Political Methodology II
Section: Panel data II, Instrumental Variables Basics

Apoorva Lal

February 24, 2022

Stanford University

Roadmap

Panel data with multiple time periods

Synthetic Control Methods

Instrumental Variables I

Roadmap

Panel data with multiple time periods

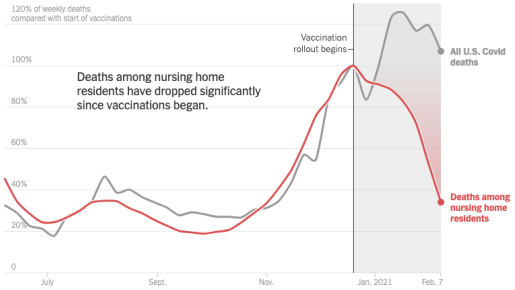
Synthetic Control Methods

Instrumental Variables I

Difference-in-Differences Example

Nursing Homes, Once Hotspots, Far Outpace U.S. in Covid Declines

By Matthew Conlen, Sarah Mervosh and Danielle Ivory Feb. 25, 2021



Source: New York Times database; U.S. Department of Health and Human Services • Data shown is normalized compared with the weekly deaths for the week ending Dec. 20, 2020 and is through Feb. 7.

In Practice

- Parallel trends are most easily satisfied when the intervention is truly exogenous
- You can try and condition on pre-trends
- Our running example so far has been a two-period 2-by-2 DiD
  - **Things becomes more difficult when design becomes complicated**
  - staggered treatment
  - continuous treatment
  - treatment reversals

Motivation

Why do we like panel data?

- It allows us to relax some of the assumptions from the selection on observables world.
- Specifically, we can account for some specific types of unobservable confounders!
- In fixed effects/diff-in-diff: we can account for time-invariant unit fixed effects and for time-specific shocks that affect all units equally (as well as some more complicated trends).

Accounting for Time-invariant Unobserved Effects

Imagine that the data are actually generated according to the model

$$y_{it} = \mathbf{x}_{it}\beta + c_i + \varepsilon_{it}, \quad i = 1, \ldots, N; \; t = 1, \ldots, T.$$

Accounting for Time-invariant Unobserved Effects

Imagine that the data are actually generated according to the model
$y_{it} = \mathbf{x}_{it}\beta + c_i + \varepsilon_{it}, \quad i = 1, \ldots, N; \; t = 1, \ldots, T.$

How should we interpret each component?

- y_{it}: outcome variable for unit i in time t.
- $\mathbf{x}_{it}$: $1 \times k$ vector of covariates for unit i in year t.
- β: $k \times 1$ vector of marginal effects.
- c_i: unit-specific effect; all of the unobserved features affecting y_{it} equally in every time period.
- ϵ_{it}: time-varying unobserved factors affecting y_{it}; idiosyncratic errors.

What happens if we regress y_{it} on $\mathbf{x}_{it}$?

Pooled OLS

- If we regress y_{it} on $\mathbf{x}_{it}$, we are estimating an equation of the form:

$$y_{it} = \mathbf{x}_{it}\beta + v_{it}, \quad t = 1, \ldots, T$$

with the *composite error* $v_{it} \equiv c_i + \epsilon_{it}$. When is the OLS estimator of this equation $\hat{\beta}$ unbiased for β?

Pooled OLS

- If we regress y_{it} on $\mathbf{x}_{it}$, we are estimating an equation of the form:

$$y_{it} = \mathbf{x}_{it}\beta + v_{it}, \quad t = 1, \ldots, T$$

  with the *composite error* $v_{it} \equiv c_i + \epsilon_{it}$. When is the OLS estimator of this equation $\hat{\beta}$ unbiased for β?

- We know from last quarter that OLS is unbiased when the errors are uncorrelated with the regressors, so we need

$$\begin{aligned} &E[v_{it} \mid \mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{iT}] = E[v_{it} \mid \mathbf{x}_{it}] = 0 \\ \Longrightarrow &E[c_i + \epsilon_{it} \mid \mathbf{x}_{it}] = E[c_i \mid \mathbf{x}_{it}] = E[\epsilon_{it} \mid \mathbf{x}_{it}] = 0 \end{aligned}$$

Pooled OLS

- If we regress y_{it} on $\mathbf{x}_{it}$, we are estimating an equation of the form:

$$y_{it} = \mathbf{x}_{it}\beta + v_{it}, \quad t = 1, \ldots, T$$

  with the *composite error* $v_{it} \equiv c_i + \epsilon_{it}$. When is the OLS estimator of this equation $\hat{\beta}$ unbiased for β?

- We know from last quarter that OLS is unbiased when the errors are uncorrelated with the regressors, so we need

$$\begin{aligned} &E[v_{it} \mid \mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{iT}] = E[v_{it} \mid \mathbf{x}_{it}] = 0 \\ \implies &E[c_i + \epsilon_{it} \mid \mathbf{x}_{it}] = E[c_i \mid \mathbf{x}_{it}] = E[\epsilon_{it} \mid \mathbf{x}_{it}] = 0 \end{aligned}$$

- We need the composite error to be mean-independent from the covariates, but this often isn't a reasonable assumption: the value of unit fixed effects c_i by definition varied by unit.

Example: Determining the Effect of a Law

Say we want to know the effect of a new public financing law on election competitiveness. Suppose we code a variable $law_{it} = 1$ if state i had the law in year t and 0 otherwise. What happens if we regress y_{it} on law_{it}?

Example: Determining the Effect of a Law

Say we want to know the effect of a new public financing law on election competitiveness. Suppose we code a variable $law_{it} = 1$ if state i had the law in year t and 0 otherwise. What happens if we regress y_{it} on law_{it}?

- Both the law and the outcome are likely to be correlated with an unobserved factor c_i.
- We might think that some unobserved state-specific factors (say, proportion of people who don't like corporate money in politics) might affect *both* the introduction of public financing law and the competitiveness of elections.
- The result: $\hat{\beta}$ is not consistent for the true effect of the law.
- Also likely another violation of the standard OLS framework: $v_{it} \equiv c_i + \epsilon_{it}$ are serially correlated, making the standard errors invalid.

Fixed Effects Regression

How might we solve this problem? Estimate a model with a separate intercept for each unit. Recall the model $y_{it} = \mathbf{x}_{it}\beta + c_i + \epsilon_{it}$. We can estimate the c_i's in the model as unit-specific intercepts. Then:

$$(\widehat{\beta}, \widehat{c}_1, \ldots, \widehat{c}_N) = \underset{\beta, c_1, \ldots, c_N}{\arg\min} \sum_{i=1}^{N} \sum_{t=1}^{T} (y_{it} - \mathbf{x}_{it}\beta - c_i)^2 \tag{1}$$

Solving this minimization problem yields the estimator

$$\hat{\beta} = \left(\sum_{i=1}^{N} \sum_{t=1}^{T} (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)'(\mathbf{x}_{it} - \bar{\mathbf{x}}_i) \right)^{-1} \left(\sum_{i=1}^{N} \sum_{t=1}^{T} (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)'(y_{it} - \bar{y}_i) \right) \tag{2}$$

This is the same as regressing a demeaned $\tilde{y}_{it} \equiv y_{it} - \bar{y}_i$ on a demeaned $\tilde{\mathbf{x}}_{it} \equiv \mathbf{x}_{it} - \bar{\mathbf{x}}_i$.

