One-stage robust difference-in-differences regression

John Gardner
University of Mississippi

Midwestern Econometrics Group
Nov, 2024

Introduction

- DD estimates based on TWFE regressions recover difficult-to-interpret measures of average treatment effects when adoption is staggered and treatment effects are heterogeneous (de Chaisemartin and Haultfœuille, 2020; Sun Abraham, 2021; Goodman-Bacon, 2022; Borusyak, Jaravel and Spiess, 2024)
- Several robust alternative estimators have been developed (Borusyak, et al., 2021; Callaway and Sant'Anna, 2021; de Chaisemartin and D'Haultfœuille, 2020; Dube, Jordà and Taylor, 2023; Gardner, 2021; Gardner, Thakral, Tô, and Yap, 2023; Liu, Wang and Xu, 2023; Sun and Abraham, 2021; Wooldridge, 2021; Deb, Norton, Wooldridge, and Zabel, 2024)

- All of these robust estimators are multi-stage
  - Some require estimating group×period specific treatment effects, then aggregating them
  - All require standard-error adjustments, and most can only be implemented using specialized software
- I develop an alternative approach that produces robust DD estimates, along with approximately valid standard errors, from a single regression
  - Motivated by the literature on matching and selection on observables
  - Extends to event-study analyses and testing parallel trends
- Another DD estimator? No, the point estimates are numerically identical to the two-stage difference in differences, imputation, or fixed-effects counterfactual estimators (which themselves are numerically identical, Gardner, 2021; Gardner et al., 2023; Borusyak et al. 2021; Liu, Wang and Xu, 2023)
  - In other words, this is a different way of obtaining widely used existing estimators (cf. Deb et al., 2024)

Setup

- We observe the random panel $\{Y_{it}, D_{it}, X_{it}\}$ for $t = 1, \ldots, T$ and $i = 1, \ldots, N$
- The treatment is irreversible, unanticipated, and does not affect the covariates
- $C_i \in \mathcal{C} = \{2, \ldots, T, \infty\}$ denotes i's treatment cohort and $C_i^j \in \{0, 1\}$ is an indicator for whether i belongs to treatment cohort $j \in \mathcal{C}$ (TEs not identified for first cohort, "∞" means never treated)
- $\{Y_{it}^0(x), Y_{it}^1(x)\}$ denote the counterfactual outcomes that i would receive in t given covariates x, conditional on their observed cohort
  - $\beta_{it}(x) = Y_{it}^1(x) - Y_{it}^0(x)$ is the effect of the treatment for i
  - $\beta_{ct}(x) = E[\beta_{it}(x)|C_i, x]$ is the cohort-time-covariate ATT
- Untreated outcomes satisfy parallel trends:

$$E(Y_{i,t+1}^0 - Y_{it}^0|X, C^j) = E(Y_{i,t+1}^0 - Y_{it}^0|X, C^k) = \Delta\gamma_t + \Delta X_{it}'\delta$$

for all $j, k \in \mathcal{C}$

Review of 2SDD

- The overall average ATT $\beta = E[\beta_{ct}(X_{it})|D_{it} = 1]$ (over cohorts, time and covariates) can be estimated by:
  1. Estimating the model

$$Y_{it} = \lambda_c + \gamma_t + X_{it}'\delta + \varepsilon_{it}$$

     in the sample of untreated observations
  2. Regressing adjusted outcomes $Y_{it} - \hat{\lambda}_c - \hat{\gamma}_t - X_{it}'\hat{\delta}$ on treatment status Details
- Extends to individual fixed effects, event studies, and tests of parallel trends
- Standard errors can be adjusted for estimation of first-stage parameters (Kyle Butts' `did2s` packages handle this automagically)

Proposition

$\hat{\beta}^{2SDD}$ is consistent and asymptotically normal.

