

An Information Theoretic Extension to conditional differences-in-differences

Notes

- (1) Simulations when $N=25$ had no issues with either ML or GME Logit
- (2) There were issues with "rare selection into treatment". When only 1% of obs was receiving treatment, there were some datasets where no obs received treatment ($\text{so } D_{ij} = 0 \text{ always}$).
I changed the "rare treatment" case to 5%.
- (3) All estimations have an intercept term (although the true intercept might be zero).
- (4) I suppress discussion of multinomial logit. I have removed equations related to $J > 2$ case.
- (5) Eq. 16 has been changed to show its a definition.
- (6) Added more references

The difference-in-differences approach (sometimes abbreviated DiD or diff-in-diff) to causal identification has a long history, perhaps tracing back to work of physician John Snow in identifying the cause of the 1854 London Cholera outbreak (Cunningham 2021). In economics, precursors to diff-in-diff methods began to appear in the literature in the 1950s (for example (Lester 1946) as recounted in (Halkiewicz 2024)). The modern diff-in-diff approach was popularized by Ashenfelter (1978), with further crucial work by Ashenfelter and Card (1984), Angrist (1990; 1991), Angrist and Kruger(1991; 1999), and Imbens (1995). This development was part of the “credibility revolution” taking place in economics in the 1990s, which saw a variety of methods for identifying causal effects (also known as identification strategies), such as regression discontinuity designs (RDD), instrumental variables (IV), natural experiments and randomized control trials (RCT), becoming popular in the field (Angrist and Pischke 2010).

Each identification strategy relies on some assumptions about the underlying data generation process (DGP) and causal mechanism to make a valid causal inference. The identification issue arises from the “missing counterfactual” problem: we cannot observe the same units both receive and not receive the treatment (the intervention or policy variable whose effects we are interested in studying). We try to fill in this missing counterfactual with some stand-in. In the case of diff-in-diff, we identify an alternative group that never receives the treatment during our period of study. We then assume that the groups would have followed “parallel trends” for the outcome variable (that is, the gap between the treated and untreated groups would have remained constant) in the absence of any intervention. Thus, any change in the gap between the outcome variables after treatment can be attributed to the intervention.

For a concrete example, consider a diff-in-diff study that attempts to identify the effect of a job training program on wages (for example, as in (LaLonde 1986)). We observe the wages of those in the training program before and after they receive training. However, we cannot plausibly identify the impact of the training program on participants’ wages based on this information alone, since we do not observe what their wages would have been had they not received training. After all, other variables (such as the business cycle) might have caused changes in wages, and if we simply compared the before and after training wages of this group, we would be mistakenly attributing the change in wages to job training, when another variable could have been the cause. We try to get around this missing counterfactual problem by identifying a group that never receives

job training during the period of study. Let's say that before the job training program started, those not in the training program on average made \$800 more a year than those in the training program (job training is often targeted to those who are unemployed, underemployed or low income). We now make the crucial assumption that this gap would have remained constant if not for the job training program. That is, although wages for both groups might have increased or declined during the period of study (e.g. due to the business cycle), the average gap in wages between the groups would have remained constant without the job training program. If we can plausibly make this claim, then any changes in the gap after training can be attributed to the program. For example, if after the program, the gap in wages between the two groups narrowed to \$300, then the program can be attributed to an average \$500 increase on the earnings of those who received it.

The diff-in-diff methodology requires the crucial and restrictive parallel trends assumption. This assumption is sometimes implausible. In the job training example, the comparison group might differ in terms of age, education, gender and racial composition from those receiving training. Let us assume that those in the job training program are on average, younger than those in the comparison group. If younger workers experienced higher wage growth throughout the economy during our period of study, then we could be falsely attributing some of the wage gains to the job training program when the gains are really driven by the difference in ages between the two groups. To make a valid causal inference in this situation, we might adopt the less restrictive "conditional parallel trends" assumption. That is, we could assume that the groups would have followed parallel trends conditionally on some values of the covariates.

