One-stage robust difference-in-differences regression

John Gardner*

This version: 20 March, 2024

Previous versions: 11 Jan, 2023, 1 Dec., 2023

Abstract

I develop a simple method of combining regression with difference-in-difference and event-study designs to obtain treatment effect estimates that are robust to the presence of average treatment-effect heterogeneity under staggered rollout. The resulting estimator, which can be viewed as a form of matching via regression adjustment, can be obtained via a single regression, which automatically produces approximately valid asymptotic standard errors. This one-stage estimator is numerically equivalent to the two-stage difference-in-differences estimator developed in Gardner (2021) and Gardner, Thakral, Tô, and Yap (2023), which is also the same as the estimators developed in Borusyak, Jaravel and Spiess (2021) and Liu, Wang and Xu (2023), and therefore inherits the robustness (and other) properties of those estimators. The estimator can also be extended to identify duration-specific and other treatment-effect measures and implement placebo tests of parallel trends. I illustrate the properties and application of this approach using simulations and an application from the literature.

Keywords: Differences-in-differences, event-studies, treatment effects, treatment-effect heterogeneity, matching, regression, causal inference.

JEL codes: C01, C10, C21, C22, C23.

*Department of Economics, University of Mississippi, jrgardne@olemiss.edu.

1 Introduction

It is now widely known that difference-in-differences and event-study estimates based on traditional two-way-fixed-effects regression specifications do not always identify sensible measures of average treatment effects when the adoption of a treatment is staggered over time and duration-specific average treatment effects are heterogeneous across treatment cohorts (see Borusyak, Jaravel and Spiess, 2021; de Chaisemartin and D'Haultfœuille, 2020; Goodman-Bacon, 2022; Sun and Abraham, 2021). These observations have spawned a proliferation of alternative estimators that are robust to the problems facing traditional regression specifications in the staggered and heterogeneous setting (see, e.g., Borusyak, Jaravel and Spiess, 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).

In this paper, I develop a new approach to robust identification of average treatment effects using difference-in-difference designs in this setting. A key advantage of this new approach is that only requires the estimation of a single regression, which automatically produces approximately valid asymptotic standard errors. Moreover, this new approach does not really add to the growing list of robust estimators. Instead, point estimates obtained using this approach are identical to those from the two-stage difference-in-differences estimator developed in Gardner (2021) and discussed in greater detail in Gardner, Thakral, Tô, and Yap (2023), which is also the same as the imputation estimator developed in Borusyak, Jaravel and Spiess (2021) and the fixed-effects counterfactual estimator developed in Liu, Wang and Xu (2023). As a consequence, the one-stage estimator developed in this paper automatically inherits the advantages of these estimators, including robustness to treatment effect heterogeneity under staggered adoption, efficiency (under homoskedasticity, see Borusyak, Jaravel and Spiess, 2021), the ability to control for time-varying covariates (that evolve exogeneously, see Caetano, Callaway, Payne and Rodrigues, 2022), and arbitrary dependence of treatment effects on those covariates.

This one-stage approach to estimation developed in this paper is flexible. In addition to the overall average effect of the treatment on the treated, variations on the basic one-stage regression specification can be used to identify other objects of interest, including dynamic treatment effects, cohort- and time-specific average treatment effects, and coefficients that represent placebo tests of parallel trends. Furthermore, since it ultimately amounts to estimating a single regression (or perhaps a series of regressions), the one-stage approach is also fast, and can be implemented by anyone using any standard statistical software package, without the assistance of a specialized estimation routine.

The one-stage robust difference-in-differences regression approach can motivated intuitively by analogy to matching and regression methods for identification under selection on observables. Abstracting away from covariates, if it were possible to observe the same unit in the same time period in both treated and untreated states, the average effect of the treatment on the the treated could be identified nonparametrically by matching treated units to untreated versions of themselves in every time period. If, an addition, counterfactual mean outcomes were linear in unit and time indicators, this matching could also be implemented by estimating treatment-status-specific regressions of outcomes onto those indicators, then averaging the differences between predicted treated and untreated outcomes over the treated population. In fact, since these regressions would allow for extrapolation between units in their treated and untreated states, this regression approach to matching could be implemented even if individuals were never observed in both treatment states at the same time.

All difference-in-difference methods rely on some type of extrapolation under a parallel trends assumption, using a combination of outcomes for contemporaneously untreated observations and past values of treated observations as counterfactuals for treated observations. While parallel trends implies that untreated outcomes are linear in unit and time indicators, arbitrarily heterogeneous treatment effects may not be. However, treated outcomes can always be decomposed into unit- and time-specific mean components, where the remaining variation has mean zero. The average effect of the treatment on the treated is therefore

identified from treatment-status-specific regressions of outcomes onto unit and time indicators. As I show below, the averaging step of this identification procedure can be avoided by judicious choice of regression specification, allowing the average effect of the treatment on the treated to be identified from a single regression.

In Section 2, below, I outline the environment in which the one-stage regression-matching difference-in-differences estimator developed in this paper applies. Because I formally establish the consistency of this estimator by showing its equivalence to the two-stage difference-in-differences estimator, I also briefly review the properties of the latter estimator in that section. Refining the intuition presented above, I develop the one-stage robust difference-in-difference regression estimator in Section 3. There, I also I show that the regression-matching estimator is equivalent to two-stage differences in differences, and introduce several useful variations on the methodology. In Section 4, I present evidence on the performance of the one-stage estimator using Monte Carlo simulations. In Section 5, I illustrate the use of the estimator, and compare it to the two-stage estimator, in the context of an applied example from the literature. I offer some concluding remarks in Section 6.

2 Setup, and review of the two-stage approach

Suppose that the data consist of observations on outcomes Y_{it}, treatment status $D_{it} \in \{0, 1\}$, and a set of time-varying control variables X_{it} for $i = 1, \ldots, N$ units and $t = 1, \ldots, T$ time periods. Further suppose that the treatment is irreversible and unanticipated.[1]

Let $T_i \in \{2, \ldots, T, \infty\}$ be the date at which unit i adopts the treatment, and set $T_i = \infty$ if i is always treated during that period (drop any observations that are always treated during the sample period, since no treatment effects are identified for always-treated units). Also define cohort dummies $C_i^j = 1(T_i = j)$, $j \in \mathcal{C} = \{2, \ldots, T, \infty\}$. Let (Y_{it}^0, Y_{it}^1) be the counterfactual untreated and treated outcomes that i would experience at time t, conditional

[1] It is possible to allow for potential anticipation by redefining D_{it} to be an indicator for whether unit i adopts the treatment in k periods after t.

on their observed membership in treatment cohort T_i. Let $\beta_{it} = Y_{it}^1 - Y_{it}^0$ be the time-t causal effect of the treatment for unit i, once again conditional on i's observed treatment cohort. I assume for simplicity that the data $\{Y_{it}, X_{it}, C_i^j\}$, $i = 1, \ldots, N$, $t = 1, \ldots, T$, $j \in \mathcal{C}$, consist of a random panel, although all of the results in this paper also apply to repeated-cross-sectional data.[2] Finally, suppose that counterfactual outcomes follow parallel trends, in the sense that they can be expressed as

$$Y_{it}^d = \lambda_i + \gamma_t + X_{it}'\delta + \beta_{it}d + u_{it}, \quad d \in \{0, 1\}, \tag{1}$$

where $E(u_{it}|D_{is}, X_{is}, W_{is}) = 0$ for all $s \in \{1, \ldots, T\}$, and W_{is} is a vector of unit and time-period indicators.

