Fused Extended Two-Way Fixed Effects for Difference-in-Differences With Staggered Adoptions

Greg Faletto

Data Scientist, VideoAmp

USC Causal Inference Reading Group

February 8th, 2024

Access the paper

USC
School of Business

University of Southern California

Outline

1 **Difference-in-Differences Background**

2 Extended Two-Way Fixed Effects

3 FETWFE

4 Theory

5 Simulation Studies

Access the paper

Difference-in-Differences

Suppose we observe units at two times.

First time period: no units receive treatment.

Second time period: some do (not randomly assigned).

Let $y_{it}(0)$ be the potential outcome at time t for units who are never treated, and $y_{it}(2)$ be the potential outcome at time t for units who are treated at time $t = 2$.

Notice: we define potential outcomes in terms of (time-invariant) treatment group assignment, *not* treatment status.

Goal: estimate the effect of treatment on the treated units,

$$\tau := \mathbb{E}[y_{i2}(2) - y_{i2}(0) \mid W_i = 2],$$

where $W_i = 2$ denotes that unit i began treatment at time 2.

Difference-in-Differences Estimator

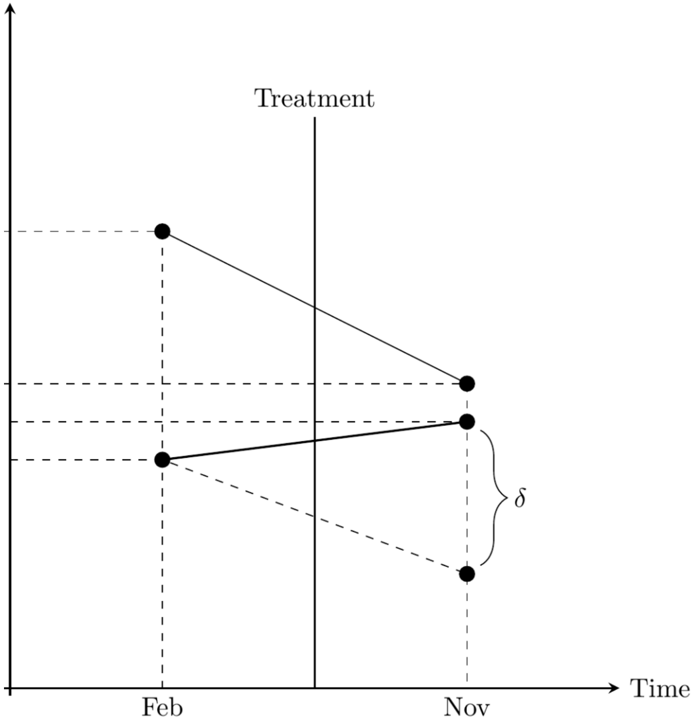
Observed responses of never-treated units give us information about changes in external conditions from $t = 1$ to 2.

Parallel trends assumption: change in untreated potential outcomes from $t = 1$ to 2 would have been same for both groups:

$$\mathbb{E}[y_{i2}(0) - y_{i1}(0) \mid W_i = 2] = \mathbb{E}[y_{i2}(0) - y_{i1}(0) \mid W_i = 0]\,.$$

Parallel trends **allows for selection bias** as long as bias is **time-invariant**:

Image source: Cunningham, Scott. *Causal inference: The mixtape*. Yale university press, 2021. Figure 9.2.



$$\boxed{\mathbb{E}[y_{i2}(0) \mid W_i = 2] - \mathbb{E}[y_{i2}(0) \mid W_i = 0]} = \boxed{\mathbb{E}[y_{i1}(0) \mid W_i = 2] - \mathbb{E}[y_{i1}(0) \mid W_i = 0]\,.}$$

Selection bias at $t = 2$

Selection bias at $t = 1$

Difference-in-Differences Estimator

$$\mathbb{E}[y_{i2}(0) - y_{i1}(0) \mid W_i = 2] = \mathbb{E}[y_{i2}(0) - y_{i1}(0) \mid W_i = 0]$$

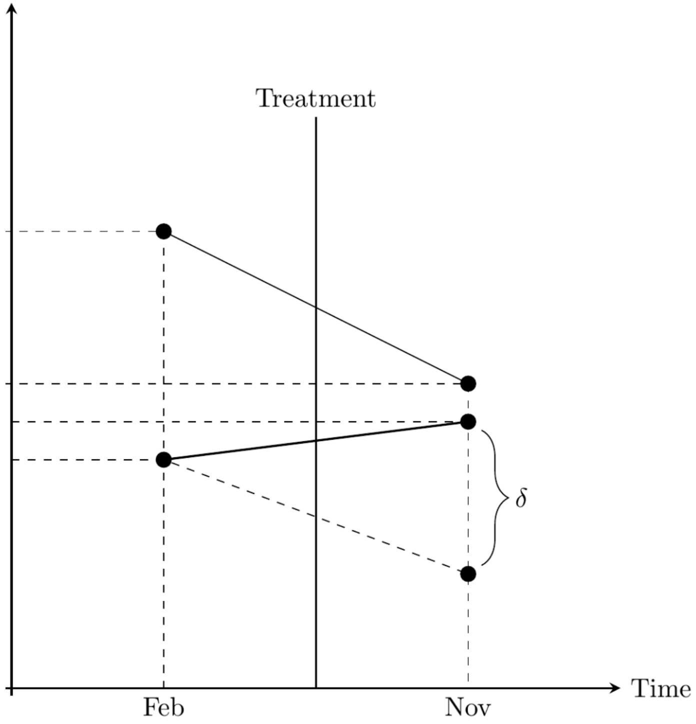
Under parallel trends, we can estimate the treatment effect by adjusting the sample mean difference in outcomes at $t = 2$ using the pre-treatment difference:

$$\hat{\tau} = \underset{\tau+\text{selection bias}}{\boxed{\bar{y}_2(2) - \bar{y}_2(0)}} - \underset{\text{selection bias}}{\boxed{\left[\bar{y}_1(2) - \bar{y}_1(0)\right]}}$$

This is the canonical **difference-in-differences** estimator.

