Targeted Difference-in-Differences Estimation with Staggered Treatment Designs

Yuhao Deng

University of Michigan

Jan 3, 2026

Outline

Background: Difference-in-differences and staggered adoption

Model-free estimands of treatment effects: From group-periods to the overall

Identifiability, estimation, and inference

Simulation

Application to NCEE: Effect of parallel admission on justified envy

Two groups and two periods difference-in-differences

- Potential outcomes: $Y_t(g)$
- Individual treatment effect: $Y_1(1) - Y_1(0)$
- Problem: unmeasured confounding between $Y_1(g)$ and G
- Assuming no anticipation and parallel trends, ATT is identifiable,

$$E\{Y_1(1) - Y_1(0) \mid G = 1\} = E\{\Delta Y_t \mid G = 1\} - E\{\Delta Y_t \mid G = 0\}$$

	$t = 0$	$t = 1$	
$G = 1$	$D_t = 0$	$D_t = 1$	$\Delta Y_t(1) = Y_1 - Y_0$
$G = 0$	$D_t = 0$	$D_t = 0$	$\Delta Y_t(0) = Y_1 - Y_0$

Staggered treatment adoption

- Potential outcomes: $Y_t(g)$
- Individual treatment effect in group g and period t: $Y_t(g) - Y_t(\infty)$

	$t = 0$	$t = 1$	$t = 2$	$t = 3$
$G = 1$	$D_t = 0$	$D_t = 1$	$D_t = 1$	$D_t = 1$
$G = 2$	$D_t = 0$	$D_t = 0$	$D_t = 1$	$D_t = 1$
$G = 3$	$D_t = 0$	$D_t = 0$	$D_t = 0$	$D_t = 1$
$G = \infty$	$D_t = 0$	$D_t = 0$	$D_t = 0$	$D_t = 0$

Staggered treatment adoption

- Question 1: How to define the treatment effects? — Science
- Question 2: How to identify the treatment effects? — Learnability
- Question 3: How to estimate the treatment effects? — Tool

	$t=0$	$t=1$	$t=2$	$t=3$
$G=1$	$D_t=0$	$D_t=1$	$D_t=1$	$D_t=1$
$G=2$	$D_t=0$	$D_t=0$	$D_t=1$	$D_t=1$
$G=3$	$D_t=0$	$D_t=0$	$D_t=0$	$D_t=1$
$G=\infty$	$D_t=0$	$D_t=0$	$D_t=0$	$D_t=0$

Two-way fixed effects model and event study

Two-way fixed effects (TWFE) model

$$Y_t = \alpha D_t + \lambda_t + \gamma_G + \beta X + \varepsilon_t$$

- α is an overall treatment effect
- α is an average of group-period treatment effects with negative weights

Event study

$$Y_t = \sum_{k=0}^{T-1} \alpha_k I(t - G = k) + \lambda_t + \gamma_G + \beta X + \varepsilon_t$$

- α_k is the dynamic treatment effect

Problem: Model-based methods mix up "science" and "tool"

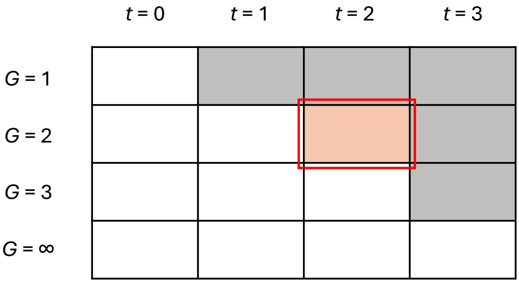
Notations

- $T + 1$ periods: $t = 0, \dots, T$
- Treatment $(D_0 = 0, D_1, \dots, D_T)$, equivalently characterized by $G \in \{1, \dots, T, \infty\}$
- Potential outcome $Y_t(g) := Y_t(d_0 = 0, d_1, \dots, d_T)$, with $g = \min\{t : d_t = 1\}$
- Time-varying covariates X_t
- Observed data: $O = (G, X_0, Y_0, \dots, X_T, Y_T)$
- Sample: $\{O_i : i = 1, \dots, n\}$