The two-stage difference-in-differences estimator is based on the implication of parallel trends that

$$Y_{it} - \lambda_i - \gamma_t - X_{it}'\gamma = \beta_{it}D_{it} + u_{it} = \beta D_{it} + (\beta_{it} - \beta)D_{it} + u_{it} \equiv \beta D_{it} + \varepsilon_{it},$$

where $\beta = E(\beta_{it}|D_{it} = 1)$ is the overall ATT (the population average treatment effect over all time periods) and $E[(\beta_{it} - \beta)D_{it}|D_{it}] = D_{it}E(\beta_{it}|D_{it}) - \beta D_{it} = 0$. Thus, if λ_i, γ_t, and δ were known, the overall ATT could be estimated from a regression of adjusted outcomes $Y_{it} - \lambda_i - \gamma_t - X'\delta$ on treatment status D_{it}. Although they are not known, as long as (i) there are untreated observations in every period, and (ii) there are pre-treatment observations for every eventually-treated unit, λ_i, γ_t, and δ can be estimated from a regression of outcomes on unit fixed effects, time fixed effects, and time-varying controls using the sample of untreated observations (although this regression does not consistently estimate the unit fixed effects, those effects can be eliminated by a variation on the usual within transformation; see Appendix A).[3] The two-stage difference-in-differences estimate $\hat{\beta}^{2SDD}$

[2] Since all of the estimands discussed in this paper are conditional on $D_{it} = 1$, causal effects for members of the never-treated cohort can be normalized to zero without loss of generality.

[3] Since this regression is estimated on the sample of untreated observations, it is not subject to the biases

of the overall ATT β is the estimated coefficient on treatment status from a regression of $Y_{it} - \hat{\lambda}_i - \hat{\gamma}_t - X'_{it}\hat{\delta}$ on D_{it}, where $\hat{\lambda}_i$, $\hat{\gamma}_t$, $\hat{\delta}$ are estimates of those parameters obtained from a first-stage regression of Y_{it} on W_{it} and X_{it} in the sample of untreated observations.

Proposition 1. *Suppose that (i)* $E\left[\sum_t (1 - D_{it})\ddot{X}^0_{it}\ddot{X}^{0\prime}_{it}\right]$ *is invertible, where* $\ddot{X}^0_{it}$ *denote the vector of deviations in the covariates from their means among all untreated observations, and (ii)* $E\left(\sum_t D_{it}\right) \neq 0$*. Then under parallel trends,* $\hat{\beta}^{2SDD} \xrightarrow{p} \beta$ *and* $\sqrt{N}(\hat{\beta}^{2SDD} - \beta) \overset{a}{\sim} N\left(0, A_0^{-1}B_0A_0^{-1}\right)$*, where* A_0 *and* B_0 *are defined in Appendix A.*

The proof is given in Appendix A.[4]

3 A robust one-stage regression approach

3.1 Motivation

The motivation for the estimator that I develop in this paper comes from the literature on matching and selection on observables. If counterfactual outcomes (Y_0, Y_1) are independent of treatment status $D \in \{0, 1\}$ conditional on a set X of covariates, then the conditional counterfactual mean outcome functions $E(Y_d|X = x)$, $d \in \{0, 1\}$, can be estimated from separate regressions of outcomes on covariates for the treated and untreated samples, or from a pooled regression of outcomes on the covariates and their interaction with treatment status:

$$Y = X'\delta_0 + D \cdot X'\delta_1 + q, \tag{2}$$

where $E(q|X, D) = 0$. If the counterfactual mean outcome functions are indeed linear in the covariates, then they are identified by these regressions, and the Average Effect of the

associated with two-way fixed-effects models.

[4]Butts and Gardner (2022) and Gardner, Thakral, Tô, and Yap (2023) derive the asymptotic distribution of the estimator (for the dummy-variable and within-transformation cases, respectively) by treating the first- and second-stages as a joint GMM estimator (see the online appendix to Gardner, Thakral, Tô, and Yap (2023) for a proof that these are equivalent. I provide an alternative, direct proof in Appendix A.

Treatment on the Treated (ATT) can be estimated as the sample analog of

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

The second, aggregation step of this procedure can be avoided by replacing the regression specification (2) with

$$Y = X'\rho_0 + D[X - E(X|D = 1)]'\rho_1 + \beta D + r. \tag{3}$$

In this case, the ATT can be estimated as the sample analog of

$$ATT = \beta + E[X - E(X|D = 1)|D = 1]'\rho_1 = \beta,$$

after replacing $E(X|D = 1)$ with its sample analog $\bar{X}^1 = \sum_i D_i X_i/(\sum_i D_i)$.[5] This modified approach has at least two practical advantages. First, it may be easier to obtain the treated means $\bar{X}^1$ than to aggregate the covariate-specific ATTs. Second, regression estimates of specification (3) will automatically produce asymptotic standard errors for the ATT, which can be used for hypothesis testing and other statistical inference (as Wooldridge, 2010, notes, the standard errors should technically account for the estimation of $\bar{X}^1$, although this unlikely to make much of a difference).

The traditional difference-in-differences estimator regresses outcomes on unit (or treatment-cohort) and time fixed effects, or equivalently, unit and time indicators (abstracting away from any other potential control variables). The nexus between the regression-adjustment matching approach described above and difference-in-differences estimation comes from pretending that these indicators are true covariates (i.e., that they are quasiexperimentally manipulable). If this were the case, the overall average effect of the treatment could be identified by matching observations for treated and untreated units belonging to the same

[5]To the best of my knowledge, this observation is due to Wooldridge (2010, Ch. 21).

unit and recorded in the same period.

Extending the regression-adjustment approach to differences in differences presents two challenges. The first is that treated units can never be matched to untreated versions of themselves in the same period. All difference-in-differences methodologies circumvent the impossibility of this thought experiment under a parallel trends assumption, which allows the evolution of outcomes for untreated units to be used in place of the counterfactual evolution of untreated outcomes for treated units. While this kind of extrapolation between units may be suspect in the general context of selection on observables, in difference-in-differences designs it is actually desirable.

The second challenge is that, while parallel trends implies that untreated mean outcomes are linear in unit and time indicators, in the presence of arbitrary heterogeneity, treated outcomes will be nonlinear in those variables if treatment effects vary at the unit$\times$time level. If the covariates W_{it} consist of a full set of unit indicators, $T-1$ relative time indicators, and time-varying controls X_{it}, parallel trends implies that untreated outcomes satisfy

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

while treated outcomes satisfy

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

Although the latter expression is nonlinear in the covariates, a closer look at the regression-adjustment approach reveals that what it really requires is that counterfactual outcomes are linear on average across the treated population. To see that the logic of this approach carries over to the case of differences in differences, express β_{it} in terms of its projection onto unit and time indicators (and time-varying controls) as

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

where, by definition, $E(\tilde{\beta}_{it}|D_{it}=1)=0$ (with the expectation taken across all time periods). Using this decomposition, we have that[6]

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

Thus, the overall ATT β can be estimated as $\bar{W}^{1'}\hat{\rho}_1$, where $\bar{W}^1 = (\sum_{it} D_{it}W_{it})/\sum_{it} D_{it}$ is the average of the covariates among treated observations and $\hat{\rho}_1$ is the estimated pooled least-squares regression coefficient vector on $D_{it}W_{it}$ from the specification

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

Moreover, the aggregation step (calculating $\bar{W}^{1'}\hat{\rho}_1$) of this procedure is obviated by using the alternative specification

$$Y_{it} = W_{it}'\rho_0 + D_{it}(W_{it} - \bar{W}^1) + \beta D_{it} + r_{it}, \tag{4}$$

from which β represents the overall ATT. Intuitively, including D_{it} forces the unit effects to measure deviations from a reference unit, while demeaning W_{it} forces this unit to be the overall average.[7]

In other words, the overall ATT can be estimated as the coefficient on treatment status from a regression of outcomes on

(i) unit and and time-period indicators, as well as any time-varying control variables,

(ii) interactions between treatment status and deviations in unit indicators, time indica-

[6]Note that expressions of the form $E(Z_{it}|D_{it}=1)$ implicitly treat Z_{it} as a single random variable whose distribution varies across time, implying that $E(Z_{it}|D_{it}=1) = E(\sum_t Z_{it}D_{it})/E(\sum_t D_{it})$ [this follows because $E(Z|D=1) = E(ZD)/E(D)$, where $E(ZD) = \sum_t E(ZD|t)/T$, and similarly for $E(D)$], which is also the probability limit of $\bar{Z}^1 = (\sum_i \sum_t Z_{it}D_{it})/(\sum_i \sum_t D_{it})$ under panel or repeated-cross section random sampling. Alternatively, we could define the estimand of interest to be $E[\sum_t (Y_{it}^1 - Y_{it}^0)]/E(\sum_t D_{it})$. I prefer the former approach since it makes clear the connection between the overall ATT and the ATT in the cross-sectional case, although both lead to the same estimator.

