Lecture 12 – Difference-in-Differences and related methods

Economics 8379
George Washington University

Instructor: Prof. Ben Williams

Difference-in-Differences

- Consider a panel data setting where we observe entitites $i = 1, \ldots, n$ in periods $t = 1, \ldots, T$.
- Consider a binary treatment, D_{it}, and outcome Y_{it}

Difference-in-Differences

- Consider a panel data setting where we observe entitites $i = 1, \ldots, n$ in periods $t = 1, \ldots, T$.
- Consider a binary treatment, D_{it}, and outcome Y_{it}
  - potential outcomes in period t: Y_{0it} and Y_{1it}
  - then $Y_{it} = Y_{0it} + D_{it}(Y_{1it} - Y_{0it})$
- in the typical DD setup, $D_{it} = G_i \times \mathbf{1}(t \geq t_0)$

Difference-in-Differences

- A cross-section estimator:
  $E(Y_{it} \mid D_{it} = 1) - E(Y_{it} \mid D_{it} = 0)$

Difference-in-Differences

- A cross-section estimator:
  $E(Y_{it} \mid D_{it} = 1) - E(Y_{it} \mid D_{it} = 0)$
  - identifying assumption: $E(Y_{0it} \mid D_{it} = 1) = E(Y_{0it} \mid D_{it} = 0)$

Difference-in-Differences

- A cross-section estimator: $E(Y_{it} \mid D_{it} = 1) - E(Y_{it} \mid D_{it} = 0)$
  - identifying assumption: $E(Y_{0it} \mid D_{it} = 1) = E(Y_{0it} \mid D_{it} = 0)$
- A before-after estimator: $E(Y_{it} \mid D_{it} = 1, D_{i(t-1)} = 0) - E(Y_{i(t-1)} \mid D_{it} = 1, D_{i(t-1)} = 0)$
  - identifying assumption: $E(Y_{0it} \mid D_{it} = 1, D_{i(t-1)} = 0) = E(Y_{0i(t-1)} \mid D_{it} = 1, D_{i(t-1)} = 0)$

Difference-in-Differences

The problems addressed by DiD model:

- A cross-sectional estimator misses selection into "treatment" group.

Difference-in-Differences

The problems addressed by DiD model:

- A cross-sectional estimator misses selection into "treatment" group.
- A before-after estimator will be biased
  - if there are time trends
  - if selection is based on $Y_{i,t-1}$

Difference-in-Differences

A DiD estimator:

$$
\begin{aligned}
DD = {} & E(Y_{it} - Y_{i(t-1)} \mid D_{it} = 1, D_{i,t-1} = 0) \\
& - E(Y_{it} - Y_{i(t-1)} \mid D_{it} = 0, D_{i,t-1} = 0)
\end{aligned}
$$

Difference-in-Differences

A DiD estimator:

$$DD = E(Y_{it} - Y_{i(t-1)} \mid D_{it} = 1, D_{i,t-1} = 0) \\ - E(Y_{it} - Y_{i(t-1)} \mid D_{it} = 0, D_{i,t-1} = 0)$$

- identifying assumption:

$$E(Y_{0it} - Y_{0i(t-1)} \mid D_{it} = 1, D_{i(t-1)} = 0) \\ = E(Y_{0it} - Y_{0i(t-1)} \mid D_{it} = 0, D_{i(t-1)} = 0)$$

Difference-in-Differences

- Keys to assessing the identifying assumptions:
  - dependence/nonstationarity in Y_{0it}
  - a model for D_{it} – heterogeneity in returns? independence of costs? opportunity costs? information available to agent?
  - can additional controls help?
  - lagged Y_{it} as a control?
- Heckman, LaLonde, Smith (1999), especially section 6, give a good summary of these issues...evaluation of a job training program

Difference-in-Differences

- MHE motivate DD with the assumption:

$$E(Y_{0it}) = \gamma_i + \lambda_t$$

Difference-in-Differences

- MHE motivate DD with the assumption:

$$E(Y_{0it}) = \gamma_i + \lambda_t$$

- Under this and the previous assumption, $DD = E(Y_{1it} - Y_{0it} \mid D_{it} = 1, D_{i(t-1)} = 0)$

Difference-in-Differences

- MHE motivate DD with the assumption:

$$E(Y_{0it}) = \gamma_i + \lambda_t$$

- Under this and the previous assumption, $DD = E(Y_{1it} - Y_{0it} \mid D_{it} = 1, D_{i(t-1)} = 0)$
- It is typically to additionally assume (as in MHE) that $Y_{1it} - Y_{0it} = \delta$ is constant.

Differences-in-Differences

The John Snow cholera example:

- Two districts in London serviced by two different water companies.
- One company moved its waterworks upriver to avoid sewage contamination.
- The district serviced by that company experienced a relative drop in cholera incidence.

