Difference-in-Differences Estimators with Continuous Treatments and no Stayers

By Clément de Chaisemartin, Xavier D'Haultfœuille and Gonzalo Vazquez-Bare*

Draft: February 8, 2024

Many treatments or policy interventions are continuous in nature. Examples include prices, taxes or temperatures. Empirical researchers have usually relied on two-way fixed effect regressions to estimate treatment effects in such cases, see e.g. Deschênes and Greenstone (2012). However, such estimators are not robust to heterogeneous treatment effects in general (De Chaisemartin and D'Haultfœuille, 2020); they also rely on the linearity of treatment effects. We propose estimators for continuous treatments that do not impose those restrictions, and that can be used when there are no stayers: the treatment of all units changes from one period to the next. This is for instance the case when the treatment is precipitations or temperatures: for instance, temperatures of all US counties change, if ever so slightly, between two consecutive years. We start by extending the nonparametric results of de Chaisemartin et al. (2023) to cases without stayers. We also present a parametric estimator, and use it to revisit Deschênes and Greenstone (2012).

I. Set-up, assumptions and parameter of interest

A representative unit is drawn from an infinite super population, and observed at two time periods. All expectations below are taken with respect to the distribution of variables in the super population. We are interested in the effect of a continuous and scalar treatment variable on that unit's outcome.

* Chaisemartin: Sciences Po Paris, clement.dechaisemartin@sciencespo.fr. D'Haultfœuille: CREST-ENSAE, xavier.dhaultfoeuille@ensae.fr. Vazquez-Bare: University of California, Santa Barbara, gvazquez@econ.ucsb.edu.

Let D_t denote the unit's treatment at period $t \in \{1,2\}$ and let $\mathcal{D}_t$ denote its support; let also $\mathcal{D}$ denote the support of (D_1, D_2). For any $(d_1, d_2) \in \mathcal{D}$, let $Y_t(d_1, d_2)$ denote the unit's potential outcome at t with treatment d, and let Y_t denote their observed outcomes: $Y_t = Y_t(D_1, D_2)$. Finally, for any random variables $(X_t)_{t=1,2}$, let $\Delta X = X_2 - X_1$. We impose the following assumptions:

ASSUMPTION 1: *(Static model) For all $t \in \{1,2\}$ and $(d_1, d_2) \in \mathcal{D}$, $Y_t(d_1, d_2)$ only depends on d_t; we denote it by $Y_t(d_t)$.*

ASSUMPTION 2: *(Parallel trends) $\forall d \in \mathcal{D}_1$, $E(\Delta Y(d)|D_1 = d, D_2) = E(\Delta Y(d)|D_1 = d)$.*

ASSUMPTION 3: *(Bounded treatment, bounded-lipschitz potential outcomes)*

1) $\mathcal{D}_1$ and $\mathcal{D}_2$ are bounded subsets of $\mathbb{R}$.

2) $\exists \overline{Y} \geq 0$: $\sup_{(d_1,d_2)\in\mathcal{D}} E[\overline{Y}|D_1 = d_1, D_2 = d_2] < \infty$, and $\forall (t, d, d') \in \{1,2\} \times \mathcal{D}_t^2$, $|Y_t(d) - Y_t(d')| \leq \overline{Y}|d - d'|$.

Assumptions 2-3 are also imposed by de Chaisemartin et al. (2023), and are discussed therein.

ASSUMPTION 4: *(No stayers but quasi-stayers) $P(\Delta D = 0) = 0$, $P(|\Delta D| \leq \eta) > 0 \; \forall \eta > 0$.*

First, Assumption 4 states that there are no "stayers", namely units for which $D_1 = D_2$. This is in contrast with de Chaisemartin et al. (2023), who assume throughout that there are stayers. Second, Assumption 4 states that there are "quasi-stayers", namely

units whose treatment change may be infinitesimally small. This assumption is realistic when the treatment is, say, temperatures: some counties may have very similar temperatures from one year to the next, though no county has exactly the same temperatures.