[7]This requires dropping one of the unit indicators from W_{it} to avoid introducing perfect collinearity, although most statistical packages will handle this automatically.

tors, and time-varying controls from their means among treated observations, and

(iii) treatment status.

3.2 Properties

When there are no control covariates, or those covariates only vary at the level of treatment-cohort×time, the Frisch-Waugh-Lovell theorem implies that unit indicators can be replaced with cohort indicators in specification (4) without changing the resulting estimates. In this case, it is easy to see that the one-stage robust difference-in-differences regression estimator is consistent for the overall ATT: under these conditions, since $E(Y_{it}|W_{it}) = W_{it}'\rho_0 + D_{it}[W_{it} - E(W_{it}|D_{it}=1)]'\rho_1 + \beta D_{it}$, and the vector $\bar{W}^1$ of average cohort indicators, time indicators, and controls among the treated converges to its population analog $E(W_{it}|D_{it}=1)$ by a law of large numbers, $\hat{\beta}^{1SDD}$ is consistent by an application of the continuous mapping theorem and standard pooled OLS arguments.[8]

In the general case, the consistency and asymptotic distribution of the one-stage estimator can be established by the following result.

Proposition 2. *The one-stage robust difference-in-differences regression estimator is numerically equivalent to the two-stage difference-in-differences estimator:* $\hat{\beta}^{1SDD} = \hat{\beta}^{2SDD}$.

Proof. Let $\hat{\lambda}_i^0$, $i = 1, \ldots .N$, $\hat{\gamma}_t^0$, $t = 1, \ldots, T$, and $\hat{\delta}^0$ denote the estimated unit fixed effects, time fixed effects, and coefficients on time-varying controls from a first-stage regression of outcomes on those variables, obtained from the sample of untreated observations. The two stage difference-in-differences estimator is the coefficient on D_{it} from a second-stage regression of $Y_{it} - \hat{\lambda}_i^0 - \hat{\gamma}_t^0 - X_{it}'\hat{\delta}^0$ on D_{it} (with no constant term). Since D_{it} and $D_{it}(W_{it} - \bar{W}^1)$ are orthogonal, the term $D_{it}(W_{it} - \bar{W}^1)$ can be added to the second-stage regression without

[8] Alternatively, one can appeal to general consistency results for two-stage estimators (see Newey and McFadden, 1994; Wooldridge, 2010, chapter 12). Identification for pooled OLS also requires that $E\left(\sum_t Q_{it}Q_{it}'\right)$ is invertible, where $Q_{it} = [W_{it}, D_{it}(W_{it} - \bar{W}^1), D_{it}]$ is the vector of observations on all covariates for unit i at time t, which in turn requires that the number of untreated units in every period and the number of treated observations grow without bound with N (this is analogous to the identification requirements for two-stage differences in differences).

changing the estimated coefficient on D_{it}. Now, if $\hat{\lambda}_i^0$, $\hat{\lambda}_t^0$, and $\hat{\delta}^0$ were the same as the estimated unit effects, time effects, and control coefficients $\hat{\lambda}_i^{1SDD}$, $\hat{\lambda}_t^{1SDD}$, and $\hat{\delta}^{1SDD}$ from the one-stage regression specification of Y_{it} on W_{it}, $D_{it}(W_{it}-\bar{W}_{it})$, and D_{it}, then the estimated coefficient $\hat{\beta}^{1SDD}$ on D_{it} from the one-stage specification would be identical to $\hat{\beta}^{2SDD}$ (this is an exercise in partitioned regression mechanics; see, for example, Greene, 2018, Ch. 3) .

By the Frisch-Waugh-Lovell theorem, $\hat{\lambda}_i^{1SDD}$, $\hat{\lambda}_t^{1SDD}$, and $\hat{\delta}^{1SDD}$ can be obtained from a regression of Y_{it} on the residuals from auxiliary regressions of the elements of W_{it} on D_{it} and $D_{it}(W_{it} - \bar{W})$[1]. However, since D_{it} and $D_{it}(W_{it} - \bar{W})$ perfectly predict W_{it} for treated observations (if $D_{it} = 1$, we can always write the kth element of W_{it} as $W_{kit} = D_{it}(W_{kit} - \bar{W}_k) + \bar{W}_k D_{it}$), these residuals will be zero for all treated observations. Therefore, $\hat{\lambda}_i^{1SDD}$, $\hat{\lambda}_t^{1SDD}$, and $\hat{\delta}^{1SDD}$ can also be obtained by regressing Y_{it} on W_{it} in the sample of untreated observations. That is, $\hat{\lambda}_i^{1SDD}$, $\hat{\lambda}_t^{1SDD}$, and $\hat{\delta}^{1SDD}$ equal $\hat{\lambda}_i^0$, $\hat{\lambda}_t^0$, and $\hat{\delta}^0$. □

Thus, the one-stage robust regression estimator is identical to the two-stage difference-in-differences estimator, and the consistency of the former follows formally from that of the latter. The implication of Proposition 2 is that the one-stage approach is another way of obtaining two-stage difference-in-differences estimates. The primary advantage of the one-stage approach is that regression estimates of specification (4) automatically produce standard error estimates that do not need to be adjusted to account for the first-stage estimation of the fixed effects and control coefficients (as Wooldridge, 2010, Ch. 21, notes, the standard errors should technically be adjusted for the use of $\bar{W}^1$ in place of $E(W|D = 1)$, although this likely has a small effect on the resulting standard errors). Thus, the estimator can easily be implemented in any statistical package, without any specialized estimation routine.

3.3 Averaging and aggregated specifications

An apparent drawback of the one-stage robust approach is that, when the control covariates vary at the individual level, it may require the inclusion of a large number of regressors (in-

teractions between treatment status and the deviations of unit indicators from their averages among treated observations), which may be computationally impractical when the number of units is large.[9] In many applications, this will not be a binding constraint. A typical use of difference-in-difference analysis is to examine the effect of a policy change across, say, 50 US states. If the outcome variable is measured at an aggregate level (e.g., a state-level average or count), the one-stage approach will easily be computationally tractable. Since, in microdata settings, treatment status usually varies at more coarse a level than the outcome variable, treatment effects can often be estimated without controlling for individual fixed effects. For example, if the data consist of individuals grouped into states s, then taking expectations conditional on s and time in expression (1) for parallel trends gives

$$\begin{aligned} E(Y_{it}^d|s,t) &= E(\lambda_i|s) + \gamma_t + E(X_{it}|s,t)'\delta + E(\beta_{it}|s,t)D_{st} + E(u_{it}|s,t) \\ &\equiv \lambda_s + \gamma_t + E(X_{it}|s,t)'\delta + \beta_{st}D_{st} + u_{st}, \end{aligned}$$

where D_{st} is an indicator for whether members of state s are treated in period t. Thus, the identification argument underlying the one-stage approach can be applied to state×time-level means in order to identify the overall ATT β after replacing $E(Y_{it}|s,t)$ and $E(X_{it}|s,t)$ with their sample analogs $\bar{Y}_{st}$ and $\bar{X}_{st}$. Moreover, since a regression of $\bar{Y}_{st}$ on state indicators, time indicators, state×time-average controls $\bar{X}_{st}$ and the interaction between state-level treatment status D_{st} and the deviations of those variables from their treated means, it is only necessary to aggregate the individual-level controls to the state×time level. In other words, the robust one-stage procedure can amended by replacing unit fixed effects with state fixed effects and individual×time-level controls with state×time-average controls.

[9]The unit effects themselves (i.e., those that are not converted to deviations from treated means and interacted with treatment status) can be removed using a within transformation, although in many practical cases even this is unnecessary.