Science: Treatment effects (model-free)

- Group-period ATT

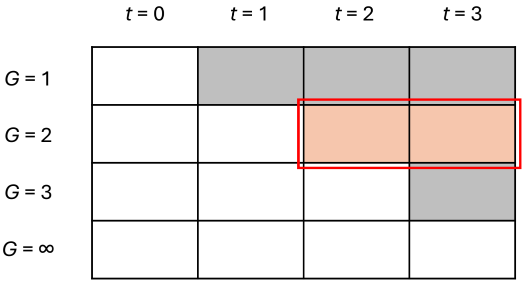
$$\tau_{g,t} = E\{Y_t(g) - Y_t(\infty) \mid G = g\}$$



Science: Aggregating treatment effects

- Groupwise ATT

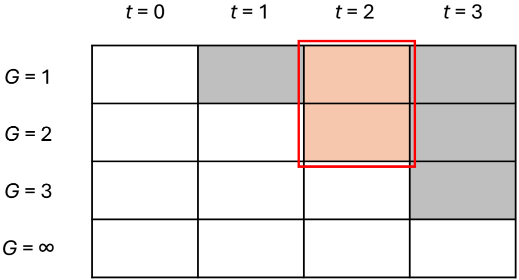
$$\tau_g = \frac{1}{T - g + 1} \sum_{t=g}^{T} \tau_{g,t}$$



Science: Aggregating treatment effects

- Periodwise ATT

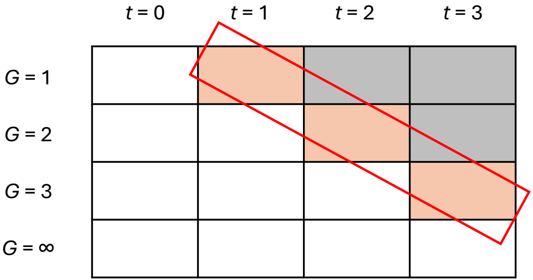
$$\tau_t = \frac{1}{\sum_{g=1}^{t} P(G=g)} \sum_{g=1}^{t} P(G=g)\tau_{g,t}$$



Science: Aggregating treatment effects

- Dynamic ATT

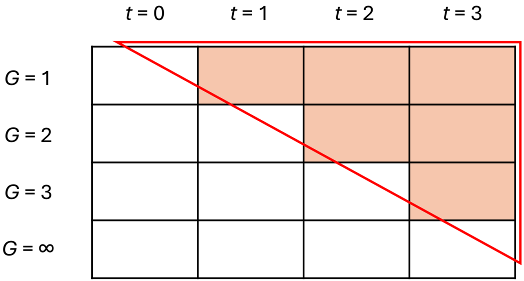
$$\tau_s = \frac{1}{\sum_{t=s+1}^{T} P(G = t - s)} \sum_{t=s+1}^{T} P(G = t - s)\tau_{t-s,t}$$



Science: Aggregating treatment effects

- Overall ATT

$$\tau = \frac{1}{\sum_{g=1}^{T}(T-g+1)P(G=g)} \sum_{g=1}^{T}\sum_{t=g}^{T} P(G=g)\tau_{g,t}$$



