DIFFERENCE IN DIFFERENCES

Michal Kolesár*

April 22, 2024

1. 2 BY 2 DIFFERENCE-IN-DIFFERENCES

Let us first briefly review difference-in-differences (DiD) designs with 2 groups $g \in \{0,1\}$ and 2 time periods $t \in \{0,1\}$. Treatment is a deterministic function of time and group membership. We denote by $D_{gt} = \mathbb{1}\{g = 1, t = 1\}$ the treatment of group g at time t. Let F_{gt} denote the distribution of observed outcomes in group g at time t. Let $Y_{gt} \sim F_{gt}$ denote a random variable with this distribution. We focus on identification, and therefore omit the unit-level subscripts. Let $Y_{gt}(0)$ and $Y_{gt}(1)$ denote potential outcomes, so that $Y_{gt} = Y_{gt}(D_{gt})$. For simplicity, there are no covariates. We may have panel data (we draw samples from the joint distribution of (Y_{g0}, Y_{g1})), or a repeated cross-section (we draw samples from the marginal distribution F_{gt}).

The DiD estimand is given by

$$\beta = E[Y_{11} - Y_{10}] - E[Y_{01} - Y_{00}].$$

In the panel case, this can be estimated as a regression of the difference $Y_{g1} - Y_{g0}$ onto a constant and $D_{g1} = D_{g1} - D_{g0}$. Since there are only two periods, this is equivalent to running a fixed effects regression with unit and time fixed effects. In the repeated cross-section case, we regress the outcome on the group dummy interacted with a time dummy.

The estimand can be written as the average treatment effect for the treated (ATT) plus a selection bias term coming from differential trends among the treated and control units:

$$\beta = E[Y_{11}(1) - Y_{11}(0)] + \underbrace{E[Y_{11}(0) - Y_{10}(0)] - E[Y_{01}(0) - Y_{00}(0)]}_{\text{Differential trend}}. \tag{1}$$

This follows directly from the definition of the estimand and the fact that $Y_{gt} = Y_{gt}(1)$ if $g = t = 1$, and $Y_{gt} = Y_{gt}(0)$ otherwise.

Therefore, the crucial assumption underlying DiD designs is that the treated and

*Email: mkolesar@princeton.edu.

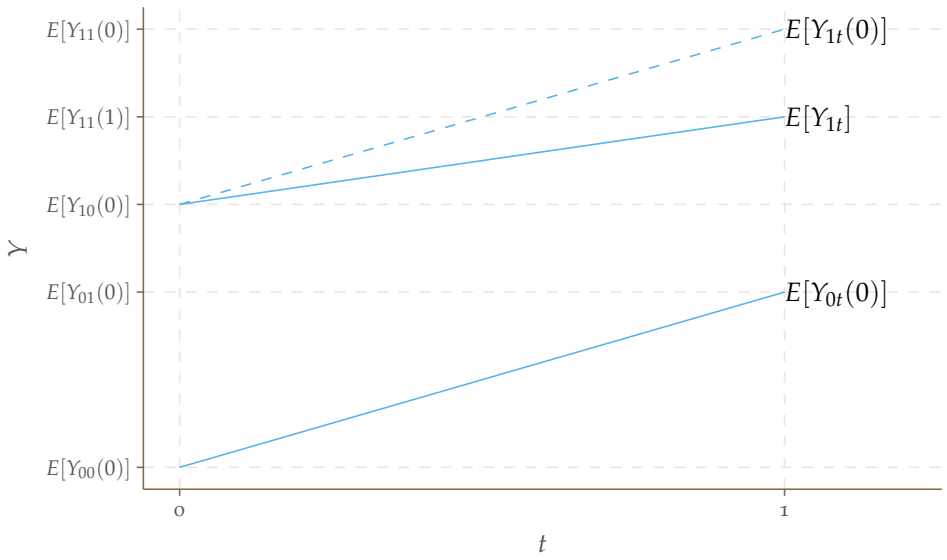

Figure 1: Common trends assumption in DiD designs. The solid lines are observed, the dotted line is imputed by shifting $E[Y_{10}]$ by $E[Y_{01}] - E[Y_{00}]$.

control units have *common trends* (or *parallel trends*):

$$E[Y_{1t}(0) - Y_{10}(0)] = E[Y_{0t}(0) - Y_{00}(0)] \tag{2}$$

for $t = 1$. This allows us to impute the counterfactual mean $E[Y_{11}(0)]$, as illustrated in Figure 1.

Remark 1. For the DiD method to make sense, we only need a notion of a potential outcome $Y_{11}(0)$. It is not necessary to define potential outcomes $Y_{0t}(1)$. In some examples, conceptualizing a treated outcome for the control units can be difficult—but we do not need to do so.

Remark 2 (Ashenfelter's dip). The parallel trends assumption is fragile—it allows selection on levels, but not on differences. It rules out the "dip" observed in Ashenfelter (1978, p. 51): "earnings of trainees [in a labor market program] tend to fall, both absolutely and relative to the comparison group, in the year prior to training.". In other words, the treated group experiences a downward "dip" relative to the control group in earnings prior to treatment. In such case DiD methods will overstate the treatment effect if there is reversion to the mean.

1.1. *Falsification tests*

There are two falsification tests commonly used in practice to check the common trends assumption: estimate the treatment for groups not affected, and, with multiple pre-

treatment periods, estimate pretrends.

EFFECTS ON GROUPS NOT AFFECTED For example, Gruber (1994) uses DiD to evaluate the effect of the passage of mandatory maternity benefits in some US states on the wages of married women of childbearing age. In this context, the common trends assumption implies that in the absence of the intervention, married women of childbearing age would have experienced the same increase in log wages in states that adopted mandated maternity benefits and states that did not. To evaluate this assumption, Gruber (1994) compares the changes in log wages in adopting and non-adopting states for single men and for women over 40 years old.

PRETRENDS TEST Suppose we have data on multiple periods, running from $-T_0 \leq 0$ to $T_1 \geq 1$, but there are still just two groups, with the treatment path given by $D_{gt} = \mathbb{1}\{g = 1, t \geq 0\}$. Write

$$E[Y_{gt}] = \mathbb{1}\{t \neq 0\}\lambda_t + \alpha_g + \mathbb{1}\{g = 1, t \neq 0\}\beta_t, \tag{3}$$

where $\alpha_g = E[Y_{g0}]$, $\lambda_t = E[Y_{0t} - Y_{00}]$, and $\beta_t = E[Y_{1t} - Y_{10}] - E[Y_{0t} - Y_{00}]$. We normalize $\lambda_0 = \beta_0 = 0$ (with two groups, the number of coefficients we can identify is two times the number of time periods, so if we keep the group effects α_g, we need to drop on β_t and one λ_t). Under this normalization, β_t for $t < 0$ measures violations from common trends, since $\beta_t = E[Y_{1t}(0) - Y_{10}(0)] - E[Y_{0t}(0) - Y_{00}(0)]$, which equals zero under eq. (2). For $t \geq 1$, β_t measures the dynamic effect of the treatment.

We can implement eq. (3) by simply running a regression of the outcome on group fixed effects (with repeated cross-sections, with panel data we can use unit fixed effects), time fixed effects with $\mathbb{1}\{t = 0\}$ excluded, and their interaction. We then test the joint significance of the coefficients on the leads to treatment adoption, β_t for $t < 0$, using an F-test. It is also common to show the estimates graphically to visually assess the common trends assumption.

