Large Scale Longitudinal Experiments: Estimation and Inference

October 25, 2024

Apoorva Lal
Paper (joint with Alex Fischer, Matthew Wardrop): arXiv:2410.09952

Why study panel data in recommender systems?

- We often have longitudinal data for users/devices/... : $\{\mathbf{Z}_{i,t}\}_{i=1,t=1}^{N,T}$
- In typical experiments, we *flatten the time dimension* and compute a difference in means in the post-treatment window (optionally adjusting for average outcome in pre-treatment window)
- This leads us to miss out on treatment-effect dynamics, which are intrinsic to recommender systems

Why study panel data in recommender systems?

- We often have longitudinal data for users/devices/... : $\{\mathbf{Z}_{i,t}\}_{i=1,t=1}^{N,T}$
- In typical experiments, we *flatten the time dimension* and compute a difference in means in the post-treatment window (optionally adjusting for average outcome in pre-treatment window)
- This leads us to miss out on treatment-effect dynamics, which are intrinsic to recommender systems
    1. Compliance is often not immediate: typically experimental treatments are rolled out but compliance isn't measured; analyzed as *intent to treat* (ITT)

Why study panel data in recommender systems?

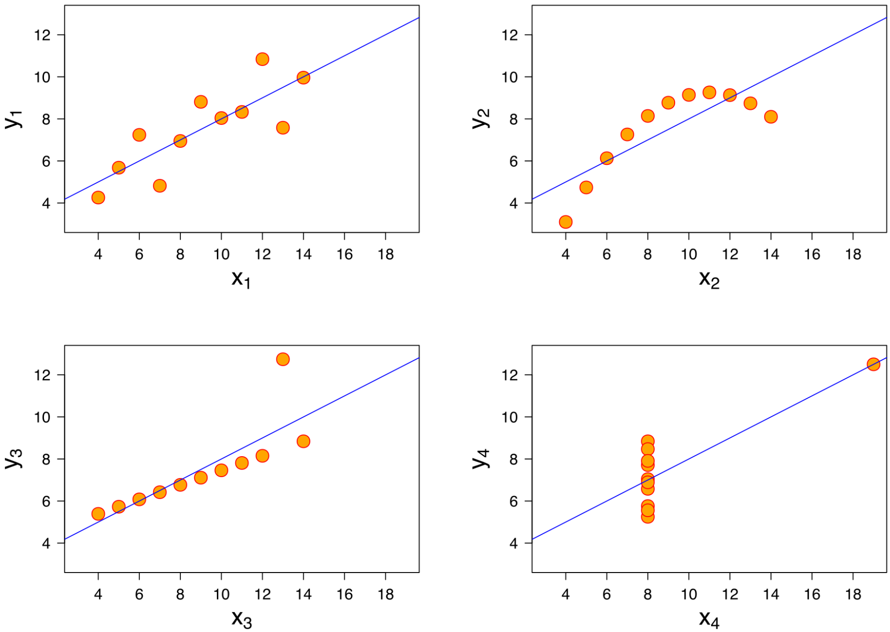
- We often have longitudinal data for users/devices/... : $\{\mathbf{Z}_{i,t}\}_{i=1,t=1}^{N,T}$
- In typical experiments, we *flatten the time dimension* and compute a difference in means in the post-treatment window (optionally adjusting for average outcome in pre-treatment window)
- This leads us to miss out on treatment-effect dynamics, which are intrinsic to recommender systems
    1. Compliance is often not immediate: typically experimental treatments are rolled out but compliance isn't measured; analyzed as *intent to treat* (ITT)
    2. Timing of compliance is informative: **activity bias :=** estimates for early compliers might not be externally valid

Why study panel data in recommender systems?

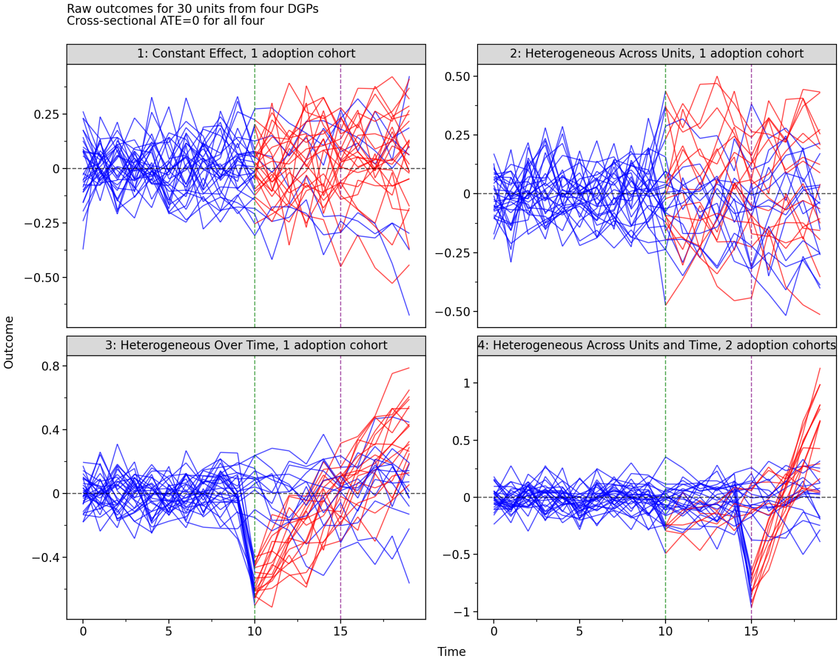
- We often have longitudinal data for users/devices/... : $\{\mathbf{Z}_{i,t}\}_{i=1,t=1}^{N,T}$
- In typical experiments, we *flatten the time dimension* and compute a difference in means in the post-treatment window (optionally adjusting for average outcome in pre-treatment window)
- This leads us to miss out on treatment-effect dynamics, which are intrinsic to recommender systems
  1. Compliance is often not immediate: typically experimental treatments are rolled out but compliance isn't measured; analyzed as *intent to treat* (ITT)
  2. Timing of compliance is informative: **activity bias :=** estimates for early compliers might not be externally valid
  3. Dynamic effects are important: **novelty bias :=** estimates of early time periods might not be externally valid

Why study panel data in recommender systems?