Notice: the crucial **parallel trends assumption is untestable**—depends on unobserved potential outcomes.

Image source: Cunningham, Scott. *Causal inference: The mixtape*. Yale university press, 2021. Figure 9.2.



Two-Way Fixed Effects

Equivalent way to calculate the difference-in-differences estimator: estimate $\hat{\tau}$ in the linear regression

$$y_{it} = \alpha_i + \gamma_2 \cdot 1\{t = 2\} + \tau \cdot 1\{t = 2\}1\{W_i = 2\} + \epsilon_{it},$$

where:

α_i is a separate intercept for each unit

γ_2 is a separate intercept for $t = 2$.

This is called a **two-way fixed effects** regression.

(Notice: we got here **not by assuming linearity**, but instead by making some reasonable assumptions about the potential outcomes—a linear model **just happens to have the right estimand**.)

Two-Way Fixed Effects

$$y_{it} = \alpha_i + \gamma_2 \cdot 1\{t = 2\} + \tau \cdot 1\{t = 2\}1\{W_i = 2\} + \epsilon_{it}$$

Now that we're calculating a linear regression…

Can we allow an **arbitrary number of times** instead of just 2?

Allow units to **start treatment at arbitrary times** after time 1?

(Example application: effect of a law that is passed state-by-state over time, like public smoking bans or unilateral ["no-fault"] divorce.)

In general: **not with this model!** With arbitrary T and staggered adoptions, this model will be **biased if treatment effects vary over time and/or between cohorts**.

How do we need to change the model?

Outline

1 Difference-in-Differences Background

2 Extended Two-Way Fixed Effects

3 FETWFE

4 Theory

5 Simulation Studies

Access the paper →

Diff-in-Diff With Staggered Adoptions

Notation: suppose we observe T time periods. We have **cohorts** that begin treatment at times $r \in \mathscr{R} \subseteq \{2, 3, \ldots, T\}$ and continue receiving treatment until T. (*Staggered adoptions*)

Wooldridge (2021): **We can achieve unbiased linear regression** with time-invariant (pre-treatment) covariates! We just need more parameters.

Key ingredients:

Estimate **separate treatment effects** for each cohort and time, τ_{rt}

Add lots of parameters that are linear in X_i (see Assumption LINS in the paper, and the slide after next).

Wooldridge calls this the **extended two-way fixed effects** model.

Building Towards Extended Two-Way Fixed Effects

Wooldridge (2021): in two-way fixed effects, we don't need fixed effects for every unit, we can just estimate the coefficient $\hat{\tau}$ in the linear regression

$$y_{it} = \boxed{\eta} + \boxed{\nu_2} + \boxed{\gamma_2} \cdot 1\{t = 2\} + \boxed{\tau} \cdot 1\{t = 2\} 1\{W_i = 2\} + \epsilon_{it},$$

where:

η is the expected response for the never-treated units at $t = 1$

$\eta + \nu_2$ is the expected response for the treated units at $t = 1$ (ν_2 is the selection bias)

$\eta + \gamma_2$ is the expected response for the never-treated units at $t = 2$ (γ_2 is the expected *trend*)

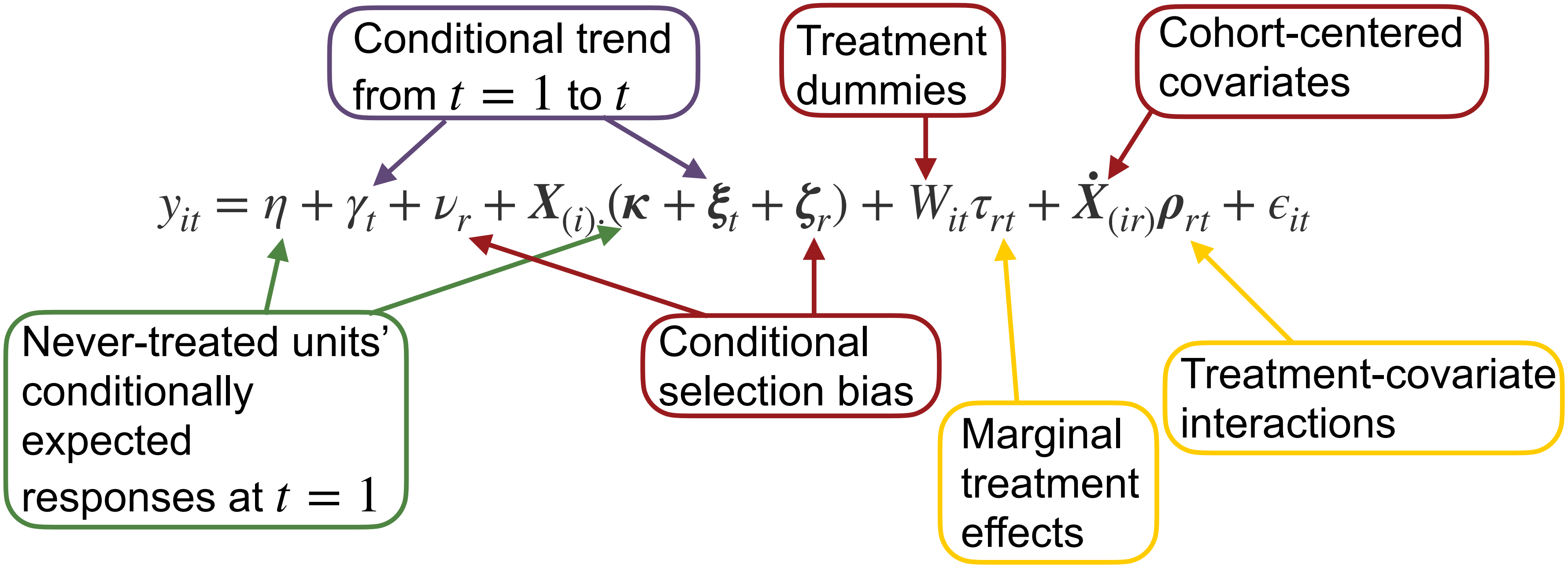
τ corresponds to the average treatment effect on the treated units at $t = 2$ under parallel trends.