Differences-in-Differences

The John Snow cholera example:

TABLE XII.

Sub-Districts.	Deaths from Cholera in 1849.	Deaths from Cholera in 1854.	Water Supply.
St. Saviour, Southwark	283	371	
St. Olave	157	161	
St. John, Horsleydown	[illegible]	[illegible]	
Norwood	2	10	Lambeth Company only.
Streatham	154	15	
Dulwich	1	—	
Sydenham	5	12	
First 12 sub-districts	2261	2458	Southwk. & Vauxhall.
Next 16 sub-districts	3905	2547	Both Companies.
Last 4 sub-districts	162	37	Lambeth Company.

More examples

- Card and Krueger (1994)
  - employer-level data
  - policy is minimum wage, which is at the state level
- LaLonde (1986)
  - worker-level data
  - “treatment” is participation in the program; at the individual level

No regressors

- Let G_i be a dummy for treatment/control group.
- Let Y_{itd} be a potential outcome for $d \in \{0, 1\}$.
- **Common trend:**
  $E(Y_{i20} - Y_{i10} \mid G_i = 1) = E(Y_{i20} - Y_{i10} \mid G_i = 0)$

No regressors

- Let G_i be a dummy for treatment/control group.
- Let Y_{itd} be a potential outcome for $d \in \{0, 1\}$.
- **Common trend:** $E(Y_{i20} - Y_{i10} \mid G_i = 1) = E(Y_{i20} - Y_{i10} \mid G_i = 0)$
- Then

$$\begin{aligned} & E(Y_{i2} \mid G_i = 1) - E(Y_{i1} \mid G_i = 1) \\ & - (E(Y_{i2} \mid G_i = 0) - E(Y_{i1} \mid G_i = 0)) \\ & = E(Y_{i21} - Y_{i20} \mid G_i = 1) \end{aligned}$$

("treatment on the treated")

No regressors

- Let P_t be a dummy for post-/pre-treatment time period.
- The simple DiD regression model is

$$Y_{it} = \beta_0 + \beta_1 G_i + \beta_2 P_t + \beta_3 G_i P_t + \varepsilon_{it}$$

- Here, β_3 is the treatment on the treated under the common trend assumption.
- Even if there is treatment effect heterogeneity (so the regression equation is misspecified) estimating this regression is a valid estimate of the TT.

With regressors

- With additional regressors,

$$Y_{it} = \beta_0 + \beta_1 G_i + \beta_2 P_t + \beta_3 G_i P_t + \beta' X_{it} + \varepsilon_{it}$$

- **Common trend:** $E(Y_{i20} - Y_{i10} \mid G_i = 1, X_{i1}, X_{i2}) = E(Y_{i20} - Y_{i10} \mid G_i = 0, X_{i1}, X_{i2})$
- it can be shown that this estimates a weighted average of “conditional on X treatment on the treated” parameters
- alternatively, we can compute a diff-in-diff matching (PS matching) estimator.

the right control group

- define G_i in a way that is not affected by treatment...
  - for example, migration...

the right control group

- define G_i in a way that is not affected by treatment...
  - for example, migration...
  - then it is an intent to treat effect.

control variables

- control variables should be included to address common trend
  - include variables X_{it} that explain relative changes over time in treatment vs control
  - include $W_i \times \mathbf{1}(t = \tau)$ that explain relative changes over time in treatment vs control

Event study

- Dynamic treatment effects/assessing common trends when $T > 2$.
- Suppose treatment starts in period t_0.
- Let P_{tk} be a dummy variable equal to 1 in period $t_0 + k$ and 0 in other periods.
- We can estimate the regression model

$$Y_{it} = \beta_1 G_i + \delta_t + \sum_{k=-t_0+2}^{T-t_0} \gamma_k P_{tk}$$

  - The untestable assumption now is that the trend between period 1 and 2 is the same for the two groups.
  - Given this assumption, we can test whether there is a common trend in the entire pre-treatment period by looking at γ_k for $k < 0$.

Event study

- Dynamic treatment effects/assessing common trends when $T > 2$.
- Suppose treatment starts in period t_0.
- Let P_{tk} be a dummy variable equal to 1 in period $t_0 + k$ and 0 in other periods.
- We can estimate the regression model

$$Y_{it} = \beta_1 G_i + \delta_t + \sum_{k=-t_0+2}^{T-t_0} \gamma_k P_{tk}$$

  - The untestable assumption now is that the trend between period 1 and 2 is the same for the two groups.
  - Given this assumption, we can test whether there is a common trend in the entire pre-treatment period by looking at γ_k for $k < 0$.
  - If we assume that the TT is constant, i.e., $E(Y_{it1} - Y_{it0} \mid G_i = 1) = TT$ for all $t \geq t_0$, then we can average γ_k for $k \geq 0$ to get an estimate of TT.

