | 0.239 | 0.521 | 0.394 | 0.220 |
| | | | (0.044) | (0.034) | (0.038) | (0.088) |
| Second and third exposure year (seventh and eighth grades) | Math (N = 1,321) | 0.061 | 0.294 | 0.921 | 0.665 | 0.242 |
| | | | (0.021) | (0.088) | (0.091) | (0.054) |
| | ELA (N = 1,323) | 0.129 | 0.131 | 0.918 | 0.668 | 0.083 |
| | | | (0.020) | (0.088) | (0.091) | (0.047) |

# Looking Ahead

- We've now seen the basic design-based logic for regression/IV
  - Main practical takeaway: be clear on what variation in $x_i$ or $z_i$ you want to use, and pick controls appropriately for extracting it

# Looking Ahead

- We've now seen the basic design-based logic for regression/IV
  - Main practical takeaway: be clear on what variation in $x_i$ or $z_i$ you want to use, and pick controls appropriately for extracting it

- On Wednesday, we'll expand the discussion to other features of design
  1. Heterogeneous effects $\implies$ no worries about "negative weights"
  2. Inference $\implies$ clearer guidance on how to cluster standard errors

# Looking Ahead

- We've now seen the basic design-based logic for regression/IV
  - Main practical takeaway: be clear on what variation in $x_i$ or $z_i$ you want to use, and pick controls appropriately for extracting it
- On Wednesday, we'll expand the discussion to other features of design
  1. Heterogeneous effects $\Longrightarrow$ no worries about "negative weights"
  2. Inference $\Longrightarrow$ clearer guidance on how to cluster standard errors
- Before then, you have the chance to play with a real-world application