In some cases, we may want to exclude or collapse some of the pre-treatment indicators $\mathbb{1}\{g = 1, t \neq 0\}$. Say with $T_0 = -3$, we may include $\mathbb{1}\{g = 1, t < -1\}$ and $\mathbb{1}\{g = 1, t = 1\}$, or perhaps only include $\mathbb{1}\{g = 1, t = 1\}$. Similarly, if we think that the dynamic effect of the treatment is constant after some time period, we can collapse the post-treatment indicators. Say with $T_1 = 3$, we may include $\mathbb{1}\{g = 1, t = 1\}$ and $\mathbb{1}\{g = 1, t > 1\}$.

The nice thing about the pretrends test is that it can be done using pre-treatment data alone.

Table 1: Table 12 from Snow (1855, p. 90).

	Deaths from Cholera	
Water supply	1849	1854
Southwark & Vauxhall	2261	2458
Lamberth	162	37

Total deaths in sub-districts served by different water companies during two Cholera outbreaks in London, in 1849 and 1854.

1.2. *Examples*

Example 1. The first DiD design appears as Table 12 in Snow (1855), reproduced in Table 1. Snow challenged the conventional wisdom that cholera spreads by "bad air": he noted that Lamberth changed its water source away from the Thames river in 1852, while Southwark & Vauxhall did not. Unfortunately, there are no standard errors. ⊠

Example 2 (Card and Krueger 1994). On April 1, 1992, New Jersey (NJ) raised the state minimum wage from $4.25 to $5.05 per hour (the law was passed in 1989). In the meantime, in Pennsylvania (PA) the minimum wage remained constant and equal to $4.25. To study the effect of the minimum wage increase, Card and Krueger (1994) surveyed 410 fast-food restaurants in NJ and 7 counties in eastern PA (Burger King, Wendy's, KFC, and Roy Rogers), in February 1992, and again in November 1992. Employment contracted in NJ, but so it did in PA, due to an upward trend in unemployment in 1991–1993 in the mid-Atlantic. Table 2 shows the results from regressing various outcomes of interest on a NJ dummy, with the primary outcome being full-time equivalent (FTE) employment. One can tell a coherent story based on these results. However, these results were subsequently questioned in a comment by Neumark and Wascher (2000), who find the opposite effect on employment using administrative payroll data obtained from a sample of 235 restaurants in NJ and PA, drawn from the same geographic areas and the same chains (and hence most overlapping substantially with the Card and Krueger sample). In a response to this comment, Card and Krueger (2000), use Bureau of Labor Statistics data to plot employment in fast-food restaurants in NJ and PA relative to February 1992 levels, reproduced in Figure 2. More pre-intervention data would have been useful, but, nonetheless, the figure is not encouraging. ⊠

Example 3 (Meyer, Viscusi, and Durbin 1995). On July 15, 1980, Kentucky (KY) raised the maximum benefit paid by workers' compensation (medical and cash benefits due to temporary total disability as a result of on-the-job injury) from $131 to $217 per week. A similar change was passed in Michigan (MI) on January 1, 1982, with the maximum benefit going from $181 to $307 per week. Low earners, who were not affected by this change, provide a natural control group for high earners, who were (see Figure 3). Meyer, Viscusi, and Durbin (1995) use a repeated cross-section to look at the effect of

Table 2: Replication of Card and Krueger (1994).

	FTE (1)		Months to raise (2)		Price of entrée (3)	
NJ	2.75	(1.34)	2.51	(2.15)	0.08	(0.04)
Constant	−2.28	(1.25)	1.26	(1.96)	−0.03	(0.04)
Observations	384		321		375	
R^2	0.015		0.005		0.007	

Notes: FTE: Change in FTE employment. Months to raise: Months to first salary raise.

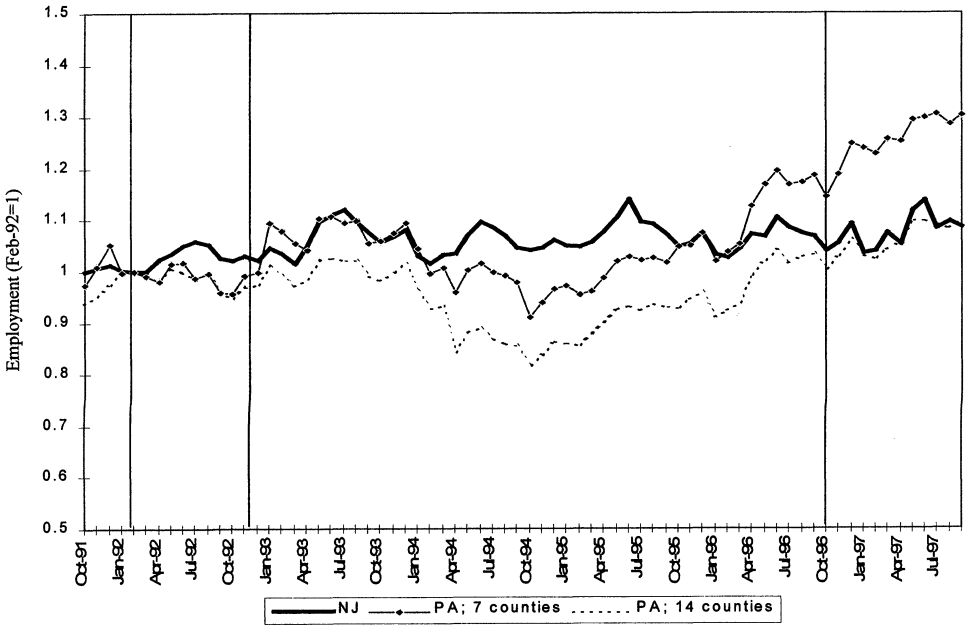

Figure 2: Figure 2 from Card and Krueger (2000). Vertical lines indicate dates of original Card and Krueger survey and another minimum wage increase in October 1996

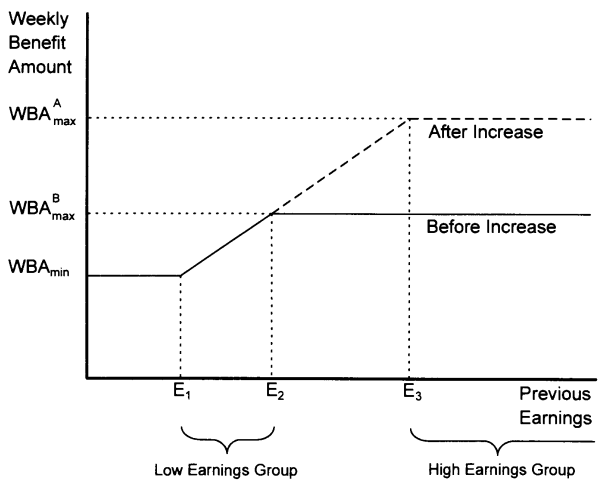

Figure 3: Figure 1 from Meyer, Viscusi, and Durbin (1995). Temporary total benefit schedule before and after an increase in the maximum weekly benefit.

this change on the duration of injury claims benefit. The key results are replicated in Table 3. We see that there is a significant effect on log duration in the larger KY sample, but not much action in terms of effect of duration. ⊠

Example 4 (Donohue and Wolfers 2005). Compelling DiD analyses show observations for a period long enough to discern the underlying trends, with attention focused on how deviations from trend relate to changes in policy. A telling graph in this vein is Figure 4, reproduced from Donohue and Wolfers (2005), showing that changes in death penalty have little effect on trends in homicide rates. ⊠

1.3. *Changes-in-changes*

The common trend assumption not invariant to nonlinear transformations of the dependent variable: if it holds in levels, it doesn't hold in logs[1]. This makes it important that we choose the outcome variable carefully.