- We often have longitudinal data for users/devices/... : $\{\mathbf{Z}_{i,t}\}_{i=1,t=1}^{N,T}$
- In typical experiments, we *flatten the time dimension* and compute a difference in means in the post-treatment window (optionally adjusting for average outcome in pre-treatment window)
- This leads us to miss out on treatment-effect dynamics, which are intrinsic to recommender systems
  1. Compliance is often not immediate: typically experimental treatments are rolled out but compliance isn't measured; analyzed as *intent to treat* (ITT)
  2. Timing of compliance is informative: **activity bias :=** estimates for early compliers might not be externally valid
  3. Dynamic effects are important: **novelty bias :=** estimates of early time periods might not be externally valid
- **this project:** propose scalable panel data estimators that help identify these
- temporal and cohort-level granularity is informative and important - don't flatten them with a T-test



y_1
x_1
y_2
x_2
y_3
x_3
y_4
x_4


Raw outcomes for 30 units from four DGPs
Cross-sectional ATE=0 for all four
1: Constant Effect, 1 adoption cohort
2: Heterogeneous Across Units, 1 adoption cohort
3: Heterogeneous Over Time, 1 adoption cohort
4: Heterogeneous Across Units and Time, 2 adoption cohorts
Outcome
Time
0.25
0
−0.25
−0.50
0.50
0.8
0.4
−0.4
1
0.5
−0.5
−1
0
5
10
15

Compressing Least Squares

$$\widehat{\boldsymbol{\beta}} = \Bigg(\underbrace{\mathbf{X}^\top}_{P \times N}\underbrace{\mathbf{X}}_{N \times P}\Bigg)^{-1}\underbrace{\mathbf{X}^\top}_{P \times N}\underbrace{\mathbf{y}}_{N \times 1}$$

$$\mathbb{V}\left[\widehat{\beta}\right] = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\Omega\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}$$

Compressing Least Squares

$$\widehat{\boldsymbol{\beta}} = \Bigg(\underbrace{\mathbf{X}^\top}_{P \times N}\underbrace{\mathbf{X}}_{N \times P}\Bigg)^{-1} \underbrace{\mathbf{X}^\top}_{P \times N}\underbrace{\mathbf{y}}_{N \times 1}$$

$$\mathbb{V}\left[\widehat{\beta}\right] = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \Omega \mathbf{X} \left(\mathbf{X}^\top \mathbf{X}\right)^{-1}$$

- $\mathbf{X}$ includes intercept and binary treatment indicator W_i
- Suppose $\mathbf{X}$ takes on values in $\mathcal{X}$ with finite cardinality C
- Then, we compute sufficient statistics by strata, run $\widetilde{\mathbf{y}}/\widetilde{\mathbf{n}} \sim \widetilde{\mathbf{X}}\boldsymbol{\beta} + \varepsilon$: C observations

Compressing Least Squares

$$\widehat{\boldsymbol{\beta}} = \Big(\underbrace{\mathbf{X}^\top}_{P \times N}\underbrace{\mathbf{X}}_{N \times P}\Big)^{-1}\underbrace{\mathbf{X}^\top}_{P \times N}\underbrace{\mathbf{y}}_{N \times 1}$$

$$\mathbb{V}\left[\widehat{\beta}\right] = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\Omega\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}$$

- $\mathbf{X}$ includes intercept and binary treatment indicator W_i
- Suppose $\mathbf{X}$ takes on values in $\mathcal{X}$ with finite cardinality C
- Then, we compute sufficient statistics by strata, run $\widetilde{\mathbf{y}}/\widetilde{\mathbf{n}} \sim \widetilde{\mathbf{X}}\boldsymbol{\beta} + \varepsilon$: C observations

Table 1: Example dataset and its compressed versions.

(a)

$\mathbf{M}$	$\mathbf{y}$
A	1
A	1
A	2
B	3
B	4
C	5

(b)

$\dot{\mathbf{M}}$	$\dot{\mathbf{y}}$	$\dot{\mathbf{n}}$
A	1	2
A	2	1
B	3	1
B	4	1
C	5	1

(c)

$\bar{\mathbf{M}}$	$\bar{\mathbf{y}}$	$\bar{\mathbf{n}}$
A	1.33	3
B	3.5	2
C	5	1

(d)

$\tilde{\mathbf{M}}$	$\tilde{\mathbf{y}}'$	$\tilde{\mathbf{y}}''$	$\tilde{\mathbf{n}}$
A	4	6	3
B	7	25	2
C	5	25	1

(a) Uncompressed data. (b) f-weights: $(\mathbf{y}, \mathbf{M})$-compressed records.
(c) Groups: $(\mathbf{M})$-compressed records. (d) Sufficient Statistics: $(\mathbf{M})$-compressed records.