Extended Two-Way Fixed Effects

Two-way fixed effects (TWFE):

$$y_{it} = \boxed{\eta} + \boxed{\nu_2} + \boxed{\gamma_2} \cdot 1\{t = 2\} + \boxed{\tau} \cdot 1\{t = 2\}1\{W_i = 2\} + \epsilon_{it},$$

Extended two-way fixed effects (ETWFE; Wooldridge 2021):



$$y_{it} = \eta + \gamma_t + \nu_r + \boldsymbol{X}_{(i)}(\boldsymbol{\kappa} + \boldsymbol{\xi}_t + \boldsymbol{\zeta}_r) + W_{it}\tau_{rt} + \dot{\boldsymbol{X}}_{(ir)}\boldsymbol{\rho}_{rt} + \epsilon_{it}$$

Conditional trend from $t = 1$ to t

Treatment dummies

Cohort-centered covariates

Never-treated units' conditionally expected responses at $t = 1$

Conditional selection bias

Marginal treatment effects

Treatment-covariate interactions

Extended Two-Way Fixed Effects

What exactly have we gained?

Unbiased estimation of treatment effects in much more general setting

Parallel trends assumption is replaced with **conditional parallel trends:** generally more plausible than marginal parallel trends.

What price do we pay?

We have a lot of parameters to estimate! If there are lots of time periods or covariates, our parameter estimates may be too noisy to be useful.

In other words, ETWFE may be too flexible. (Bad bias/variance tradeoff.) Flexibility avoids bias (**very important in causal inference!**), but maybe this model is a bit much.

Extended Two-Way Fixed Effects

Wooldridge proposes an *ad hoc* remedy for the problem of too many parameters: **assume some of the parameters are equal** (“restrictions”).

Example: maybe treatment effects don’t actual differ in time since treatment. Assume $\tau_{rt} = \tau_r$ for all cohorts at all post-treatment times.

Or, weaken this: assume there is an “early treatment” and “late treatment” effect, so that we only have two treatment effects to estimate for each cohort instead of $T - r + 1$.

Problem: **is this wishful thinking?**

Probably some restrictions like this are true. But unless we know the exact correct restrictions, we risk re-introducing the bias we removed by adding these parameters in the first place.

Can we use the fact that some of these restrictions probably exist in the data without putting our “thumb on the scale” by selecting the restrictions by hand?

Outline

1 Difference-in-Differences Background

2 Extended Two-Way Fixed Effects

3 FETWFE

4 Theory

5 Simulation Studies

Access the paper

Fused Extended Two-Way Fixed Effects

Idea: let's reduce the number of parameters to estimate using sparse (bridge) fusion regularization on ETWFE.

Encode our expectations about which parameters might be equal in the way that we estimate the model. Then allow the model to learn the restrictions from the data automatically.

> Example: for each cohort, bridge penalty on the differences between adjacent treatments: $|\tau_{r,t+1} - \tau_{rt}|^q$ for $q \in (0,2]$.

This encodes our belief that some of the treatment effects in adjacent times are probably close together, so **we can set them equal unless the data give us a good reason to estimate separate values.**