Two-Way Fixed Effects Regression: Motivation

- We often have access to panel data, wherein each individual $i \in \{1, \ldots, N\}$ is observed for $T \geq 2$ time periods
- Stipulate following potential outcomes $Y_{it}^{(d)}$
  - $\mathbb{E}\left[Y_{it}^{0} | \alpha_i, t, D_{it}\right] = \alpha_i + \lambda_t$
    - α_i is a unit fixed-effect: each individual has an intercept α_i - absorbs time-invariant unit-specific confounders
    - λ_t is a time fixed-effect: each time period has an intercept λ_t - absorbs unit-invariant time-specific confounders
  - Suppose D_{it} is as-good as randomly assigned conditional on α_i
  - Stipulate constant, additive effect of treatment. Then, $\mathbb{E}\left[Y_{it}^{1}\right] = \mathbb{E}\left[Y_{it}^{0}\right] + \tau$
- This motivates the popular **two-way fixed-effects** regression

$$Y_{it} = \tau D_i + \alpha_i + \gamma_t + \varepsilon_{it}$$

Within Estimator

- With large datasets, estimating individual α_is can involve inverting a very large matrix
  - With short panels, the estimates of α_is are inconsistent anyway := incidental parameters problem (Neyman-Scott)
- Instead, we can use Frisch-Waugh-Lovell (again!) and partial out FEs
- Calculate individual averages of the 2wFE equation

$$\overline{Y}_i = \alpha_i + \overline{\lambda} + \tau \overline{D}_i + \overline{\varepsilon}$$

Within Estimator

- With large datasets, estimating individual α_is can involve inverting a very large matrix
  - With short panels, the estimates of α_is are inconsistent anyway := incidental parameters problem (Neyman-Scott)
- Instead, we can use Frisch-Waugh-Lovell (again!) and partial out FEs
- Calculate individual averages of the 2wFE equation

$$\overline{Y}_i = \alpha_i + \overline{\lambda} + \tau \overline{D}_i + \overline{\varepsilon}$$

- Subtract this from the FE equation

$$Y_{it} - \overline{Y}_i = \lambda_t - \overline{\lambda} + \tau(D_{it} - \overline{D}_i) + (\varepsilon_{it} - \overline{\varepsilon})$$

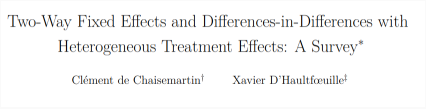
Staggered Adoption, Treatment Reversals, and other complications

- Recall that we stipulated a **constant, additive treatment effect** and **treatment timing as-good-as-random (conditional on FEs)**
- Last 5 years of methods literature on panel data studies what happens when we relax these parametric assumptions
- Weird weights redux: 2WFE no longer consistent for ATT


What's Trending in Difference-in-Differences?
A Synthesis of the Recent Econometrics Literature

Jonathan Roth* Pedro H. C. Sant'Anna† Alyssa Bilinski‡ John Poe§

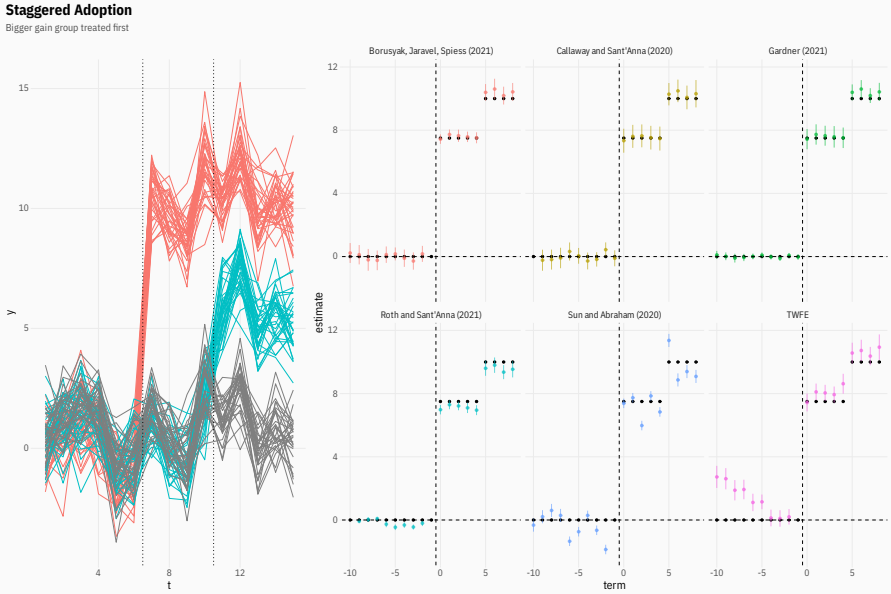
January 3, 2022


Two-Way Fixed Effects and Differences-in-Differences with Heterogeneous Treatment Effects: A Survey*

Clément de Chaisemartin† Xavier D'Haultfœuille‡

- review paper 1
- review paper 2

Introducing Heterogeneity



Introducing Heterogeneity II

Roadmap

Panel data with multiple time periods

Synthetic Control Methods

Instrumental Variables I

Synth Setup: Doudchenko and Imbens (2016)

- Panel methods can be characterised into 3 broad groups (as of 2016):
  - Difference-in-differences : $\Delta Y^{\text{post}} - \Delta Y^{\text{pre}}$
  - Matching: on both pre-treatment outcomes and other covariates
  - Synthetic Control: For each treated unit, a 'synthetic control' is constructed as a weighted average of control units s.t. the weighted average matches pre-treatment outcomes and covariates
- This paper: framework to nest existing approaches + estimator that relaxes some assumptions.
  - Main contribution: framework to clarify assumptions
  - Resting WP; Cannibalised by later papers (esp. Arkhangelsky et al 2020)?

Notation

- $N + 1$ units observed for T periods, with a subset of treated units (for simplicity - unit 0) treated from T_0 onwards

Notation

- $N+1$ units observed for T periods, with a subset of treated units (for simplicity - unit 0) treated from T_0 onwards
- Treatment : $D_{i,t} = \mathbb{1}_{i=0 \, \wedge \, t \in T_0+1,\ldots,T}$

Notation

- $N+1$ units observed for T periods, with a subset of treated units (for simplicity - unit 0) treated from T_0 onwards
- Treatment : $D_{i,t} = \mathbb{1}_{i=0 \, \wedge \, t \in T_0+1,\ldots,T}$
- Potential outcomes for unit 0 define the treatment effect: $\tau_{0,t} := Y_{0,t}(1) - Y_{0,t}(0)$ for $t = T_0 + 1, \ldots, T$

Notation

- $N+1$ units observed for T periods, with a subset of treated units (for simplicity - unit 0) treated from T_0 onwards
- Treatment : $D_{i,t} = \mathbb{1}_{i=0 \, \wedge \, t \in T_0+1,\ldots,T}$
- Potential outcomes for unit 0 define the treatment effect: $\tau_{0,t} := Y_{0,t}(1) - Y_{0,t}(0)$ for $t = T_0 + 1, \ldots, T$
- Observed outcome: $Y_{i,t}^{obs} = Y_{i,t}(D_{i,t})$

Notation

- $N+1$ units observed for T periods, with a subset of treated units (for simplicity - unit 0) treated from T_0 onwards
- Treatment : $D_{i,t} = \mathbb{1}_{i=0 \wedge t \in T_0+1,\ldots,T}$
- Potential outcomes for unit 0 define the treatment effect: $\tau_{0,t} := Y_{0,t}(1) - Y_{0,t}(0)$ for $t = T_0 + 1, \ldots, T$
- Observed outcome: $Y_{i,t}^{obs} = Y_{i,t}(D_{i,t})$
- Time-invariant characteristics $X_i := (X_{i,1}, \ldots, X_{i,M})^\top$ for each unit, which may include lagged outcomes $Y_{i,t}^{obs}$ for $t \leq T_0$