non-common trends

- suppose we fail the pre-trend test (or anticipate doing so...)
  - entity-specific trends
    - of order q if $T \geq q + 2$
  - interactive fixed effects model: $\gamma_{0i} + \lambda_{0t} + \gamma_{1i}\lambda_{1t}$
    - also requires more time periods - more on this next class

intensity of treatment

- in some cases everyone is treated but some are treated more intensely
    - federal minimum wage increase...youth employment affected more/less depending on what the old minimum wage was in each state

triple difference

- in some cases treatment intensity varies within treated groups
  - for example, some states implement a new policy but this new policy only affects some groups within the state
  - diff-in-diff-in-diff is implemented by including state-by-year, state-by-group, and year-by-group fixed effects

Assessing the common trend assumption

Other concerns:

- how to use past/future values of dependent variable?
  - differences or controls?
  - over what window do you expect common trend assumption to hold?
  - permanent or temporary effect?
  - length of exposure to treatment?

Ashenfelter's dip

- One concern is that participation in treatment or a policy change happens in response to recent outcomes experienced.
- More generally, the time series patterns in outcomes can be a source of bias even in the individual-level treatment framework.
    - Ashenfelter (1978)
    - Heckman and Smith (1999)
    - Heckman, LaLonde and Smith (1999)

Ashenfelter's dip

- Ashenfelter's dip

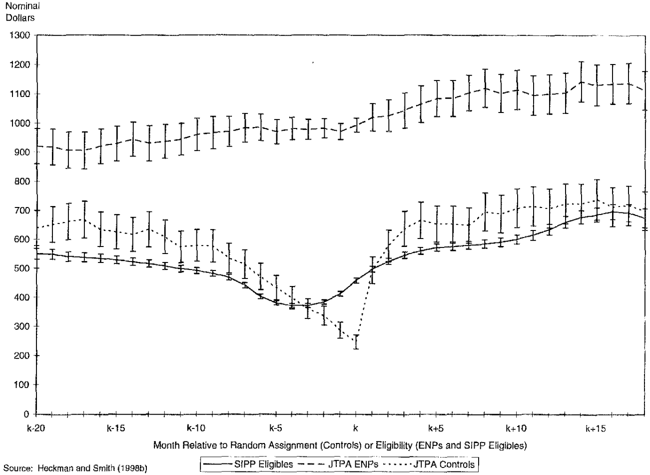


Fig. 1. Mean self-reported monthly earnings: National JTPA Study controls and eligible non-participants (ENPs) and SIPP eligibles (male adults). Source: Heckman and Smith (1999).

Ashenfelter's dip

- is the dip permanent or transitory?
- it's important to consider the dynamic process before treatment
- and the determinants of treatment choice
- Ashenfelter's dip can be present in state-level DiD application as well.

The linear panel model

- Basic model and assumptions:

$$y_{it} = \beta' x_{it} + \eta_i + \nu_{it}$$

A1 $E(\nu_{i1}, \ldots, \nu_{iT} \mid x_{i1}, \ldots, x_{iT}, \eta_i) = 0$
A2 $Var(\nu_{i1}, \ldots, \nu_{iT} \mid x_{i1}, \ldots, x_{iT}, \eta_i) = \sigma^2 I_T$

- These assumptions can be replaced by weaker but harder to interpret assumptions.

Differencing and within variation

- Some notation first:
  - $y_i = (y_{i1}, \ldots, y_{iT})'$
  - $x_i = (x_{i1}, \ldots, x_{iT})'$
  - $\nu_i = (\nu_{i1}, \ldots, \nu_{iT})'$

- The basic idea you've seen before:

$$\Delta y_{it} = \beta' \Delta x_{it} + \Delta \nu_{it}$$
$$\text{and } E(\Delta \nu_{it} \mid \Delta x_{it}) = 0$$

- In matrix notation,

$$D y_i = D x_i \beta + D \nu_i$$

where D is the $(T-1) \times T$ first difference operator.

Differencing and within variation

- The fixed effects regression is *not* $(\sum_{i=1}^{n} x_i' D' D x_i)^{-1} \sum_{i=1}^{n} x_i' D' D y_i$, though this *first differences* estimator would be consistent under assumption A1.

Differencing and within variation

- The fixed effects regression is *not* $(\sum_{i=1}^{n} x_i' D' D x_i)^{-1} \sum_{i=1}^{n} x_i' D' D y_i$, though this *first differences* estimator would be consistent under assumption A1.
- Because $Var(D\nu_i \mid x_i) = \sigma^2 DD'$, the GLS estimator is more efficient,

$$\hat{\beta}_{fe} := (\sum_{i=1}^{n} x_i' D' (DD')^{-1} D x_i)^{-1} \sum_{i=1}^{n} x_i' D' (DD')^{-1} D y_i$$

Differencing and within variation

- But $Q = D'(DD')^{-1}D$ is idempotent and equal to $I_T - \iota\iota'/T$. This is the within-group operator.
  - The fixed effects estimator is based on *within* variation.
  - The fixed effects estimator is equivalent to including entity dummies.

