Large Scale Longitudinal Experiments: Estimation and Inference

Apoorva Lal, Alex Fischer, and Matthew Wardrop

Netflix, Trivago, Netflix

Introduction

- A/B tests often analyzed with simple methods (t-tests, linear regression - CUPED)
- These methods flatten time-dimension into single 'post-treatment' outcome
- In presence of effect heterogeneity, post-treatment average may not be good summary statistic for decisionmaking

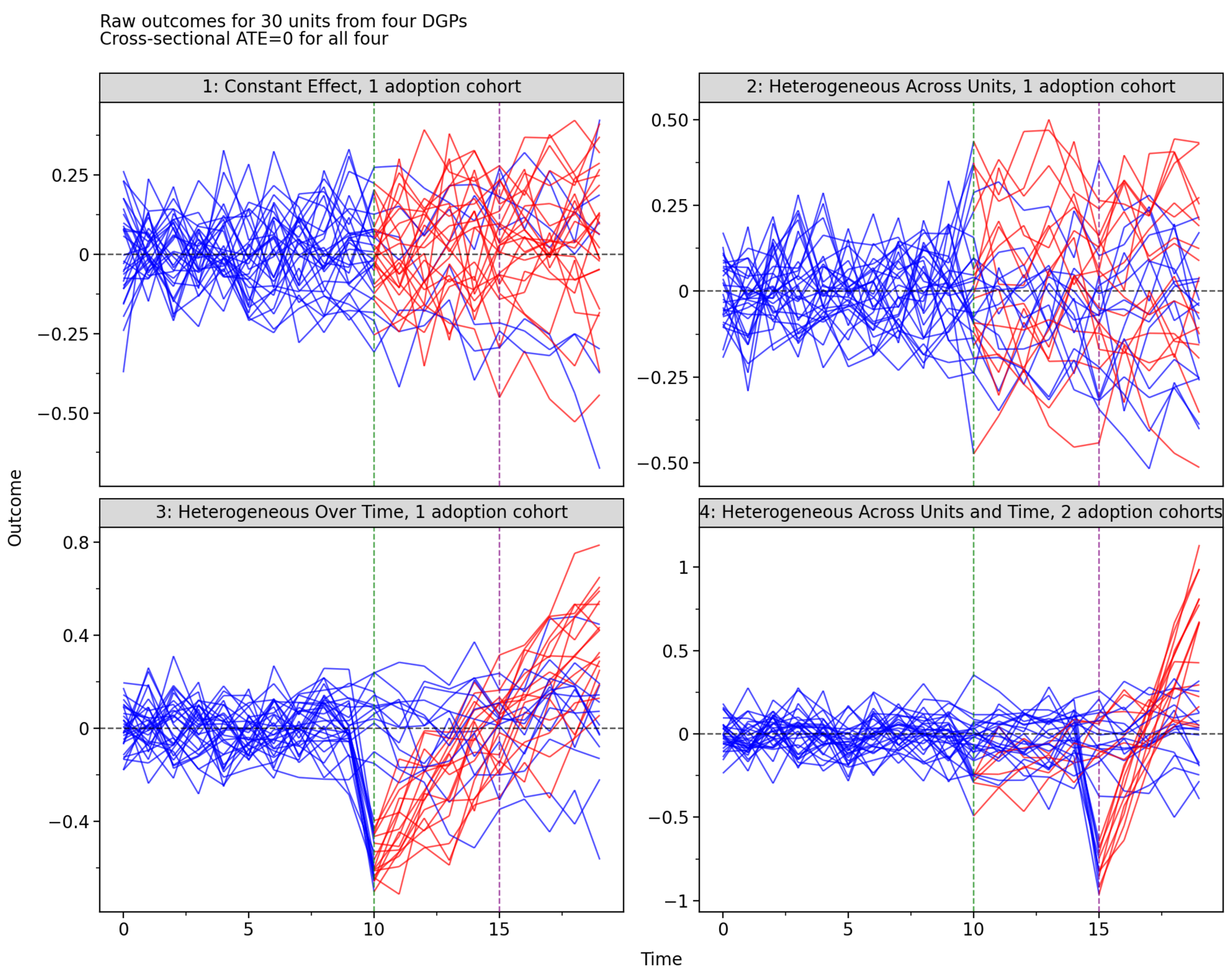


Figure 1. A Panel Data Anscombe's Quartet

Contribution

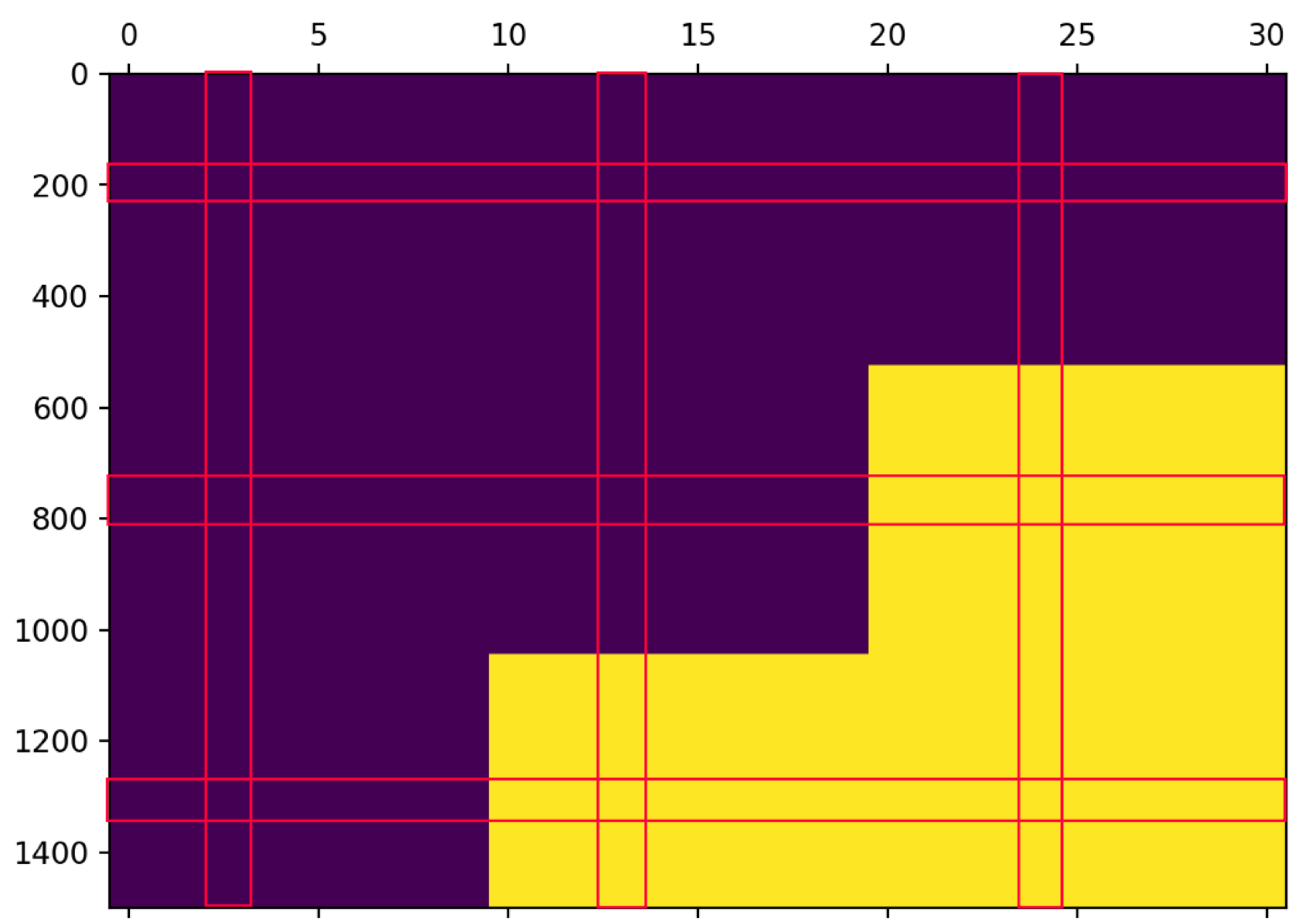
- Propose scalable panel-regression methods using reparametrization + compression
  - Reparametrization: Mundlak trick - replace intercepts with regressor averages
  - Compression: Mundlak specification is stratified and has much lower cardinality than FE specification - Weighted Least Squares with Frequency Weights
- open-source Python libraries for in-memory and out-of-memory computation
  - out-of-memory: `duckreg` (powered by DuckDB)
  - in-memory: `pyfixest`
- Compression performed in SQL: scales to arbitrarily large data

The Mundlak Representation

- Mundlak (1978) insight: unit intercepts can be eliminated using covariate averages.
- Extends to arbitrary stratified regressions (2WFE is a special case) [Arkhangelsky and Imbens (2023)]