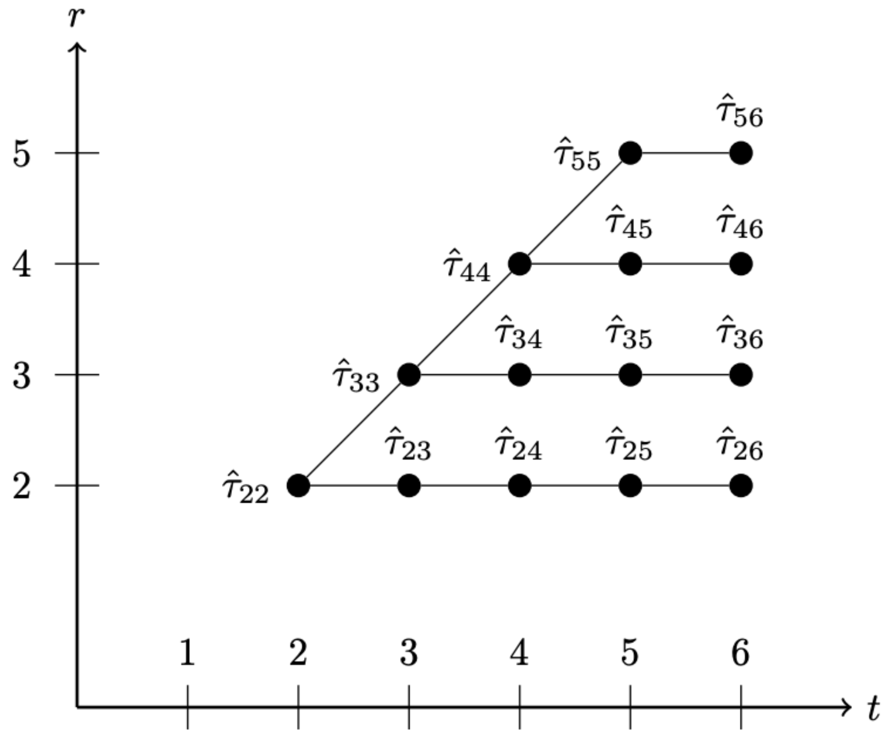


Figure 1: Visualization of which of the estimated marginal average treatment effect terms $\hat{\tau}_{rt}$ from regression (4) (which estimate the average treatment effects $\tau_{ATT}(r, t)$ from Equation 1) we penalize towards each other in the FETWFE penalty (26). In this setting, $T = 6$ and $\mathcal{R} = \{2, \ldots, 5\}$. The horizontal axis depicts time and the vertical axis depicts cohorts. FETWFE works well under an assumption that the linked treatment effects tend to be close together, and at least some of them are exactly equal. See further details in Section 5.

Fused Extended Two-Way Fixed Effects

We collect all of the ETWFE coefficients in a vector $\boldsymbol{\beta}^* \in \mathbb{R}^{p_N}$ and create a matrix $\boldsymbol{D}_N \in \mathbb{R}^{p_N \times p_N}$ that constructs these restrictions: the vector $\boldsymbol{D}_N \boldsymbol{\beta}^*$ contains entries with terms like $\tau_{r,t+1} - \tau_{rt}$.

We also construct a design matrix $\boldsymbol{Z} \in \mathbb{R}^{NT \times p_N}$ containing all of the fixed effects, covariates, etc.

We estimate the ETWFE regression with an added penalty term $\lambda_N ||\boldsymbol{D}_N \boldsymbol{\beta}||_q^q$, with tuning parameter λ_N chosen by BIC (for example).

This is the **fused extended two-way fixed effects model:** for $q > 0$,

$$\hat{\boldsymbol{\beta}}^{(q)} := \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\arg\min} \left\{ ||\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta}||_2^2 + \lambda_N ||\boldsymbol{D}_N \boldsymbol{\beta}||_q^q \right\}.$$

What can we estimate with FETWFE?

Lots of things! The building blocks are:

Average treatment effect for treated units in cohort r at time t:
$\tau_{\text{ATT}}(r, t) := \mathbb{E}[y_{it}(r) - y_{it}(0) \mid W_i = r]$.

Average treatment effect for treated units in cohort r at time t conditional on covariates $\boldsymbol{x}$:
$\tau_{\text{CATT}}(r, t, \boldsymbol{x}) := \mathbb{E}[y_{it}(r) - y_{it}(0) \mid W_i = r, \boldsymbol{X}_i = \boldsymbol{x}]$.

Can construct linear combinations of these: given any set of constants $\{\psi_{rt}\}_{r \in \mathscr{R}, t \geq r}$, we can consider the estimand

$$\sum_{r \in \mathscr{R}} \sum_{t=r}^{T} \psi_{rt} \tau_{\text{ATT}}(r, t)$$

and likewise for $\tau_{\text{CATT}}(r, t, \boldsymbol{x})$.

What can we estimate with FETWFE?

Can construct linear combinations of these: given any set of constants $\{\psi_{rt}\}_{r \in \mathscr{R}, t \geq r}$, we can consider the estimand

$$\sum_{r \in \mathscr{R}} \sum_{t=r}^{T} \psi_{rt} \tau_{\text{ATT}}(r, t)$$

and likewise for $\tau_{\text{CATT}}(r, t, \boldsymbol{x})$.

Given the right choice of $\{\psi_{rt}\}_{r \in \mathscr{R}, t \geq r}$, we can use these to estimate (for example):

Average treatment effects for a cohort (take a simple average across time)

Average treatment effect across all cohorts and all times

Average treatment effect at a fixed time since treatment began

How exactly will we estimate these?

$$y_{it} = \eta + \gamma_t + \nu_r + \boldsymbol{X}_{(i)\cdot}(\boldsymbol{\kappa} + \boldsymbol{\xi}_t + \boldsymbol{\zeta}_r) + W_{it}\boxed{\tau_{rt}} + \dot{\boldsymbol{X}}_{(ir)}\boxed{\boldsymbol{\rho}_{rt}} + \epsilon_{it}$$

With the estimated FETWFE regression coefficients:

Average treatment effect for treated units in cohort r at time t:
$\hat{\tau}_{\text{ATT}}(r, t) := \hat{\tau}_{rt}.$

Average treatment effect for treated units in cohort r at time t conditional on covariates $\boldsymbol{x}$:
$\hat{\tau}_{\text{CATT}}(r, t, \boldsymbol{x}) := \hat{\tau}_{rt} + \hat{\boldsymbol{\rho}}_{rt}^\top \left(\boldsymbol{x} - \overline{\boldsymbol{X}}_r\right)$, where
$\overline{\boldsymbol{X}}_r := \dfrac{1}{N_r} \sum_{i=1}^{N} 1\{W_i = r\} \cdot \boldsymbol{X}_i$ is the sample mean for units in cohort r.