Notation

- $N+1$ units observed for T periods, with a subset of treated units (for simplicity - unit 0) treated from T_0 onwards
- Treatment : $D_{i,t} = \mathbb{1}_{i=0 \wedge t \in T_0+1,\ldots,T}$
- Potential outcomes for unit 0 define the treatment effect: $\tau_{0,t} := Y_{0,t}(1) - Y_{0,t}(0)$ for $t = T_0 + 1, \ldots, T$
- Observed outcome: $Y_{i,t}^{obs} = Y_{i,t}(D_{i,t})$
- Time-invariant characteristics $X_i := (X_{i,1}, \ldots, X_{i,M})^\top$ for each unit, which may include lagged outcomes $Y_{i,t}^{obs}$ for $t \leq T_0$
  - $\mathbf{X}_c$ is $N \times M$ matrix that stacks Xs for control units

Notation

- $N+1$ units observed for T periods, with a subset of treated units (for simplicity - unit 0) treated from T_0 onwards
- Treatment : $D_{i,t} = \mathbb{1}_{i=0 \wedge t \in T_0+1,\ldots,T}$
- Potential outcomes for unit 0 define the treatment effect: $\tau_{0,t} := Y_{0,t}(1) - Y_{0,t}(0)$ for $t = T_0 + 1, \ldots, T$
- Observed outcome: $Y_{i,t}^{obs} = Y_{i,t}(D_{i,t})$
- Time-invariant characteristics $X_i := (X_{i,1}, \ldots, X_{i,M})^\top$ for each unit, which may include lagged outcomes $Y_{i,t}^{obs}$ for $t \leq T_0$
  - $\mathbf{X}_c$ is $N \times M$ matrix that stacks Xs for control units
  - $\mathbf{X}_t$ is $M-$ row vector of covariates for control

Notation

- $N+1$ units observed for T periods, with a subset of treated units (for simplicity - unit 0) treated from T_0 onwards
- Treatment : $D_{i,t} = \mathbb{1}_{i=0 \wedge t \in T_0+1,\ldots,T}$
- Potential outcomes for unit 0 define the treatment effect: $\tau_{0,t} := Y_{0,t}(1) - Y_{0,t}(0)$ for $t = T_0 + 1, \ldots, T$
- Observed outcome: $Y_{i,t}^{obs} = Y_{i,t}(D_{i,t})$
- Time-invariant characteristics $X_i := (X_{i,1}, \ldots, X_{i,M})^\top$ for each unit, which may include lagged outcomes $Y_{i,t}^{obs}$ for $t \leq T_0$
  - $\mathbf{X}_c$ is $N \times M$ matrix that stacks Xs for control units
  - $\mathbf{X}_t$ is $M-$ row vector of covariates for control
  - stack them to get $\mathbf{X}$

Outcome Matrices

$$\mathbf{Y}^{\text{obs}} = \begin{bmatrix} \mathbf{Y}^{\text{obs}}_{\text{t, post}} & \mathbf{Y}^{\text{obs}}_{\text{c, post}} \\ \mathbf{Y}^{\text{obs}}_{\text{t, pre}} & \mathbf{Y}^{\text{obs}}_{\text{c, pre}} \end{bmatrix} = \begin{bmatrix} \mathbf{Y}_{\text{t, post}}(1) & \mathbf{Y}_{\text{c, post}}(0) \\ \mathbf{Y}_{\text{t, pre}}(0) & \mathbf{Y}_{\text{c, pre}}(0) \end{bmatrix} \qquad T \times (N+1)$$

$$\mathbf{Y}(0) = \begin{bmatrix} ? & \mathbf{Y}_{\text{c, post}}(0) \\ \mathbf{Y}_{\text{t, pre}}(0) & \mathbf{Y}_{\text{c, pre}}(0) \end{bmatrix} = \begin{bmatrix} ? & \mathbf{Y}_{\text{c, post}}(0) \\ \mathbf{Y}_{\text{t, pre}}(0) & \mathbf{Y}_{\text{c, pre}}(0) \end{bmatrix}$$

- relative magnitudes of T and N might dictate whether we impute the missing potential outcome $?$ using this or this comparison
  - **Many Units and Multiple Periods**: $N >> T_0$, $\mathbf{Y}(0)$ is 'fat', and red comparison becomes challenging relative to blue. So matching methods are attractive.

Outcome Matrices

$$\mathbf{Y}^{\text{obs}} = \begin{bmatrix} \mathbf{Y}^{\text{obs}}_{\text{t, post}} & \mathbf{Y}^{\text{obs}}_{\text{c, post}} \\ \mathbf{Y}^{\text{obs}}_{\text{t, pre}} & \mathbf{Y}^{\text{obs}}_{\text{c, pre}} \end{bmatrix} = \begin{bmatrix} \mathbf{Y}_{\text{t, post}}(1) & \mathbf{Y}_{\text{c, post}}(0) \\ \mathbf{Y}_{\text{t, pre}}(0) & \mathbf{Y}_{\text{c, pre}}(0) \end{bmatrix} \qquad T \times (N+1)$$

$$\mathbf{Y}(0) = \begin{bmatrix} ? & \mathbf{Y}_{\text{c, post}}(0) \\ \mathbf{Y}_{\text{t, pre}}(0) & \mathbf{Y}_{\text{c, pre}}(0) \end{bmatrix} = \begin{bmatrix} ? & \mathbf{Y}_{\text{c, post}}(0) \\ \mathbf{Y}_{\text{t, pre}}(0) & \mathbf{Y}_{\text{c, pre}}(0) \end{bmatrix}$$

- relative magnitudes of T and N might dictate whether we impute the missing potential outcome $?$ using this or this comparison
  - **Many Units and Multiple Periods**: $N >> T_0$, $\mathbf{Y}(0)$ is 'fat', and red comparison becomes challenging relative to blue. So matching methods are attractive.
  - $T_0 >> N$, $\mathbf{Y}(0)$ is 'tall', and matching becomes infeasible. So it might be easier to estimate blue dependence structure.

Outcome Matrices

$$\mathbf{Y}^{\text{obs}} = \begin{bmatrix} \mathbf{Y}^{\text{obs}}_{\text{t, post}} & \mathbf{Y}^{\text{obs}}_{\text{c, post}} \\ \mathbf{Y}^{\text{obs}}_{\text{t, pre}} & \mathbf{Y}^{\text{obs}}_{\text{c, pre}} \end{bmatrix} = \begin{bmatrix} \mathbf{Y}_{\text{t, post}}(1) & \mathbf{Y}_{\text{c, post}}(0) \\ \mathbf{Y}_{\text{t, pre}}(0) & \mathbf{Y}_{\text{c, pre}}(0) \end{bmatrix} \qquad T \times (N+1)$$

$$\mathbf{Y}(0) = \begin{bmatrix} ? & \mathbf{Y}_{\text{c, post}}(0) \\ \mathbf{Y}_{\text{t, pre}}(0) & \mathbf{Y}_{\text{c, pre}}(0) \end{bmatrix} = \begin{bmatrix} ? & \mathbf{Y}_{\text{c, post}}(0) \\ \mathbf{Y}_{\text{t, pre}}(0) & \mathbf{Y}_{\text{c, pre}}(0) \end{bmatrix}$$

- relative magnitudes of T and N might dictate whether we impute the missing potential outcome $?$ using this or this comparison
  - **Many Units and Multiple Periods**: $N >> T_0$, $\mathbf{Y}(0)$ is 'fat', and red comparison becomes challenging relative to blue. So matching methods are attractive.
  - $T_0 >> N$, $\mathbf{Y}(0)$ is 'tall', and matching becomes infeasible. So it might be easier to estimate blue dependence structure.
  - Finally, if $T_0 \approx N$, regularization strategy for limiting the number of control units that enter into the estimation of $Y_{0,T_0+1}(0)$ may be important