	(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

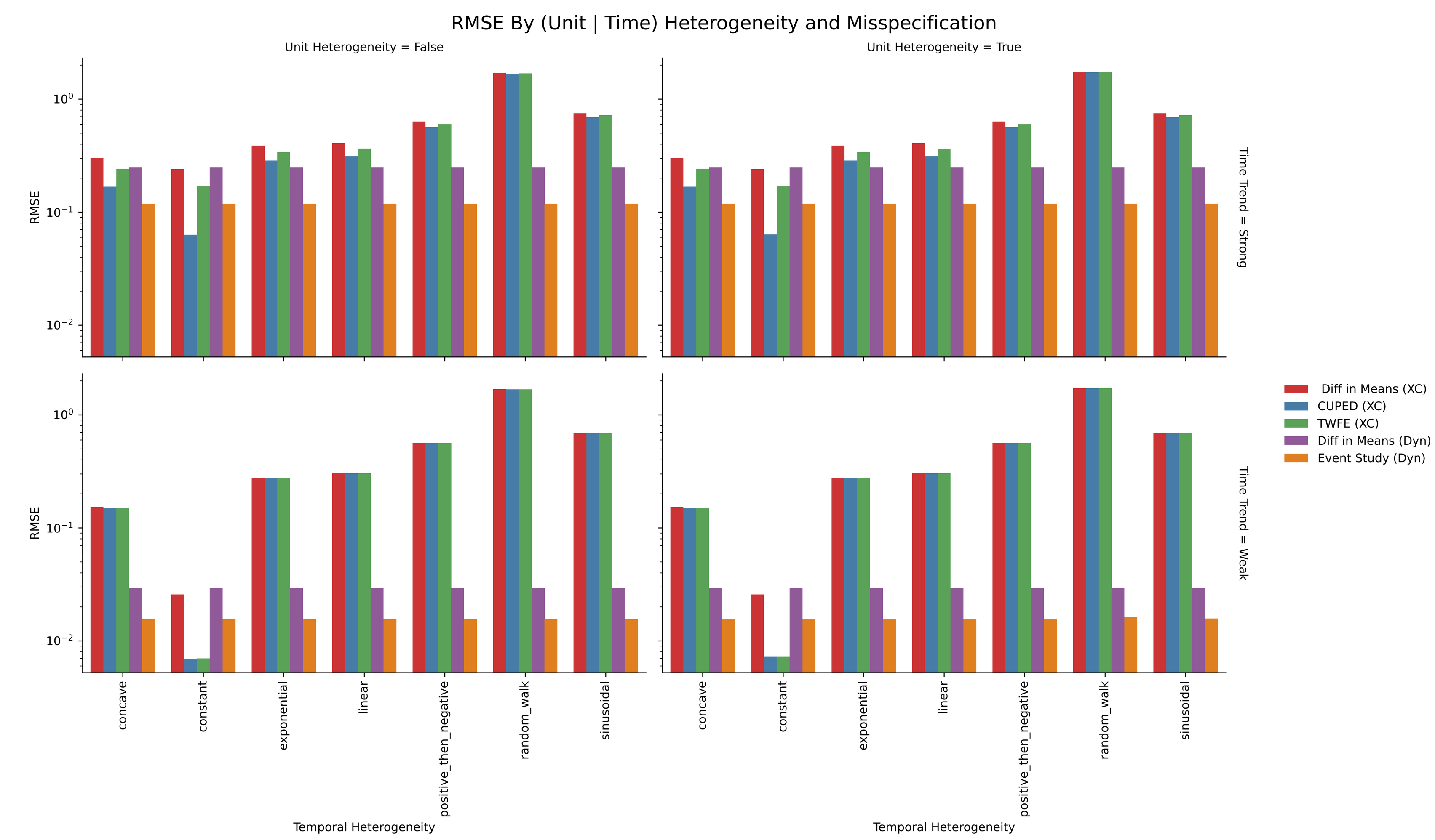
- N units, T time periods, C treatment cohorts; $M, \tilde{M}$ size of design matrix
- RHS of (1) unique by $W_{it}, \alpha_i + \gamma_t \rightarrow$ cannot be compressed; infeasible at large scale $N >> T$; $20m$ obs $\times 90$ days $= 1.8$ billion obs
- RHS of (2) unique by $W_{it}, \bar{W}_{i,\cdot}, \bar{W}_{\cdot,t}$, which is compressible
- For regular A/B: TWM has $\tilde{N}$ =4 observations



- coefs, HC(0-3) SEs computable in closed-form from summary stats (Wong et al)
- Clustered SEs with cluster bootstrap, or closed-form via distributed computing

Numerical Experiments

- DGP: $Y_{it} = \alpha_i + \gamma_t + \beta_i t + \tau_{it} W_{it} + \varepsilon_{it}$
- Time trend piece is unmodeled; variance of β_i controls degree of misspecification



- Timing: `duckreg:pyfixest:statsmodels` runtimes scale $1:40:600$ for cross-sectional regressions
- panel simulations: 14K to 140M observations
  - for $N, T = 140M, 42$, duckreg (OOM) is between 4-6x faster than pyfixest (in-memory)
  - duckreg scales arbitrarily well
  - statsmodels: Repeated OOM errors

References

- Mundlak, Y. (1978). "On the pooling of time series and cross section data". Econometrica.
- Wong, Jeffrey, Eskil Forsell, Randall Lewis, Tobias Mao, and Matthew Wardrop. 2021. "You Only Compress Once: Optimal Data Compression for Estimating Linear Models." arXiv [Cs.LG]. arXiv. http://arxiv.org/abs/2102.11297.
- Wooldridge, J. M. (2021). "Two-Way Fixed Effects, the Two-Way Mundlak Regression, and Difference-in-Differences Estimators". Working paper.
- Dmitry Arkhangelsky, Guido W Imbens, Fixed Effects and the Generalized Mundlak Estimator, The Review of Economic Studies, Volume 91, Issue 5, October 2024

Netflix Tech Blog

Netflix Jobs