To address this issue, Athey and Imbens (2006) provide an alternative to the basic DiD model, called the changes-in-changes (CiC) model. Suppose that $Y_{gt}(0) = h_t(U_g)$, with (i) h strictly increasing in the unobservable u; (ii) the distribution of the unobservable doesn't vary over time within groups $U_g \mid t = 0 \sim U_g \mid t = 1$; an (iii) and overlap condition holds: the support of U_1 is contained in its support given U_0. Together, these assumptions imply that the relative ranking of treatment and control units is stable over time: if control units account for 95% of top earners in period 0, they would also have

1. This is not quite right: the common trends is insensitive to functional form assumptions if we can partition the population into a fraction θ for whom the untreated potential outcome depends only on group membership, but not time, and a fraction $1 - \theta$ for whom it depends only on time, but not group membership. Arguably, this is a rather special case. See Roth and Sant'Anna (2023).

Table 3: Replication of Meyer, Viscusi, and Durbin (1995).

	Mean duration (weeks)				Mean of log duration			
	KY		MI		KY		MI	
	(1)		(2)		(3)		(4)	
Panel A: DiD								
High × After	0.95	(1.28)	1.96	(3.97)	0.19	(0.07)	0.19	(0.16)
High	4.91	(0.88)	3.82	(2.50)	0.26	(0.05)	0.17	(0.11)
After	0.77	(0.51)	2.69	(1.90)	0.01	(0.04)	0.10	(0.08)
Constant	6.27	(0.30)	10.96	(1.09)	1.13	(0.03)	1.41	(0.06)
Panel B: CiC								
ATT (upper bound)	1.08	(1.61)	2.37	(4.43)	0.58	(0.16)	0.34	(0.16)
ATT (lower bound)	0.07	(1.60)	1.30	(4.48)	0.14	(0.13)	0.02	(0.16)
Observations	5626		1524		5626		1524	

Notes: FTE: Change in FTE employment. Months to raise: Months to first salary raise. CiC refers to the changes-in-changes estimator discussed in Section 1.3.

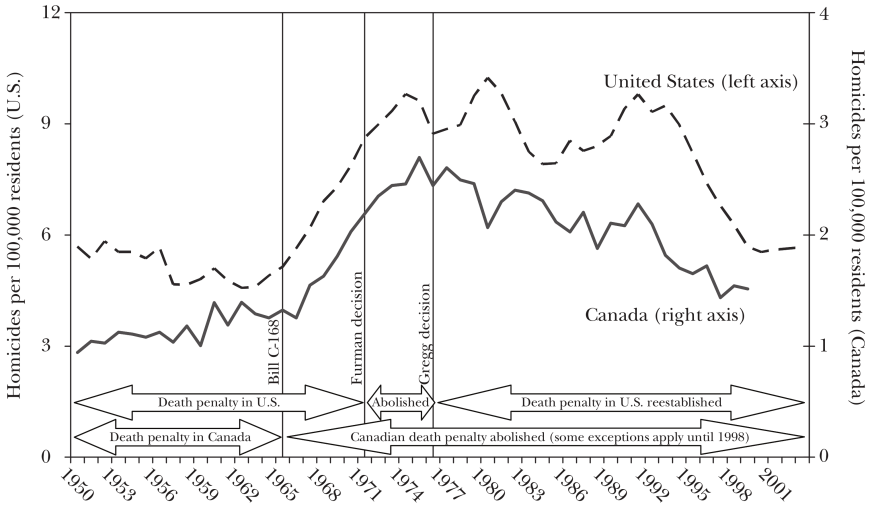

Figure 4: Figure 2 from Donohue and Wolfers (2005). Homicide rates and the death penalty in the United States and Canada.

accounted for 95% of them in period 1 in absence of the intervention (this is similar to the rank invariance assumption that Chernozhukov and Hansen (2005) make in a different context).

The advantage of this setup is that we can now devise an identification strategy that's invariant to transformations of the outcome; the cost is that we're restricting the whole distribution of $Y_{gt}(0)$, not just its mean. The CiC model is useful as a robustness check in cases where the means of the treated and control groups are very different, when we worry about the fragility of the mean restriction imposed by the DiD model.

Theorem 3 (Athey and Imbens 2006). Let F_{gt} denote the cumulative distribution function (CDF) of Y_{gt}. Then, under assumptions (i), (ii), and (iii) above,

$$E[Y_{11}(0)] = E[F_{01}^{-1}(F_{00}(Y_{10}))], \tag{4}$$

Proof. We have

$$F_{0t}(h_t(u)) = P(h_t(U_0) \leq h_t(u)) = P(U_0 \leq u),$$

where the second equality uses the strict monotonicity of u. Thus, $F_{01}(h_1(u)) = F_{00}(h_0(u))$, so that

$$h_1(u) = F_{01}^{-1}(F_{0,0}(h_0(u))), \qquad h_0(u) \in \text{support}(Y_{00}).$$

This is the period 1 outcome for someone with period 0 outcome equal to $y = h_0(u)$. Since the distribution of U_g is time-invariant, this implies (4) as claimed, provided that $\text{support}(Y_{10}) \subseteq \text{support}(Y_{00})$, which holds by assumption (iii). □

We can estimate eq. (4) using empirical distributions and sample averages. In particular, let $\mathbf{Y}_{gt}$ denote the vector of outcomes for group g and time t, and let $\hat{F}_{gt}$ denote its empirical CDF. Define

$$\hat{F}_{01}^{-1}(q) = \min\{y \in \text{support}(\mathbf{Y}_{01}) \colon \hat{F}_{01}(y) \geq q\}.$$

(I am giving an explicit formula here, because this is not the default way of computing the sample quantile function in many software packages). Then we can estimate the ATT as

$$\frac{1}{n_{11}} \sum_{y \in \mathbf{Y}_{11}} y - \frac{1}{n_{10}} \sum_{y \in \mathbf{Y}_{10}} \hat{F}_{01}^{-1}(\hat{F}_{00}(y)), \tag{5}$$

where n_{gt} is the length of the $\mathbf{Y}_{gt}$ vector.

Here the transform $F_{01}^{-1}(F_{00}(y))$ gives the second-period outcome for an individual with an unobserved component u such that $h_0(u) = y$. The logic is as follows. Take an individual from group 1 in period 0 with outcome equal to y. Compute their quantile $q = F_{00}(y)$ if they were in the control group. This quantile wouldn't change in period 1, so their counterfactual outcome in period 1 would need to correspond to this outcome: $y^c = F_{01}^{-1}(F_{00}(y))$. See Figure 5. Since we're matching quantiles, the exercise is invariant to monotone transformations of the outcome. In contrast, in a DiD model, the counterfactual outcome is estimated as $y^c = y + E[Y_{01}] - E[Y_{00}]$, which is not invariant to monotone transformations of Y.