Outline

1 Difference-in-Differences Background
2 Extended Two-Way Fixed Effects
3 FETWFE
4 Theory
5 Simulation Studies

Access the paper

Behind the Curtain: Intuition Behind the Theory

We've constructed an asymptotically unbiased regression (under parallel trends, linearity assumption, etc.).

We've also added a penalty term that enforces plausible restrictions:

$$\hat{\boldsymbol{\beta}}^{(q)} := \underset{\boldsymbol{\beta} \in \mathbb{R}^{p_N}}{\arg\min} \left\{ \|\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta}\|_2^2 + \lambda_N \|\boldsymbol{D}_N \boldsymbol{\beta}\|_q^q \right\} .$$

Consider the change in coordinates $\boldsymbol{\theta} := \boldsymbol{D}_N \boldsymbol{\beta}$. If the differences matrix $\boldsymbol{D}_N$ is invertible, we have $\boldsymbol{\beta} = \boldsymbol{D}_N^{-1} \boldsymbol{\theta}$, and we can solve the equivalent problem

$$\hat{\boldsymbol{\theta}}^{(q)} := \underset{\boldsymbol{\theta} \in \mathbb{R}^{p_N}}{\arg\min} \left\{ \|\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{D_N}^{-1}\boldsymbol{\theta}\|_2^2 + \lambda_N \|\boldsymbol{\theta}\|_q^q \right\}$$

using bridge regression on the modified matrix $\boldsymbol{Z}\boldsymbol{D}_N^{-1}$, then get $\hat{\boldsymbol{\beta}}^{(q)} = \boldsymbol{D}_N^{-1} \hat{\boldsymbol{\theta}}^{(q)}$.

If a subset of the restrictions hold exactly (sparsity of $\boldsymbol{\theta}$), with some "plumbing" to connect the coordinates **we can use off-the-shelf bridge regression theory on this modified optimization problem to get theory for FETWFE.**

Why bridge regression?

There's nothing special about bridge regression except that it allows us to get asymptotic normality (in a single step!) for $q \in (0,1)$. Important for economists—**want to construct confidence intervals**, not just get good point estimates.

I considered other options that allow for inference and/or asymptotic normality. Some of these could be reasonable too, and you could develop similar theory.

Adaptive lasso: requires two steps (isn't this complicated enough already?)

De-biased lasso: point estimates are no longer sparse, which is a nice property for interpretability.

Post-selection inference for the lasso: stepwise procedure (isn't this complicated enough already?)

Why bridge regression?

The price we pay with bridge regression for $q \in (0,1)$ is that the optimization is non-convex, so FETWFE might not scale well with large data.

Also, my theory requires $p_N \leq NT$, which could be restrictive.

> (You could prove consistency of FETWFE with existing lasso theory for $p_N \gg NT$ and $q = 1$. But not asymptotic normality unless you switched to one of the strategies from the last slide.)

Theoretical Guarantees: Main Assumptions

Conditional parallel trends (time-invariant conditional selection bias): $\mathbb{E}[y_{it}(0) - y_{i1}(0) \mid W_i, \boldsymbol{X}_i] = \mathbb{E}[y_{it}(0) - y_{i1}(0) \mid \boldsymbol{X}_i]$ a.s. for all $t \geq 2$.

(Trend is mean-independent of treatment conditional on $\boldsymbol{X}_i$.)

Conditional no anticipation: $\mathbb{E}[y_{it}(r) - y_{it}(0) \mid W_i = r, \boldsymbol{X}_i] = 0$ almost surely ("a.s.") for all $r \in \{2, \ldots, T\}$, $t < r$.

Linearity: the conditionally expected trends, selection biases, untreated potential outcomes, and treatment effects are all linear in $\boldsymbol{X}_i$. (By the way, $\boldsymbol{X}_i \in \mathbb{R}^{d_N}$, and d_N may tend to infinity with N, but not too quickly.)

Sparsity: The vector $\boldsymbol{\theta}^* = \boldsymbol{D}\boldsymbol{\beta}^*$ is sparse, with $s_N < p_N$ nonzero entries. (s_N can increase with N subject to regularity conditions for first two results.)

Full-column rank design matrix: minimum eigenvalue of $\frac{1}{NT}\boldsymbol{Z}^\top \boldsymbol{Z}$ is positive—**requires** $p_N \leq NT$.

Theorem 6.1: Consistency

Theorem 6.1: Under the previous assumptions and some mild regularity conditions on the distributions of $\boldsymbol{X}_i$, $\boldsymbol{y}_i$, and W_i, for any $q \in (0,2]$ and, it holds that

$$\left| \hat{\tau}_{\text{ATT}}(r,t) - \tau_{\text{ATT}}(r,t) \right|$$

converges in probability to 0 at a rate at least as fast as $\sqrt{p_N/N}$.