Hereafter, we focus on the following effect:

$$(1)\quad \theta_0 = E\left(\frac{|\Delta D|}{E(|\Delta D|} \times \frac{Y_2(D_2) - Y_2(D_1)}{D_2 - D_1}\right) = \frac{E\left(\text{sgn}(\Delta D)(Y_2(D_2) - Y_2(D_1))\right)}{E(|\Delta D|)}.$$

θ_0 is a weighted average of the slopes of units' potential-outcome functions, from their period-one to their period-two treatment, the so-called WAOSS in de Chaisemartin et al. (2023). It follows from the mean-value theorem that it may be seen as a weighted average marginal effect.

II. Nonparametric identification and estimation

THEOREM 1: *If Assumptions 1-4 hold,*

$$\theta_0 = [E\left(S\Delta Y\right) - \zeta_0]/E[|\Delta D|],$$

with $S := sgn(\Delta D)$ *and*

$$\zeta_0 := E\left[S \lim_{\eta \downarrow 0} E(\Delta Y | D_1, |D_2 - D_1| \leq \eta)\right].$$

Theorem 1 shows that without stayers, θ_0 is identified by the limit (as $\eta \downarrow 0$) of a difference-in-difference comparing the ΔY of all units and of quasi-stayers.

We now discuss estimation of θ_0. Only the estimation of ζ_0 raises difficulties. We show in the proof of Theorem 1 that under our assumptions, $g(d_1, \delta) := E[\Delta Y | D_1 = d_1, \Delta D = \delta]$ is well-defined and continuous at $(d_1, 0)$, for any $d_1 \in \mathcal{D}_1$. Hence, ζ_0 satisfies $\zeta_0 = E\left[Sg(D_1, 0)\right]$. This formulation links our problem to the estimation of nonparametric additive models. To see this, suppose that the variables $(W, X) \in \mathbb{R} \times \mathbb{R}^k$ satisfy $h(x) := E[W|X = x] = \sum_{j=1}^{k} h_j(x_j)$ for some unknown functions $(h_j)_{j=1,...,k}$. Then, under the normalization $E[h_j(X_j)] = 0$ for $j < k$, we can identify and estimate h_k by remarking that

$$(2)\quad h_k(x_k) = E[h(X_1, ..., X_{k-1}, x_k)].$$

We can then estimate $h_k(x_k)$ by first estimating h with any usual nonparametric estimator, and second plugging it in the sample counterpart of the expectation in (2). As Linton and Nielsen (1995) and Kong, Linton and Xia (2010) show, the corresponding estimator is, under regularity conditions, asymptotically normal and converges at the standard univariate nonparametric rate (namely, $n^{2/5}$, with n the sample size). This rate is also the optimal convergence rate for this problem (Stone, 1985). Up to minor changes (in ζ_0, g plays the role of h in (2) and ζ_0 also includes S), our parameter ζ_0 can be obtained in the same way as $h_k(x_k)$, so we can also obtain an asymptotically normal estimator converging at the $n^{2/5}$ rate.

This contrasts with the standard ($n^{1/2}$) rate obtained for the estimators of the WAOSS in the presence of stayers, as shown by de Chaisemartin et al. (2023). To understand the difference, note that with stayers, the proportion of units used as controls to reconstruct switchers' counterfactual outcome evolution remains positive as $n \to \infty$. On the other hand, it tends to zero here, since we need to consider quasi-stayers, with $\eta \to 0$ as $n \to \infty$ to avoid any bias. This results in a lower rate of convergence.

Finally, in applications with no stayers, it is more difficult to propose placebo estimators of the parallel trends assumption. When a third period of data, period zero, is available, a placebo mimics the actual estimator, replacing ΔY by units' period-zero-to-one outcome evolution. However, as units' treatments may have changed from period zero to one, one would need to restrict the sample to period-zero-to-one quasi-stayers, to avoid that the placebo differs from zero due to the treatment's effect. Thus, the placebo would compare the period-zero-to-one outcome evolution of period-one-to-two switchers and quasi-stayers, restricting the sample to period-zero-to-one quasi-stayers. Then, we conjecture that the number of units used

as controls by the placebo may tend to zero faster than the number of units used as controls by the actual estimator, for instance if being a period-zero-to-one and a period-one-to-two quasi-stayer are independent events. Then, the placebo may converge at an even slower rate than the actual estimator.