Differencing and within variation

- Properties of the fixed effects (or within-group) estimator:
    - For a fixed T, $\hat{\beta}_{fe}$ is unbiased and optimal[1], and as $n \to \infty$ it is consistent and asymptotically normal.
    - Estimates of η_i are unbiased but only consistent if $T \to \infty$.
    - If $T \to \infty$ then $\hat{\beta}_{fe}$ is consistent, even if n is fixed.

Differencing and within variation

- Robust standard errors:
  - If A2 does not hold then the usual standard error formula for OLS on the transformed data is inconsistent.
  - If T is fixed and n is large then the clustered (on entity) standard error formula provides a HAC estimator.
  - If T is large and n is fixed then a Newey West type std error estimator is required for consistency under serial correlation.

Differencing and within variation

- Under serial correlation in ν_{it}, the fixed effects estimator is not optimal. Let $\nu_i^* = D\nu_i$.
  - Generally, if $Var(\nu_i^* \mid x_i) = \Omega(x_i)$ then the GLS estimator is

$$\left(\sum_{i=1}^{n} x_i' D' \Omega(x_i) D x_i\right)^{-1} \sum_{i=1}^{n} x_i' D' \Omega(x_i) D y_i$$

  - In the special case where $Var(\nu_i^* \mid x_i) = \Omega$, replace $\Omega(x_i)$ with

$$\hat{\Omega} = n^{-1} \sum_{i=1}^{n} \hat{\nu}_i^* \hat{\nu}_i^{*\prime}$$

  to get a feasible GLS estimator.

Random effects

- Pooled OLS estimator is

$$\hat{\beta}_{pooled} = \left(\sum_{i=1}^{n} x_i' x_i\right) \sum_{i=1}^{n} x_i' y_i$$

- It's unbiased and consistent only under the assumption that $E(\eta_i x_{it}) = 0$.
- Under assumption A2 and $Var(\eta_i \mid x_i) = \sigma_\eta^2$,

$$Var(\eta_i \iota + \nu_i \mid x_i) = \sigma_\eta^2 \iota \iota' + \sigma^2 I_T$$

Random effects

- The GLS estimator is then

$$\hat{\beta}_{GLS} = \left(\sum_{i=1}^{n} x_i V^{-1} x_i' \right) \sum_{i=1}^{n} x_i V^{-1} y_i$$

where $V^{-1} = \sigma^{-2} \left(I_T - \sigma_\eta^2 \iota\iota' / (\sigma^2 + T\sigma_\eta^2) \right)$.
  - This is the *random effects* estimator.
  - When $T \to \infty$, this becomes the fixed effects estimator.
  - More generally, if $\psi = \sigma_\eta^2/(\sigma^2 + T\sigma_\eta^2)$ goes to 0 we get fixed effects and if ψ goes to 1 we get pooled OLS.

Random effects

- Feasible GLS
  - Estimate ψ in first stage to get estimate of $\hat{V}$.
  - Several ways to estimate ψ.
  - This is what `xtreg ..., re` in Stata does.
- An alternative is the maximum likelihood estimator that will estimate β and σ and σ_η^2 simultaneously.
  - the usual MLE assumes that $\eta_i \sim N(0, \sigma_\eta^2)$ though different distributions can be used.

Random effects vs fixed effects

- *The primary difference between the two is that random effects assumes η_i is uncorrelated with x_{it}.*
- The idea of fixed (non-random) versus random effects is not the real distinction.
- Mundlak (1978) showed that the fixed effects estimator is equivalent to a random effects type (GLS) estimator of the model where $\eta_i = a'\bar{x}_i + \omega_i$ where ω_i is independent of x_i.
  - Not true in nonlinear models!

Measurement error

- Motivating example – Bover and Watson (2000)
  - consider a simplified version of the model from Arellano (2003)
  - Conditional money demand equation:
    - y_{it} denotes cash holdings (real money balances) of firm i in year t
    - x_{it} denotes sales
    - $\eta_i = -log(a_i)$ where a_i denotes a firm's "financial sophistication"

Measurement error

- Suppose $\tilde{x}_{it} = x_{it} + \varepsilon_{it}$ and the true regressor values, x_{it} are unobserved.
- Fixed effects can exacerbate measurement error bias:

Measurement error

- Suppose $\tilde{x}_{it} = x_{it} + \varepsilon_{it}$ and the true regressor values, x_{it} are unobserved.
- Fixed effects can exacerbate measurement error bias:
  - The measurement error bias in the FE estimator when $T = 2$ is $\beta\left(1 - \frac{1}{1+\lambda}\right)$ where

$$\lambda = Var(\Delta\varepsilon_{it})/Var(\Delta x_{it})$$

    - If ε_{it} and x_{it} are both iid then this attentuation bias is identical to the cross-sectional bias.
    - If ε_{it} is iid but x_{it} is positively serially correlated then the bias is *larger* than in the cross-section.