ABlaze is built around this: Wong et al(2021)

Compressing Least Squares

$$\widehat{\boldsymbol{\beta}} = \Big(\underbrace{\mathbf{X}^\top}_{P\times N}\underbrace{\mathbf{X}}_{N\times P}\Big)^{-1}\underbrace{\mathbf{X}^\top}_{P\times N}\underbrace{\mathbf{y}}_{N\times 1}$$

$$\mathbb{V}\left[\widehat{\beta}\right] = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\Omega\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}$$

- $\mathbf{X}$ includes intercept and binary treatment indicator W_i
- Suppose $\mathbf{X}$ takes on values in $\mathcal{X}$ with finite cardinality C
- Then, we compute sufficient statistics by strata, run $\widetilde{\mathbf{y}}/\widetilde{\mathbf{n}} \sim \widetilde{\mathbf{X}}\boldsymbol{\beta} + \varepsilon$: C observations

Table 1: Example dataset and its compressed versions.

(a)

$\mathbf{M}$	$\mathbf{y}$
A	1
A	1
A	2
B	3
B	4
C	5

(b)

$\dot{\mathbf{M}}$	$\dot{\mathbf{y}}$	$\dot{\mathbf{n}}$
A	1	2
A	2	1
B	3	1
B	4	1
C	5	1

(c)

$\bar{\mathbf{M}}$	$\bar{\mathbf{y}}$	$\bar{\mathbf{n}}$
A	1.33	3
B	3.5	2
C	5	1

(d)

$\tilde{\mathbf{M}}$	$\tilde{\mathbf{y}}'$	$\tilde{\mathbf{y}}''$	$\tilde{\mathbf{n}}$
A	4	6	3
B	7	25	2
C	5	25	1

(a) Uncompressed data. (b) f-weights: $(\mathbf{y}, \mathbf{M})$-compressed records.
(c) Groups: $(\mathbf{M})$-compressed records. (d) Sufficient Statistics: $(\mathbf{M})$-compressed records.

ABlaze is built around this: Wong et al(2021)

$$\widehat{\mathbb{V}}(\boldsymbol{\beta}) = \overbrace{(\widetilde{\mathbf{X}}^\top \mathrm{diag}(\widetilde{\mathbf{n}})\widetilde{\mathbf{X}})^{-1}}^{\text{Bread}} \overbrace{\widetilde{\mathbf{X}}^\top \mathrm{diag}\Big(\sum_j \underbrace{(\widetilde{y}_j^2\widetilde{n}_j - 2\widetilde{y}_j\widetilde{y}_j' + \widetilde{y}_j'')}_{\text{RSS in } j}\Big)\widetilde{\mathbf{X}}}^{\text{Meat}} \overbrace{(\widetilde{\mathbf{X}}^\top \mathrm{diag}(\widetilde{\mathbf{n}})\widetilde{\mathbf{X}})^{-1}}^{\text{Bread}}$$

Building Blocks

- Base untreated potential outcome : $Y_{it}^0 = \alpha_i + \gamma_t + \epsilon_{it}$
- Treated potential outcome under static, constant effects: $Y_{it}^1 = Y_{it}^0 + \tau$

Building Blocks

- Base untreated potential outcome : $Y_{it}^0 = \alpha_i + \gamma_t + \epsilon_{it}$
- Treated potential outcome under static, constant effects: $Y_{it}^1 = Y_{it}^0 + \tau$
- Treated potential outcome under unit heterogeneity: $Y_{it}^1 = Y_{it}^0 + \tau_i$
- Treated potential outcome under time heterogeneity: $Y_{it}^1 = Y_{it}^0 + \sum_{k\geq 0} \tau_k 1\{t - T_i = k\}$

Building Blocks

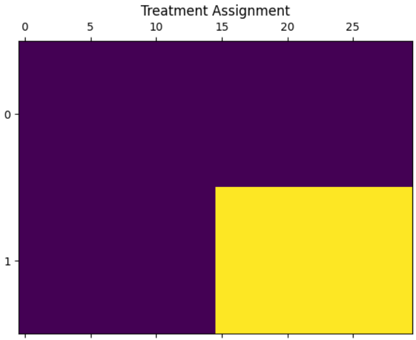
- Base untreated potential outcome : $Y_{it}^{0} = \alpha_i + \gamma_t + \epsilon_{it}$
- Treated potential outcome under static, constant effects: $Y_{it}^{1} = Y_{it}^{0} + \tau$
- Treated potential outcome under unit heterogeneity: $Y_{it}^{1} = Y_{it}^{0} + \tau_i$
- Treated potential outcome under time heterogeneity: $Y_{it}^{1} = Y_{it}^{0} + \sum_{k \geq 0} \tau_k 1\{t - T_i = k\}$
- Treated potential outcome under time heterogeneity: $Y_{it}^{1} = Y_{it}^{0} + \sum_{k \geq 0} \tau_{ik} 1\{t - T_i = k\}$

Building Blocks

- Base untreated potential outcome : $Y_{it}^0 = \alpha_i + \gamma_t + \epsilon_{it}$
- Treated potential outcome under static, constant effects: $Y_{it}^1 = Y_{it}^0 + \tau$
- Treated potential outcome under unit heterogeneity: $Y_{it}^1 = Y_{it}^0 + \tau_i$
- Treated potential outcome under time heterogeneity: $Y_{it}^1 = Y_{it}^0 + \sum_{k \geq 0} \tau_k 1\{t - T_i = k\}$
- Treated potential outcome under time heterogeneity: $Y_{it}^1 = Y_{it}^0 + \sum_{k \geq 0} \tau_{ik} 1\{t - T_i = k\}$
- Observed outcome:

$$Y_{it} = Y_{it}^0(1 - W_{it}) + Y_{it}^1 W_{it}$$

- Under random assignment or parallel trends, we can form estimates of Y_{it}^0 and construct estimates of (reasonable averages of) τ_{it}