3.4 Extensions

Difference-in-differences analyses are usually accompanied by event-study estimates that indicate the dynamic path of the treatment over time and placebo tests for the plausibility of parallel trends. Let $D_{it}^r = 1(t - T_i = r + 1)$ for $r \in \{-(T-2), \ldots 0, 1, \ldots, T-1\}$ be $r+1$-period leads (for $r \leq 0$) or r-period lags (for $r \geq 1$) of treatment status. Also let Y_{it}^{1r} be the counterfactual effect that i would experience at time t after being treated for r periods (holding cohort membership fixed at its observed value), and define $\beta^r = E(Y_{it}^{1r} - Y_{it}^0 | D_{it}^r = 1)$.

First, consider the case where $r \geq 1$. Under the two-stage difference-in-differences methodology, regressing adjusted outcomes $Y_{it} - \hat{\lambda}_i - \hat{\gamma}_t$ on the D_{it}^r, $r \geq 0$, produces estimates of the average effect of being treated for r periods (on units treated for r periods). The one-stage robust estimator can be adapted to incorporate these dynamic estimands by replacing treatment status D_{it} with a set of r-period treatment status indicators D_{it}^r, $r \geq 1$. Following the logic of Section 3.1, under parallel trends

$$Y_{it} = W_{it}'\rho_0 + \sum_{r \geq 1} \{D_{it}^r [W_{it} - E(W_{it} | D_{it}^r = 1)]' \rho_1 + \beta^r D_{it}^r\} + v_{it}, \tag{5}$$

so that

$$E[E(Y_{it}^{1r} | W_{it}) - E(Y_{it}^0 | W_{it}) | D_{it}^r = 1] = \beta^r + E[W_{it} - E(W_{it} | D_{it}^r = 1) | D_{it}^r = 1]' \rho_1 = \beta^r,$$

and hence β^r can estimated consistently as the coefficient on D_{it}^r from a feasible version of (5) that regresses outcomes on W_{it}, $D_{it}^r(W_{it} - \bar{W}^{1r})$, and D_{it}^r, for all $r \geq 1$, where $\bar{W}^{1r}$ is the vector of average covariates among units treated for r units.

To establish the consistency of the dynamic effects, first note that since the D_{it}^r are mutually orthogonal, the two-stage estimates of β^r can be obtained from a version of the two-stage procedure for the overall ATT that replaces D_{it} in the second stage equation

with D_{it}^r (but still estimates the first stage using the subsample of untreated observations). Similarly, the coefficient on D_{it}^r from the dynamic one-stage regression specification (i.e., the sample analog of (5)) is identical to the coefficient that would obtain from first subsetting the data to contain only observations on untreated units and those treated for exactly r periods, then estimating a version of the one-stage specification (4) for the overall ATT that replaces overall treatment status D_{it} with r-period treatment status D_{it}^r. Hence, the consistency of the one-stage dynamic-effect estimates follows from the consistency of the overall ATT estimates.

Placebo testing for parallel trends with the one-stage approach works slightly differently. Choose some pre-treatment period $k \leq 0$, and consider a feasible version of (5) (i.e., replacing $E(W_{it}|D_{it}^{1r} = 1)$ with $\bar{W}^{1r}$) that sums over all $r \geq k$ (rather than over all $r \geq 1$), for some $k \leq 0$. By the preceding argument, this specification is equivalent to estimating the dynamic specification after redefining the onset of the treatment as being $k+1$ periods *before* its actual onset. Consequently, the estimated coefficients on D^r, $r \in \{k, \ldots, 0\}$, represent consistent estimates of $k+1$ pre-treatment placebo ATTs, which can be used to test the plausibility of parallel trends (under which these pre-treatment ATTs should be zero). The consistency of these estimates follows from the preceding discussion, which implies that this procedure is equivalent to estimating dynamic treatment effects by two-stage differences in differences, after redefining treatment status as being $k+1$ periods before the adoption of the treatment.[10]

The estimated coefficients on D^r, $r \geq 1$, from a version of specification (5) that sums over all $r \geq k$ represent consistent estimates of the r-period post-treatment ATTs β^r. However, because this specification only uses observations that are more than k periods away from the actual onset of the treatment as the untreated sample, estimates of β^r, $r \geq 1$, from this specification will differ from those based on a version of (5) that only sums over $r \geq 1$. The dynamic ATT estimates β^r obtained from the original specification may therefore be

[10]This is one of several approaches to testing parallel trends using two-stage differences in differences. For other approaches, see Gardner, Thakral, Tô, and Yap (2023), Borusyak, Jaravel and Spiess (2021) and Liu, Wang and Xu (2023).

preferable when the data are consistent with parallel trends, since those estimates compare the same treated observations to a larger sample of untreated observations.

To summarize, the dynamic and placebo effects of the treatment are identified from regressions of outcomes on

(i) unit and time indicators and control variables, collected into the vector of covariates W_{it},

(ii) interactions $D_{it}^r W_{it}$ between these covariates and duration-specific treatment-status indicators D_{it}^r, for all $r \geq k$ and some $k \leq 0$, and

(iii) duration-specific treatment status indicators D_r, for all $r \geq k$ and some $k \leq 0$.

The coefficients on D_r represent r-period ATTs for $r \geq 1$ and placebo tests of parallel trends for $r \leq 0$ (to estimate the dynamic ATTs using as many observations as possible, set $k = 1$).

Difference-in-differences analyses sometimes also include estimates of cohort-specific ATTs $\beta^j = E(Y_{it}^1 - Y_{it}^0 | D_{it} = 1, C_i^j = 1)$, $c \in \mathcal{C}$. These ATTs are easy to estimate using the robust one-stage approach: simply replace D_{it} and $D_{it}(W_{it} - \bar{W}^1)$ with $D_{it}C_i^j$ and $D_{it}C_{it}(W_{it} - \bar{W}^{1j})$, $j \in \mathcal{C}$, where $\bar{W}^{1j}$ is the average of the covariates among treated observations corresponding to cohort j. These cohort-specific ATTs can also be averaged (perhaps weighting by relative cohort sizes) using commands available in standard software.[11] An analogous variation on the one-stage approach can be used to estimate calendar-time specific average treatment effects.

4 Simulations

4.1 Rejection rates

In order to illustrate the properties of the robust one-stage estimator, I present results from a number of Monte Carlo simulations. To begin, I present rejection rates from a series of

[11] For example, using Stata's `lincom` command.

simulations in which the treatment has no effect, so that

$$Y_{it}^1 = Y_{it}^0 = \lambda_i + \gamma_t + \varepsilon_{it},$$

where $\lambda_i \sim N(T_i, 1)$, $\gamma_t \sim N(0, 1)$, and $\varepsilon_{it} \sim N(0, 3)$. For each of these simulations, $T = 5$, no units are treated in the first period, and adoption times T_i are drawn from a discrete uniform distribution with support $\{2, \ldots, 6\}$ (units with $T_i = 6$ are never treated). In each simulated dataset, I estimate the effect of the treatment using both two-stage difference in differences and one-stage robust differences in differences, each with standard errors clustered at the individual level.[12] For each variation on the simulation exercise, I report results from 1,000 simulated datasets.

The top left panel of Table 1 summarizes the results from estimates that include individual (i.e., unit) fixed effects in simulations in which all variables are re-drawn in each simulated dataset. In this setting, tests based on both the one- and two-stage estimators have a slight tendency to over-reject (at the 5% level) in smaller samples, although this tendency is greater for the one-stage estimator and for estimates of treatment effects that occur longer after the treatment is adopted (this latter tendency spills over, affecting the rejection rate for the overall ATT). As the sample size increases from 50 to 100 to 500 units, the two-stage rejection rate decreases to the appropriate .05, while the one-stage rate remains slightly above this level at .066.[13] The greater tendency of the one-stage estimator to over-reject has less to do with the one-stage approach itself and more to do with the performance of the cluster-robust variance estimator in the presence of many fixed effects.[14] As evidence of

[12] I obtained the 2SDD estimates using Kyle Butts' Stata package `did2s` (Butts, 2021).

[13] In samples of 500, I only estimate the overall ATT for models that include unit fixed effects, since the one-stage specification for duration-specific ATTs includes over 2,000 covariates.