Measurement error

- When $T > 2$, ε_{it} is iid and x_{it} is positively serially correlated – Griliches and Hausman (1986) show that the bias of the fixed effects estimator lies between the bias of pooled OLS and that of OLS in first-differences.

Measurement error

- When $T > 2$, ε_{it} is iid and x_{it} is positively serially correlated – Griliches and Hausman (1986) show that the bias of the fixed effects estimator lies between the bias of pooled OLS and that of OLS in first-differences.
- Panel IV can be a solution to the measurement error problem when ε_{it} is not serially correlated and x_{it} is.
  - If η_i is independent (random effects/pooled OLS model) then

$$E(\tilde{x}_{is}(y_{it} - \beta' \tilde{x}_{it})) = 0$$

    for $s \neq t$

Measurement error

- If η_i is correlated with x_{it}, one solution is to take first differences and use the moment conditions

$$E(\tilde{x}_{is}(\Delta y_{it} - \beta' \Delta \tilde{x}_{it})) = 0$$

  for $s = 1, \ldots, t-2, t+1, \ldots, T$
  - This requires $T \geq 3$.
  - Also, the rank condition should be considered carefully. What if x_{it} is white noise? What is x_{it} is a random walk? What if $x_{it} = \alpha_i + \xi_{it}$?
- With larger T, there is a tradeoff between allowing serial correlation in ε_{it} and needing serial correlation in x_{it}.

Measurement error

- Table from Bover and Watson (2000):

Table 4.1
Firm Money Demand Estimates
Sample period 1986–1996

	OLS Levels	OLS Orthogonal deviations	OLS 1st-diff.	GMM 1st-diff.	GMM 1st-diff. m. error	GMM Levels m. error
Log sales	.72	.56	.45	.49	.99	.75
	(30.)	(16.)	(12.)	(16.)	(7.5)	(35.)
Log sales ×trend	−.02	−.03	−.03	−.03	−.03	−.03
	(3.2)	(9.7)	(4.9)	(5.3)	(5.0)	(4.0)
Log sales ×trend^2	.001	.002	.001	.001	.001	.001
	(1.2)	(6.6)	(1.9)	(2.0)	(2.3)	(1.4)
Sargan (p-value)				.12	.39	.00

All estimates include year dummies, and those in levels also include industry dummies. t-ratios in brackets robust to heteroskedasticity & serial correlation. N=5649. Source: Bover and Watson (2000).

Measurement error

- The relationship among the pooled OLS, FE, and first difference estimators is consistent with measurement error in sales.
- Column (4) is GMM on first differences using other time periods as instruments.
  - The Sargan test here is also marginally suggestive of measurement error.
- Columns (5) and (6) seem to correct for measurement error and are consistent with the expectation that pooled OLS should be downward biased.

AR model with fixed effects

- Consider as a simple example the autoregressive model: $y_{it} = \alpha y_{i(t-1)} + \eta_i + \nu_{it}$
  - B1 $E(\nu_{it} \mid y_i^{t-1}, \eta_i) = 0$
  - B2 $E(\nu_{it}^2 \mid y_i^{t-1}, \eta_i) = \sigma^2$
  - B3 (mean stationarity) $E(y_{i0} \mid \eta_i) = \eta_i/(1-\alpha)$
  - B4 (covariance stationarity) $Var(y_{i0} \mid \eta_i) = \sigma^2/(1-\alpha^2)$
- The fixed effects estimator has a bias that is
  - equal to $-(1+\alpha)/2$ when $T = 2$
  - approximately $-(1+\alpha)/T$ for large T
- This is called the Nickell bias due to pioneering work of Nickell (1981).

AR model with fixed effects

- Without assumptions B3 and B4 the bias is more complicated.
  - E.g., if $T = 2$ and $\sigma_\eta^2 / Var(\nu_{i1})$ is large then the bias is very small.
- What if T is large but the same order of magnitude as n?
  - Formally, if $n/T \to c > 0$ then

$$\sqrt{nT}(\hat{\alpha}_{fe} - \alpha) \approx N(-c(1+\alpha), (1-\alpha^2)/(nT))$$

  - For moderate values of T, a bias-corrected estimator:

$$\hat{\alpha}_{fe,bc} = \hat{\alpha}_{fe} + \frac{1 + \hat{\alpha}_{fe}}{T}$$

IV solution

- Anderson and Hsiao (1981, 1982) suggested using an IV estimator that uses $y_{i(t-2)}$ or $\Delta y_{i(t-2)}$ as an instrument for Δy_{it} when $T \geq 3$ or $T \geq 4$.