III. A parametric approach

We now consider a parametric root-n consistent estimator, that avoids issues related to nonparametric estimation and inference, while still allowing for heterogeneous and nonlinear effects. Specifically, we impose that $g(d_1, \delta) = g_{\lambda_0}(d_1, \delta)$, where the family $(g_\lambda)_{\lambda \in \mathbb{R}^p}$ is known (but λ_0 is not). By definition of g and Assumption 2,

$$g(d_1, \delta) = E[Y_2(d_1) - Y_1(d_1)|D_1 = d_1] + \delta \times E\left[\frac{Y_2(d_1 + \delta) - Y_2(d_1)}{\delta} \Big| D_1 = d_1, \Delta D = \delta\right].$$

Thus, the parametric assumption amounts to imposing restrictions on both $d_1 \mapsto E[Y_2(d_1) - Y_1(d_1)|D_1 = d_1]$ and the average slope $(d_1, \delta) \mapsto E[(Y_2(d_1 + \delta) - Y_2(d_1))/\delta | D_1 = d_1, \Delta D = \delta]$. For instance, if $g_\lambda(d_1, \delta)$ is linear, we assume that the former function is linear, and the latter is constant. Similarly, g is a polynomial if both functions are polynomial. Note that we can test that $E[\Delta Y|D_1 = d_1, \Delta D = \delta] = g_{\lambda_0}(d_1, \delta)$ for some λ_0 by a parametric specification test, see e.g. Bierens (1982) or Hong and White (1995).

We consider a simple two-step estimator based on this parametric restriction and an i.i.d. sample $(D_{1i}, \Delta D_i, \Delta Y_i)_{i=1,\dots,n}$. In the first step, we estimate λ_0 by (linear or nonlinear) least squares or, more generally, a GMM estimator $\widehat{\lambda}$. In the second step, we estimate θ_0 by

$$\widehat{\theta} = \frac{\sum_{i=1}^n S_i(\Delta Y_i - g_{\widehat{\lambda}}(D_{1i}, 0))}{\sum_{i=1}^n |\Delta D_i|}.$$

Since $\widehat{\theta}$ may be seen as a two-step GMM estimator, we obtain, under Assumptions 1-4 and standard regularity conditions on $\lambda \mapsto g_\lambda(d_1, \delta)$,

$$\sqrt{n}\left(\widehat{\theta} - \theta_0\right) \xrightarrow{d} \mathcal{N}\left(0, V(\psi)\right),$$

where the influence function ψ satisfies

$$\psi = \frac{1}{E[|\Delta D|]}\left[S\left(\Delta Y - g_{\lambda_0}(D_1, 0)\right) - E\left[S\frac{\partial g}{\partial \lambda}(D_1, 0)_{|\lambda=\lambda_0}\right] \times \xi - \theta_0 |\Delta D|\right],$$

with ξ the influence function of $\widehat{\lambda}$. We can thus simply estimate $V(\psi)$ by a plug-in estimator, using an initial estimator of ξ.

IV. Application

We use the data from Deschênes and Greenstone (2012) to compute our parametric estimator. The authors use a balanced panel of 2,342 US counties in years 1987, 1992, 1997, and 2002, and consider TWFE regressions, weighted by counties' farmland acres, of annual agricultural profits in county c and year t on four treatment variables: growing season degree days, growing season degree days squared, precipitations, and precipitations squared. To fit in the two-periods-one-treatment case we consider, we restrict the data to years 1997 and 2002, and we focus on the growing season degree days treatment. The coefficient of that treatment in a TWFE regression estimated on years 1997 and 2002 and weighted by counties' farmland acres is equal to -0.024 (s.e. clustered at the county level: 0.007), which is close to the corresponding TWFE coefficient keeping the four years and all treatments (-0.015, s.e. clustered at the county level: 0.005). Assuming that

$$E[Y_2(d_1) - Y_1(d_1)|D_1 = d_1] = \lambda_{0,1} + \lambda_{0,2} d_1$$

and

$$E\left[\frac{Y_2(d_1 + \delta) - Y_2(d_1)}{\delta} \Big| D_1 = d_1, \Delta D = \delta\right] = \lambda_{0,3} + \lambda_{0,4} d_1 + \lambda_{0,5} \delta,$$

we find that $\widehat{\theta}$, weighted by counties' farmland acres as well, is equal to -0.018 (s.e.: 0.011) Thus, the conclusion from the TWFE

regression seems robust to allowing for some effect heterogeneity, even though the estimated effect is less significant. While arguably restrictive, our model for the conditional expectation function of slopes allows for some non-linearity and heterogeneity in the effects of temperatures on agricultural output.