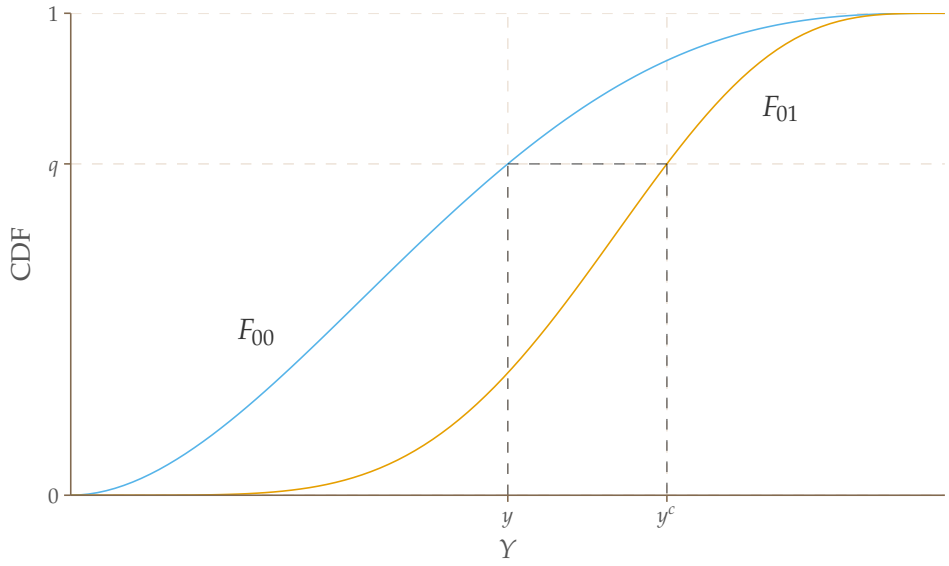

Figure 5: Computation of the counterfactual outcome y^c in period 1 for an individual in group 1 with outcome y in period 0.

Things to note:

- See the paper for inference results, for an extension to a multi-period and multi-group case (here the paper suggests estimating the potentially many treatment effects $E[Y_{gt}(1) - Y_{gt}(0)]$).
- Since we only need the distribution of U_g to be time-invariant, with panel data, this accommodates a fixed effect structure, $Y_{it}(0) = h_t(\eta_i + \epsilon_{it})$, where ϵ_{it} is an idiosyncratic shock with the same distribution across periods. Estimation is the same as in the repeated cross-section case, inference will be different.
- If the outcome is discrete, then the model doesn't quite make sense, since h can't be strictly increasing in u, if u is continuously distributed. If we weaken assumption (i) to only require h to be weakly increasing, then it turns out that the ATT is only partially identified, and eq. (5) estimates the lower endpoint of the identified set. To estimate the upper endpoint, replace $\hat{F}_{00}(y)$ in eq. (5) with $1 - \hat{F}_{-00}(-y)$, where $\hat{F}_{-00}$ is the empirical CDF of $-\mathbf{Y}_{00}$.

  Question 1. Can you see this result from Figure 5? (imagine discrete F_{00} that jumps at y)

- The support condition ensures that the quantile $q' = F_{00}(y)$ is well-defined. If $\mathbf{Y}_{01}$ is not in the support of Y_{00}, one option is to trim, similarly to trimming in average treatment effect (ATE) estimation under unconfoundedness, or to bound the ATT (**Homework question:** what are the bounds if the outcome is not bounded?)

- Athey and Imbens (2006) propose to accommodate covariates via the model $y = h_t(u) + x'\beta$.

Example 3 (continued). Panel B in Table 3 applies the estimator to the Meyer, Viscusi, and Durbin (1995) data, accounting for the discreteness in the outcome variable. Even though there are over 130 support points, the bounds are quite wide, and the DiD estimates from both a levels and a logs specification are included in the bounds. ⊠

1.4. *Conditional common trends*

In some cases, it may be more plausible to assume conditional common trends,

$$E[Y_{11}(0) - Y_{10}(0) \mid X_1 = x] = E[Y_{01}(0) - Y_{00}(0) \mid X_0 = x],$$

conditional on some time-invariant covariate X (Abadie 2005, about 2k citations). Here X just needs to be known at time 0; it could, in principle, correspond to past outcomes. Note this assumption does not imply the common trends assumption unless the distribution of X is the same between the treated and untreated (i.e. $X_1 \sim X_0$); and common trends obviously doesn't imply conditional common trends. So it's neither a weaker nor a stronger assumption. Under this assumption, one could estimate DiD conditional on X, and average the estimates, similarly to a regression approach to estimating the ATT under unconfoundedness.

Remark 4 (Treatment effects under unconfoundedness). Recall that, with cross-section data, if we assume a version of unconfoundedness, $E[Y(0) \mid D, X] = E[Y(0) \mid X] =: \mu_0(X)$, and let $p(X) = P(D = 1 \mid X)$ denote the propensity score, then

$$\begin{aligned} E[Y(1) - Y(0) \mid D = 1] &= E[Y(1) \mid D = 1] - E[E[Y \mid D = 0, X] \mid D = 1] \\ &= E[Y(1) \mid D = 1] - \frac{1-p}{p} E\left[\frac{p(X)Y}{1-p(X)} \mid D = 0\right]. \end{aligned}$$

So we can either use regression or propensity score weighting to estimate the ATT.

Proof. Here the first equality uses iterated expectations, and the fact that by unconfoundedness, $E[Y(0) \mid D = 1, X] = E[Y(0) \mid D = 0, X] = E[Y \mid D = 1, X] = \mu_0(X)$. The second uses the fact that by applying Bayes formula twice, $f_{X|D_1=1}(x) = p(x) f_X(x) / P(D_1 = 1) = \frac{1-p}{p} \frac{p(x)}{1-p(x)} f_{X|D_1=0}(x)$, so that for any function g, we have $E[g(X) \mid D = 1] = \frac{1-p}{p} E[g(X)p(X)/(1-p(X)) \mid D = 0]$. Thus, letting $g(x) = E[Y \mid X = x, D = 0]$, we have

$$\begin{aligned} E[Y(0) \mid D = 1] &= E[E[Y \mid X, D = 0] \mid D = 1] \\ &= \frac{1-p}{p} E\left[E[Y \mid X, D = 0] \frac{p(X)}{1-p(X)} \mid D = 0\right] = \frac{1-p}{p} E\left[\frac{p(X)Y}{1-p(X)} \mid D = 0\right]. \end{aligned}$$ □

In our setting, we can apply Remark 4 with $Y_{g1} - Y_{g0}$ playing the role of Y, and with

$p(X) = P(G = 1 \mid X)$. This delivers an analogous result:

$$\begin{aligned} E[Y_{11}(1) - Y_{11}(0)] &= E[Y_{11} - Y_{10}] - E[E[Y_{01} - Y_{00} \mid X_0 = X_1]] \\ &= E[Y_{11} - Y_{10}] - E\left[\frac{(Y_{01} - Y_{00})p(X_0)/p}{(1 - p(X_0))/(1 - p)}\right]. \end{aligned}$$

Intuitively, the propensity score weighting gives more weight to observations in the control group that look more like treatment group units. To implement it, we need to either estimate $p(X)$ nonparametrically, or estimate the conditional mean $E[Y_{01} - Y_{00} \mid X_0]$ nonparametrically.

- Note that adding the covariates linearly in the regression does not estimate the ATT in general (the equation in the display after eq. (8) in Abadie (2005)). **Homework question:** what do we estimate? What assumptions ensure the estimate has a causal interpretation? (Hint: think back to notes on ordinary least squares (OLS)).
- By the analogy with estimation of the ATT under unconfoundedness, one could also use other estimation strategies, such as matching.