Learnability: Assumptions for identifiability

- Assumption 1: No anticipation if unexposed

$$Y_t(g) = Y_t(\infty) \;\text{ for every }\; t < g$$

- Assumption 2: Parallel trends for the counterfactual increase in potential outcomes

$$E\{\Delta Y_t(\infty) \mid X_t, G\} = E\{\Delta Y_t(\infty) \mid X_t\}$$

- Assumption 3: Positivity

$$\eta < P(G = g \mid X_t) < 1 - \eta \;\text{ for some constant }\; \eta > 0$$

- Assumption 4: Consistency

$$Y_t(G) = Y_t$$

An example

- Baseline covariate Z, time-varying covariate Z_t
- Time-invariant unmeasured confounder U
- Causal structural model
$$Y_t(\infty) = f_t(Z, Z_t) + U + \varepsilon_t$$
where ε_t is exogenous random error
- Then
$$E\{\Delta Y_t(\infty) \mid Z, Z_t, Z_{t-1}, G\} = f_t(Z, Z_t) - f_{t-1}(Z, Z_{t-1})$$
does not depend on G
- $X_t = (Z, Z_t, Z_{t-1})$ adjusts the parallel trends
- In general, X_t can include either baseline or history

Identifiability

- Covariates shift from control groups to treated groups
- Under these assumptions,

$$\begin{aligned}\tau_{g,t} &= E\{Y_t(g) - Y_t(\infty) \mid G = g\} \\ &= E(Y_t - Y_{g-1} \mid G = g) - \sum_{k=g}^{t} E\{E(Y_k - Y_{k-1} \mid X_k, G = \infty) \mid G = g\}\end{aligned}$$

- As a result, group-period, groupwise, periodwise, dynamic, and overall ATTs are all identifiable

Estimation based on regression

- Model: conditional change in outcomes $\delta_{g,t}(X_t) = E(\Delta Y_t \mid X_t, G = g)$
- Motivated by the identification formula,

$$\widetilde{\tau}^{\text{id}}_{g,t} = \frac{1}{\mathbb{P}_n\{I(G=g)\}}\mathbb{P}_n\left[I(G=g)\left\{Y_t - Y_{g-1} - \sum_{k=g}^{t}\widehat{\delta}_{\infty,k}(X_k)\right\}\right]$$

- Regression (imputation) estimator

$$\widetilde{\tau}^{\text{reg}}_{g,t} = \frac{1}{\mathbb{P}_n\{I(G=g)\}}\mathbb{P}_n\left[I(G=g)\sum_{k=g}^{t}\{\widehat{\delta}_{G,k}(X_k) - \widehat{\delta}_{\infty,k}(X_k)\}\right]$$

Estimation based on weighting

- Model: propensity score $\pi_{g,t}(X_t) = P(G = g \mid X_t)$
- Weighting estimator using the never-treated group as a reference

$$\widehat{\tau}_{g,t}^{\text{wt,nt}} = \frac{1}{\mathbb{P}_n\{I(G=g)\}}\mathbb{P}_n\left[I(G=g)(Y_t - Y_{g-1}) - \sum_{k=g}^{t}\frac{\pi_{g,k}(X_k)}{\pi_{\infty,k}(X_k)}I(G=\infty)\Delta Y_k\right]$$

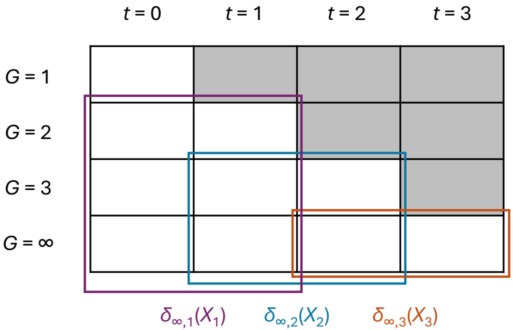
- Weighting estimator using not-yet-treated groups as a reference

$$\widehat{\tau}_{g,t}^{\text{wt,ny}} = \frac{1}{\mathbb{P}_n\{I(G=g)\}}\mathbb{P}_n\left[I(G=g)(Y_t - Y_{g-1}) - \sum_{k=g}^{t}\frac{\pi_{g,k}(X_k)}{\sum_{l>k}\pi_{l,k}(X_k)}I(G>k)\Delta Y_k\right]$$

Motivation to improve efficiency

How to improve efficiency?