When assignment time is endogenous, need generalized propensity score Arkhangelsky et al (2024)

Regressions we'd like to run

$$\bar{Y}_{i,t>T_0} = \alpha + \tau W_i + \varepsilon_i$$ Diff in Means

$$\bar{Y}_{i,t>T_0} = \alpha + \tau W_i + \beta \bar{Y}_{i,t<T_0} + \varepsilon_i$$ CUPED

- $Z_{it} := \mathbb{1}\{\text{argmin}\{t : W_{it} = 1\} - t = k\}$ (treatment indicator $\times$ event-time)

Regressions we'd like to run

$\bar{Y}_{i,t>T_0} = \alpha + \tau W_i + \varepsilon_i$	Diff in Means
$\bar{Y}_{i,t>T_0} = \alpha + \tau W_i + \beta \bar{Y}_{i,t<T_0} + \varepsilon_i$	CUPED
$Y_{it} = \alpha_i + \gamma_t + \tau W_{it} + \varepsilon_{it}$	Static TWFE
$Y_{it} = \alpha_i + \gamma_t + \sum_{k=0,k\neq -1}^{T} \tau_k Z_{it}^k + \varepsilon_{it}$	Dynamic TWFE (Event Study)
$Y_{it} = \alpha + \gamma_t + \sum_{k\geq 0}^{T} \tau_k Z_{it}^k + \varepsilon_{it}$	Dynamic DiM
$Y_{it} = \alpha_i + \gamma_t + \sum_{c=2}^{C} \sum_{k=0,k\neq -1}^{T} \mathbb{1}\{G_i = c\} \tau_{kc} Z_{it}^k + \varepsilon_{it}$	Disagg Dynamic TWFE (Staggered Event Study)

- $Z_{it} := \mathbb{1}\{\text{argmin}\{t : W_{it} = 1\} - t = k\}$ (treatment indicator $\times$ event-time)

Regressions we'd like to run

$$\bar{Y}_{i,t>T_0} = \alpha + \tau W_i + \varepsilon_i$$ Diff in Means

$$\bar{Y}_{i,t>T_0} = \alpha + \tau W_i + \beta \bar{Y}_{i,t<T_0} + \varepsilon_i$$ CUPED

$$Y_{it} = \alpha_i + \gamma_t + \tau W_{it} + \varepsilon_{it}$$ Static TWFE

$$Y_{it} = \alpha_i + \gamma_t + \sum_{k=0,k\neq -1}^{T} \tau_k Z_{it}^k + \varepsilon_{it}$$ Dynamic TWFE (Event Study)

$$Y_{it} = \alpha + \gamma_t + \sum_{k\geq 0}^{T} \tau_k Z_{it}^k + \varepsilon_{it}$$ Dynamic DiM

$$Y_{it} = \alpha_i + \gamma_t + \sum_{c=2}^{C} \sum_{k=0,k\neq -1}^{T} \mathbb{1}\{G_i = c\} \tau_{kc} Z_{it}^k + \varepsilon_{it}$$ Disagg Dynamic TWFE (Staggered Event Study)

- $Z_{it} := \mathbb{1}\{\text{argmin}\{t : W_{it} = 1\} - t = k\}$ (treatment indicator $\times$ event-time)
- $N \times T \approx 50m \times 95 = 4.75$b obs

Regressions we'd like to run

$$\bar{Y}_{i,t>T_0} = \alpha + \tau W_i + \varepsilon_i$$ Diff in Means

$$\bar{Y}_{i,t>T_0} = \alpha + \tau W_i + \beta \bar{Y}_{i,t<T_0} + \varepsilon_i$$ CUPED

$$Y_{it} = \alpha_i + \gamma_t + \tau W_{it} + \varepsilon_{it}$$ Static TWFE

$$Y_{it} = \alpha_i + \gamma_t + \sum_{k=0, k \neq -1}^{T} \tau_k Z_{it}^k + \varepsilon_{it}$$ Dynamic TWFE (Event Study)

$$Y_{it} = \alpha + \gamma_t + \sum_{k \geq 0}^{T} \tau_k Z_{it}^k + \varepsilon_{it}$$ Dynamic DiM

$$Y_{it} = \alpha_i + \gamma_t + \sum_{c=2}^{C} \sum_{k=0, k \neq -1}^{T} \mathbb{1}\{G_i = c\} \tau_{kc} Z_{it}^k + \varepsilon_{it}$$ Disagg Dynamic TWFE (Staggered Event Study)

- $Z_{it} := \mathbb{1}\{\text{argmin}\{t : W_{it} = 1\} - t = k\}$ (treatment indicator $\times$ event-time)
- $N \times T \approx 50m \times 95 = 4.75$b obs
- W_i, α_i, γ_t jointly identify a single observation - cannot compress
- Unit intercepts α_i: millions of distinct values

Trick 1: Partialling out - the within estimator

- We don't inherently care about α_i, γ_t; they are nuisance parameters
- Partial them out (i.e. kick them out of $\mathbf{X}'\mathbf{X}$)
  - Frisch-Waugh-Lovell Theorem / Gram-Schmidt Process