A robust one-stage approach

- Motivation comes from literature on cross-sectional matching and selection on observables
- Temporarily forget the DD context, and suppose that $(Y_0, Y_1) \perp\!\!\!\perp D|X$ and $E(Y_d|X) = X'\delta_d$, $d \in \{0, 1\}$
- Counterfactual mean outcomes functions can be estimated from treatment-status-specific regressions, or from the pooled regression

$$Y = X'\delta_0 + D \cdot X'\delta_1 + q$$

- The ATT is identified as

$$E[E(Y_1|X) - E(Y_0|X)|D = 1] = E(X|D = 1)'\delta_1$$

- The "aggregation" step can be avoided by estimating ATT as $\hat{\rho}_1$ from the regression (replacing $E(X|D = 1)$ with its sample analog; Wooldridge, 2010)

$$Y = X'\rho_0 + D[X - E(X|D = 1)]'\rho_1 + \beta D + r$$

Extending this approach to DD: Challenge #1

- Traditional DD methods regress outcomes on treatment status, "controlling" for cohort and time FEs
- If cohort and time were quasiexperimentally manipulable, we could estimate the ATT by matching treated to untreated units in the same cohort at the same time (*but they aren't*)
- *All* DD methods overcome this challenge using a parallel trends assumption, which allows us to extrapolate counterfactual untreated outcomes from untreated to treated units
- Unlike selection-on-observables contexts, *extrapolation is desirable* for DD

Extending this approach to DD: Challenge #2

- Let W_{it} be a vector of cohort indicators, time indicators, and time-varying controls
- Under parallel trends,

$$E(Y_{it}^0|W_{it}) = \lambda_c + \gamma_t + X_{it}'\delta \equiv W_{it}'\rho_0$$

and

$$E(Y_{it}^1|W_{it}) = \lambda_c + \gamma_t + X_{it}'\delta + \beta_{ct}(X_{it}) \equiv W_{it}'\rho_0 + \beta_{ct}(X_{it}).$$

- If treatment effects vary at the cohort$\times$time level, *treated outcomes are not linear in cohort and time*
- Take the linear projection of $\beta_{ct}(X_{it})$ onto W_{it} in the treated population:

$$\beta_{ct}(X_{it}) = \beta_c + \beta_t + X_{it}'\beta_x + \tilde{\beta}_{ct} = W_{it}'\rho_1 + \tilde{\beta}_{ct},$$

where $E(\tilde{\beta}_{ct}|D_{it} = 1) = 0$ by definition

- Now,

$$E[(Y_{it}^1 - Y_{it}^0|W_{it})|D_{it} = 1] = E(\beta_c + \beta_t + X_{it}'\beta_x + \tilde{\beta_{ct}}|D_{it} = 1)$$
$$= E(W_{it}|D_{it} = 1)'\rho_1$$

- Hence, we could estimate the overall ATT (β) by estimating the regression

$$Y_{it} = W_{it}'\rho_0 + D_{it}W_{it}'\rho_1 + s_{it}$$

and taking the average $\bar{W}^{1'}\hat{\rho}_1$, where $\bar{W}^1 = (\sum_{it} D_{it}W_{it})/\sum_{it} D_{it}$

- Or we could directly estimate it as $\hat{\beta}$ from the specification

$$Y_{it} = W_{it}'\rho_0 + D_{it}(W_{it} - \bar{W}^1)'\rho_2 + \beta D_{it} + r_{it}$$

1SDD and its properties

- To summarize, the overall ATT can be estimated by regression outcomes on
  1. Cohort and time-period indicators, as well as any time-varying controls (W_{it})
  2. Interactions between treatment status and deviations in cohort indicators, time indicators, and time-varying controls from their means among treated units $[D_{it}(W_{it} - \bar{W}^1)]$
  3. Treatment status (D_{it})
- Courtesy of the Frisch-Waugh-Lovell theorem, we also have:

Proposition

$\hat{\beta}^{1SDD} = \hat{\beta}^{2SDD}$.

- The equivalence result can be used to prove consistency (which can also be obtained directly by appealing to the preceding motivation)
- The regression can be implemented without any specialized software
- Will automatically produce approximately valid standard errors (up to the sampling error in $\bar{W}^1$)
- *A note on individual FEs*: The equivalence result still holds if cohort FEs are replaced with unit FEs, but clustered standard errors will be mechanically biased (because OLS FOCs require each unit's residuals to sum to zero in this case)