Common Structure: 4 assumptions

- Focus on last period for now: $\tau_{0,T} = Y_{0,T}(1) - Y_{0,T}(0) = Y_{0,T}^{\text{obs}} - Y_{0,T}(0)$

Common Structure: 4 assumptions

- Focus on last period for now: $\tau_{0,T} = Y_{0,T}(1) - Y_{0,T}(0) = Y_{0,T}^{\text{obs}} - Y_{0,T}(0)$
- Many estimators impute $Y_{0,T}(0)$ with the linear structure $\widehat{Y}_{0,T}(0) = \mu + \sum_{i=1}^{n} \omega_i \cdot Y_{i,T}^{\text{obs}}$

Common Structure: 4 assumptions

- Focus on last period for now: $\tau_{0,T} = Y_{0,T}(1) - Y_{0,T}(0) = Y_{0,T}^{\text{obs}} - Y_{0,T}(0)$
- Many estimators impute $Y_{0,T}(0)$ with the linear structure $\widehat{Y}_{0,T}(0) = \mu + \sum_{i=1}^{n} \omega_i \cdot Y_{i,T}^{\text{obs}}$
  - Methods differ in how μ and ω are chosen as a function of $\mathbf{Y}_{\text{c, post}}^{\text{obs}}$, $\mathbf{Y}_{\text{t, pre}}^{\text{obs}}$, $\mathbf{Y}_{\text{c, pre}}^{\text{obs}}$

Common Structure: 4 assumptions

- Focus on last period for now: $\tau_{0,T} = Y_{0,T}(1) - Y_{0,T}(0) = Y_{0,T}^{\text{obs}} - Y_{0,T}(0)$
- Many estimators impute $Y_{0,T}(0)$ with the linear structure $\widehat{Y}_{0,T}(0) = \mu + \sum_{i=1}^{n} \omega_i \cdot Y_{i,T}^{\text{obs}}$
  - Methods differ in how μ and ω are chosen as a function of $\mathbf{Y}_{\text{c, post}}^{\text{obs}}$, $\mathbf{Y}_{\text{t, pre}}^{\text{obs}}$, $\mathbf{Y}_{\text{c, pre}}^{\text{obs}}$
- Impose four constraints

Common Structure: 4 assumptions

- Focus on last period for now: $\tau_{0,T} = Y_{0,T}(1) - Y_{0,T}(0) = Y_{0,T}^{\text{obs}} - Y_{0,T}(0)$
- Many estimators impute $Y_{0,T}(0)$ with the linear structure $\widehat{Y}_{0,T}(0) = \mu + \sum_{i=1}^{n} \omega_i \cdot Y_{i,T}^{\text{obs}}$
  - Methods differ in how μ and ω are chosen as a function of $\mathbf{Y}_{\text{c, post}}^{\text{obs}}$, $\mathbf{Y}_{\text{t, pre}}^{\text{obs}}$, $\mathbf{Y}_{\text{c, pre}}^{\text{obs}}$
- Impose four constraints
  1. **No Intercept**: $\mu = 0$. Stronger than Parallel trends in DiD.

Common Structure: 4 assumptions

- Focus on last period for now: $\tau_{0,T} = Y_{0,T}(1) - Y_{0,T}(0) = Y_{0,T}^{\text{obs}} - Y_{0,T}(0)$
- Many estimators impute $Y_{0,T}(0)$ with the linear structure $\widehat{Y}_{0,T}(0) = \mu + \sum_{i=1}^{n} \omega_i \cdot Y_{i,T}^{\text{obs}}$
  - Methods differ in how μ and ω are chosen as a function of $\mathbf{Y}_{\text{c, post}}^{\text{obs}}$, $\mathbf{Y}_{\text{t, pre}}^{\text{obs}}$, $\mathbf{Y}_{\text{c, pre}}^{\text{obs}}$
- Impose four constraints
  1. **No Intercept**: $\mu = 0$. Stronger than Parallel trends in DiD.
  2. **Adding up** : $\sum_{i=1}^{n} \omega_i = 1$. Common to DiD, SC.

Common Structure: 4 assumptions

- Focus on last period for now: $\tau_{0,T} = Y_{0,T}(1) - Y_{0,T}(0) = Y_{0,T}^{\text{obs}} - Y_{0,T}(0)$
- Many estimators impute $Y_{0,T}(0)$ with the linear structure $\widehat{Y}_{0,T}(0) = \mu + \sum_{i=1}^{n} \omega_i \cdot Y_{i,T}^{\text{obs}}$
  - Methods differ in how μ and ω are chosen as a function of $\mathbf{Y}_{\text{c, post}}^{\text{obs}}$, $\mathbf{Y}_{\text{t, pre}}^{\text{obs}}$, $\mathbf{Y}_{\text{c, pre}}^{\text{obs}}$
- Impose four constraints
  1. **No Intercept**: $\mu = 0$. Stronger than Parallel trends in DiD.
  2. **Adding up** : $\sum_{i=1}^{n} \omega_i = 1$. Common to DiD, SC.
  3. **Non-negativity**: $\omega_i \geq 0 \; \forall \; i$. Ensures uniqueness via 'coarse' regularisation + precision control. Negative weights may improve out-of-sample prediction.

Common Structure: 4 assumptions

- Focus on last period for now: $\tau_{0,T} = Y_{0,T}(1) - Y_{0,T}(0) = Y_{0,T}^{\text{obs}} - Y_{0,T}(0)$
- Many estimators impute $Y_{0,T}(0)$ with the linear structure $\widehat{Y}_{0,T}(0) = \mu + \sum_{i=1}^{n} \omega_i \cdot Y_{i,T}^{\text{obs}}$
  - Methods differ in how μ and ω are chosen as a function of $\mathbf{Y}_{\text{c, post}}^{\text{obs}}$, $\mathbf{Y}_{\text{t, pre}}^{\text{obs}}$, $\mathbf{Y}_{\text{c, pre}}^{\text{obs}}$
- Impose four constraints
  1. **No Intercept**: $\mu = 0$. Stronger than Parallel trends in DiD.
  2. **Adding up** : $\sum_{i=1}^{n} \omega_i = 1$. Common to DiD, SC.
  3. **Non-negativity**: $\omega_i \geq 0 \; \forall \; i$. Ensures uniqueness via 'coarse' regularisation + precision control. Negative weights may improve out-of-sample prediction.
  4. **Constant Weights**: $\omega_i = \overline{\omega} \; \forall \; i$

Common Structure: 4 assumptions

- Focus on last period for now: $\tau_{0,T} = Y_{0,T}(1) - Y_{0,T}(0) = Y_{0,T}^{\text{obs}} - Y_{0,T}(0)$
- Many estimators impute $Y_{0,T}(0)$ with the linear structure $\widehat{Y}_{0,T}(0) = \mu + \sum_{i=1}^{n} \omega_i \cdot Y_{i,T}^{\text{obs}}$
  - Methods differ in how μ and ω are chosen as a function of $\mathbf{Y}_{\text{c, post}}^{\text{obs}}$, $\mathbf{Y}_{\text{t, pre}}^{\text{obs}}$, $\mathbf{Y}_{\text{c, pre}}^{\text{obs}}$
- Impose four constraints
  1. **No Intercept**: $\mu = 0$. Stronger than Parallel trends in DiD.
  2. **Adding up** : $\sum_{i=1}^{n} \omega_i = 1$. Common to DiD, SC.
  3. **Non-negativity**: $\omega_i \geq 0 \; \forall \; i$. Ensures uniqueness via 'coarse' regularisation + precision control. Negative weights may improve out-of-sample prediction.
  4. **Constant Weights**: $\omega_i = \overline{\omega} \; \forall \; i$
- DiD imposes 2-4.