1.5. *Fuzzy designs*

In many applications (about 10% of DiD papers according to a count in de Chaisemartin and D'Haultfœuille 2018), the treatment can be thought of as an encouragement or a subsidy. One could of course estimate the effect of the encouragement, which corresponds to the intent-to-treat (ITT) effect. Can we divide the ITT estimate by the first stage as in fuzzy regression discontinuity (RD), or instrumental variables (IV) to estimate the causal effect of a treatment of interest?

Example 5 (Duflo 2001). In 1973–74, the Indonesian government launched a major primary school construction program. Because the program intensity was related to 1972 enrollment rates, which vary across regions, we classify individuals into two groups $g = 0, 1$ depending on whether the individual was born in a region with high construction. There are two cohorts: cohort $t = 0$ consists of men aged 12–17 in 1974, who were out of primary school by the time the program launched. Cohort $t = 1$ consists of men aged 2–6 in 1974. Simple DiD estimates suggest an effect of 0.12 on years of education, and 0.026 on log wages (Table 3). Using the $\{t = 1\} \times \{g = 1\}$ interaction as an instrument for education, the implied effect on the return to education is quite high, about 19.5% (In Table 7 in the paper, it's quite a bit lower, 0.0752, I am not sure why). Assuming school construction affects wages only through years of schooling, under what conditions can we interpret these estimates as causal effects? ⊠

Denote the treatment of interest by D, and denote the encouragement design by $Z_{gt} = \mathbb{1}\{g = t = 1\}$. Let $Y_{gt}(d)$ denote the potential outcomes, and $D_{gt}(z)$ the potential

treatments.[2] The IV estimand is given by

$$\beta_W = \frac{E[Y_{11} - Y_{10}] - E[Y_{01} - Y_{00}]}{\pi}, \qquad \pi = E[D_{11} - D_{10}] - E[D_{01} - D_{00}].$$

What is this estimating? de Chaisemartin and D'Haultfœuille (2018) make the common trends assumption (2), in analogy to the sharp case where $D = Z$. Let $\tau_{gt} = Y_{gt}(1) - Y_{gt}(0)$, so that $Y_{gt} = Y_{gt}(0) + D_{gt}\tau_{gt}$. The $Y_{gt}(0)$ intercepts then cancel out by the common trends assumption, and we obtain

$$\beta_W = \frac{E[\tau_{11}D_{11}] - E[\tau_{10}D_{10}] - (E[\tau_{01}D_{01}] - E[\tau_{00}D_{00}])}{\pi}$$

Since $E[\tau_{gt}D_{dt}] = E[\tau_{gt} \mid D_{gt} = 1]E[D_{gt}]$, this is a non-convex combination of ATTs for different groups: the weights sum to one by definition of π, but some of them are negative. Bad news if there is heterogeneity in treatment effects. de Chaisemartin and D'Haultfœuille (2018) try to save the day by restricting heterogeneity in treatment effects, and by making no-defier type assumptions $D_{11}(1) \geq D_{10}(0)$, but we have to be able to conceptualize moving individuals through time to make sense of such assumptions. Regardless of the interpretation issues, they have limited success with this strategy.

de Chaisemartin and D'Haultfœuille (2018) propose an alternative approach based on the assumption that $E[Y_{g1}(d) - Y_{g0}(d) \mid D_{g0}(0) = d]$ doesn't depend on g. However, this approach only leads to bounds when the share of treated in the control group changes. They also consider an approach based on adapting the CiC model.

Remark 5. One could reach a very different conclusion about the attractiveness of the β_W estimand if we set things up differently. Instead of assuming (2), let us impose common trends on the treatment, $E[D_{11}(0) - D_{10}(0)] = E[D_{01}(0) - D_{00}(0)]$ Then $\pi = E[D_{11}(1) - D_{11}(0)]$ by arguments as in eq. (1). Also, make a common trends assumption on the outcomes in absence of the subsidy,

$$E[Y_{11}(D_{11}(0)) - Y_{10}(D_{10}(0))] = E[Y_{01}(D_{01}(0)) - Y_{00}(D_{00}(0))].$$

This allows us to interpret the ITT DiD regression as a causal effect. By arguments analogous to (1),

$$E[Y_{11} - Y_{10}] - E[Y_{01} - Y_{00}] = E[Y_{11}(D_{11}(1)) - Y_{11}(D_{11}(0))].$$

Make the monotonicity assumption that $D_{11}(1) \geq D_{11}(0)$ (now it's an assumption about what would happen if we canceled the subsidy, rather than a no defiers assumption across time). Then $E[Y_{11}(D_{11}(1)) - Y_{11}(D_{11}(0))] = E[(D_{11}(1) - D_{11}(0))\tau_{11}] = E[\tau_{11} \mid D_{11}(1) > D_{11}(0)]E[D_{11}(1) - D_{11}(0)]$, so that

$$\beta_W = E[\tau_{11} \mid D_{11}(1) > D_{11}(0)].$$

2. The setup in de Chaisemartin and D'Haultfœuille (2018) instead works with potential treatments that consider moving an individual to period t. That does not seem feasible...

So there is no issue...

Question 2. What is the takeaway from all this?

Research Question. This setup is a special case of a model-based approach to IV. There are presumably things we can say in general here. What are they? ⊠

2. MULTIPLE PERIODS (AND GROUPS)

Suppose now that there are more than two periods. To simplify the discussion, we analyze the data at the individual level, $i = 1, \ldots, n$. Let $t = 1, \ldots, T$ index time. We abstract from individual-level controls. The notation $Y_{it}(d)$ and D_{it} is as before. Unlike the previous section, we don't make any restrictions on when the D_{it} may be zero or non-zero. Later, we'll consider particular designs for D_{it}. As in the case with two groups, we consider D_{it} to be a deterministic function of the individual and time. Equivalently, we can think of the exercise as conditioning on D_{it}. Make the common trends assumption:

$$E[Y_{it}(0) - Y_{i0}(0)] = \lambda_t, \qquad t \geq 1, i = 1, \ldots, n. \tag{6}$$

Remark 6. Since the treatment is non-random, the treatment assignment is exogenous, in the sense discussed in our OLS lecture. To see the connection, let us map the notation to the usual notation for estimating treatment effects under unconfoundedness. Let Y denote the outcome of a randomly picked observation, and let α and λ denote the observation's fixed effect and time effect. Let $Y(d)$ denote the potential outcomes. Think of (α, λ) as the controls/covariates. What makes this setup somewhat special is that the treatment D is a deterministic function of the controls. Using this notation, eq. (6) is equivalent to the model-based assumption (Assumption 6 in the lecture note on OLS)

$$E[Y(d) \mid D, \alpha, \lambda] = E[Y(d) \mid \alpha, \lambda] = \alpha + \lambda \tag{7}$$

In our setup, the first equality ("exogeneity") is trivial (why?), and the content comes from the second equality, that it suffices to control for the individual effect and the time effect in an additive manner using two-way fixed effects (2WFE): we don't need to control for a more complicated, non-linear function of the group and time dummies.