Appendix: proof of theorem 1

It suffices to show that a.s.,

$$\lim_{\eta \downarrow 0} E\left(\Delta Y | D_1, |\Delta D| \leq \eta\right) = E\left(Y_2(D_1) - Y_1(D_1) | D_1, D_2\right). \tag{3}$$

Fix $\eta > 0$. By Assumption 4, $P(|\Delta D| \leq \eta | D_1) > 0$. Thus, $E\left(\Delta Y | D_1, |\Delta D| \leq \eta\right)$ is well-defined. Moreover,

$$\begin{aligned} E\left(\Delta Y | D_1, |\Delta D| \leq \eta\right) = & \\ E\left(Y_2(D_2) - Y_2(D_1) | D_1, |\Delta D| \leq \eta\right) & \\ + E\left(Y_2(D_1) - Y_1(D_1) | D_1, |\Delta D| \leq \eta\right). & \end{aligned} \tag{4}$$

Now, by Jensen's inequality and Point 2 of Assumption 3,

$$\begin{aligned} & \left| E\left[Y_2(D_2) - Y_2(D_1) | D_1, |\Delta D| \leq \eta\right] \right| \\ \leq & E\left(\left|Y_2(D_2) - Y_2(D_1)\right| \ | D_1, |\Delta D| \leq \eta\right) \\ \leq & E\left(\overline{Y} |D_2 - D_1| \ | D_1, |\Delta D| \leq \eta\right) \\ \leq & \eta E\Bigg[\sup_{(d_1, d_2) \in \mathcal{D}} E\left(\overline{Y} | D_1 = d_1, D_2 = d_2\right) | \\ & D_1, |\Delta D| \leq \eta \Bigg] \\ \leq & \overline{K} \eta \end{aligned} \tag{5}$$

for some $\overline{K} < \infty$. Next, by Assumption 2,

$$\begin{aligned} E\left(Y_2(D_1) - Y_1(D_1) | D_1, |\Delta D| \leq \eta\right) = & \\ E\left(Y_2(D_1) - Y_1(D_1) | D_1\right) = & \\ E\left(Y_2(D_1) - Y_1(D_1) | D_1, D_2\right). & \end{aligned}$$

Combined with (4)-(5), this yields (3) □

REFERENCES

Bierens, Herman J. 1982. "Consistent model specification tests." *Journal of Econometrics*, 20(1): 105–134.

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

de Chaisemartin, Clément, Xavier D'Haultfœuille, Félix Pasquier, and Gonzalo Vazquez-Bare. 2023. "Difference-in-differences estimators for treatments continuously distributed at every period." arXiv preprint arXiv:2201.06898.

Deschênes, Olivier, and Michael Greenstone. 2012. "The economic impacts of climate change: evidence from agricultural output and random fluctuations in weather: reply." *American Economic Review*, 102(7): 3761–3773.

Hong, Yongmiao, and Halbert White. 1995. "Consistent specification testing via nonparametric series regression." *Econometrica: Journal of the Econometric Society*, 63: 1133–1159.

Kong, Efang, Oliver Linton, and Yingcun Xia. 2010. "Uniform Bahadur representation for local polynomial estimates of M-regression and its application to the additive model." *Econometric Theory*, 26(5): 1529–1564.

Linton, Oliver, and Jens Perch Nielsen. 1995. "A kernel method of estimating structured nonparametric regression based on marginal integration." *Biometrika*, 82(1): 93–100.

Stone, Charles J. 1985. "Additive regression and other nonparametric models." *The Annals of Statistics*, 13(2): 689–705.