Dynamic and placebo effects

- Let $\{D_{it}^r\}$ be r-period lags (if $r \geq 0$) or leads (if $r < 0$) of treatment adoption, for all $r \geq k$, where $k \leq 0$
- Consider a regression of outcomes on
  1. Cohort indicators, time indicators, and time-varying controls (W_{it}),
  2. Interactions between treatment status and deviations in cohort indicators, time indicators, and covariates from *duration-specific* means [$D_{it}(W_{it} - \bar{W}_{it}^r)$ for all $r \geq k$]
  3. Leads and lags of treatment status (D_{it}^r, $r \geq k$)
- If $k = 0$, the coefficients on the D_{it}^r identify the r-period ATTs
- If $k < 0$, the coefficients on the D_{it}^r for $r < 0$ identify r-period *placebo ATTs* which can be used to test the validity of parallel trends (for $r \geq 0$, they still identify r-period ATTs)
- Can identify group- and time-averaged effects similarly

Simulations

- I first simulate a setting where the treatment effect is zero, so that

$$Y_{it} = \lambda_i + \gamma_t + \varepsilon_{it},$$

  with $\lambda_i \sim N(C_i, 1)$, $\gamma_t \sim N(0, 1)$, $\varepsilon_{it} \sim N(0, 3)$, $T = 5$, and the C_i are drawn from a discrete uniform distribution
- I estimate the treatment effect using 1SDD and 2SDD, as well as by "manually" aggregating $\bar{W}^{1\prime}\hat{\rho}_1$ (which is numerically equivalent)
- I draw samples of size $N \in \{50, 100, 500\}$ for 1K simulations, and record the rejection rates
- To illustrate the effect of the sampling error in using $\bar{W}^1$ for 1SDD, I repeat this exercise in a fixed design where cohort membership is fixed in repeated samples
- The rejection rates for 1SDD and 2SDD are similar, even in smaller samples, regardless of whether the design is random

		Random			Fixed		
	N	50	100	500	50	100	500
2SDD	D	0.074	0.059	0.05	0.08	0.053	0.051
	D^1	0.05	0.058	0.047	0.059	0.062	0.063
	D^2	0.068	0.058	0.046	0.075	0.043	0.049
	D^3	0.072	0.071	0.047	0.064	0.05	0.05
	D^4	0.099	0.071	0.043	0.095	0.074	0.048
1SDD	D	0.069	0.059	0.05	0.068	0.053	0.05
	D^1	0.051	0.058	0.047	0.056	0.058	0.064
	D^2	0.062	0.057	0.047	0.069	0.046	0.049
	D^3	0.066	0.069	0.047	0.061	0.049	0.05
	D^4	0.085	0.064	0.042	0.077	0.067	0.048
Manual	D	0.069	0.059	0.05	0.068	0.053	0.05

- I also simulate settings where 500 units are organized into 50 states, and λ_i, γ_t, and ε_{it} are drawn as before, and

$$Y_{it} = \lambda_i + \gamma_t + \delta X_{it} + \beta_{it} D_{it} + \varepsilon_{it}$$

- I run four simulations:
  1. TE independent of covariate: $X_{it} \sim N(1, 1)$ and $\beta_{it} \sim N(2, 1)$
  2. TE depends on covariate linearly: $X_{it} \sim N(1, 1)$ and $\beta_{it} = t - C_i + 1 + X_{it}/4$
  3. TE depends on covariate multiplicatively: $X_{it} \sim N(\lambda_i/25, 1)$ and $\beta_{it} = (t - C_i + 1) \cdot X_{it}/4$
  4. Covariates correlated within cohorts: $X_{it} \sim N(C_i/6, 1)$ and $\beta_{it} = (t - C_i + 1) \cdot X_{it}/4$

- I estimate the treatment effect several ways:
  - 2SDD, 2SDD using state-average covariates, 2SDD collapsing to state average outcomes and covariates
  - 1SDD, using these same variations
- Here, there is more variation within and between 1SDD and 2SDD (both tend to perform better when the data are collapsed to the state level)