- Key 1: Using not-yet-treated groups to determine parallel trends period by period
- Key 2: Using control groups that have smaller variation to determine time trends



Proposed AIVW estimator

- Variance of conditional change $\sigma^2_{g,t}(X_t) = \text{var}(\Delta Y_t \mid X_t, G = g)$.
- Weight by the inverse of variance

$$W_{l,k}(X_k) = \left[\sum_{s>k} \frac{\pi_{s,k}(X_k)}{\sigma^2_{s,k}(X_k)}\right]^{-1} \frac{\pi_{l,k}(X_k)}{\sigma^2_{l,k}(X_k)}$$

- Augmented inverse variance weighting (AIVW) estimator

$$\widehat{\tau}_{g,t} = \frac{1}{\mathbb{P}_n\{I(G=g)\}}\mathbb{P}_n\Bigg[I(G=g)\sum_{k=g}^{t}\{\Delta Y_k - \widehat{\delta}_{\infty,k}(X_k)\} - \sum_{k=g}^{t} I(G>k)\frac{\widehat{\pi}_{g,k}(X_k)}{\widehat{\pi}_{G,k}(X_k)}\widehat{W}_{G,k}(X_k)\{\Delta Y_k - \widehat{\delta}_{\infty,k}(X_k)\}\Bigg]$$

Proposed AIPW estimator

- Homoskedastic working model: $\sigma^2_{g,t}(X_t)$ is a constant
- Augmented inverse probability weighting (AIPW) estimator

$$\widehat{\tau}_{g,t} = \frac{1}{\mathbb{P}_n\{I(G=g)\}}\mathbb{P}_n\Bigg[I(G=g)\sum_{k=g}^{t}\{\Delta Y_k - \widehat{\delta}_{\infty,k}(X_k)\} \\ - \sum_{k=g}^{t} I(G>k)\frac{\widehat{\pi}_{g,k}(X_k)}{\sum_{l>k}\widehat{\pi}_{l,k}(X_k)}\{\Delta Y_k - \widehat{\delta}_{\infty,k}(X_k)\}\Bigg]$$

Proposed estimators for aggregated ATTs

- Aggregating to groupwise, periodwise, dynamic, and overall ATTs

$$\widehat{\tau}_g = \frac{1}{T-g+1}\sum_{t=g}^{T}\widehat{\tau}_{g,t}$$

$$\widehat{\tau}_t = \frac{1}{\mathbb{P}_n\{I(G \leq t)\}}\sum_{g=1}^{t}\mathbb{P}_n\{I(G=g)\}\widehat{\tau}_{g,t}$$

$$\widehat{\tau}_s = \frac{1}{\sum_{t=s+1}^{T}\mathbb{P}_n\{I(G=t-s)\}}\sum_{t=s+1}^{T}\mathbb{P}_n\{I(G=t-s)\}\widehat{\tau}_{t-s,t}$$

$$\widehat{\tau} = \frac{1}{\sum_{g=1}^{T}(T-g+1)\mathbb{P}_n\{I(G=g)\}}\sum_{g=1}^{T}\sum_{t=g}^{T}\mathbb{P}_n\{I(G=g)\}\widehat{\tau}_{g,t}$$

Consistency and double robustness

- Suppose that either the $\{\widehat{\delta}_{\cdot,\cdot}(\cdot)\}$ or $\{\widehat{\pi}_{\cdot,\cdot}(\cdot)\}$ is correctly specified (in L_1-norm)
- In addition, suppose the models belong to a Glivenko–Cantelli class
- Then $\widehat{\tau}_{g,t}$ is consistent for $\tau_{g,t}$ (similar for other ATTs)
- Consistent estimation does not require correct specification of $\sigma^2_{\cdot,\cdot}(\cdot)$
- Double robustness for both AIVW and AIPW