The literature suggests several ways to estimate this type of conditional diff-in-diff model. Firstly, if we have a good model of how the outcome variables would have evolved conditional on the covariates (for example, if we know how age affects wages), we can incorporate this into the two-way fixed-effects (TWFE) regression commonly used to estimate the classical diff-in-diff model (Baker et al. 2025). Other methods include matching and inverse-probability weighting (IPW) (Abadie 2005; Abadie and Imbens 2006), regression adjustment (J. J. Heckman, Ichimura, and Todd 1997) or doubly robust diff-in-diff (Sant'Anna and Zhao 2020), which incorporates both regression adjustment and inverse probability weighting.

This essay applies and expands on a weighting scheme suggested by Abadie (2005). Specifically, it uses the generalized maximum entropy logistic regression (GME Logit) as proposed by Golan

Judge and Perloff (1996) to obtain the probability of selection into treatment and then uses the weighting scheme developed by Abadie to obtain the treatment effect. I show that the information theory (IT) based estimation of the selection equation outperforms against a classical logistic regression (Logit or ML Logit) for estimating selection probabilities in certain situations. This is because the IT-based estimator is known to outperform ML Logit in all finite samples. The more complex and ill-behaved the data, the larger the advantage of GME Logit.

Researchers are often interested in studying the treatment effect for a subpopulation (that is, for specific values of a covariate). Abadie (2005) also develops a semiparametric method to estimate this treatment effect for subpopulations, using the weights discussed above. Abadie's method replaces the true (usually high dimensional) conditional average treatment effect with a parametric approximation. If the parametric approximation is a linear model (often the case in practice), Abadie's method becomes weighted least squares (WLS) with the weights obtained from estimating the selection equation in stage 1. This essay expands on Abadie's method in this area as well, replacing the WLS-based estimator he suggests with a (weighted) generalized maximum entropy-based estimator. I also show that this method outperforms against the WLS-based estimator under certain circumstances.

The essay proceeds as follows: section II reviews the literature on diff-in-diff and conditional diff-in-diff; section III develops the theory; section IV compares the GME-based and classical methods in a series of simulations; section V applies the method to a well-known dataset and section VI concludes.

II. Review of Literature

The history of canonical diff-in-diff was briefly discussed earlier. I do not recap it here but discuss developments since then. The diff-in-diff literature developed in two distinct waves (Cunningham 2023). The first wave, which started in the 1980s (and discussed earlier), introduced the term and saw the publication of several influential papers. It peaked around 2007. The second wave began in 2011 and was characterized by “an exponential growth in popularity” (Cunningham 2023) and consistent use of terminology (such as parallel-trends and event study) in many papers. This second wave continues today, with diff-in-diff becoming one of the most popular methods of causal identification.

The validity of what we now call the parallel trends assumption was under scrutiny very early in the history of the literature. For example, Ashenfelter (1978) discussed that participants in job training programs are often selected after they receive a temporary dip in their earnings or employment status. In such cases, wages could be expected to bounce back even in the absence of the training program. LaLonde (1986) showed that econometric methods such as diff-in-diff fail to match results obtained from a randomized control trial (RCT), and attributed this to mismatches in outcome trends between the treatment group and potential comparison groups (e.g. drawn from the current population survey, panel study of income dynamics or similar data sources for the general labor force) researchers might use. Heckman, et. al. (1998; 1997) proposed a method called regression adjustment that can recover the ATT when parallel trends are violated if the researcher correctly specifies the model for conditional outcome evolution. Recent research has progressed in three directions: (1) testing (conditional) parallel trends assumptions (e.g. (Roth 2022; Rambachan and Roth 2023)) (2) extending the method to staggered/multi-period treatment adoption (e.g. (Goodman-Bacon 2021; Sun and Abraham 2021; Callaway and Sant'Anna 2021)) or where the level/intensity of treatment can be varied continuously (e.g. (Callaway, Goodman-Bacon, and Sant'Anna 2024)) and (3) relaxing the parallel trends assumption (e.g. (Abadie 2005; Athey and Imbens 2006)). This essay is an expansion of methods that relax the parallel trends assumption.