If d_N (and therefore p_N) is fixed, so does

$$\left| \hat{\tau}_{\text{CATT}}(r,t,\boldsymbol{x}) - \tau_{\text{CATT}}(r,t,\boldsymbol{x}) \right|$$

(treating $\boldsymbol{x}$ as fixed).

Theorem 6.1: Consistency

Since T is fixed, given any set of constants $\{\psi_{rt}\}_{r \in \mathscr{R}, t \geq r}$ it follows that the same holds for the estimator $\sum_{r \in \mathscr{R}} \sum_{t=r}^{T} \psi_{rt} \, \hat{\tau}_{\text{ATT}}(r, t)$ and its corresponding estimand.

$q = 1$ is the lasso and $q = 2$ is ridge regression. So FETWFE is consistent for these convex optimization problems.

Theorem 6.2: Restriction Selection Consistency

Theorem 6.2: In addition to the previous assumptions, suppose that the largest eigenvalue of $\frac{1}{NT}Z^\top Z$ is almost surely less than a finite constant, and choose any $q \in (0,1)$. Then as $N \to \infty$, FETWFE identifies the correct restrictions with probability tending to 1.

Takeaway: **FETWFE successfully selects the restrictions for us**—we don't have to choose them by hand.

Can't use the lasso or ridge regression anymore, though (for this and the next result)—have to use nonconvex bridge regression.

> (In exchange, we don't require something like the "irrepresentable condition" or "neighborhood stability condition" needed for lasso selection consistency.)

Theorem 6.3: Oracle Efficiency

Theorem 6.3: In addition to assumptions of Theorem 6.2, assume that the sparsity $s_N = s$ is fixed as $N \to \infty$ and some additional regularity conditions. Then if $\tau_{\text{ATT}}(r, t) \neq 0$, the sequence of random variables

$$\sqrt{NT}(\hat{\tau}_{\text{ATT}}(r, t) - \tau_{\text{ATT}}(r, t))$$

converges in distribution to a mean 0 Gaussian random variable with asymptotic variance that depends only on the model with all of the correct restrictions identified.

(If $\tau_{\text{ATT}}(r, t) = 0$, sequence converges in probability to 0.)

Again, we can make the same statement about the estimator

$$\sum_{r \in \mathscr{R}} \sum_{t=r}^{T} \psi_{rt} \, \hat{\tau}_{\text{ATT}}(r, t).$$

Theorem 6.3: Oracle Efficiency

FETWFE is an **oracle procedure**:

Even if the number of covariates grows asymptotically, under the assumptions of Theorem 6.3 FETWFE converges at the same $1/\sqrt{N}$ rate as an ordinary least squares model estimated using all of the correct restrictions.

The asymptotic variance of FETWFE is not affected by the parameters we didn’t need to separately estimate.

Theorem 6.4: Asymptotic Confidence Intervals
(With Feasible Variance Estimator)

Maintain the assumptions from Theorem 6.3. Then for any $\alpha \in (0,1)$,

$$\lim_{N\to\infty} \mathbb{P}\left(\sum_{r\in\mathscr{R}}\sum_{t=r}^{T}\psi_{rt}\tau_{\mathsf{ATT}}(r,t) \in \mathsf{CI}_N(\alpha)\right) = 1-\alpha,$$

where

$$\mathsf{CI}_N(\alpha) := \left[\sum_{r\in\mathscr{R}}\sum_{t=r}^{T}\psi_{rt}\hat{\tau}_{\mathsf{ATT}}(r,t) \pm \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\sqrt{\frac{\hat{v}_N^{(\mathsf{C})}}{NT}}\right],$$

$\Phi(\,\cdot\,)$ is the cdf of the standard Gaussian distribution, and $\hat{v}_N^{(\mathsf{C})}$ is a a finite sample variance estimator (explicitly defined in the paper).

Theorem 6.4: Asymptotic Normality With Feasible Variance Estimator

Not in paper yet, but we also get oracle efficiency and asymptotical normality for conditional average treatment effects $\hat{\tau}_{\text{CATT}}(r, t, \boldsymbol{x})$ under fixed d_N if we **split the data**:

Half is used for estimating model,

Half used for estimating cohort sample means of covariates

$$\overline{\boldsymbol{X}}_r = \frac{1}{N_r} \sum_{i=1}^{N} 1\{W_i = r\} \cdot \boldsymbol{X}_i$$

(I think this can probably also be done using **cross-fitting,** which would use the data more efficiently.)

A Peek at Some Other Results

Probability-weighted average treatment effects

Consider the estimands

$$\sum_{r\in\mathscr{R}} \frac{\mathbb{P}(W_i = r)}{\mathbb{P}(W_i \neq 0)} \sum_{t=r}^{T} \psi_{rt}\tau_{\text{ATT}}(r,t).$$

These are probability-weighted treatment effects.

Example: we can define the overall average treatment effect for units at the time they begin treatment as

$$\sum_{r\in\mathscr{R}} \frac{\mathbb{P}(W_i = r)}{\mathbb{P}(W_i \neq 0)} \tau_{\text{ATT}}(r,r).$$

A Peek at Some Other Results

Probability-weighted average treatment effects

Consider the estimands

$$\sum_{r\in\mathscr{R}} \frac{\mathbb{P}(W_i = r)}{\mathbb{P}(W_i \neq 0)} \sum_{t=r}^{T} \psi_{rt}\tau_{\mathsf{ATT}}(r,t).$$

These are probability-weighted treatment effects.