Asymptotic normality

- Suppose that the estimated models $\{\widehat{\delta}_{\cdot,\cdot}(\cdot), \widehat{\pi}_{\cdot,\cdot}(\cdot), \widehat{\sigma}^2_{\cdot,\cdot}(\cdot)\}$ converge to the true value at a rate of $o_p(n^{-1/4})$ (in L_2-norm)
- In addition, suppose the models belong to a Donsker class
- Then $\sqrt{n}(\widehat{\tau}_{g,t} - \tau_{g,t}) \xrightarrow{d} N(0, E\varphi^2_{g,t})$ (similar for other ATTs)
- Influence function of $\tau_{g,t}$ by AIVW

$$
\begin{aligned}
\varphi_{g,t} = \frac{1}{P(G=g)} \Bigg[& I(G=g) \sum_{k=g}^{t} \{\Delta Y_k - \delta_{\infty,k}(X_k)\} - I(G=g)\tau_{g,t} \\
& - \sum_{k=g}^{t} I(G>k) \frac{\pi_{g,k}(X_k)}{\pi_{G,k}(X_k)} W_{G,k}(X_k) \{\Delta Y_k - \delta_{\infty,k}(X_k)\} \Bigg]
\end{aligned}
$$

- Similar asymptotic normality for AIPW

Semiparametric efficiency

When is the AIVW estimator most efficient?

- If $X_t = H_t$ is the entire history (including historical time-varying covariates and observed outcomes), then the AIVW estimator is semiparametrically most efficient
- If ΔY_t is generated based solely on (G, X_t), then the AIVW estimator is semiparametrically most efficient

Estimation for AIPW connected with outcome regression

- Linear model with interaction for outcome regression

$$Y_t = \sum_{k=0}^{T-1} \alpha_k I(t - G = k) + \lambda_t + \gamma_G + \beta_1 X_t + \beta_2 X_t D_t + \beta_3 X_t t + \varepsilon_t$$

- Counterfactual mean outcome under control

$$\mu_{g,t}^0(X_t) = \lambda_t + \gamma_g + \beta_1 X_t + \beta_3 X_t t.$$

- Ordinal logistic regression for propensity score

$$OR_{k,t}(X_t) = \frac{P(G \leq k \mid X_t)}{P(G > k \mid X_t)} = \frac{\sum_{s \leq k} \pi_{s,t}(X_t)}{\sum_{s > k} \pi_{s,t}(X_t)} = \exp(\zeta_{kt0} + \zeta_{kt} X_t),$$

Influence function as weighted sum of residuals

- Define

$$H_{G,t} = D_t - (T - G + 1)I(t = G - 1) \\ - OR_{G,t}(X_t)(T - t + 1)I(G > t) + OR_{G,t+1}(X_{t+1})(T - t)I(G > t + 1)$$

- The influence function of τ is

$$\varphi = \frac{1}{(T+1)P(D_t = 1)} \sum_{t=0}^{T} H_{G,t}\{Y_t - \mu^0_{G,t}(X_t) - D_t\tau\}$$

- Solving $\mathbb{P}_n\widehat{\varphi} = 0$, the AIPW estimator $\widehat{\tau}$ is the average of $\widehat{H}_{G,t}\{Y_t - \widehat{\mu}^0_{G,t}(X_t)\}$ in the sample $\{(i, t) : D_{ti} = 1\}$

Simulation settings

Data generation

- Two baseline covariates Z_1 and Z_2
- A time-varying covariate $Z_{3,t}$
- Potential outcome $Y_t(g)$ generated based on Z_1, Z_2, $Z_{3,t}$, G, plus random error
- $X_t = (Z_1, Z_2, Z_{3,t}, Z_{3,t-1})$

Methods under comparison

- Two-way fixed effects model
- Doubly robust estimators (never-treated, not-yet-treated) by Callway and Sant'Anna (2021)
- Proposed AIPW and AIVW