Common Structure: 4 assumptions

- Focus on last period for now: $\tau_{0,T} = Y_{0,T}(1) - Y_{0,T}(0) = Y_{0,T}^{\text{obs}} - Y_{0,T}(0)$
- Many estimators impute $Y_{0,T}(0)$ with the linear structure $\widehat{Y}_{0,T}(0) = \mu + \sum_{i=1}^{n} \omega_i \cdot Y_{i,T}^{\text{obs}}$
  - Methods differ in how μ and ω are chosen as a function of $\mathbf{Y}_{\text{c, post}}^{\text{obs}}$, $\mathbf{Y}_{\text{t, pre}}^{\text{obs}}$, $\mathbf{Y}_{\text{c, pre}}^{\text{obs}}$
- Impose four constraints
  1. **No Intercept**: $\mu = 0$. Stronger than Parallel trends in DiD.
  2. **Adding up** : $\sum_{i=1}^{n} \omega_i = 1$. Common to DiD, SC.
  3. **Non-negativity**: $\omega_i \geq 0 \; \forall \; i$. Ensures uniqueness via 'coarse' regularisation + precision control. Negative weights may improve out-of-sample prediction.
  4. **Constant Weights**: $\omega_i = \overline{\omega} \; \forall \; i$
- DiD imposes 2-4.
- ADH(2010, 2014) impose 1-3

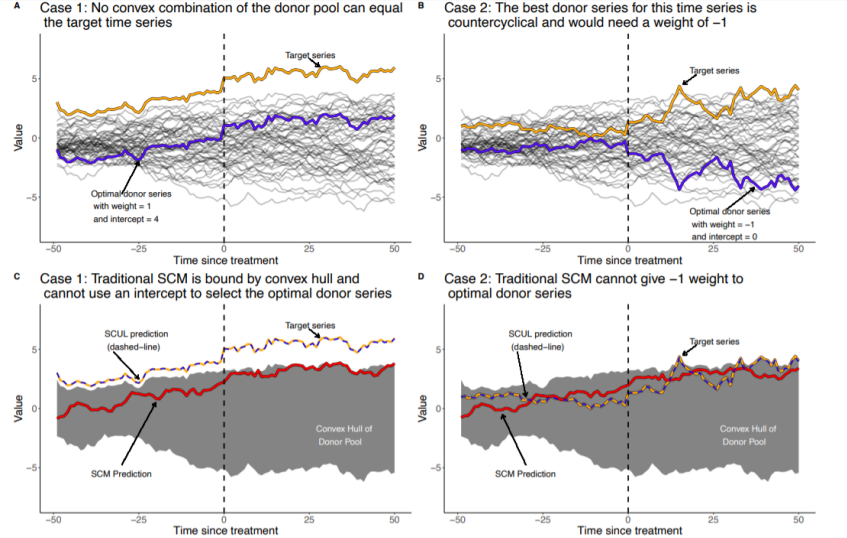
Common Structure: 4 assumptions

- Focus on last period for now: $\tau_{0,T} = Y_{0,T}(1) - Y_{0,T}(0) = Y_{0,T}^{\text{obs}} - Y_{0,T}(0)$
- Many estimators impute $Y_{0,T}(0)$ with the linear structure $\widehat{Y}_{0,T}(0) = \mu + \sum_{i=1}^{n} \omega_i \cdot Y_{i,T}^{\text{obs}}$
  - Methods differ in how μ and ω are chosen as a function of $\mathbf{Y}_{\text{c, post}}^{\text{obs}}$, $\mathbf{Y}_{\text{t, pre}}^{\text{obs}}$, $\mathbf{Y}_{\text{c, pre}}^{\text{obs}}$
- Impose four constraints
  1. **No Intercept**: $\mu = 0$. Stronger than Parallel trends in DiD.
  2. **Adding up** : $\sum_{i=1}^{n} \omega_i = 1$. Common to DiD, SC.
  3. **Non-negativity**: $\omega_i \geq 0 \; \forall \; i$. Ensures uniqueness via 'coarse' regularisation + precision control. Negative weights may improve out-of-sample prediction.
  4. **Constant Weights**: $\omega_i = \overline{\omega} \; \forall \; i$
- DiD imposes 2-4.
- ADH(2010, 2014) impose 1-3
  - 1 + 2 imply 'No Extrapolation'.

Relaxing the assumptions

- Negative weights
  - If treated units are outliers on important covariates, negative weights might improve fit
  - Bias reduction - negative weights increase bias-reduction rate
- When $N >> T_0$, (1-3) alone might not result in a unique solution. Choose by
  - Matching on pre-treatment outcomes : one good control unit is better than synthetic one comprised of disparate units
  - Constant weights - implicit in DiD
- Given many pairs of (μ, ω)
- prefer values s.t. synthetic control unit is similar to treated units in terms of lagged outcomes
- low dispersion of weights
- few control units with non-zero weights

Case for nonconvex or negative Weights : Hollingworth and Wing (2021)



The optimisation problem: general case

Ingredients of objective function

- **Balance**: difference between pre-treatment outcomes for treated and linear-combination of pre-treatment outcomes for control
  - $\left\|\mathbf{Y}_{\text{t, pre}} - \mu - \omega^\top \mathbf{Y}_{\text{c, pre}}\right\|_2^2 = (\mathbf{Y}_{\text{t, pre}} - \mu - \omega^\top \mathbf{Y}_{\text{c, pre}})^\top (\mathbf{Y}_{\text{t, pre}} - \mu - \omega^\top \mathbf{Y}_{\text{c, pre}})$
- **Sparse and small weights**:
  - sparsity : $\|\omega\|_1$
  - magnitude: $\|\omega\|_2$

$$(\widehat{\mu}^{en}(\lambda,\alpha), \widehat{\omega}^{en}(\lambda,\alpha)) = \underset{\mu,\omega}{\arg\min}\ Q(\mu,\omega|\mathbf{Y}_{\text{t, pre}}, \mathbf{Y}_{\text{c, pre}}; \lambda, \alpha)$$

where $$Q(\mu,\omega|\mathbf{Y}_{\text{t, pre}}, \mathbf{Y}_{\text{c, pre}}; \lambda, \alpha) = \left\|\mathbf{Y}_{\text{t, pre}} - \mu - \omega^\top \mathbf{Y}_{\text{c, pre}}\right\|_2^2 + \lambda\left(\frac{1-\alpha}{2}\|\omega\|_2^2 + \alpha\|\omega\|_1\right)$$

Choosing α, λ: Tailored regularisation

- don't want to scale covariates $\mathbf{Y}_{\text{c, pre}}$ to preserve interpretability of weights
- Instead, treat each control unit as a 'pseudo-treated' unit and compute $\widehat{Y}_{j,T}(0) = \widehat{\mu}^{\text{en}}(j;\alpha,\lambda) + \sum_{i \neq j} \widehat{\omega}_i(j;\alpha,\lambda) \cdot Y_{i,T}^{\text{obs}}$ where

$$(\widehat{\mu}^{en}(j;\lambda,\alpha), \widehat{\omega}^{en}(j;\lambda,\alpha)) = \underset{\mu,\omega}{\arg\min} \sum_{t=1}^{T_0} \left(Y_{j,t} - \mu - \sum_{i \neq 0,j} \omega_i Y_{i,t} \right)^2 + \lambda \left(\frac{1-\alpha}{2} \|\omega\|_2^2 + \alpha \|\omega\|_1 \right)$$

pick the value of the tuning parameters $(\alpha_{opt}^{en}, \lambda_{opt}^{en})$ that minimises

$$CV^{en}(\alpha,\lambda) = \frac{1}{N} \sum_{j=1}^{N} (Y_{j,T} - \overbrace{\widehat{\mu}^{en}(j;\alpha,\lambda) - \sum_{i \neq 0,j} \widehat{\omega}_i^{en}(j;\alpha,\lambda) \cdot Y_{i,T}}^{\widehat{Y}_{j,T}(0)})$$