	(1)	(2)	(3)	(4)
2SDD	1.9997	2.2421	.0538	.2281
	[0.034]	[0.013]	[0.068]	[0.063]
2SDD, avg. X	1.9997	2.2421	.0539	.2282
	[0.036]	[0.011]	[0.064]	[0.061]
2SDD, avg.	1.9997	2.2421	.0539	.2282
	[0.036]	[0.011]	[0.064]	[0.061]
1SDD	1.9997	2.2421	.0538	.2281
	[0.039]	[0.076]	[0.077]	[0.074]
1SDD, avg. X	1.9997	2.2421	.0539	.2282
	[0.046]	[0.076]	[0.073]	[0.074]
1SDD, avg.	1.9997	2.2421	.0539	0.2282
	[0.034]	[0.064]	[0.064]	[0.065]
ATT	2.0	2.2424	.0542	0.2282

Empirical application

- I also revisit Cheng and Hoekstra's (2013) analysis of the effects of "stand your ground" laws on violent crime (log violent crimes per 10,000 people)
- I estimate three specifications: Overall, dynamic, and dynamic with three leads (for placebo testing)
  - For 2SDD, I also implement an alternative test of PT (using all untreated obs. in first stage and including leads in second stage)
- The 1SDD SEs tend to be slightly larger (although not in every case), although this doesn't change the practical conclusions of the exercise
- The tests of parallel trends also agree (as do estimates in models with and without placebos)

		One stage			Two stage		
		(1)	(2)	(3)	(4)	(5)	(6)
Overall ATT	D	0.0706**			0.0706**		
		(0.0279)			(0.0347)		
Dynamic effects	D^1		0.0852***	0.0811**		0.0852***	0.0811**
			(0.0267)	(0.0397)		(0.0291)	(0.0406)
	D^2		0.0727**	0.0657		0.0727*	0.0657
			(0.0304)	(0.0401)		(0.0386)	(0.0454)
	D^3		0.0658*	0.0497		0.0658	0.0497
			(0.0363)	(0.0456)		(0.0450)	(0.0520)
	D^4		0.0373	0.0145		0.0373	0.0145
			(0.0397)	(0.0478)		(0.0502)	(0.0579)
	D^5		0.126***	0.105**		0.126***	0.105**
			(0.0465)	(0.0463)		(0.0447)	(0.0437)
	N	550	550	550	550	550	550

		One stage			Two stage		
		(1)	(2)	(3)	(4)	(5)	(6)
Placebo effects	D^0			-0.000905		0.00841	-0.000905
				(0.0360)		(0.0181)	(0.0378)
	D^{-1}			-0.0533		-0.0356**	-0.0533
				(0.0335)		(0.0176)	(0.0351)
	D^{-2}			-0.0165		0.00493	-0.0165
				(0.0333)		(0.0146)	(0.0320)
	D^{-3}			-0.00194		0.0176	-0.00194
				(0.0256)		(0.0185)	(0.0292)
	D^{-4}					-0.0174	
						(0.0201)	
	D^{-5}					0.0263	
						(0.0169)	
	D^{-6}					0.0464**	
						(0.0184)	
	D^{-7}					-0.0605	
						(0.0369)	
	D^{-8}					-0.153***	
						(0.0370)	
	D^{-9}					-0.252***	
						(0.0268)	
	N	550	550	550	550	550	550

Conclusion

- 1SDD is a simple alternative way to obtain 2SDD (aka imputation or FEct) estimates
- Motivated by regression methods for matching under selection on observables
- Produces approximately valid SEs with standard statistical software
- Performs well in simulations and empirical application

2SDD details

Back

- This implies that outcomes satisfy

$$Y_{it} = \lambda_c + \gamma_t + X_{it}'\delta + \beta_{ct}(X_{it})D_{it} + u_{it},$$

  where $E(u_{it}|C_i, X_{it}) = 0$
- 2SDD is based on the realization that

$$\begin{aligned} Y_{it} - \lambda_c - \gamma_t - X_{it}'\gamma &= \beta_{ct}(X_{it})D_{it} + u_{it} \\ &= \beta D_{it} + [\beta_{ct}(X_{it}) - \beta]D_{it} + u_{it}, \end{aligned}$$

  where
  - $\beta = E[\beta_{ct}(X_{it})|D_{it} = 1]$ is the overall average ATT (over cohorts, time, and covariates), and
  - $E\{[\beta_{ct}(X_{it}) - \beta]D_{it}|D_{it}\} = D_{it}E[\beta_{ct}(X_{it})|D_{it}] - \beta D_{it} = 0$