Simulation results for ATT: Homoskedastic errors

(A)		Scenario 1: Homogeneous effects Homoskedastic error terms					Scenario 2: Heterogeneous effects Homoskedastic error terms				
n		TWFE	DRnt	DRny	AIPW	AIVW	TWFE	DRnt	DRny	AIPW	AIVW
100	Bias	0.187	0.005	0.007	0.005	0.006	-0.265	-0.186	-0.185	-0.006	-0.003
	SD	0.229	0.303	0.290	0.256	0.259	0.237	0.314	0.300	0.264	0.268
	SE	0.269	0.291	0.279	0.253	0.255	0.284	0.301	0.290	0.263	0.265
	CP	0.930	0.939	0.935	0.944	0.942	0.889	0.881	0.887	0.942	0.937
500	Bias	0.181	0.004	0.004	0.001	0.001	-0.241	-0.154	-0.154	0.021	0.021
	SD	0.105	0.123	0.122	0.110	0.110	0.109	0.128	0.127	0.114	0.115
	SE	0.121	0.125	0.123	0.111	0.111	0.127	0.131	0.128	0.116	0.116
	CP	0.702	0.952	0.946	0.964	0.961	0.516	0.782	0.783	0.952	0.953
2000	Bias	0.184	0.004	0.004	0.004	0.004	-0.274	-0.189	-0.189	-0.011	-0.010
	SD	0.055	0.062	0.062	0.056	0.056	0.057	0.065	0.064	0.057	0.057
	SE	0.060	0.062	0.061	0.056	0.056	0.063	0.065	0.064	0.058	0.058
	CP	0.121	0.948	0.946	0.946	0.949	0.006	0.171	0.161	0.942	0.945

Simulation results for ATT: Heteroskedastic errors

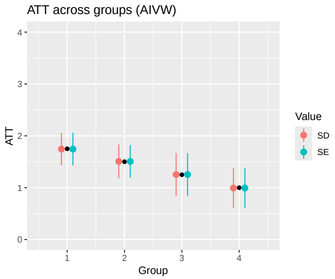
(B)		Scenario 3: Homogeneous effects Heteroskedastic error terms					Scenario 4: Heterogeneous effects Heteroskedastic error terms				
n		TWFE	DRnt	DRny	AIPW	AIVW	TWFE	DRnt	DRny	AIPW	AIVW
100	Bias	0.184	-0.001	0.000	-0.001	-0.000	-0.268	-0.192	-0.191	-0.011	-0.010
	SD	0.221	0.253	0.257	0.237	0.227	0.230	0.265	0.269	0.245	0.236
	SE	0.277	0.242	0.245	0.235	0.223	0.291	0.254	0.258	0.245	0.234
	CP	0.952	0.931	0.933	0.939	0.939	0.908	0.855	0.860	0.939	0.943
500	Bias	0.181	0.003	0.003	-0.001	-0.001	-0.241	-0.156	-0.155	0.019	0.019
	SD	0.101	0.110	0.112	0.103	0.097	0.105	0.116	0.118	0.108	0.102
	SE	0.124	0.110	0.112	0.104	0.099	0.130	0.115	0.117	0.109	0.104
	CP	0.721	0.946	0.940	0.955	0.953	0.553	0.732	0.737	0.951	0.952
2000	Bias	0.184	0.003	0.003	0.003	0.003	-0.274	-0.190	-0.190	-0.012	-0.012
	SD	0.053	0.055	0.057	0.052	0.049	0.055	0.058	0.060	0.053	0.051
	SE	0.062	0.055	0.056	0.052	0.049	0.065	0.058	0.059	0.054	0.052
	CP	0.126	0.956	0.953	0.951	0.946	0.003	0.096	0.107	0.950	0.958

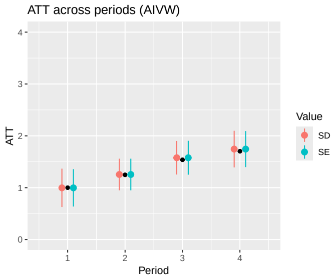
Simulation results for ATT: Cumulative errors