IV solution

- Anderson and Hsiao (1981, 1982) suggested using an IV estimator that uses $y_{i(t-2)}$ or $\Delta y_{i(t-2)}$ as an instrument for Δy_{it} when $T \geq 3$ or $T \geq 4$.
- There are potentially many more moment conditions under assumption B1:

$$E(y_i^{t-1}(\Delta y_{it} - \alpha \Delta y_{i(t-1)})) = 0, \quad t = 2, \ldots, T$$

IV solution

- Holtz-Eakin, Newey, and Rosen (1988) and Arellano and Bond (1991) suggest implementing a GMM estimator that uses all $(T-1)T/2$ moment conditions.
- The Arellano Bond estimator uses a one-step optimal weighting matrix that accounts for serial correlation due to differencing,

$$\hat{V} = \sum_{i=1}^{n} z_i' D D' z_i$$

- There is a bias however when $n \approx T$ that is proportional to $1/n$.

IV solution

- Advice:
  - When T is larger than n, use FE.
  - When n is larger than T, use Arellano-Bond.
  - When n is similar in magnitude to T, use bias-correction or limited number of instruments/moments.

A factor model

- Suppose that

$$Y_{it} = \lambda_t' \alpha_i + \varepsilon_{it}$$

  - The α_i is a vector of common factors.
  - The ε_{it} are idiosyncratic factors.
  - The λ_t are factor loadings.

A factor model

- Identification based on:

$$Var(Y_i) = \Lambda Var(\alpha_i)\Lambda' + \Delta$$

  under restrictions on Δ
  - if T is small, Δ diagonal is typical restriction
  - if T is large, we can do better

A factor model

- Identification based on:

$$Var(Y_i) = \Lambda Var(\alpha_i)\Lambda' + \Delta$$

  under restrictions on Δ
  - if T is small, Δ diagonal is typical restriction
  - if T is large, we can do better
- Normalizations needed:
  - For example, $E(\alpha_i) = 0$ and $Var(\alpha_i) = I$ and Λ is lower triangular.
- See Anderson and Rubin (1954) and Williams (forthcoming, Ect. Rev.).

The “interactive fixed effects” model

- An extension of the twoway FE model:

$$Y_{it} = \beta' X_{it} + \lambda_t' \alpha_i + \varepsilon_{it}$$

- Often a time FE is explicitly included,

$$Y_{it} = \beta' X_{it} + \lambda_{0t} + \lambda_t' \alpha_i + \varepsilon_{it}$$

- This is more general, more flexible than the “entity-specific trend” modelling approach.

The “interactive fixed effects” model

- We will talk about several ways to estimate this model.
  - Bai (2009)
  - Ahn, Lee, and Schmidt (2013)
  - A new approach that Bob Phillips and I have been working on.
  - The synthetic control method.

Application

- Divorce rates and divorce law reforms.
  - Friedberg (1998) – reforms lead to increased divorce rate, using FE/DD with state-specific quadratic trends
  - Wolfers (2006) cast doubt on these results, arguing in part that the state-specific quadratic trend method is not very robust
  - Kim and Oka (2014) applied Bai (2009)'s IFE estimator and found that results are more robust.

Bai (2009)'s "interactive fixed effects" estimator

- If n and T are both large then we can treat λ_t and α_i as parameters to be estimated.

Bai (2009)'s "interactive fixed effects" estimator

- If n and T are both large then we can treat λ_t and α_i as parameters to be estimated.
- The problem is to minimize

$$\sum_{i=1}^{n}\sum_{t=1}^{T}\left(Y_{it} - \beta' X_{it} - \lambda_t' \alpha_i\right)^2$$

Bai (2009)'s "interactive fixed effects" estimator

- Bai (2009) suggests doing this by iterating the following two steps.
  1. Given $\{\lambda_t^{(s)}\}$ and $\{\alpha_i^{(s)}\}$, choose $\beta = \beta^{(s+1)}$ to minimize

$$\sum_{i=1}^{n}\sum_{t=1}^{T}\left(Y_{it} - \beta' X_{it} - \lambda_t^{(s)\prime}\alpha_i^{(s)}\right)^2$$

  2. Given $\beta = \beta^{(s+1)}$, choose $\lambda_t = \lambda_t^{(s+1)}$ and $\alpha = \alpha_i^{(s+1)}$ to minimize

$$\sum_{i=1}^{n}\sum_{t=1}^{T}\left(Y_{it} - \beta^{(s+1)\prime} X_{it} - \lambda_t'\alpha_i\right)^2$$