Trick 1: Partialling out - the within estimator

- We don't inherently care about α_i, γ_t; they are nuisance parameters
- Partial them out (i.e. kick them out of $\mathbf{X}'\mathbf{X}$)
  - Frisch-Waugh-Lovell Theorem / Gram-Schmidt Process
  - specialized methods for high dimensional categorical covariates: Method of alternating projections / Kaczmarz method (implemented in `areg`/`reghdfe`/`fixest`/`felm`/`pyfixest`/`...`)
- Residualise RHS $W_{it} - \overline{W}_{i,\cdot} - \overline{W}_{\cdot,t} =: \ddot{W}_{it}$

Trick 1: Partialling out - the within estimator

- We don't inherently care about α_i, γ_t; they are nuisance parameters
- Partial them out (i.e. kick them out of $\mathbf{X}'\mathbf{X}$)
  - Frisch-Waugh-Lovell Theorem / Gram-Schmidt Process
  - specialized methods for high dimensional categorical covariates: Method of alternating projections / Kaczmarz method (implemented in `areg`/`reghdfe`/`fixest`/`felm`/`pyfixest`/`...`)
- Residualise RHS $W_{it} - \overline{W}_{i,\cdot} - \overline{W}_{\cdot,t} =: \ddot{W}_{it}$
- Obviates the need to adjust for time-invariant member characteristics (all absorbed in FEs)

$$Y_{it} = \tau \ddot{W}_{it} + \varepsilon_{it}$$

This regression *can* be compressed. However, we don't have an equivalent representation for dynamic regressions (event studies, disaggregated event studies), etc.

Trick 2: Mundlak (1978) + Wooldridge (2021) Trick

$$Y_{it} = \alpha_i + \gamma_t + \tau W_{it} + \varepsilon_{it}$$ **Static TWFE**

$$Y_{it} = \delta + \tau W_{it} + \phi \bar{W}_{i,\cdot} + \psi \bar{W}_{\cdot,t} + \varepsilon_{it}$$ **Mundlak-ed Static TWFE**

Trick 2: Mundlak (1978) + Wooldridge (2021) Trick

$$Y_{it} = \alpha_i + \gamma_t + \tau W_{it} + \varepsilon_{it}$$ **Static TWFE**

$$Y_{it} = \delta + \tau W_{it} + \phi \bar{W}_{i,\cdot} + \psi \bar{W}_{\cdot,t} + \varepsilon_{it}$$ **Mundlak-ed Static TWFE**

$$Y_{it} = \alpha_i + \gamma_t + \sum_{k=0, k\neq -1}^{T} \tau_k Z_{it}^k + \varepsilon_{it}$$ **Dynamic TWFE**

$$Y_{it} = \delta + \psi D_i + \sum_{k=1}^{T} \phi_t \mathbb{1}\{t = k\} + \sum_{k=1}^{T} \tau_k D_i \mathbb{1}\{t = k\} + \varepsilon_{it}$$

Trick 2: Mundlak (1978) + Wooldridge (2021) Trick

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

Static TWFE

$$Y_{it} = \delta + \tau W_{it} + \phi \bar{W}_{i,\cdot} + \psi \bar{W}_{\cdot,t} + \varepsilon_{it}$$

Mundlak-ed Static TWFE

$$Y_{it} = \alpha_i + \gamma_t + \sum_{k=0, k\neq -1}^{T} \tau_k Z_{it}^k + \varepsilon_{it}$$

Dynamic TWFE

$$Y_{it} = \delta + \psi D_i + \sum_{k=1}^{T} \phi_t \mathbb{1}\{t = k\} + \sum_{k=1}^{T} \tau_k D_i \mathbb{1}\{t = k\} + \varepsilon_{it}$$

$$Y_{it} = \alpha_i + \gamma_t + \sum_{c=2}^{C} \sum_{k=0, k\neq -1}^{T} \mathbb{1}\{G_i = c\} \tau_{kc} Z_{it}^k + \varepsilon_{it}$$

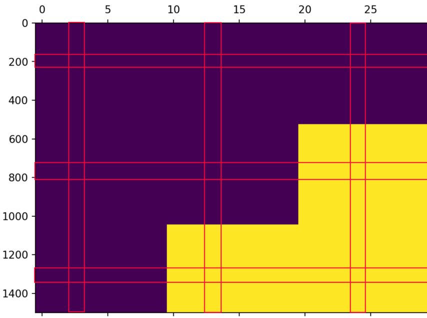
Disagg Dynamic TWFE

$$Y_{it} = \delta + \underbrace{\sum_{c=1}^{C} \psi_c \mathbb{1}\{D_i = k\}}_{\text{Cohort Dummies}} + \underbrace{\sum_{k=1}^{T} \phi_t \mathbb{1}\{t = k\}}_{\text{Time Dummies}} + \underbrace{\sum_{c=1}^{C} \sum_{k=1}^{T} \tau_{kc} \mathbb{1}\{D_i = c\} \mathbb{1}\{t = k\}}_{\text{Cohort} \times \text{Time interactions}} + \varepsilon_{it}$$