[14] More specifically, the issue arises because the unit-specific sums of the residuals from a model that includes unit fixed effects must be zero, but these sums are a component of the cluster-robust estimate of the variance of those fixed effects. To see this, note that (as I discuss in Appendix B), the one-stage estimator can also be obtained by estimating treatment-status-specific regressions of outcomes on unit and time fixed effects, taking treated-untreated differences in those effects, then averaging the sum of differential unit and time effects over the treated sample. From the OLS first-order conditions, the estimated unit fixed effects are (ignoring treatment status for the sake of simplicity) $\hat{\lambda}_i = \bar{Y}_i - \bar{X}_i'\hat{\delta}$, where $\hat{\delta}$ includes the coefficients on the time effects (as well as any other covariates). Hence, the conditional variance of $\hat{\lambda}_i$ is

this, the top right panel of the table summarizes results from a reprise of these simulations in which treatment status and the unit and time fixed effects are fixed across simulated datasets (only the error term is re-drawn in every dataset), so that the terms $\bar{W}^1$ that appear in the one-stage specification are equal to their population analogues. In this case, both the one- and two-stage estimators have slightly larger tendencies to over-reject than in the random case (at least with respect to the overall ATT; the one-stage estimator appears to perform better for some of the duration-specific ATTs). However, the one-stage estimator still over-rejects more frequently than its two-stage counterpart, and its rate of over-rejection diminishes more slowly as the sample size grows.[15]

The bottom panels of the Table 1 repeat these simulations, replacing unit fixed effects with treatment-cohort fixed effects (recall that in this case with no covariates, replacing unit with cohort fixed effects does not affect the point estimates for either method). Here, regardless of the nature of the simulations, the performance of the one-stage estimator is much closer to, and in some cases better than, that of the two-stage estimator. In particular, when the sample size increases to 500, tests based on either estimator achieve the desired size. The relative performance of the one-stage estimator, and how that performances changes when replacing unit with cohort fixed effects or using a homoskedastic variance estimator, suggests that while tests based on the one-stage estimator and its standard errors may over-reject slightly more than their two-stage counterparts, this over-rejection is not driven by the one-stage estimator's use of estimated treated-sample means, and is unlikely to be much different than those for other regressions that include a large number of fixed effects.

The final row of Table 1 illustrates another numerical equivalence with the one-stage regression approach. The results in the row labelled "manual" report the simulated results

$Var(\bar{\varepsilon}_i|X) - \bar{X}_i' Var(\hat{\delta}|X)\bar{X}_i$. Under clustering at any level more coarse than the unit, the natural estimate of the first component (and as a tedious calculation shows, the actual estimate) is $\left(\sum_t e_{it}\right)^2 / T^2$, which is zero by the first-order conditions for OLS, underestimating the variance of the unit effect (and, in the one-stage differences in differences context, the estimated ATT).

[15] As further evidence that the greater tendency of the one-stage estimator to over-reject is due to the cluster-robust variance estimator (rather than the use of estimated $\bar{W}^1$), when I repeat these simulations using a homoskedastic standard error estimator, the rejection rates are .057, .055, and .048 for samples of size 50, 100, and 500.

from regressing outcomes on unit and time indicators (collected into the vector W_{it}) and the interactions $D_{it}W_{it}$ between these variables and treatment status (i.e., without demeaning relative the treated sample), then estimating the overall ATT as the average product $\bar{W}^1\hat{\rho}_1$ of the interaction terms with their coefficients among the treated sample (and calculating delta-method standard errors for this average). The resulting point estimates and standard errors, and hence rejection rates, are identical to those obtained from the one-stage regression specification.

4.2 Aggregation

The discussion in Section 3.3 shows how the one-stage estimator can be applied with averaged covariates and group fixed effects, or to aggregated data, in order to apply the procedure with fewer regressors. Table 2 presents a second series of simulations designed to illustrate the properties of the one- and two-stage estimators in the presence of covariates and under various forms of aggregation. Here, as in the previous simulations, $T = 5$ and the time T_i of treatment adoption is distributed uniformly over periods 2 through 6, where $T_i = 6$ represents never-treated units. For this simulation exercise, each simulated dataset consists of 500 units organized into 50 "states." In the simulations described below, all variables are re-drawn for every simulated dataset.

For each simulation exercise and simulated dataset, I estimate a number of variations on the one- and two-stage specifications. In Table 2, the row labelled "2SDD" reports the average over 1,000 simulated datasets of two-stage estimates that use treatment-cohort fixed effects and control for the individual-level covariate X_{it}. The row labelled "2SDD, avg. X" reports the results from a variant of 2SDD that uses cohort fixed effects, but replaces the individual covariate X_{it} with the state×time average $\bar{X}_{st}$. The row labelled "2SDD, avg." reports the results from two-stage estimates applied to data that have been aggregated to the state×time-level (i.e., using cohort fixed effects and replacing both Y_{it} and X_{it} with their state-average counterparts). Similarly, the row labelled "1SDD" reports results from a variant

of the one-stage specification that uses cohort fixed effects and controls for the individual-level covariate.[16] The row labelled "1SDD, avg. X" reports results that are estimated on the individual microdata using a model with cohort fixed effects and state×time-average covariates. The row labelled "1SDD, avg." reports results from the one-stage specification applied to state×time aggregate data (that is, using cohort fixed effects and state×time-averaged outcomes and covariates). The final row summarizes the average true ATT.

In all of the simulations, outcomes are determined by

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

where λ_i, γ_t, and ε_{it} are drawn as in the previous simulations. In simulation (1), $X_{it} \sim N(1,1)$ and the treatment effect $\beta_{it} \sim N(2,1)$ is independent of the covariate. As the results show, all of the variants of the one- and two-stage estimates covariates similar point estimates and rejection rates.[17] In simulation (2), $X_{it} \sim N(1,1)$ as before, but the treatment effect depends on X_{it} according to $\beta_{it} = t - T_i + 1 + X_{it}/4$. Here, again, all variants of both the one- and two-stage estimates are similar to each other and to the average true ATT. In this case, 2SDD tends to under-reject at the 5% level, while 1SDD has a (somewhat smaller) tendency to over-reject, this tendency being less pronounced when the procedure is applied to state×year averaged data.[18] In simulation (3), the covariates themselves are correlated with the unit fixed effects, being drawn according to $X_{it} \sim N(\lambda_i/25, 1)$, and the treatment effect depends on X_{it} according to $\beta_{it} = (t - T_i + 1)X_{it}/4$.[19] Illustrating the numerical

[16] Using state fixed effects in these models would lead to the variance underestimation issue described above (see footnote 14). One purpose of these simulations is to show that models with cohort, rather than state or individual, fixed effects can still produce consistent ATT estimates.

[17] Although it is difficult to show without reporting results to a ridiculous level of numerical procedure, the one- and two-stage results that using individual covariates are numerically identical, as are the results from all of variants of the one- and two-stage procedures that use state-averaged covariates.

[18] Part of the difference in rejection rates for 1SDD applied to micro and aggregate data can be explained by the fact that the finite-sample adjustment that Stata applied, which is proportional to $(N-1)/(N-K)$ (where K is the number of regressors), is more influential for the aggregated estimates.

[19] When observations are grouped into coarser units such as cohorts (or states), parallel trends as defined in (1) can always be re-expressed as

$$Y_{it} = \lambda_c + \gamma_t + X_{it}'\delta + \beta_{it} D_{it} + (\lambda_i - \lambda_c) + \varepsilon_{it}.$$

equivalencies, the pattern is the same as before: all variants of the one-stage estimator are identical to their two-stage equivalents, and all are close to the average true ATT. Here, all procedures tend to over-reject, although this tendency is slightly larger for the one-stage approach applied to microdata. Finally, simulation (4) is identical to simulation (3), with the exception that the covariates are drawn as $X_{ut} \sim N(T_i/6, 1)$, so that they are correlated with cohort membership. The results are similar to those above: all variants are similar to each other and the average true ATT, with rejection rates near the target of 5%, although the one-stage estimates obtained using microdata have a marginally larger tendency to over reject.