Under this setup, a DiD comparison using any two groups and two time periods comparison yields a causal comparison. In particular, since $Y_{it} = \tau_{it} D_{it} + Y_{it}(0)$, where $\tau_{it} = Y_{it}(1) - Y_{it}(0)$, the $Y_{it}(0)$ terms cancel, and we obtain:

$$E[Y_{it} - Y_{is} - (Y_{jt} - Y_{jt})] = D_{it} E[\tau_{it}] - D_{is} E[\tau_{is}] - (D_{jt} E[\tau_{jt}] - D_{js} E[\tau_{js}]), \tag{8}$$

So if $D_{it} = 1$ and $D_{is} = D_{jt} = D_{js} = 0$, then this comparison estimates an ATT. The question is how we should aggregate these ATTs. The standard approach is to realize

that eq. (6) implies

$$E[Y_{it}] = \alpha_i + \lambda_t + D_{it}E[\tau_{it}], \quad \alpha_i = E[Y_{i0}(0)], \tag{9}$$

$\lambda_0 = 0$, and $\lambda_t = E[Y_{0t} - Y_{00}]$. If the treatment effect $\tau_{it} = \tau$ is constant, we can estimate τ using a 2WFE regression. If the variance of the residual $Y_{it} - \alpha_i - \lambda_t - D_{it}\tau$ is homoskedastic, then this estimator has the usual optimality properties by the Gauss-Markov theorem.

2.1. *Heterogeneous treatment effects*

Suppose now that τ_{it} is not constant. What is the 2WFE regression estimating? By Remark 6, we can think of the problem as a problem of inference under heterogeneous treatment effects using the model-based approach to identification. By the results from the lecture on OLS, we estimate a weighted average of treatment effects $E[\tau_{it}]$, with weights

$$\lambda_{it} = \frac{\ddot{D}_{it}D_{it}}{\sum_{i,t}\ddot{D}_{it}D_{it}}, \tag{10}$$

where

$$\ddot{D}_{it} = D_{it} - \overline{D}_i - \frac{1}{n}\sum_j (D_{jt} - \overline{D}_j), \tag{11}$$

is the residual from regressing D_{it} onto unit and time fixed effects. Here where $\overline{D}_i = \frac{1}{T}\sum_{t=1}^{T} D_{it}$. This is Theorem 1 in de Chaisemartin and D'Haultfœuille (2020).

If $T = 2$, and no units are treated initially, we get $\ddot{D}_{i2}D_{i2} = D_{it}(1 - 1/2 - \pi_1 + \pi_1/2) = D_{i2}(1 - \pi_1)/2$, where π_1 is the fraction of units treated in period 2, so we estimate the ATT. The trouble is that, as discussed in the OLS lecture, the weights λ_{it} are in general negative for some groups and time periods. So we are estimating a linear, but not a convex combination of the (i,t)-level ATTs. The first differences estimator suffers from a similar problem. Again, the issue stems from the fact that the model-based assumption in eq. (7) does not generally guarantee that controlling for the covariates linearly in a regression will yield an estimate of a convex combination of treatment effects. This also implies, for instance, that using, say an interactive fixed effects specification will not fix the problem.

To provide additional intuition, Goodman-Bacon (2018, Theorem 1) decomposes the 2WFE estimator $\hat{\beta}_{\text{2WFE}}$ into a weighted average of different 2×2 DiD estimands, assuming an event study design, or, equivalently a staggered adoption design (i.e. the treatments never switch off). The next result generalizes this decomposition to a setting where groups may drop the treatment, and considerably simplifies resulting expression.

Lemma 7. Let $\pi_{ij}^{ab} = \frac{1}{T}\sum_t \mathbb{1}\{D_{it} = a\}\mathbb{1}\{D_{jt} = b\}$ *denote the fraction of time* $D_{it} = a$ *and*

$D_{jt} = b$, *and let* $\Delta_{ij}^{ab} = \frac{1}{T}\sum_t \mathbb{1}\{D_{it} = a\}\,\mathbb{1}\{D_{jt} = b\}(Y_{it} - Y_{jt})/\pi_{ij}^{ab}$. *Then*

$$\hat{\beta}_{\text{2WFE}} = \frac{\sum_{ij}(\pi_{ij}^{10}\pi_{ij}^{11}(\Delta_{ij}^{10} - \Delta_{ij}^{11}) + (\pi_{ij}^{00} + 2\pi_{ij}^{01})\pi_{ij}^{10}(\Delta_{ij}^{10} - \Delta_{ij}^{00}))}{\sum_{ij}(\pi_{ij}^{01}\pi_{ij}^{11} + (\pi_{ij}^{00} + 2\pi_{ij}^{01})\pi_{ij}^{10})}. \tag{12}$$

Proof. By the Frisch–Waugh–Lovell (FWL) theorem, the 2WFE estimator may be written as

$$\hat{\beta}_{\text{2WFE}} = \frac{\sum_{it}\ddot{D}_{it}Y_{it}}{\sum_{it}\ddot{D}_{it}^2} = \frac{\frac{1}{N}\sum_{ijt}(D_{it} - \overline{D}_i)(Y_{it} - Y_{jt})}{\sum_{it}\ddot{D}_{it}^2},$$

where the second equality follows by switching the order of summation in the identity

$$\sum_{it}\ddot{D}_{it}Y_{it} = \sum_{it}(D_{it} - \overline{D}_i)Y_{it} - \frac{1}{N}\sum_{ijt}(D_{jt} - \overline{D}_j)Y_{it}.$$

Since $\overline{D}_i = \pi_{ij}^{10} + \pi_{ij}^{11}$, the numerator of $\hat{\beta}_{\text{2WFE}}$ may be written as

$$\begin{aligned}
\sum_{it}\ddot{D}_{it}Y_{it} &= \frac{1}{N}\sum_{ijt}(D_{it} - \pi_{ij}^{10} - \pi_{ij}^{11})(Y_{it} - Y_{jt}) \\
&= \frac{1}{N}\sum_{ijt}(D_{it}(\pi_{ij}^{00} + \pi_{ij}^{01}) - (1 - D_{it})(\pi_{ij}^{10} + \pi_{ij}^{11}))(Y_{it} - Y_{jt}) \\
&= \frac{T}{N}\sum_{ij}((\pi_{ij}^{00} + \pi_{ij}^{01})(\pi_{ij}^{11}\Delta_{ij}^{11} + \pi_{ij}^{10}\Delta_{ij}^{10}) - (\pi_{ij}^{10} + \pi_{ij}^{11})(\pi_{ij}^{01}\Delta_{ij}^{01} + \pi_{ij}^{00}\Delta_{ij}^{00})),
\end{aligned}$$

where the second line uses $1 = \pi_{ij}^{00} + \pi_{ij}^{01} + \pi_{ij}^{10} + \pi_{ij}^{11}$. Rearranging the expression yields

$$\begin{aligned}
\sum_{it}\ddot{D}_{it}Y_{it} &= \frac{T}{N}\sum_{ij}(\pi_{ij}^{00}\pi_{ij}^{11}(\Delta_{ij}^{11} - \Delta_{ij}^{00}) + \pi_{ij}^{01}\pi_{ij}^{11}(\Delta_{ij}^{11} - \Delta_{ij}^{01})) \\
&\quad + \frac{T}{N}\sum_{ij}(\pi_{ij}^{00}\pi_{ij}^{10}(\Delta_{ij}^{10} - \Delta_{ij}^{00}) + \pi_{ij}^{01}\pi_{ij}^{10}(\Delta_{ij}^{10} - \Delta_{ij}^{01})), \\
&= \frac{T}{N}\sum_{ij}(\pi_{ij}^{00}\pi_{ij}^{11}(\Delta_{ij}^{11} - \Delta_{ij}^{00}) + \pi_{ij}^{01}\pi_{ij}^{11}(\Delta_{ij}^{11} - \Delta_{ij}^{01}) + (\pi_{ij}^{00} + 2\pi_{ij}^{01})\pi_{ij}^{10}(\Delta_{ij}^{10} - \Delta_{ij}^{00})) \\
&= \frac{T}{N}\sum_{ij}(\pi_{ij}^{01}\pi_{ij}^{11}(\Delta_{ij}^{11} - \Delta_{ij}^{01}) + (\pi_{ij}^{00} + 2\pi_{ij}^{01})\pi_{ij}^{10}(\Delta_{ij}^{10} - \Delta_{ij}^{00})),
\end{aligned}$$

where the second line uses $\Delta_{ij}^{10} - \Delta_{ij}^{01} = \Delta_{ij}^{10} - \Delta_{ij}^{00} + (\Delta_{ji}^{10} - \Delta_{ji}^{00})$, and the third line uses and uses the symmetry property $\Delta_{ij}^{11} - \Delta_{ij}^{00} = -(\Delta_{ji}^{11} - \Delta_{ji}^{00})$. Applying the same arguments to the denominator of $\hat{\beta}_{\text{2WFE}}$ and switching the i and j index in the first sum yields the result. □