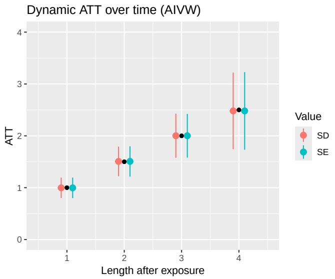
(C)		Scenario 5: Homogeneous effects Cumulative error terms					Scenario 6: Heterogeneous effects Cumulative error terms				
n		TWFE	DRnt	DRny	AIPW	AIVW	TWFE	DRnt	DRny	AIPW	AIVW
100	Bias	0.183	-0.018	-0.018	-0.011	-0.010	-0.269	-0.210	-0.210	-0.021	-0.020
	SD	0.241	0.316	0.305	0.286	0.284	0.249	0.327	0.316	0.293	0.292
	SE	0.321	0.282	0.277	0.278	0.275	0.333	0.293	0.288	0.287	0.284
	CP	0.962	0.908	0.915	0.936	0.939	0.942	0.840	0.842	0.943	0.936
500	Bias	0.180	-0.001	-0.001	-0.001	-0.002	-0.243	-0.160	-0.160	0.018	0.018
	SD	0.107	0.130	0.127	0.122	0.120	0.110	0.135	0.132	0.125	0.124
	SE	0.145	0.131	0.128	0.125	0.124	0.150	0.135	0.133	0.129	0.128
	CP	0.833	0.946	0.946	0.958	0.957	0.687	0.770	0.777	0.953	0.953
2000	Bias	0.183	0.003	0.003	0.003	0.003	-0.275	-0.191	-0.191	-0.012	-0.012
	SD	0.055	0.067	0.065	0.062	0.062	0.056	0.070	0.068	0.064	0.063
	SE	0.073	0.065	0.064	0.063	0.062	0.075	0.068	0.066	0.065	0.064
	CP	0.234	0.945	0.944	0.951	0.955	0.017	0.206	0.175	0.950	0.949

Simulation results for other ATTs

- Scenario 4: heterogeneous effects, heteroskedastic errors (AIVW)

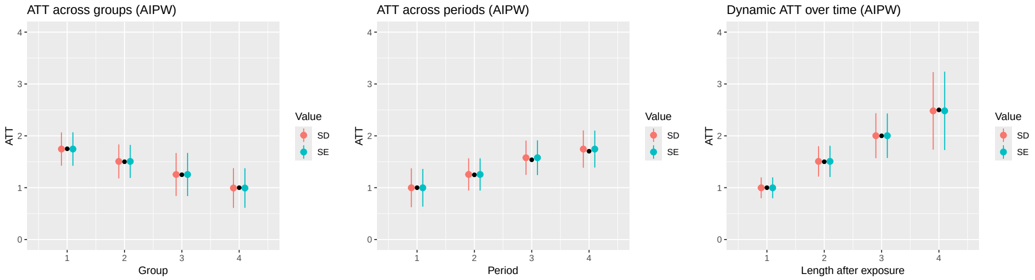






Simulation results for other ATTs

- Scenario 4: heterogeneous effects, heteroskedastic errors (AIPW)



Data: National College Entrance Examination (Gaokao)

Staggered adoption of the parallel mechanism across provinces

- In 2007, all 27 provinces used immediate admission ($n = 27 \times 2$, stem and non-stem)
- In 2008, 3 provinces switched to parallel admission ($G = 1$)
- In 2009, 10 provinces switched to parallel admission ($G = 2$)
- In 2010, 6 provinces switched to parallel admission ($G = 3$)
- In 2011, 2 provinces switched to parallel admission ($G = 4$)
- 6 provinces had not been reformed by 2011 ($G = \infty$)

Covariates

- log GDP per capita, population, track (stem or non-stem)
- Game size, when to submit preference (after the exam, after knowing the score)