Theory

I begin by defining the diff-in-diff methodology in the canonical setting. The setup is described in many textbooks such as (Cunningham 2021) or (Angrist and Pischke 2009). The researcher selects a causal quantity of interest (target parameter) for study. Adapting the potential outcomes framework of Rubin (1976), let $Y_{i,t}(0,0)$ denote unit i 's potential outcome at time t if it remains untreated in both periods, and $Y_{i,t}(0,1)$ if it is untreated in period $t = 0$ and treated in period $t = 1$. Since all units are untreated in period $t = 0$, we can simplify notation to: $Y_{i,t}(0) = Y_{i,t}(0,0)$ and $Y_{i,t}(1) = Y_{i,t}(0,1)$. Ideally, we would like to observe $Y_{i,t}(0)$ and $Y_{i,t}(1)$ for the same units. Then the treatment effect for unit i would be $Y_{i,t}(1) - Y_{i,t}(0)$. In practice, we can only observe $Y_{i,t}(1)$ for the treated units and $Y_{j,t}(0)$ for the untreated units. If D_i is an indicator variable that takes a

 \xrightarrow{j} to indicate different unit

Different than $D_{i,t}$ defined later

If value $D_i = 1$ if unit i is treated in period $t = 1$, and $D_i = 0$ otherwise, then the observed values are:

$$Y_{i,t} = (1 - D_i)Y_{i,t}(0) + D_i Y_{i,t}(1) \quad (1)$$

The potential outcomes framework defines a treatment effect $Y_{i,t}(1) - Y_{i,t}(0)$ for every unit i .

In practice, researchers are often interested in finding the average treatment effect on the treated:

$$ATT = \mathbb{E}[Y_{i,t}(1) - Y_{i,t}(0)|D_i = 1] = \mathbb{E}[Y_{i,t}(1)|D_i = 1] - \mathbb{E}[Y_{i,t}(0)|D_i = 1] \quad (2)$$

Without additional assumptions, the ATT is not identified since we are unable to observe $\mathbb{E}[Y_{i,t}(0)|D = 1]$ (the expected potential outcome if unit i had not received treatment, given that it actually did receive treatment). If we are willing to make additional assumptions, we could potentially identify the ATT. Different research designs make different assumptions for identifying the missing counterfactual. In the case of diff-in-diff, the crucial assumption (known as the parallel trends assumption) is:

$$\mathbb{E}[Y_{i,t=1}(0)|D_i = 1] - \mathbb{E}[Y_{i,t=0}(0)|D_i = 1] = \mathbb{E}[Y_{i,t=1}(0)|D_i = 0] - \mathbb{E}[Y_{i,t=0}(0)|D_i = 0] \quad (3)$$

In words, the expected change in outcome between pre-treatment and post-treatment periods ($\mathbb{E}[Y_{i,t=1}(0) - Y_{i,t=0}(1)]$) is the same for treated and comparison groups. Under these assumptions, the ATT is:

$$ATT = \mathbb{E}[Y_{i,t=1}|D_i = 1] - (\mathbb{E}[Y_{i,t=0}|D_i = 1] + (\mathbb{E}[Y_{i,t=1}|D_i = 0] - \mathbb{E}[Y_{i,t=0}|D_i = 0])) \quad (4)$$

where the term in parenthesis on the right-hand side is our counterfactual $\mathbb{E}[Y_{i,t}(0)|D_i = 1]$. The above can be rewritten as:

$$ATT = ([Y_{i,t=1}|D_i = 1] - \mathbb{E}[Y_{i,t=0}|D_i = 1]) - (\mathbb{E}[Y_{i,t=1}|D_i = 0] - \mathbb{E}[Y_{i,t=0}|D_i = 0]) \quad (5)$$

where the first term in the parenthesis is the change in the average outcome for treated units between period $t = 0$ and $t = 1$ and the second term is the change in the average outcome for the untreated units during the same period. Thus, the average treatment effect (on the treated) has a

natural interpretation as the “additional” average change in the outcome of the treated units on top of the change experienced by the untreated units during the period of study.