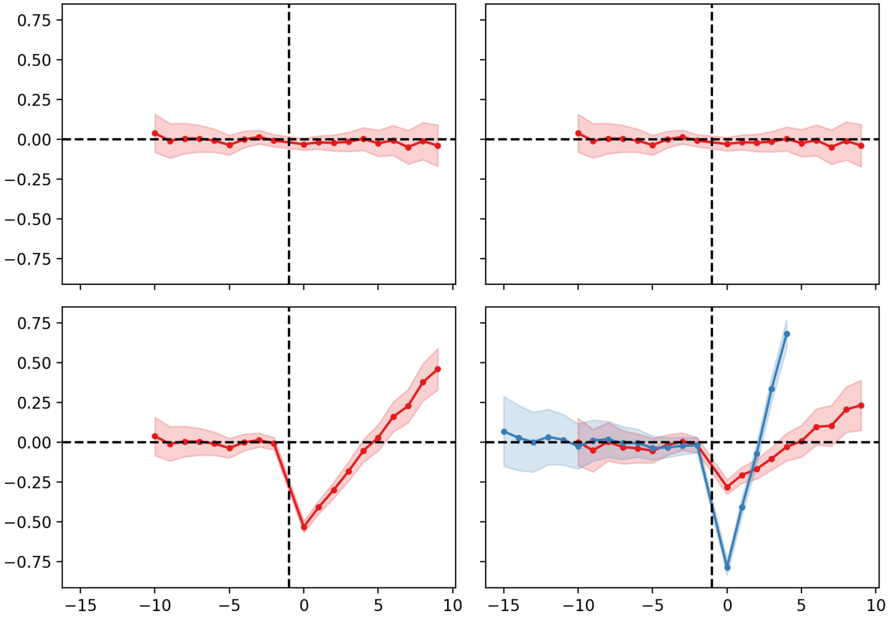
Design Matrix Dimensions

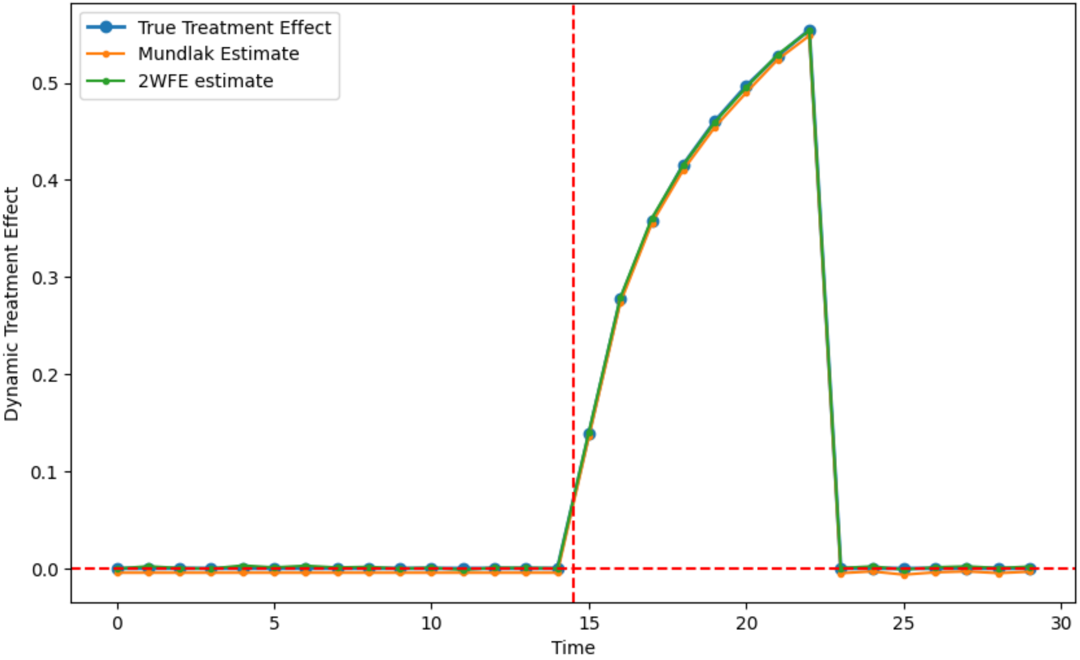
	(1) Standard	M	(2) Mundlak	$\tilde{M}$
Static	$Y_{it} = \alpha_i + \gamma_t + \tau W_{it} + \varepsilon_{it}$	NT	$Y_{it} = \alpha + \tau W_{it} + \psi \overline{W}_{i,\cdot} + \phi \overline{W}_{\cdot,t} + \varepsilon_{it}$	2+(C+1)
Dyn	$Y_{it} = \alpha_i + \gamma_t + \sum_{k \neq -1} \tau_k Z_{it}^k + \varepsilon_{it}$	NT	$Y_{it} = \alpha + \psi D_i + \sum_{k=1}^{T} \phi_t 1_{t=k} + \sum_{k=1}^{T} \tau_k D_i 1_{t=k} + \varepsilon_{it}$	2T
Dyn+Stagg	$Y_{it} = \alpha_i + \gamma_t + \sum_{c=1}^{C} \sum_{k \neq -1} \tau_{kc} 1_{G_i = c} Z_{it}^k + \varepsilon_{it}$	NT	$Y_{it} = \alpha + \sum_{c=1}^{C} \psi_c 1_{D_i = c} + \sum_{k=1}^{T} \phi_t 1_{t=k} + \sum_{c=1}^{C} \sum_{k=1}^{T} \tau_{kc} 1_{D_i = c} 1_{t=k} + \varepsilon_{it}$	CT