Justified envy

Justified envy

- Student i justifiably envies student j for school s if i would rather be assigned to school s, where some student j, who has a lower priority (i.e., lower score) than i, is assigned
- In this case, student i is a blocking student; student i and school s are a blocking pair; student i, j, and school s are a blocking triplet

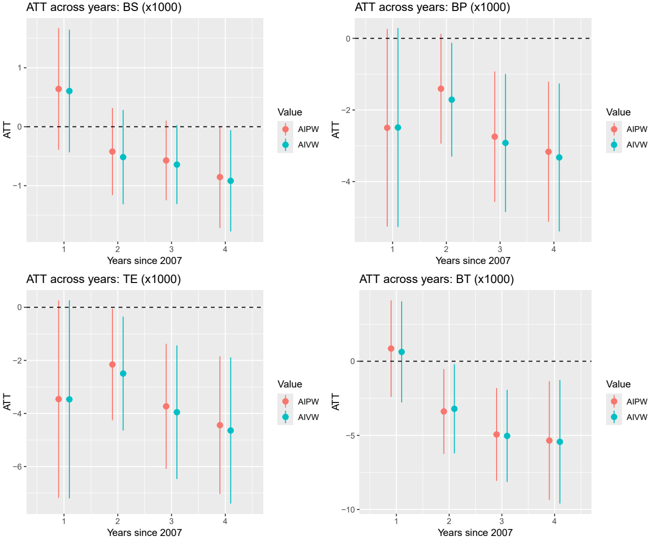
Four measures of justified envy

- Blocking students (BS)
- Blocking pairs (BP)
- Total tridimensional envy of blocking triplets (TE)
- Blocking triplets (BT)

Effect of parallel admission on justified envy

Outcome	BS ($\times$1000)			BP ($\times$1000)		
	ATT	(SE)	P	ATT	(SE)	P
TWFE	-0.739	(0.276)	0.008**	-2.460	(0.833)	0.003**
DRnt	-1.512	(0.542)	0.005**	-0.651	(0.728)	0.371
DRny	-1.313	(0.545)	0.016*	-1.542	(0.698)	0.027*
AIPW	-0.578	(0.321)	0.072	-2.580	(0.829)	0.002**
AIVW	-0.649	(0.322)	0.044*	-2.769	(0.882)	0.002**
Outcome	**TE ($\times$1000)**			**BT ($\times$1000)**		
	ATT	(SE)	P	ATT	(SE)	P
TWFE	-3.513	(1.101)	0.002**	-4.510	(1.158)	0.000***
DRnt	-1.580	(0.941)	0.093	-8.826	(2.510)	0.000***
DRny	-2.511	(0.908)	0.006**	-7.170	(2.041)	0.000***
AIPW	-3.617	(1.096)	0.001**	-4.422	(1.474)	0.003**
AIVW	-3.847	(1.175)	0.001**	-4.458	(1.497)	0.003**

Periodwise effect of parallel admission on justified envy (AIVW)



Concluding remarks: Contribution

Science

- We define treatment effects that are model-free
- Group-period ATTs are aggregated to groupwise, periodwise, dynamic, and overall ATTs

Learnability

- We allow time-varying covariates to adjust for parallel trends

Tool

- We propose doubly robust AIVW and AIPW estimators for ATTs
- Estimated by the empirical average of weighted residuals on the target population
- Semiparametrically efficient under specific cases

Concluding remarks: Limitation and future research

- AIVW is not universally efficient
- AIVW has higher finite-sample variability due to more fitted models involved
- AIPW is computationally simpler, and has comparative performance to AIVW

- Over-identification of ATTs—parallel trends are not necessary for identifiability
- How to estimate ATTs efficiently
- Targeted minimum-loss-based estimation (TMLE) to improve finite-sample performance

Acknowledgements

Credits to

- Le Kang (Nanjing University)
- Haoyu Wei (University of California, San Diego)
- Yixiao Sun (University of California, San Diego)