The lemma says that the estimator is a weighted average of two types of DiD comparisons:

1. $\Delta_{ij}^{10} - \Delta_{ij}^{00}$. This is the usual DiD comparison, comparing i and j over periods where i is treated and j is not, vs neither of them is treated. By eq. (8), is an unbiased estimate of the ATT for group g over the periods $D_{gt} = 1, D_{ht} = 0$.

2. $\Delta_{ij}^{10} - \Delta_{ij}^{11}$. This is a "forbidden comparison": j serves as a control for i, but we're not comparing to an untreated period, but to a period in which both are treated.

Two ways of thinking about this. To parse through them, let $\tau^{ab}_{i|j} = \frac{1}{T}\sum_t \mathbb{1}\{D_{it} = a, D_{jt} = b\}E[\tau_{it}]/\pi^{ab}_{ij}$. First, if we only assume eq. (6), then by eq. (8), this estimates $\Delta^{10}_{ij} - \Delta^{11}_{ij} = \tau^{10}_{i|j} - \tau^{11}_{i|j} + \tau^{11}_{j|i}$, and this is how we wind up with negative weights on treatment effects. Alternatively, if we assume that eq. (6) also holds for $Y(1)$, then this estimates the ATE for the untreated, $E[\tau^{01}_{j|i}]$ (show this!). However, assuming common trends for both $Y(0)$ and $Y(1)$ restricts treatment effect heterogeneity, since it implies (why?) $E[\tau_{it}] = \kappa_i + \gamma_t$.

Thus, in DiDs designs, to interpret β_{2WFE} as a weighted average of treatment effects, we need to assume that eq. (6) holds for $Y(1)$ as well. Otherwise, the 2WFE estimand doesn't have a causal interpretation whenever the fitted values from the propensity score regression exceed one.

Example 6. Suppose that there are three time periods, $t = 1, 2, 3$, and two equal-sized groups. Group 1 starts being treated at $t = 2$, but group 0 only in the last period, $t = 3$. Then $\ddot{D}_{03} = 1/6$, while $\ddot{D}_{12} = 1/3$, and $\ddot{D}_{13} = -1/6$. So by eq. (10), we're estimating $\frac{1}{2}\tau_{03} + \tau_{12} - \frac{1}{2}\tau_{13}$, where $\tau_{gt} = E[\tau_{it} \mid g(i) = g]$ are group-specific treatment effects.

Alternatively, using Lemma 7, we're estimating the average of two DiD estimators, one comparing $\overline{Y}_{12} - \overline{Y}_{02} - (\overline{Y}_{11} - \overline{Y}_{01})$, where $\overline{Y}_{gt}$ are group means, which is unbiased for τ_{12}, with weight $1/2$, and the other one comparing $\overline{Y}_{12} - \overline{Y}_{02} - (\overline{Y}_{13} - \overline{Y}_{03})$ with weight $1/2$, which is unbiased for $\tau_{12} - \tau_{13} + \tau_{03}$. If we assume parallel trends on the treatment effects, so that $\tau_{g2} - \tau_{g3}$ is constant, this second term equals τ_{02}. ⊠

What to do? There have been many proposals (e.g. Callaway and Sant'Anna 2021; Sun and Abraham 2021; Imai and Kim 2021; Borusyak, Jaravel, and Spiess 2024).

1. We can compute these weights: a good idea to always to this as a robustness check!

2. Either estimate the ATT for the largest subpopulation that we can, or else pick the weights to combine the 2×2 DiD estimates efficiently.

3. If we worry about dynamic treatment effects, we can estimate the (weighted) ATT for units who just switch into treatment (suggested by Imai and Kim 2021), switched two periods ago, etc.

4. de Chaisemartin and D'Haultfœuille (2020) suggest imposing common trends on $Y_{gt}(1)$ as well, and estimate the average treatment effect for "switchers" (ATT for units just entering treatment, plus the ATE for the untreated units who just dropped treatment). But by the discussion above, if we do make this additional assumption, then there is no need to change the estimator! This point was made in Fabre (2023).

Research Question. Is there some attractive default? ⊠

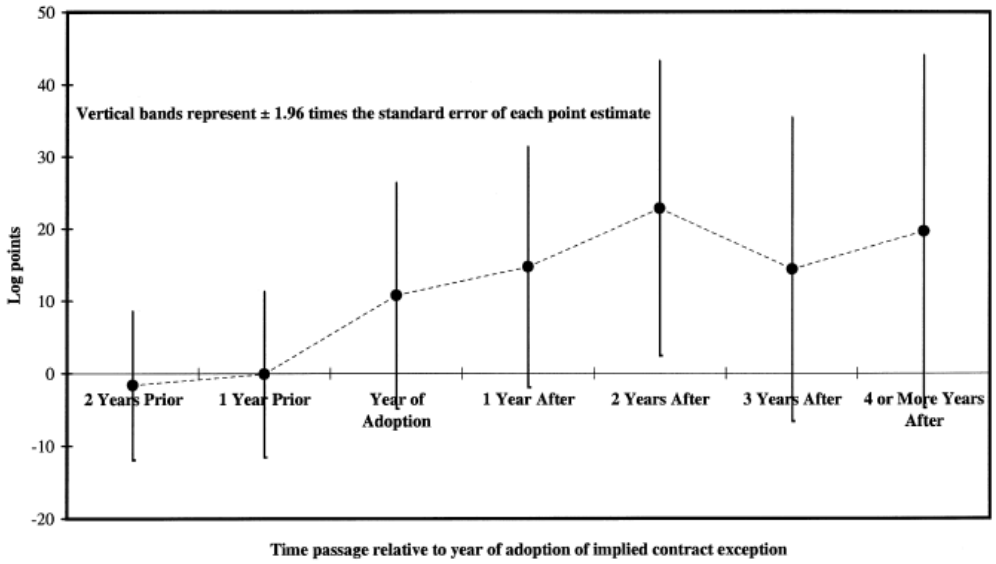

Figure 6: Figure 3 from Autor (2003), plotting coefficients from a regression of log temporary employment on a state-specific linear trend, and leads and lags of the adoption of implied contract exception by a state.

2.2. *Pretrends in event study designs*

As we discussed in Section 1, in staggered adoption designs (or equivalently, event study designs), it is common to regress Y_{it} onto leads and lags of when the treatment was adopted, and plot the coefficients to assess pretrends. Figure 6 reproduces such a plot from Autor (2003), who is interested in the effect of passing an implied contract exception to the employment at will doctrine on temporary help services.[3] Angrist and Pischke (2009) refer to this as a "Granger causality test". As discussed there, the coefficient on the period just before adoption is normalized to zero (if no units are untreated, we need to drop another leads or lag to avoid multicollinearity).