We can estimate these ratios of probabilities using the observed counts of units in each cohort:

$$\sum_{r\in\mathscr{R}} \frac{N_r}{N_\tau} \sum_{t=r}^{T} \psi_{rt}\hat{\tau}_{\mathsf{ATT}}(r,t),$$

where N_r is the number of units in cohort r and $N_\tau := \sum_{r\in\mathscr{R}} N_r$ is the total number of treated units.

A Peek at Some Other Results

Probability-weighted average treatment effects

$$\sum_{r \in \mathscr{R}} \frac{N_r}{N_\tau} \sum_{t=r}^{T} \psi_{rt} \hat{\tau}_{\mathsf{ATT}}(r, t)$$

This class of estimators enjoys all of the theoretical guarantees from earlier under the same assumptions.

For asymptotic normality: split the data into two subsamples:

Half to estimate the treatment effects $\hat{\tau}_{\mathsf{ATT}}(r, t)$,

Half to estimate the cohort probabilities N_r / N_τ.

Make sure to split with respect to units (which are assumed to be independent), keeping observations at all times corresponding to the same unit in the same split!

(Also possible to avoid sample splitting, if you’re willing to put up with asymptotic subgaussianity instead of normality and a conservative variance estimator. And again, I think cross-fitting might work too.)

A Peek at Some Other Results

Probability-weighted conditional average treatment effects

We can also consider probability-weighted conditional estimands

$$\sum_{r \in \mathscr{R}} \frac{\mathbb{P}(W_i = r \mid \boldsymbol{X}_i = \boldsymbol{x})}{\mathbb{P}(W_i \neq 0 \mid \boldsymbol{X}_i = \boldsymbol{x})} \sum_{t=r}^{T} \psi_{rt} \tau_{\text{CATT}}(r, t, \boldsymbol{x}).$$

Assume we have access to an estimator of the conditional cohort membership probabilities (generalized propensity scores) $\hat{\pi}_r(\boldsymbol{x}) := \hat{\mathbb{P}}(W_i = r \mid \boldsymbol{X}_i = \boldsymbol{x})$. Then we can estimate these quantities using

$$\sum_{r \in \mathscr{R}} \frac{\hat{\pi}_r(\boldsymbol{x})}{\hat{\pi}_\tau(\boldsymbol{x})} \sum_{t=r}^{T} \psi_{rt} \hat{\tau}_{\text{CATT}}(r, t, \boldsymbol{x}),$$

where $\hat{\pi}_\tau(\boldsymbol{x}) := \sum_{r \in \mathscr{R}} \hat{\pi}_r(\boldsymbol{x})$.

A Peek at Some Other Results

Probability-weighted conditional average treatment effects

$$\sum_{r \in \mathscr{R}} \frac{\hat{\pi}_r(\boldsymbol{x})}{\hat{\pi}_\tau(\boldsymbol{x})} \sum_{t=r}^{T} \psi_{rt} \hat{\tau}_{\text{CATT}}(r, t, \boldsymbol{x})$$

If $\hat{\pi}_r(\boldsymbol{x})$ is consistent, we get consistency of this estimator under an assumption of fixed d_N (and therefore p_N).

If $\hat{\pi}_r(\boldsymbol{x})$ is asymptotically normal, we get asymptotic normality (again, not in the paper yet) when we split data into three parts:

One third to estimate $\hat{\tau}_{\text{CATT}}(r, t, \boldsymbol{x})$,

One third to estimate cohort-specific covariate sample means $\overline{\boldsymbol{X}}_r$,

One third to estimate the $\hat{\pi}_r(\boldsymbol{x})$.

Outline

1 Difference-in-Differences Background

2 Extended Two-Way Fixed Effects

3 FETWFE

4 Theory

5 Simulation Studies

Access the paper

Simulation Studies: Setup

We generate synthetic data with $N = 120$ units, $T = 30$ time periods, 5 cohorts, and $d_N = 12$ covariates. This results in a total of $NT = 3600$ observations and $p_N = 2209$ parameters to estimate.

We generate a coefficient vector where 90% of the restrictions hold (so only about 220 parameters actually need to be estimated).

We repeat the following for 700 simulations:

Generate a random Gaussian covariate vector $\boldsymbol{X}_i$ for each unit

Generate random treatment assignments: each unit is either untreated or in one of the cohorts with equal probability

Generate a random response using the coefficient vector and added noise, as specified in the earlier assumptions

Estimate the treatment effects using four different methods

Simulation Studies: Methods

We estimate the overall average treatment effect on treated units (using probability weighting) using four different methods:

FETWFE: $q = 0.5$, λ_N selected from a set of 100 values by BIC

ETWFE

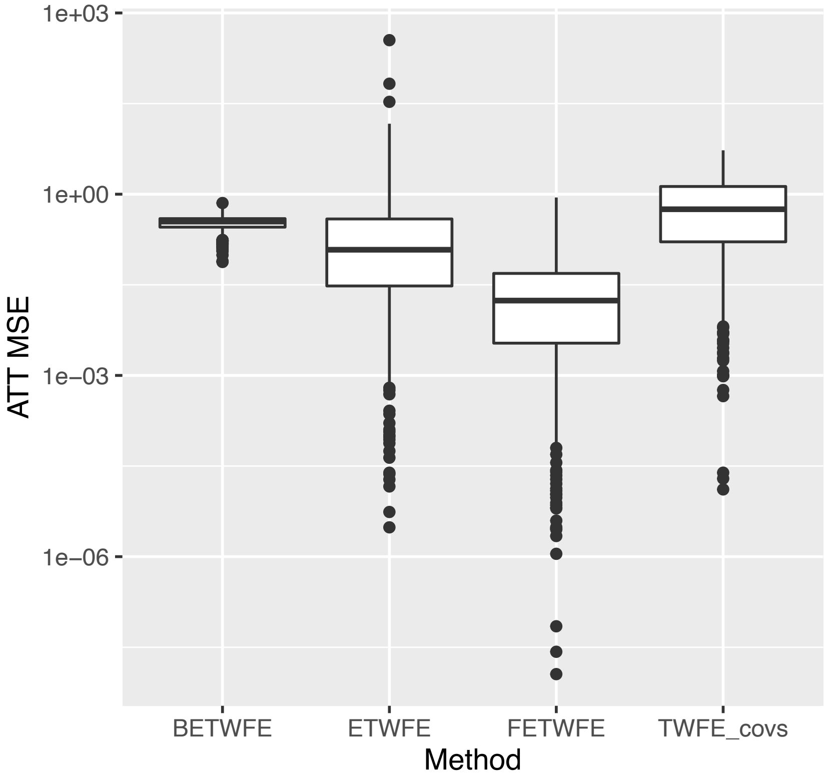
Penalized ETWFE (adding bridge regularization with $q = 0.5$ to ETWFE, but penalizing each coefficient directly rather than penalizing the coefficients towards each other)

Ordinary least squares on the oversimplified model $y_{it} = \nu_r + \gamma_t + \boldsymbol{X}_i \boldsymbol{\kappa} + \tau_r W_{it,r} + \epsilon_{it}$ (separate intercept and treatment effect for each cohort, added covariates)

Simulation Studies: Estimation error

I evaluate the squared error of each method in estimating the average treatment effect.

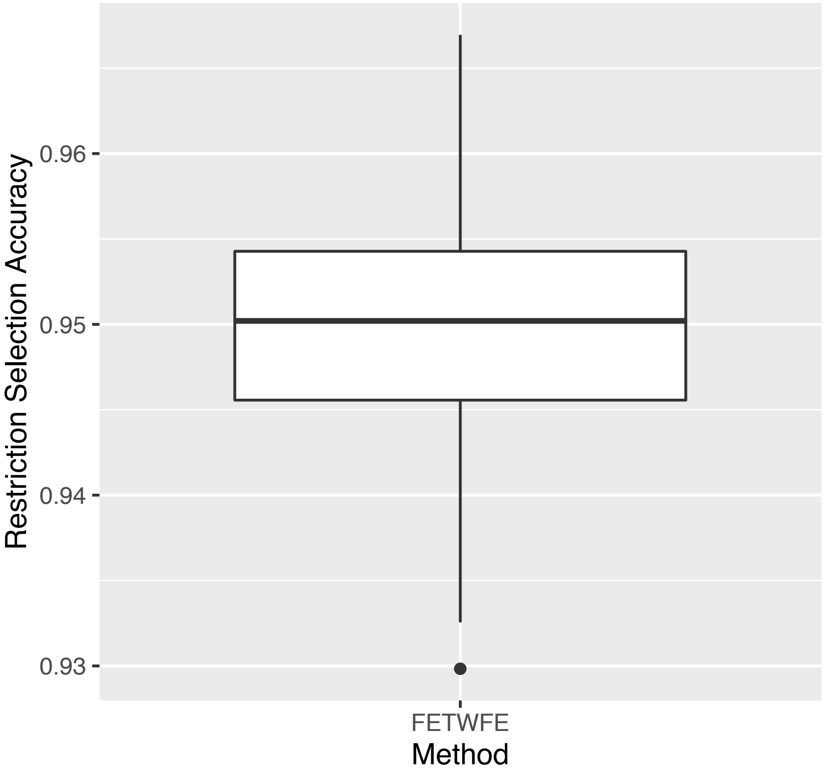
FETWFE outperforms all other methods at estimating the average treatment effect. The results are statistically significant.



Simulation Studies: Restriction Selection Consistency

On each simulation, I calculate the percentage of treatment effect restrictions that FETWFE successfully identifies.

On average, FETWFE successfully decides whether or not to choose a restriction in 95.0% of cases in each simulation (standard error: 0.025%).



Simulation Studies: Asymptotic Distribution

Second simulation study:
$N = 1200$, $T = 5$, 3 cohorts,
$d_N = 2$.

I construct nominal 95% confidence intervals for the cohort-specific average treatment effects $\tau_{\text{ATT}}(r)$ for each cohort using the recipe from Theorem 6.4.

The finite-sample confidence intervals don't quite achieve the nominal coverage level, but they're pretty close.

Cohort	Coverage	Standard Error
2	94.4%	0.9%
3	94.1%	0.9%
4	93.0%	1.0%

Summary

FETWFE solves both the high bias problem of two-way fixed effects and the high variance problem of ETWFE.

Given candidate restrictions to consider, FETWFE identifies the correct restrictions, improving estimation efficiency.

FETWFE identifies restrictions, estimates treatment effects, and allows construction of asymptotically valid confidence intervals in a single step without data splitting, for both marginal and conditional average treatment effects.

Access the paper

Thank you!

Access the paper