Re-expressing Standard Methods

Difference in Differences

- assume (2-4)
- No unique μ, ω solution for $T = 2$, so fix $\omega = \frac{1}{N}$

$$\omega_i^{\text{did}} = \frac{1}{N} \quad \forall i \in \{1, \dots N\}$$

$$\widehat{\mu}^{\text{did}} = \frac{1}{T_0} \sum_{s=1}^{T_0} Y_{0,s} - \frac{1}{NT_0} \sum_{s=1}^{T_0} \sum_{i=1}^{N} Y_{i,s}$$

Best Subset; One-to-one Matching

$(\widehat{\mu}^S, \widehat{\omega}^S) = \arg\min_{\mu,\omega} Q(\cdot; \lambda = 0, \alpha)$ with $\sum_{i=1}^N \mathbb{1}_{\omega_i \neq 0} \leq k$ (=1 for OtO)

Synthetic Control

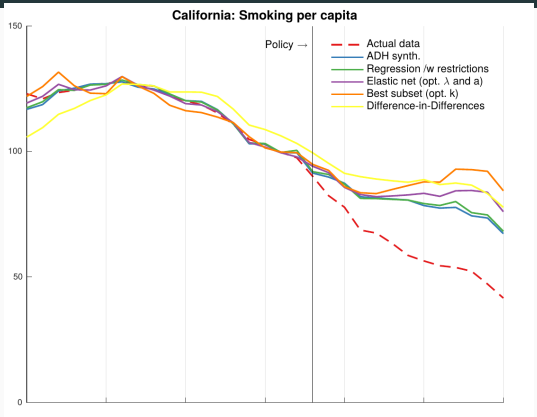
- assume (1-3) (i.e. $\mu = 0$)
- For $M \times M$ PSD diagonal matrix $\mathbf{V}$

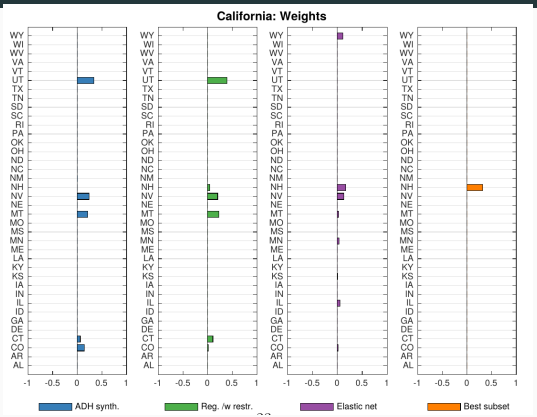
$$(\widehat{\omega}(\mathbf{V}), \widehat{\mu}(\mathbf{V})) = \arg\min_{\omega,\mu} \{(\mathbf{X}_t - \mu - \omega^\top \mathbf{X})^\top \mathbf{V} (\mathbf{X}_t - \mu - \omega^\top \mathbf{X})\}$$

$$\widehat{\mathbf{V}} = \arg\min_{\mathbf{V}=\text{diag}(v_1,\dots,v_M)} \{(\mathbf{Y}_{\text{t, pre}} - \widehat{\omega}(\mathbf{V})^\top \mathbf{Y}_{\text{c, pre}})^\top (\mathbf{Y}_{\text{t, pre}} - \widehat{\omega}(\mathbf{V})^\top \mathbf{Y}_{\text{c, pre}})\}$$

Constrained regression: When $X_i = Y_{i,t};\ 1 \leq t \leq T_0$ (Lagged Outcomes only) $\mathbf{V} = \mathbf{I}_N$ and $\lambda = 0$

Revisiting ADH California smoking example





Model	$\sum_i \omega_i$	μ	$\widehat{\tau}$	s.e.
Original Synth	1	0	-22.1	16.1
Constrained	1	0	-22.9	12.8
Elastic Net	.55	18.5	-26.9	16.8
Best Subset	.32	37.6	-31.9	20.3
Diff-in-Diff	1	-14.4	-32.4	18.9

Roadmap

Panel data with multiple time periods

Synthetic Control Methods

Instrumental Variables I

Big Picture

- IV methods can solve the problem of omitted variable bias when trying to isolate a causal effect using observational data.

Big Picture

- IV methods can solve the problem of omitted variable bias when trying to isolate a causal effect using observational data.
- The setting is selection on unobservables. Originally developed to address simultaneity in the 1920s

Big Picture

- IV methods can solve the problem of omitted variable bias when trying to isolate a causal effect using observational data.
- The setting is selection on unobservables. Originally developed to address simultaneity in the 1920s
- There are two broad types of IV frameworks: the (older) one that assumes constant treatment effects, and Potential-outcome-based one that assumes heterogeneous treatment effects.

Big Picture

- IV methods can solve the problem of omitted variable bias when trying to isolate a causal effect using observational data.
- The setting is selection on unobservables. Originally developed to address simultaneity in the 1920s
- There are two broad types of IV frameworks: the (older) one that assumes constant treatment effects, and Potential-outcome-based one that assumes heterogeneous treatment effects.
- The estimators we use in each framework are the same, but the assumptions and interpretation of the treatment effect are slightly different.

Big Picture

- IV methods can solve the problem of omitted variable bias when trying to isolate a causal effect using observational data.
- The setting is selection on unobservables. Originally developed to address simultaneity in the 1920s
- There are two broad types of IV frameworks: the (older) one that assumes constant treatment effects, and Potential-outcome-based one that assumes heterogeneous treatment effects.
- The estimators we use in each framework are the same, but the assumptions and interpretation of the treatment effect are slightly different.
- We'll start with the constant treatment effects framework to motivate the estimators, and then move on (next week) to the heterogeneous treatment effect framework, which is more common nowadays

Setup

Suppose we are interested in the effect of schooling (s_i) on wages (Y_i). Using a selection-on-observables story with constant treatment effects, we know that conditional on a vector of control variables for "ability" (A_i), the causal model is

$$Y_i = \alpha + \tau s_i + A_i'\gamma + v_i$$

Setup

Suppose we are interested in the effect of schooling (s_i) on wages (Y_i). Using a selection-on-observables story with constant treatment effects, we know that conditional on a vector of control variables for "ability" (A_i), the causal model is

$$Y_i = \alpha + \tau s_i + A_i'\gamma + v_i$$

If we could observe A_i, we would just estimate this regression and be done. But what if we can't observe A_i? If we controlled for nothing, we would estimate

$$Y_i = \alpha + \tau s_i + e_i$$

What are we worried about in the short regression?

Setup

Suppose we are interested in the effect of schooling (s_i) on wages (Y_i). Using a selection-on-observables story with constant treatment effects, we know that conditional on a vector of control variables for "ability" (A_i), the causal model is

$$Y_i = \alpha + \tau s_i + A_i'\gamma + v_i$$

If we could observe A_i, we would just estimate this regression and be done. But what if we can't observe A_i? If we controlled for nothing, we would estimate

$$Y_i = \alpha + \tau s_i + e_i$$

What are we worried about in the short regression? Correlation between s_i and e_i. We know from our selection-on-observables story that this correlation is entirely captured by A_i.