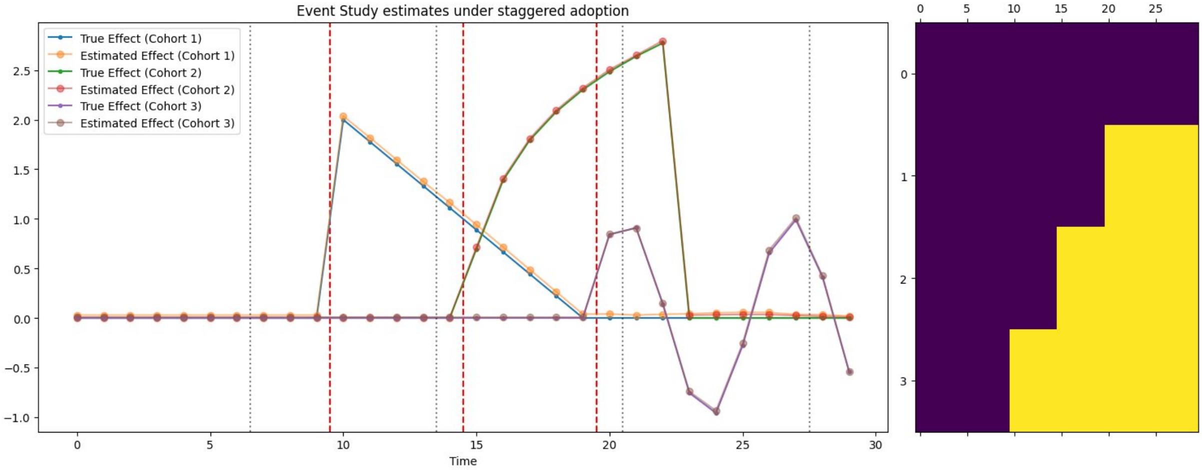
- N units, T time periods
- M, $\widetilde{M}$: number of required rows in design matrix
- C unique adoption cohorts (including control)
- $2 + (C + 1) = 4$ obs in standard A/B test
- Cluster-robust inference with cluster-bootstrap



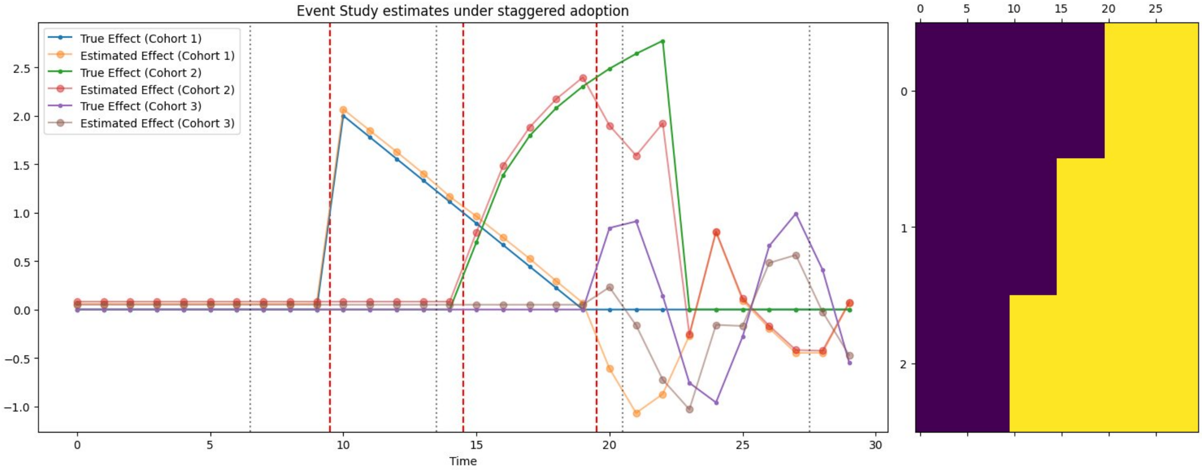
Event Study figures for the four scenarios



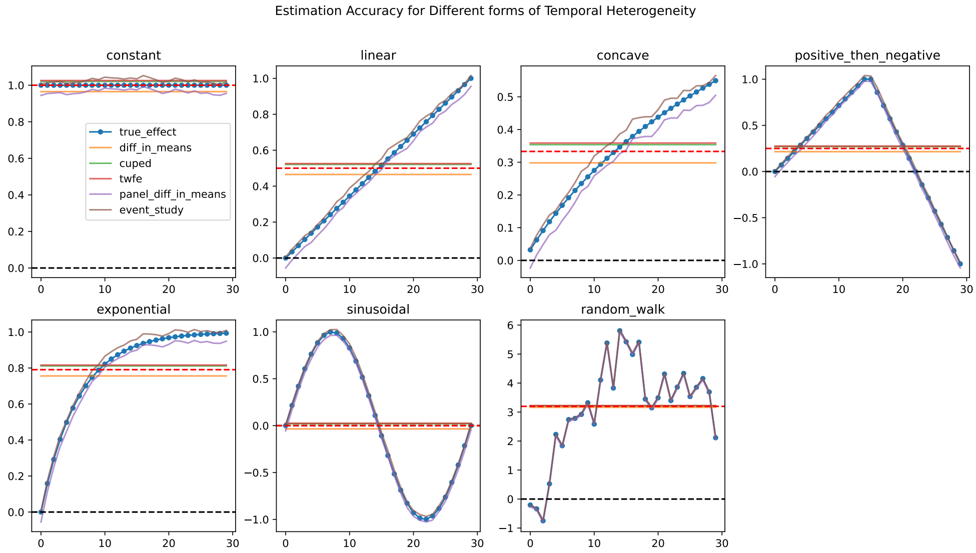

True Treatment Effect
Mundlak Estimate
2WFE estimate
Dynamic Treatment Effect
0.0
0.1
0.2
0.3
0.4
0.5
0
5
10
15
20
25
30
Time