Ahn, Lee, and Schmidt (2013)

- ALS (2013) propose a GMM estimation strategy based on quasi-differencing.
- This is easiest to see when α_i is scalar. In that case,

$$Y_{it} - \frac{\lambda_t}{\lambda_s} Y_{is} = \beta' \left(X_{it} - \frac{\lambda_t}{\lambda_s} X_{is} \right) + \tilde{u}_{it}$$

- Under various exogeneity conditions we get moments such as

$$E\left(Z_{i\tau} \left(Y_{it} - \frac{\lambda_t}{\lambda_s} Y_{is} - \beta' \left(X_{it} - \frac{\lambda_t}{\lambda_s} X_{is} \right) \right) \right) = 0$$

where $Z_{i\tau}$ can be $Y_{i\tau}$ or $X_{i\tau}$.

Ahn, Lee, and Schmidt (2013)

- The propose a two step optimal GMM estimator based on all valid moment conditions.
- Rank condition is not super transparent – need to use the moments to identify β *and* λ_t.
- But this can work with fairly small T.
- One caveat: moment conditions proliferate as T increases, as in Arellano-Bond.

Phillips and Williams

- Define the linear projection,

$$\alpha_i = \psi' X_i + \xi_i,$$

where ξ_i is uncorrelated with X_i

- Plugging this in we get

$$Y_{it} = \beta' X_{it} + \lambda_t' \psi' X_i + \lambda_t' \xi_i + \varepsilon_{it}$$

- We propose a least squares estimator that minimizes

$$\sum_{i=1}^{n} \sum_{t=1}^{T} \left(Y_{it} - \beta' X_{it} - \lambda_t' \psi' X_i \right)^2$$

- This is similar to Bai (2009) except that in "step 2" we use a method to estimate λ_t and ψ that works with small T.

Synthetic control analysis

- Similar to matching-based estimators.
- The idea is to compare the treated state to a weighted average of control states.
- The weights are chosen to match covariates and past outcomes.
- Abadie et al. (2010) argue that this works under a general interactive fixed effects and time-varying coefficient specification

Synthetic control analysis

- The method in principle:
  - Suppose states $s = 1, \ldots, S$ are controls and state $S+1$ is treated.
  - First, find nonnegative weights $w_1, \ldots, w_S$ that add up to 1 so that
$$\sum_{s=1}^{S} w_s X_s = X_{S+1}$$
and
$$\sum_{s=1}^{S} w_s Y_{st} = Y_{S+1,t}$$
for each period t before treatment occurs at T_0.
  - Then, for $t > T_0$, estimate the TT using these weights
$$Y_{S+1,t} - \sum_{s=1}^{S} w_s Y_{st}$$

Synthetic control analysis

- Suppose
$$Y_{0st} = \lambda_{0t} + \lambda'_{1t}\gamma_s + \beta'_t X_s + \varepsilon_{st}$$
- For large T_0, the above method would ensure that γ_s and X_s are equal between $S+1$ and the "synthetic control"
- So $Y_{0,S+1,T_0+1}$, $Y_{0,S+1,T_0+2}$, ... are unbiased estimates of the counterfactuals.

Synthetic control analysis

- The method in practice:
  - First, find nonnegative weights $w_1, \ldots, w_S$ that add up to 1 so that
  $$||X_1 - X_0 W||$$
  is minimized.
  - Then, for $t > T_0$, estimate the TT using these weights
  $$Y_{S+1,t} - \sum_{s=1}^{S} w_s Y_{st}$$

Synthetic control analysis

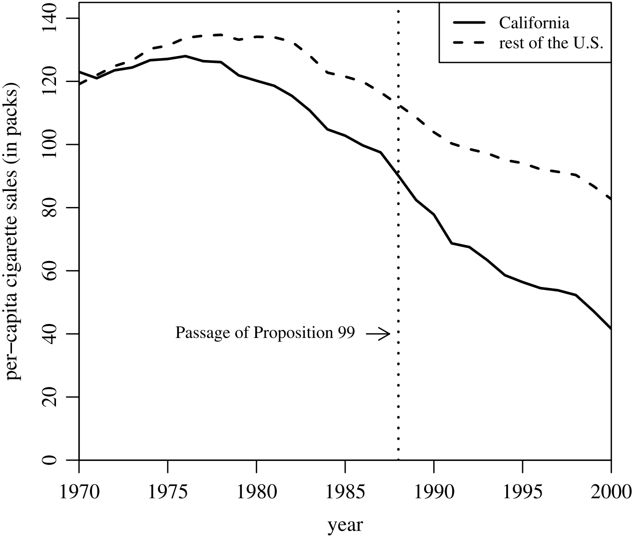
- Inference is not settled but Abadie et al. (2010) propose formalizing a placebo test as a permutation test.
- Requires large T_0.
- Ferman and Pinto (2016) show that the method is typically still biased, though it generally outperforms DiD.
- Requires the other states to be roughly comparable – convex hull assumption.
  - If we allow more general weights, this is not necessary, but then results rely on extrapolation.
- Stata command: `synth`

Synthetic control analysis

- Abadie et al. (2010)
  - Proposition 99 in California in 1988 to control tobacco consumption (increased tax and other measures).
  - Did this decrease tobacco consumption?
  - First state to do this and most states did not implement similar measures until 2000.