Omitted Variable Bias

Formally, the naive OLS estimate of τ is:

$$\begin{aligned}\hat{\tau}_{ols} &= \frac{Cov(s_i, Y_i)}{Var(s_i)} = \frac{Cov(s_i, \alpha + \tau s_i + e_i)}{Var(s_i)} \\ &= \frac{\tau Cov(s_i, s_i) + Cov(s_i, e_i)}{Var(s_i)} = \tau + \frac{Cov(s_i, e_i)}{Var(s_i)} \\ &= \tau + \frac{Cov(s_i, A_i'\gamma + v_i)}{Var(s_i)} \\ &= \tau + \gamma' \frac{Cov(s_i, A_i)}{Var(s_i)} + \frac{Cov(s_i, v_i)}{Var(s_i)}\end{aligned}$$

Omitted Variable Bias

Formally, the naive OLS estimate of τ is:

$$\begin{aligned}\hat{\tau}_{ols} &= \frac{Cov(s_i, Y_i)}{Var(s_i)} = \frac{Cov(s_i, \alpha + \tau s_i + e_i)}{Var(s_i)} \\ &= \frac{\tau Cov(s_i, s_i) + Cov(s_i, e_i)}{Var(s_i)} = \tau + \frac{Cov(s_i, e_i)}{Var(s_i)} \\ &= \tau + \frac{Cov(s_i, A_i'\gamma + v_i)}{Var(s_i)} \\ &= \tau + \gamma' \frac{Cov(s_i, A_i)}{Var(s_i)} + \frac{Cov(s_i, v_i)}{Var(s_i)}\end{aligned}$$

Our SOO story assumes $\mathbb{E}[s_i v_i] = 0$, so the expected value of $\hat{\tau}_{ols}$ is

$$\mathbb{E}[\hat{\tau}_{ols}] = \tau + \gamma' \mathbb{E}\left[\frac{Cov(s_i, A_i)}{Var(s_i)}\right]$$

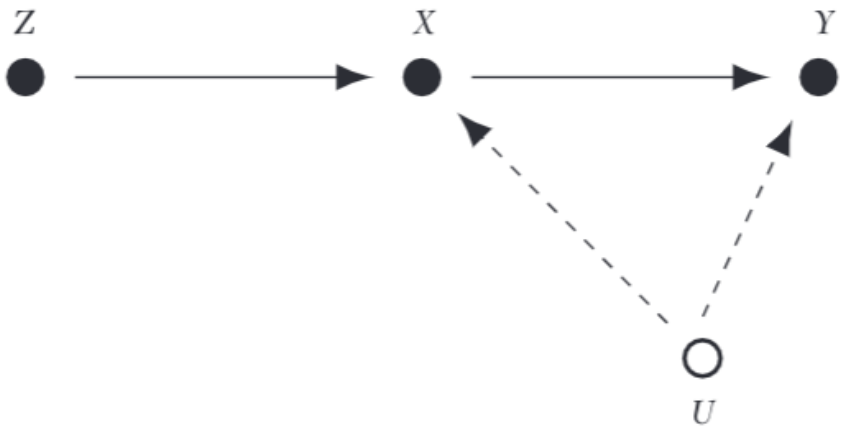
Finding an instrument

- We can see that the problem with using the schooling variable as measured in the wild is that it produces a non-zero $Cov(s_i, A_i)$. But what if we could find another variable, z_i, that is correlated with s_i but not with A_i or v_i? In other words, z_i produces as-if randomized variation in schooling, and is only correlated with wages through schooling. What could be some possibilities for this kind of instrument?

Finding an instrument

- We can see that the problem with using the schooling variable as measured in the wild is that it produces a non-zero $Cov(s_i, A_i)$. But what if we could find another variable, z_i, that is correlated with s_i but not with A_i or v_i? In other words, z_i produces as-if randomized variation in schooling, and is only correlated with wages through schooling. What could be some possibilities for this kind of instrument?
- One example used by Angrist and Krueger (1991) is the variation induced in years of schooling by the fact that most states require students to start school in the calendar year that they turn 6 years old. This means that kids born at the beginning of the calendar year are older when they start school than kids born at the end of the year, and the two groups will have had different amounts of time in school when they reach the legal dropout age at 16.

IV DAG



IV Setup

- We suppress controls without loss-of-generality since, by the FWL, one can eliminate the controls $\boldsymbol{c}_i$ in the structural equation by regressing $Y, X,$ and Z on $\boldsymbol{c}_i$ and using the residuals $\widetilde{Y}, \widetilde{X}, \widetilde{Z}$ for all subsequent computation.

IV Setup

- We suppress controls without loss-of-generality since, by the FWL, one can eliminate the controls $\boldsymbol{c}_i$ in the structural equation by regressing Y, X, and Z on $\boldsymbol{c}_i$ and using the residuals $\widetilde{Y}, \widetilde{X}, \widetilde{Z}$ for all subsequent computation.

$$\begin{aligned}\text{Structural Equation}: Y &= \alpha_0 + \beta X + \varepsilon \\ \text{First Stage}: X &= \pi_0 + \pi Z + \eta \\ \text{Reduced Form}: Y &= \gamma_0 + \gamma Z + \upsilon\end{aligned}$$

$$\begin{aligned}Y &= \alpha_0 + \beta X + \varepsilon \\ &= \alpha_0 + \beta(\pi_0 + \pi Z + \eta) + \varepsilon \qquad \text{substitute in X from first-stage} \\ &= \underbrace{(\alpha_0 + \beta\pi_0)}_{\gamma_0} + \underbrace{(\beta\pi)}_{\gamma} Z + (\beta\eta + \varepsilon) \implies \gamma = \beta\pi \ \to \beta = \frac{\gamma}{\pi}\end{aligned}$$

IV Assumptions

Assumption (Exclusion Restriction / Validity)

X: $\varepsilon_i \perp (Z_i, \boldsymbol{c_i})$

This requires that Z has no direct effect on Y except through X, where ε_i is the residual in the strictural equation. The instrument needs to be uncorrelated with unobservables in the structural equation, potentially conditional on controls $\boldsymbol{c_i}$.

Assumption (Relevance)

Z affects X, i.e. Cov $[Z, X] \neq 0$ or $\pi_1 \neq 0$.

Which of these is testable in our sample data?

IV Assumptions

Assumption (Exclusion Restriction / Validity)

X: $\varepsilon_i \perp (Z_i, \boldsymbol{c_i})$

This requires that Z has no direct effect on Y except through X, where ε_i is the residual in the strictural equation. The instrument needs to be uncorrelated with unobservables in the structural equation, potentially conditional on controls $\boldsymbol{c_i}$.

Assumption (Relevance)

Z affects X, i.e. Cov $[Z, X] \neq 0$ or $\pi_1 \neq 0$.

Which of these is testable in our sample data? Non-zero first stage. Can you think of any critiques of the quarter of birth instrument based on these assumptions?