Sun and Abraham (2021) show that this test only works under constant treatment effects—we may then have non-zero pretrends even if the common trends assumption holds (unless we only have two groups as in eq. (3)). The issue is that we're trying to estimate multiple treatments, and doing this using regression where we control for the covariates linearly. But such regressions are not generally robust to treatment effect heterogeneity (Goldsmith-Pinkham, Hull, and Kolesár 2024). How to best conduct such a test in a manner that is robust to heterogeneity in treatment effects is an open question.

In addition, one may worry about the usual power issues with such a pretrend check,

3. As a rule, US labor law allows "employment at will", which means that workers can be fired for just cause or no cause, at the employer's whim. But beginning in 1967, state courts have allowed a number of exceptions to the employment-at-will doctrine, which raised the chance that the employer may face an "unjust dismissal" lawsuit. The implied contract exception prohibits the firing of a worker after an "implied contract" has been established. Such a contract is an expectation of continued employment, which can be created in the form of oral assurances or expectations created by employer handbooks or policies.

and about what to do if the test rejects. Freyaldenhoven, Hansen, and Shapiro (2019) propose to exploit the presence of a covariate that is affected by a confounder we worry about, but not by the policy. This allows for an IV solution to the pretrend problem, under the (strong) assumption that the dynamic relationship between the treatment and the instrument is the same as that between the confounder and the treatment. Rambachan and Roth (2023) propose conducting sensitivity analysis to deviations from the parallel trends assumption.

REFERENCES

Abadie, Alberto. 2005. "Semiparametric Difference-in-Differences Estimators." *The Review of Economic Studies* 72, no. 1 (January): 1–19. https://doi.org/10.1111/0034-6527.00321.

Angrist, Joshua D., and Jorn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton, NJ: Princeton University Press. https://doi.org/10.2307/j.ctvcm4j72.

Ashenfelter, Orley. 1978. "Estimating the Effect of Training Programs on Earnings." *The Review of Economics and Statistics* 60, no. 1 (February): 47–57. https://doi.org/10.2307/1924332.

Athey, Susan, and Guido W. Imbens. 2006. "Identification and Inference in Nonlinear Difference-in-Differences Models." *Econometrica* 74, no. 2 (March): 431–497. https://doi.org/10.1111/j.1468-0262.2006.00668.x.

Autor, David H. 2003. "Outsourcing at Will: The Contribution of Unjust Dismissal Doctrine to the Growth of Employment Outsourcing." *Journal of Labor Economics* 21, no. 1 (January): 1–42. https://doi.org/10.1086/344122.

Borusyak, Kirill, Xavier Jaravel, and Jann Spiess. 2024. "Revisiting Event-Study Designs: Robust and Efficient Estimation." Forthcoming, *Review of Economic Studies* (February). https://doi.org/10.1093/restud/rdae007.

Callaway, Brantly, and Pedro H.C. Sant'Anna. 2021. "Difference-in-Differences with Multiple Time Periods." *Journal of Econometrics* 225, no. 2 (December): 200–230. https://doi.org/10.1016/j.jeconom.2020.12.001.

Card, David, and Alan B. Krueger. 1994. "Minimum Wages and Employment: A Case Study of the Fast Food Industry in New Jersey and Pennsylvania." *American Economic Review* 84, no. 4 (September): 772–793. https://www.jstor.org/stable/2118030.

Card, David, and Alan B. Krueger. 2000. "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania: Reply." *American Economic Review* 90, no. 5 (December): 1397–1420. https://doi.org/10.1257/aer.90.5.1397.

Chernozhukov, Victor, and Christian B. Hansen. 2005. "An IV Model of Quantile Treatment Effects." *Econometrica* 73, no. 1 (January): 245–261. https://doi.org/10.1111/j.1468-0262.2005.00570.x.

de Chaisemartin, Clément, and Xavier D'Haultfœuille. 2018. "Fuzzy Differences-in-Differences." *The Review of Economic Studies* 85, no. 2 (April): 999–1028. https://doi.org/10.1093/restud/rdx049.

———. 2020. "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects." *American Economic Review* 110, no. 9 (September): 2964–2996. https://doi.org/10.1257/aer.20181169.

Donohue, John J., III, and Justin Wolfers. 2005. "Uses and Abuses of Empirical Evidence in the Death Penalty Debate." *Stanford Law Review* 58, no. 3 (December): 791–845.

Duflo, Esther. 2001. "Schooling and Labor Market Consequences of School Construction in Indonesia: Evidence from an Unusual Policy Experiment." *The American Economic Review* 91, no. 4 (September): 975–813. https://doi.org/10.1257/aer.91.4.795.

Fabre, Anaıs. 2023. "Robustness of Two-Way Fixed Effects Estimators to Heterogeneous Treatment Effects." Working paper, Toulouse School of Economics, https://www.tse-fr.eu/sites/default/files/TSE/documents/doc/wp/2022/wp_tse_1362.pdf.

Freyaldenhoven, Simon, Christian B. Hansen, and Jesse M. Shapiro. 2019. "Pre-Event Trends in the Panel Event-Study Design." *American Economic Review* 109, no. 9 (September): 3307–3338. https://doi.org/10.1257/aer.20180609.

Goldsmith-Pinkham, Paul, Peter Hull, and Michal Kolesár. 2024. "On Estimating Multiple Treatment Effects with Regression" (February). arXiv: 2106.05024.

Goodman-Bacon, Andrew. 2018. *Difference-in-Differences with Variation in Treatment Timing*. Working Paper 25018. Cambridge, MA: National Bureau of Economic Research, September. https://doi.org/10.3386/w25018.

Gruber, Jonathan. 1994. "The Incidence of Mandated Maternity Benefits." *American Economic Review* 84, no. 3 (June): 622–641. https://www.jstor.org/stable/2118071.

Imai, Kosuke, and In Song Kim. 2021. "On the Use of Two-Way Fixed Effects Regression Models for Causal Inference with Panel Data." *Political Analysis* 29, no. 3 (July): 405–415. https://doi.org/10.1017/pan.2020.33.

Meyer, Bruce, W. Kip Viscusi, and David Durbin. 1995. "Workers' Compensation and Injury Duration." *American Economic Review* 85, no. 3 (June): 322–340. https://www.jstor.org/stable/2118177.

Neumark, David, and William Wascher. 2000. "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania: Comment." *American Economic Review* 90, no. 5 (December): 1362–1396. https://doi.org/10.1257/aer.90.5.1362.

Rambachan, Ashesh, and Jonathan Roth. 2023. "A More Credible Approach to Parallel Trends." *Review of Economic Studies* 90, no. 5 (September): 2555–2591. https://doi.org/10.1093/restud/rdad018.

Roth, Jonathan, and Pedro H. C. Sant'Anna. 2023. "When Is Parallel Trends Sensitive to Functional Form?" *Econometrica* 91, no. 2 (March): 737–747. https://doi.org/10.3982/ECTA19402.

Snow, John. 1855. *On the Mode of Communication of Cholera.* 2nd ed. London: John Churchhill.

Sun, Liyang, and Sarah Abraham. 2021. "Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects." *Journal of Econometrics* 225, no. 2 (December): 175–199. https://doi.org/10.1016/j.jeconom.2020.09.006.