Synthetic control analysis

- Abadie et al. (2010)



Synthetic control analysis

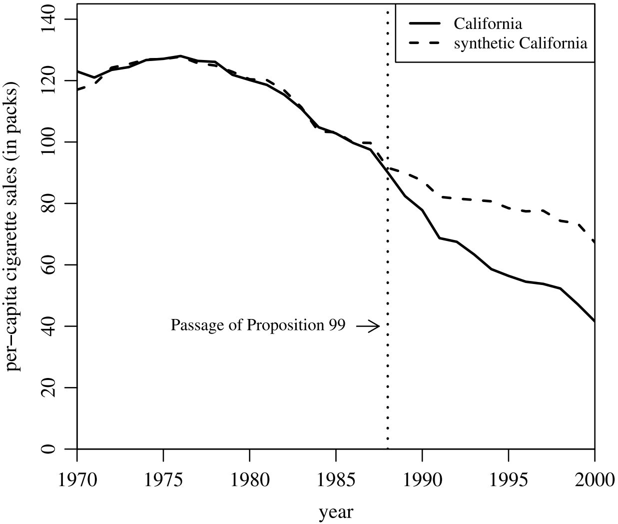
- Abadie et al. (2010)

Table 2. State weights in the synthetic California

State	Weight	State	Weight
Alabama	0	Montana	0.199
Alaska	–	Nebraska	0
Arizona	–	Nevada	0.234
Arkansas	0	New Hampshire	0
Colorado	0.164	New Jersey	–
Connecticut	0.069	New Mexico	0
Delaware	0	New York	–
District of Columbia	–	North Carolina	0
Florida	–	North Dakota	0
Georgia	0	Ohio	0
Hawaii	–	Oklahoma	0
Idaho	0	Oregon	–
Illinois	0	Pennsylvania	0
Indiana	0	Rhode Island	0
Iowa	0	South Carolina	0
Kansas	0	South Dakota	0
Kentucky	0	Tennessee	0
Louisiana	0	Texas	0
Maine	0	Utah	0.334
Maryland	–	Vermont	0
Massachusetts	–	Virginia	0
Michigan	–	Washington	–
Minnesota	0	West Virginia	0
Mississippi	0	Wisconsin	0
Missouri	0	Wyoming	0

Synthetic control analysis

- Abadie et al. (2010)



The inferential problem

- In Card and Krueger (1994), they have 410 fast food restaurants. So is their analysis based on 820 observations, 410, or just 4?
- If just 4, standard errors cannot even be calculated.
- The solution is to add more "states" and/or more observations over time
  - But how are standard errors calculated?

The inferential problem

- Let $u_{ist} = \nu_{st} + \eta_{ist}$
- One problem is correlation within state-year:

$$Cov(u_{ist}, u_{jst}) = Var(\eta_{st})$$

- An additional problem is that ν_{st} may be correlated with $\nu_{st'}$.

The inferential problem

- Let $u_{ist} = \nu_{st} + \eta_{ist}$
- One problem is correlation within state-year:

$$Cov(u_{ist}, u_{jst}) = Var(\eta_{st})$$

- An additional problem is that ν_{st} may be correlated with $\nu_{st'}$.
    - Then $Cov(u_{ist}, u_{jst'}) = Cov(\eta_{st}, \eta_{st'}) \neq 0$.
    - Serial correlation.

The Moulton factor

- Moulton (1986) and subsequent literature showed that conventional standard errors will be biased unless there is no *intraclass* correlation in the regressors.
  - When the policy is at the "state" level, this is a big problem.
  - When the policy is at a "lower" level it is not necessarily a problem.
- The (old) solution is (was) *clustered standard errors* at the state-year level.
- Bertrand et al. (2004) provide simulation evidence that clustering only on state-year can lead to massive overrejection.
  - due to serial correlation in ν_{st}

Some newer solutions

- When the number of states is sufficiently large (50 seems to be enough), cluster on the state.
- What if the number of states/groups is more like 10 or 20?
  - use critical values from Student's t with $S - K$ degrees of freedom (Bell McCaffrey; Imbens Kolesar)
  - The block bootstrap (Cameron, Gelbach and Miller, 2008)
- What if the number of states/groups is very small?
  - estimate time series model (e.g., AR(1)) for errors
  - get more data!