Two-stage least squares

The 2SLS coefficient is equivalent to the IV estimator:

$$
\begin{aligned}
\hat{\beta} &= \frac{\text{Cov}(Y_i, \hat{x}_i)}{\text{Var}(\hat{x}_i)} = \frac{\text{Cov}(Y_i, \widehat{\pi_0} + \hat{\pi} z_i)}{\text{Var}(\widehat{\pi_0} + \hat{\pi} z_i)} \\
&= \frac{\hat{\pi}\text{Cov}(Y_i, z_i)}{\hat{\pi}^2 \text{Var}(z_i)} = \frac{\text{Cov}(Y_i, z_i)}{\hat{\pi}\text{Var}(z_i)} \\
&= \frac{\text{Cov}(Y_i, z_i)}{\frac{\text{Cov}(x_i, z_i)}{\text{Var}(z_i)} \cdot \text{Var}(z_i)} = \frac{\text{Cov}(Y_i, z_i)}{\text{Cov}(x_i, z_i)} = \frac{\text{Reduced Form}}{\text{First Stage}}
\end{aligned}
$$

$$
\widehat{\beta}_{IV} = (\mathbf{Z}'\mathbf{X})^{-1} \mathbf{Z}'\boldsymbol{y}
$$

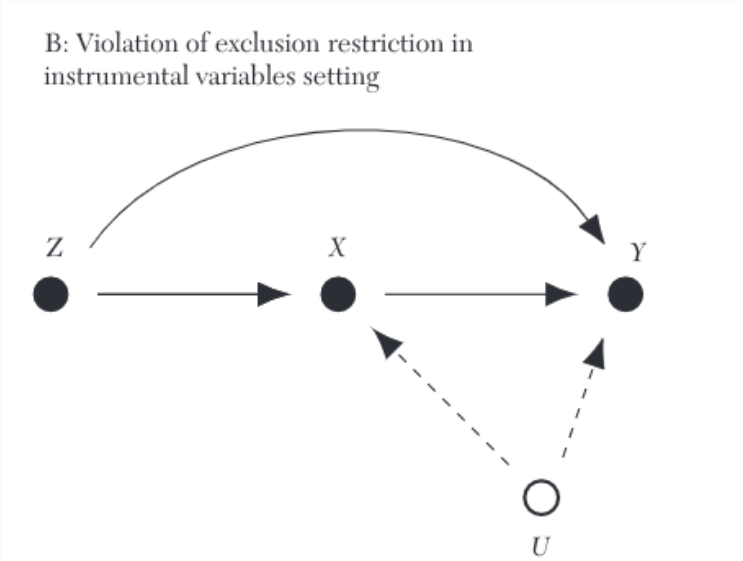
Multiple Instruments

If you have multiple valid instruments (pigs may fly), the matrix analogue is

$$\beta_{2SLS} = \left(\mathbf{X}'\mathbf{P}_z\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{P}_z\boldsymbol{y}$$

- where $\mathbf{P}_z = \mathbf{Z}\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'$ is the hat-maker matrix from the first-stage which projects the endogenous variables $\mathbf{X}$ into the column space of $\mathbf{Z}$
- this preserves only the 'clean' variation that is uncorrelated with ε.

BAD IV DAG


B: Violation of exclusion restriction in instrumental variables setting
Z
X
Y
U

Exclusion Restriction Violations

- If instrument is 'imperfect' (exclusion restriction violation) : $\text{Cov}\left[\varepsilon_i, Z_i\right] \neq 0$
  - Then $\text{Cov}\left[Y_i, Z_i\right] = \beta \text{Cov}\left[X_i, Z_i\right] + \text{Cov}\left[\varepsilon_i, Z_i\right]$.
  - Then the ratio of RF/FS is

$$\frac{\text{Cov}\left[Y_i, Z_i\right]}{\text{Cov}\left[X_i, Z_i\right]} = \beta + \frac{\text{Cov}\left[\varepsilon_i, Z_i\right]}{\text{Cov}\left[X_i, Z_i\right]} = \beta + \underbrace{\frac{\text{Cor}(\varepsilon_i, Z_i)}{\text{Cor}(X_i, Z_i)} \frac{\sigma_\varepsilon}{\sigma_x}}_{\text{Bias}}$$

  - Bias potentially very large if $\text{Cov}\left[X_i, Z_i\right] \approx 0 \implies$ problems compound each other
- To learn more about problems and potential fixes, come to Monday's departmental seminar [Lal, Lockhart, Xu, Zu 2021]

Weak Instruments

- If instrument is weak (i.e. Cov $[X_i, Z_i] \approx 0$), you're dividing by zero
- 2SLS is consistent, but is biased in small samples. The bias is worse with a weak instrument (even worse with multiple weak instruments).

Weak Instruments

- If instrument is weak (i.e. Cov $[X_i, Z_i] \approx 0$), you're dividing by zero
- 2SLS is consistent, but is biased in small samples. The bias is worse with a weak instrument (even worse with multiple weak instruments).
- In the worst-case scenario (multiple weak instruments that produce no first stage), 2SLS sampling distribution is centered on the probability limit of OLS.
- The first stage estimates reflect some of the randomness in the endogenous variable. If the population first stage is 0, then the resulting first stage coefficient just reflect randomness in the endogenous regressor.

Weak Instruments

- If instrument is weak (i.e. Cov $[X_i, Z_i] \approx 0$), you're dividing by zero
- 2SLS is consistent, but is biased in small samples. The bias is worse with a weak instrument (even worse with multiple weak instruments).
- In the worst-case scenario (multiple weak instruments that produce no first stage), 2SLS sampling distribution is centered on the probability limit of OLS.
- The first stage estimates reflect some of the randomness in the endogenous variable. If the population first stage is 0, then the resulting first stage coefficient just reflect randomness in the endogenous regressor.

$$\mathbb{E}\left[\widehat{\beta}_{2SLS} - \beta\right] \approx \overbrace{\frac{\sigma_{\eta\varepsilon}}{\sigma^2_{\varepsilon}}}^{\text{Bias of OLS}} \frac{1}{F+1}$$

- where F is the first-stage F statistic. As $F \to 0$ (i.e. the instrument is weak), the bias of the IV tends to the bias of the OLS
- Moral of the story: Always check the F-statistic for the instrument in your first stage (bigger than 10 is considered "safe" - this rule-of-thumb keeps growing; compute 'Effective F-stat').

Standard errors

- You can manually construct the 2SLS estimate using `lm`, but you can't use the OLS standard errors from the second stage regression.

Standard errors

- You can manually construct the 2SLS estimate using `lm`, but you can't use the OLS standard errors from the second stage regression.
- This is because the second stage OLS standard errors use the variance of the residuals against the fitted values from the first stage. What you really want is the variance of the residuals using the IV coefficients and the original endogenous regressor

Standard errors

- You can manually construct the 2SLS estimate using `lm`, but you can't use the OLS standard errors from the second stage regression.
- This is because the second stage OLS standard errors use the variance of the residuals against the fitted values from the first stage. What you really want is the variance of the residuals using the IV coefficients and the original endogenous regressor
- Moral of the story: use a canned routine like `ivreg` or `felm` to make the adjustment for you.

Standard errors

- You can manually construct the 2SLS estimate using `lm`, but you can't use the OLS standard errors from the second stage regression.
- This is because the second stage OLS standard errors use the variance of the residuals against the fitted values from the first stage. What you really want is the variance of the residuals using the IV coefficients and the original endogenous regressor
- Moral of the story: use a canned routine like `ivreg` or `felm` to make the adjustment for you.
- Analytic standard errors for IV are generally a mess (Alwyn Young - 'Consistency without Inference') - use the bootstrap whenever possible

Standard errors

- You can manually construct the 2SLS estimate using `lm`, but you can't use the OLS standard errors from the second stage regression.
- This is because the second stage OLS standard errors use the variance of the residuals against the fitted values from the first stage. What you really want is the variance of the residuals using the IV coefficients and the original endogenous regressor
- Moral of the story: use a canned routine like `ivreg` or `felm` to make the adjustment for you.
- Analytic standard errors for IV are generally a mess (Alwyn Young - 'Consistency without Inference') - use the bootstrap whenever possible
- More on this next week