Overall, the results from these simulation exercises demonstrate that computationally tractable variations on the robust one-stage specification can consistently estimate the effect of the treatment, even in the presence of individual level covariates that influence the effect of the treatment.

5 Empirical application

To illustrate the application of the one-stage estimator, I revisit Cheng and Hoekstra's (2013) analysis of the effects of strengthening Castle Doctrine, also known as "stand your ground" laws, on violent crime. In this analysis, the key treatment status variable is an indicator for whether a state has adopted a stand your ground law in a given state-year, and the dependent variable is the treatment-cohort average of the log of the number of homicides committed per 100,000 people in that year.

Table **??** compares the results from one- and two-stage difference-in-difference estimates of the impact of these laws on homicides.[20] All of the point estimates discussed in this

Consequently, if both treatment status and the covariates are uncorrelated with the differences $\lambda_i - \lambda_c$ between unit and cohort-average fixed effects, it is unnecessary to include unit fixed effects (and in many applications, it is reasonable to assume that treatment status is only correlated with the cohort-average effect). This simulation illustrates that using state×time-average covariates is valid even when the individual-level covariates are correlated with the unit fixed effects.

[20] All of the results presented in Table **??** are weighted by state×year populations size. As a practical

section are based on models that use cohort fixed effects, and all of the standard errors are clustered at the state level. Columns (1) and (4) report one- and two-stage estimates of the overall ATT. The point estimates are identical (for the reasons detailed above). The one-stage estimate has an estimated standard error of about .028, implying statistical significance at the 5% level, while the estimated standard error of the two-stage estimate is somewhat larger, at about .035, but implying significance at the same level. Although it is unknown which standard error provides a more accurate measure of the variation in the point estimate, from a practical perspective, the results from tests based on these methods agree with each other.[21]

Columns (2) and (5) present estimates of the dynamic effects of the treatment. The one-stage estimates are derived from a regression of outcomes on cohort and time indicators, duration-specific treatment status indicators D^r, $r \in \{1, \ldots, 5\}$, and interactions $D^r(W_{it} - \bar{W}^r)$ between duration-specific treatment status and deviations in the cohort and time indicators from their duration-specific treated means. The two-stage estimates are based on second-stage regressions of adjusted outcomes on the duration-specific treatment status variables. I also include leads of treatment status in these second-stage regressions, the coefficients on which represent tests of parallel trends, since the two-stage approach makes this easy, and their inclusion does not affect the estimated coefficients on the duration-specific treatment-status indicators (note that these placebo tests are notionally different from those described in Section 3.4, which are implemented below).[22] As before, the point estimates are identical (note that the one-stage estimates reported in column (2) do not include comparable tests of parallel trends). The one- and two-stage approaches agree on the level of significance for two of the five estimated ATTs; in the remaining cases, tests based on the one-stage approach reject at a lower level of statistical significance. Both approaches lead to

matter, when applying weights using the one-stage approach, it is necessary to use weighted regressions and deviations in the covariates from their weighted means.

[21] Although the simulation results in Table 1 suggest that the one-stage estimator is more likely to over-reject, those simulations are based on homoskedastic and serially uncorrelated errors.

[22] I obtained these estimates using the Stata package `did2s` (Butts, 2021) to obtain standard errors that reflect the first-stage estimation of the adjusted outcomes.

the same practical conclusions about the effect of the treatment.

Columns (3) and (6) present placebo tests of parallel trends that redefine treatment status to mean being four periods *before* the actual onset of the treatment. For the one-stage approach, I implement these tests using the regression specification detailed in Section 3.4, setting $k = -3$ (i.e., including four leads of treatment status, as well as interactions between these leads and de-meaned cohort and time indicators in the regression). For the two-stage approach, I implement these tests by estimating the first stage on a sample of observations that are never treated or more than four periods away from adopting the treatment, then including four leads of treatment status (in addition to the five lags) in the second-stage regression. In this case, tests based on the one- and two-stage approaches agree on the level of statistical significance for all of the placebo coefficients D^r, $r \leq 0$, although in this case most of the one-stage standard errors are larger than their two-stage counterparts. These placebo tests also produce "collateral" estimates of the duration-specific treatment effects D^r, $r > 0$ (which compare treated observations to a smaller control sample of untreated observations). For these estimates, the one- and two-stage approaches agree on the level of statistical significance of all of the coefficients, although most of the one-stage standard errors are slightly smaller. Setting comparisons between the one- and two-stage estimators aside, it is comforting that signs and significance levels of these "collateral" estimates agree with the full-sample estimates presented in columns (2) and (5), and that the placebo estimates presented in columns (3) and (6) generally agree with the alternative test of parallel trends presented for the two-stage estimator in column (6).

6 Conclusion

The problems associated with traditional regression-based difference-in-differences and event-study estimators have sparked the development of several alternative, robust estimators that reliably identify average treatment effect measures, even when adoption is staggered and

average treatment effects are heterogenous. In truth, there is no one "mother" estimator. While all of the recently devised alternative estimators offer some robustness to treatment-effect heterogeneity under staggered adoption, some also have characteristics that make them particularly well-suited for specific environments.

The one-stage regression approach to difference-in-difference and event-study analysis developed in this paper has several advantages. It is motivated intuitively by analogy to matching and regression methods for selection on observables. It is simple and easy to implement in any statistical package, using only a single regression. It is also flexible, and can be extended to identify duration-, group- and time- specific average treatment effects, as well as placebo tests for parallel trends. Because it is also identical to the two-stage difference-in-differences estimator (and its numerical equivalents), it enjoys many of the advantages of that estimator. It is, *inter alia*, robust to the presence of staggered adoption and heterogeneous average treatment effects, efficient (in some circumstances), and readily able to handle settings where parallel trends only holds conditional on time-varying covariates (and when treatment effects depend arbitrarily on those covariates). On the other hand, the two-stage approach is better-suited to models that include many unit fixed effects, offers a broader menu of options for testing whether parallel trends holds, and may be easier to adapt to some more complicated settings (triple differences in differences, to name one).

The one- and two-stage estimators perform similarly in simulation exercises and an applied example, with both leading to the same practical conclusions. In a simulation setting with homoskedastic and serially uncorrelated errors and unit fixed effects, statistical tests based on the one-stage estimator have a marginally higher tendency to over-reject than those based on the two-stage estimator. However, this appears to be because of the performance of the cluster-robust variance estimator itself, and is therefore unlikely to be much different than what one would expect from other regression specifications that include many fixed effects. In any case, the tests are close to their intended sizes in larger samples (and of course they can always be supplemented with bootstrap-based inference, which would also capture

the uncertainty associated with using treated-sample means in the robust one-stage specification). With cohort fixed effects, tests based on the one-stage approach perform similarly to, and in some cases better than, their two-stage counterparts. In the empirical example, not only do the results from the one- and two-stage estimators point to the same broad conclusions, they also exhibit a high degree of consistency: each methodology offers multiple ways of estimating average treatment effects and testing the validity of parallel trends, and the resulting estimates agree both within and across estimation approaches.

Appendix A: Large-sample properties of 2SDD

Proof of Proposition 1. Consistency. Redefine X_{it} to include time indicators, so that parallel trends implies

$$Y_{it} - \lambda_i - X_{it}'\delta = \beta D_{it} + \varepsilon_{it}.$$

Taking deviations from untreated means eliminates the λ_i, so that the second stage of the estimator can be expressed as

$$(Y_{it} - \bar{Y}_{it}^0) - (X_{it} - \bar{X}_{it}^0)'\delta = \beta D_{it} - (\varepsilon_{it} - \bar{\varepsilon}_{it}^0),$$

where $\bar{Y}_i^0 = [\sum_{t=1}^T (1 - D_{it})Y_{it}] / \sum_t (1 - D_{it})$, and similarly for the elements of $\bar{X}_i^0$. Note that since $\varepsilon_{it} = u_{it} + (\beta_{it} - \beta)D_{it}$ where D_{it} and X_{it} are strictly exogenous with respect to u_{it}, those variables are also strictly exogenous with respect to ε_{it} in all untreated periods.[23]

[23] Also note that this since $\hat{\lambda}_i = \bar{Y}_i^0 - \bar{X}_i^{0\prime}\hat{\delta}$, this form of the estimator is equivalent to regressing $Y_{it} - \hat{\lambda}_i - X_i'\hat{\delta}$ on D_{it}.