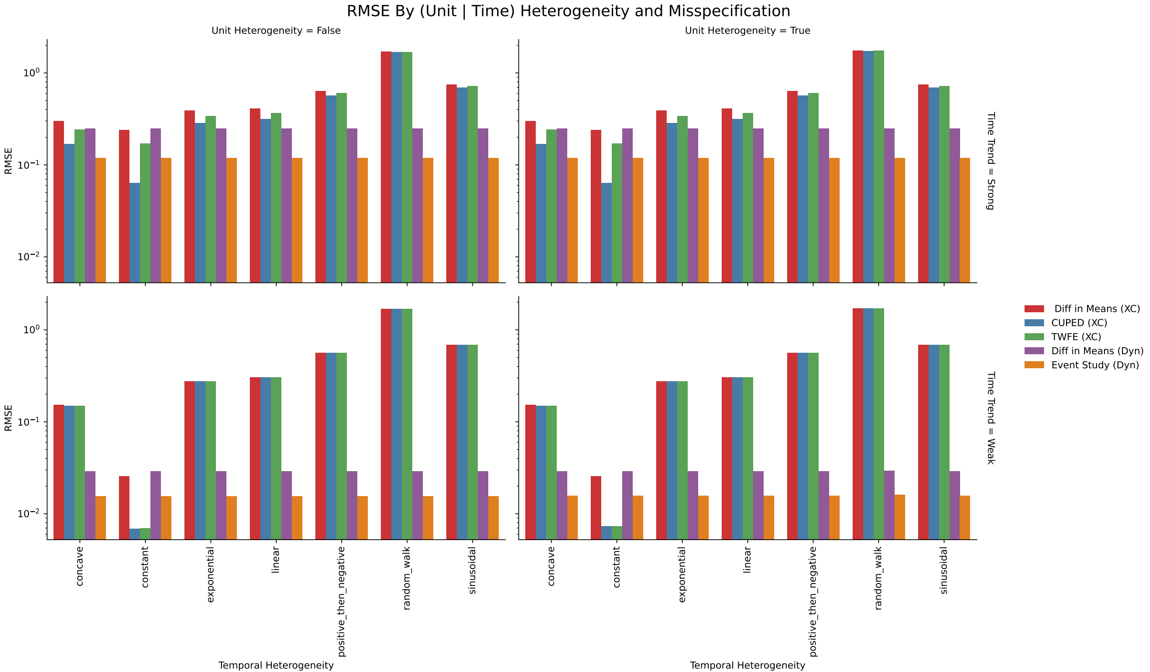
Event Study estimates under staggered adoption
True Effect (Cohort 1)
Estimated Effect (Cohort 1)
True Effect (Cohort 2)
Estimated Effect (Cohort 2)
True Effect (Cohort 3)
Estimated Effect (Cohort 3)
Time


Event Study estimates under staggered adoption
True Effect (Cohort 1)
Estimated Effect (Cohort 1)
True Effect (Cohort 2)
Estimated Effect (Cohort 2)
True Effect (Cohort 3)
Estimated Effect (Cohort 3)
2.5
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
0
5
10
15
20
25
30
Time
0
5
10
15
20
25
0
1
2

Numerical Experiments



Estimation Accuracy for Different forms of Temporal Heterogeneity
constant
linear
concave
positive_then_negative
exponential
sinusoidal
random_walk
true_effect
diff_in_means
cuped
twfe
panel_diff_in_means
event_study


RMSE By (Unit | Time) Heterogeneity and Misspecification
Unit Heterogeneity = False
Unit Heterogeneity = True
Time Trend = Strong
Time Trend = Weak
RMSE
Temporal Heterogeneity
concave
constant
exponential
linear
positive_then_negative
random_walk
sinusoidal
Diff in Means (XC)
CUPED (XC)
TWFE (XC)
Diff in Means (Dyn)
Event Study (Dyn)

Runtime Comparisons

$1000 \times \{14, 28, 42, 140, 280, 420\}$ units, $\{14, 28, 42\}$ days.

Observations	Units	Periods	duckreg	pyfixest	pyfixest compressed	statsmodels
14K	1K	14	0.03	0.18	0.17	2.75
28K	1K	28	0.03	0.19	0.18	6.27
42K	1K	42	0.04	0.20	0.21	9.43
140K	10K	14	0.05	0.22	0.26	x
280K	10K	28	0.06	0.26	0.36	x
420K	10K	42	0.04	0.31	0.47	x
1M	100K	14	0.07	0.64	1.27	x
3M	100K	28	0.21	1.00	2.28	x
4M	100K	42	0.24	1.41	3.43	x
14M	1M	14	1.03	4.88	22.12	x
28M	1M	28	3.41	8.07	62.07	x
42M	1M	42	10.92	13.60	117.63	x
140M	10M	14	19.70	123.88	x	x
Mundlak			✓	-	✓	-
Compression			✓	-	✓	-
Out-of-Memory			✓	-	-	-

Conclusion

- Methods + Software for compressed, out-of-memory computation of commonly used least-squares panel data estimators
  - `duckreg` : powered by duckDB - out of memory
  - `pyfixest`: general purpose in-memory regression package, compression in `Polars`

Conclusion

- Methods + Software for compressed, out-of-memory computation of commonly used least-squares panel data estimators
  - `duckreg` : powered by duckDB - out of memory
  - `pyfixest`: general purpose in-memory regression package, compression in `Polars`
- Applicable to both scalable estimation of event studies in experiments (focus of this talk), and difference-in-differences methods in observational settings
- Applies to any saturated regression, balanced-or-unbalanced panel data, or any high-dimensional categorical covariate
- Enables stratifying by adoption cohorts; helps mitigate against activity bias