Next, write

$$\begin{aligned}
\hat{\delta} &= \delta + \left(\frac{1}{N} \sum_i \sum_t (1 - D_{it}) \ddot{X}_{it}^0 \ddot{X}_{it}^{0\prime} \right)^{-1} \left(\frac{1}{N} \sum_i \sum_t (1 - D_{it}) \ddot{X}_{it}^0 \ddot{\varepsilon}_{it}^0 \right) \\
&\xrightarrow{p} \delta + E\left(\sum_t (1 - D_{it}) \ddot{X}_{it}^0 \ddot{X}_{it}^{0\prime} \right)^{-1} E\left(\sum_t (1 - D_{it}) \ddot{X}_{it}^0 \ddot{\varepsilon}_{it}^0 \right) \\
&= \delta,
\end{aligned}$$

where the second line follows by the weak law of large numbers and the assumption that the inverse exists, and the third from the continuous mapping theorem and the strict exogeneity of X_{it}.[24] Thus, the first stage estimates of δ (which includes the time fixed effects and the coefficients on the covariates) are consistent.

Next, write the second-stage estimate as

$$\begin{aligned}
\hat{\beta}^{2SDD} &= \beta + \left(\frac{1}{N} \sum_i \sum_t D_{it} \right)^{-1} \left(\frac{1}{N} \sum_i \sum_t D_{it} (\ddot{Y}_{it}^0 - \ddot{X}_{it}^{0\prime} \hat{\delta}) \right) \\
&= \beta + \left(\frac{1}{N} \sum_i \sum_t D_{it} \right)^{-1} \left(\frac{1}{N} \sum_i \sum_t D_{it} [\ddot{Y}_{it}^0 - \ddot{X}_{it}^{0\prime} \delta + \ddot{X}_{it}^{0\prime} (\delta - \hat{\delta})] \right) \\
&= \beta + \left(\frac{1}{N} \sum_i \sum_t D_{it} \right)^{-1} \left(\frac{1}{N} \sum_i \sum_t D_{it} \ddot{\varepsilon}_{it}^0 + \frac{1}{N} \sum_i \sum_t D_{it} \ddot{X}_{it}^{0\prime} (\delta - \hat{\delta})] \right) \\
&\xrightarrow{p} \beta + E\left(\sum_t D_{it} \right)^{-1} \left[E\left(\sum_t D_{it} \ddot{\varepsilon}_{it}^0 \right) + E\left(\sum_t D_{it} \ddot{X}_{it}^{0\prime} \right) \cdot 0 \right] = \beta,
\end{aligned}$$

where the last line follows from the weak law of large numbers, the continuous mapping theorem/Slutsky's theorem, the assumption that the inverse exists, and the strict exogeneity of D_{it}.[25]

[24] Note that the existence of the inverse implies that the size of the untreated group in every period increases without bound with N.

[25] Note here that the existence of the inverse implies that the number of treated units increases without bound with N.

Asymptotic normality. Consistency implies that

$$\sqrt{N}(\hat{\beta}^{2SDD} - \beta) = \left(\frac{1}{N}\sum_i\sum_t D_{it}\right)^{-1}\left[\frac{\sqrt{N}}{N}\sum_i\sum_t D_{it}\ddot{\varepsilon}^0_{it} + \left(\frac{1}{N}\sum_i\sum_t D_{it}\ddot{X}^{0\prime}_{it}\right)\frac{\sqrt{N}}{N}(\delta - \hat{\delta})\right]$$
$$\xrightarrow{d} N\left(0, A_0^{-1}B_0A_0^{-1}\right)$$

where $A_0 = E(\sum_t D_{tt})$ and

$$B_0 = E\left[\sum_t D_{it}\ddot{\varepsilon}^0_{it}\right] + E\left(\sum_t D_{it}\ddot{X}^{0\prime}_{it}\right)$$
$$E\left(\sum_t (1-D_{it})\ddot{X}^0_{it}\ddot{X}^{0\prime}_{it}\right)^{-1} E\left[\sum_t \ddot{X}^0_{it}\ddot{\varepsilon}^0_i(1-D_{it})\ddot{\varepsilon}^0_i\ddot{X}^{0\prime}_{it}]\right] E\left(\sum_t (1-D_{it})\ddot{X}^0_{it}\ddot{X}^{0\prime}_{it}\right)^{-1}$$
$$E\left(\sum_t D_{it}\ddot{X}^0_{it}\right).$$

In the above, the convergence in distribution comes from a weak law of large numbers, a central limit theorem, the continuous mapping theorem (for convergence in distribution), and the fact that $\hat{\delta}$ is uncorrelated with $D_{it}\ddot{\varepsilon}^0_{it}$ because they are drawn from different samples. □

Appendix B: Additional details on the large-sample properties of 1SDD

The one-stage difference-in-difference regression estimate can alternatively be obtained by (i) estimating separate regressions on samples of untreated and treated observations

$$Y_{it} = \lambda^d_i + X'_{it}\delta^d + u^d_{it}, \quad d \in \{0,1\}, \tag{6}$$

where the time fixed effects γ^d_t have been absorbed into the covariates X_{it}, (ii) forming the differences $\beta_i = \lambda^1_i - \lambda^0_i$ and $\beta_x = \delta^1 - \delta^0$, and (iii) calculating the average effect of the

treatment on the treated as[26]

$$\frac{1}{\sum_i \sum_t D_{it}} \left(\sum_i \sum_t D_{it} \hat{\beta}_i + \sum_i \sum_t D_{it} X'_{it} \hat{\beta}_x \right).$$

Moreover, each of the treatment-status-specific regressions (6) can be estimated using a within transformation, and the unit fixed effects recovered as

$$\hat{\lambda}_i^d = \bar{Y}_i^d - \bar{X}_i^{d\prime} \hat{\delta}^d,$$

where $\bar{Y}_i^d = [\sum_t Y_{it} 1(D_{it} = d)]/[\sum_t 1(D_{it} = d)]$ (and similarly for the vector $\bar{X}_i^d$). Conditional on X_i, the asymptotic variance of the estimated effects is $Avar(\hat{\lambda}_i^d) = Avar(\bar{Y}_i^d) - \bar{X}_i^{d\prime} Avar(\hat{\delta}^d) \bar{X}_i^d$, and, since $\widehat{Var}(\hat{\delta}^d|W) = \widehat{Avar}(\hat{\delta}^d)$, estimates of these asymptotic variances will be the same as the estimated finite-sample variances of dummy-variable estimates of the fixed effects.[27] This justifies the use of standard errors from dummy-variable estimates of the one-stage regression specification as estimates of the asymptotic variance of the difference-in-differences estimates. Conditional on the observed values of treatment status and the time-varying controls, the asymptotic normality of the one-stage difference-in-difference estimate (which is now seen to be a linear combination of the $\hat{\delta}^d$) then follows from the asymptotic normality of the $\hat{\delta}^d$.

This discussion also motivates an alternative proof of the consistency of the one-stage regression estimator. Since $\beta_i = \lambda_i^1 - \lambda_i^0$ where $E(\lambda^d) = E(\bar{Y}_i^d) - E(\bar{X}_i^d)'\delta^d$, a law of large numbers and the continuous mapping theorem give[28]

[26] In the case of cohort fixed effects, this can also be expressed as $\sum_{j \in \mathcal{C}} \bar{w}_j \hat{\beta}_j + \bar{X}^{1'} \hat{\beta}_x$ where $\bar{w}_j$ is the fraction of treated observations that correspond to cohort j and $\bar{X}^1$ is the vector of averages of the controls (including time indicators) among all treated observations.

[27] For example, in the homoskedastic, non-autocorrelated case, the asymptotic and finite-sample variances both equal $\sigma^2/T + \bar{x}'_j Var(\hat{\delta}^d) \bar{x}'_j$, where $\sigma^2 = Var(\varepsilon_{it}|X)$.

[28] In the case of cohort fixed effects, $\bar{w}_j = [\sum_i \sum_t D_{it} 1(T_i = j)]/(\sum_i \sum_t D_{it}) \xrightarrow{p} E[\sum_t D_t 1(T_i = j)]/E(\sum_t D_t) \equiv \pi_j$ and $\bar{X}^1 = (\sum \sum_t D_{it} X_{it})/(\sum_i \sum_t D_{it}) \xrightarrow{p} E(\sum_t D_{it} X_{it})/E(\sum_t D_{it})$, so that $\hat{\beta}^R = \sum_j \bar{w}_j \hat{\beta}_j + \bar{X}^1 \hat{\beta}_x \to \sum_j \pi_j \beta_j + E(\sum_t D_{it} X'_{it} \beta_x)/E(\sum_t D_{it})$.

$$
\begin{aligned}
\hat{\beta}^{1SDD} &= \frac{1}{\sum_i \sum_t D_{it}} \left[\sum_i \left(\sum_t D_{it} \hat{\beta}_i \right) + \sum_i \left(\sum_t D_{it} X'_{it} \hat{\beta}_x \right) \right] \\
&= \frac{N}{\sum_i \sum_t D_{it}} \left[\frac{1}{N} \sum_i \left(\sum_t D_{it} [(\bar{Y}_i^1 - \bar{X}_i^{1\prime} \hat{\delta}^d) - (\bar{Y}_i^0 - \bar{X}_i^{0\prime} \hat{\delta}^0)] \right) + \frac{1}{N} \sum_i \left(\sum_t D_{it} X'_{it} \hat{\beta}_x \right) \right] \\
&\xrightarrow{p} \frac{1}{E(\sum_t D_{it})} E \left(\sum_t D_{it} \beta_i + \sum_t D_{it} X'_{it} \beta_x \right) \\
&= E(\beta_{it} | D_{it} = 1).
\end{aligned}
$$

References

Borusyak, Kirill, Xavier Jaravel, and Jann Spiess. 2021. "Revisiting event study designs: Robust and efficient." Working paper.

Kyle Butts, 2021. "DID2S: Stata module to estimate a TWFE model using the two-stage difference-in-differences approach." Statistical Software Components S458951, Boston College Department of Economics, revised 28 Apr 2023.

Butts, Kyle and John Gardner. 2022. "did2s: Two-stage difference-in-differences." R Journal, 14 (3): 162-173.

Callaway, Brantly, and Pedro Sant'Anna. 2021. "Difference-in-differences with multiple time periods and an application on the minimum wage and employment." Journal of Econometrics, 225(2):200-230.

Caetano, Carolina, Brantly Callaway, Stroud Payne, and Hugo Sant'Anna Rodrigues. 2022. "Difference in differences with time-varying covariates." Working paper.

Cheng, Cheng and Mark Hoekstra. 2012. "Does strengthening self-defense law deter crime or escalate violence? Evidence from the expansions to the Castle Doctrine." Journal of Human Resources, 48(3):821-853.

Dube, Arandrajit, Girardi, Daniele, Jordà, Òscar and Alan M. Taylor. 2023. "A local projections approach to difference-in-differences event studies." NBER Working Paper no. 31184.

de Chaisemartin, Clément and Xavier D'Haultfœuille. 2020. "Two-way fixed effects estimators with heterogeneous treatment effects." American Economic Review, 110(9): 2964-2996.

Goodman-Bacon, Andrew. 2021. "Difference-in-differences with variation in treatment timing." Journal of Econometrics, 225(2): 254-277.

Gardner, John. 2021. "Two-stage differences in differences." Working paper.

Gardner, John, Thakral, Neil, Tô, Linh and Luther Yap. 2023. "Two-stage differences in differences." Working paper.

Greene, William. 2018. *Econometric Analysis.* 8th Edition, Pearson Education Limited, London.

Liu, Licheng, Ye Wang, and Yiqing Xu. 2023. "A practical guide to counterfactual estimators for causal inference with time-series cross-sectional data." American Journal of Political Science.

Newey, Whitney K., and Daniel McFadden. 1994. "Large sample estimation and hypothesis testing." In Handbook of Econometrics, edited by R.F. Engle and D.L. McFadden, IV:2112–2245. Elsevier Science.

Sun, Liyang and Sarah Abraham. 2021. "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects." Journal of Econometrics, 225(2):175-199.

Wooldridge, Jeffrey M. 2010. *Econometric Analysis of Cross Section and Panel Data.* 2nd ed. The MIT Press. Cambridge, Mass.: MIT Press.

Wooldridge, Jeffrey M. 2021. "Two-way fixed effects, the two-way Mundlak regression, and difference-in-differences estimators." SSRN Electronic Journal.

Tables

Table 1: Rejection rates from 1,000 simulations

		Individual fixed effects					
		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		0.059	0.062	
	D^2	0.068	0.058		0.075	0.043	
	D^3	0.072	0.071		0.064	0.05	
	D^4	0.099	0.071		0.095	0.074	
1SDD	D	0.092	0.085	0.066	0.098	0.087	0.077
	D^1	0.078	0.084		0.075	0.125	
	D^2	0.058	0.063		0.079	0.056	
	D^3	0.068	0.063		0.057	0.051	
	D^4	0.095	0.07		0.083	0.104	
		Cohort fixed effects					
		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

Notes: "Random" denotes simulations in which all variables are re-drawn in every simulated dataset; "Fixed" denotes simulations in which treatment status is drawn only once (so that only ε_{it} is re-drawn in every simulated dataset). "2SDD" is two-stage differences in differences and "1SDD" is robust one-stage difference in differences. "D" denotes the overall ATT and "D^r" denotes the r-period ATT. "Manual" denotes estimates obtained by regressing outcomes on unit and time fixed effects, controls, and interactions between treatment status and those variables (without converting them to differences from treated means), then averaging the estimated interaction terms, multiplied by their estimated coefficients, over all treated observations (i.e., $D_{it}\bar{W}_{it}'\hat{\rho}_1$ in the notation of Section 3.1).

Table 2: Simulations from individual-level and aggregate estimates

	(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

Notes: Average point estimates from 1,000 simulations, each corresponding to 500 individuals organized into 50 states. All variables are re-drawn in every simulated dataset. In simulation (1), $X_{it} \sim N(1,1)$ and the treatment effect is independent of the covariate. In simulation (2), $X_{it} \sim N(0,1)$ and the treatment effect is $\beta_{it} = t - T_i + 1 + X_{it}/4$. In simulation (3), $X_{it} \sim N(\lambda_i/25, 1)$ and $\beta_{it} = (t - T_i + 1)X_{it}/4$. In simulation (4), $X_{it} \sim N(T_i/6, 1)$ and $\beta_{it} = (t - T_i + 1)X_{it}/4$. "2SDD" refers to the standard two-stage difference-in-differences estimate, "2SDD, avg. X" refers to 2SDD with $\bar{X}_{st}$ in place of X_{it}, "2SDD, avg." refers to 2SDD after aggregating all variables to the state×time level, "1SDD" refers to the standard one-stage difference-in-differences estimate, "1SDD, avg. X" refers to the robust one-stage estimator with state fixed effects and $\bar{X}_{st}$ as a time-varying covariate, "1SDD, avg." refers to the standard one-stage estimator after aggregating all variables to the state×time level, and "ATT" is the average overall ATT across all simulations. All estimates use cohort fixed effects and are clustered on the state level.

Table 3: One- and two-stage difference-in-difference estimates

		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)
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

Notes: Columns (1) and (4) are 1SDD and 2SDD overall ATT estimates, respectively. Columns (2) and (5) contain the 1SDD and 2SDD estimates of the dynamic post-treatment effects (column (5) also contains the default 2SDD placebo tests of parallel trends. Columns (3) and (6) contain 1SDD and 2SDD placebo tests of parallel trends that assume the treatment begins four periods before its actual adoption (i.e., pretends that $D = 1$ if $r \geq -3$), as well as the post-treatment dynamic effects implied by this placebo assumption. All estimates control for cohort×year-average police per capita, use cohort fixed effects, and cluster at the state level.