It is well-known (for a textbook treatment, see (Cunningham 2021)) that estimating this model, involving four means is the same as a two-way fixed effects (TWFE) regression:

$$(6) \quad Y_{i,t} = \theta_t + \eta_i + \beta D_{i,t} + e_{i,t}$$

\rightarrow This $D_{i,t}$ is different from the D_i defined earlier

where θ_t and η_i are time and unit fixed effects respectively, $e_{i,t}$ are idiosyncratic shocks and β is the parameter of interest and $D_{i,t}$ is an indicator variable if the unit received treatment in period t . ($D_{i,t} = 0$ for all units when $t = 0$ and $D_{i,t} = 1$ for only treated units in period $t = 1$ and 0 otherwise).

Often the parallel trends assumption is not plausible. The treatment group might differ from the comparison group in systematic ways that affect the outcome (for example, workers participating in a job training program might be less educated or younger than the comparison group and would have experienced a different wage trend than older workers regardless of whether they participated in training or not). We can relax the parallel trends assumption from above with a conditional parallel-trends assumption. Let X_i be a vector of observed determinants of changes in $Y_{i,t}$. We now assume that the groups would have followed parallel trends *conditional* on the covariates:

$$(7) \quad \mathbb{E}[Y_{i,t=1}(0) - Y_{i,t=0}(0)|X_i, D_i = 1] = \mathbb{E}[Y_{i,t=1}(0) - Y_{i,t=0}(0)|X_i, D_i = 0]$$

There are several ways to estimate this conditional parallel-trends model. Firstly, the covariates can be incorporated into an augmented TWFE regression (Baker et al. 2025) either as:

$$(8) \quad Y_{i,t} = \theta_t + \eta_i + \beta D_{i,t} + X_{i,t} \beta_{covs} + e_{i,t}$$

or as

$$(9) \quad Y_{i,t} = \theta_t + \eta_i + \beta D_{i,t} + (\mathbf{1}\{t = 1\} X_{i,t=0}) \widetilde{\beta_{covs}} + e_{i,t}$$

where $X_{i,t}$ is the observed value of the covariate at time t , $\mathbf{1}\{t = 1\}$ is an indicator variable that takes a value of 1 in the post-treatment period and 0 otherwise, and β_{covs} and $\widetilde{\beta_{covs}}$ are vectors of coefficients associated with the covariates. Note that since time-invariant variables drop out in the

TWFE regression, only the effect of changes in the levels of a covariate $\Delta X_{i,t} = (X_{i,t=1} - X_{i,t=0})$ for equation (8) or differential trends related to baseline levels of the variable $\mathbf{1}\{t = 1\}X_{i,t=0}$ for equation (9) can be estimated with this type of model. In addition, researchers need to be careful about what variables are controlled for. For example, if the treatment affects the outcome by changing the value of $X_{i,t}$ (e.g. job training improving some measured skill), then controlling for this in the TWFE regression will bias our estimate of ATT.

Abadie (2005) provides another way to estimate the average treatment effect. If we are willing to make the mild assumptions that (1) at least some portion of the population receives treatment and (2) at every stratum/value of the covariate, some portion of the population remains untreated, that is:

$$\Pr(D_i = 1) > 0 \text{ and } 0 < \Pr(D_i = 1|X_i) < 1 \quad (10)$$

then Abadie shows that the ATT *conditional* on X_i is:

$$\mathbb{E}[Y_{i,t=1}(1) - E_{i,t}(0)|X_i, D_i = 1] = \mathbb{E}[\rho_0 \cdot (Y(1) - Y(0))|X_i] \quad (11)$$

where:

$$\rho_0 = \frac{D_i - \Pr(D_i = 1|X_i)}{\Pr(D_i = 1|X_i)(1 - \Pr(D_i = 1|X_i))} \quad (12)$$

The (unconditional) ATT can be recovered as:

$$\begin{aligned} \mathbb{E}[Y_{i,t}(1) - Y_{i,t}(0)|D_i = 1] &= \int \mathbb{E}[\rho_0 \cdot (Y_{i,t}(1) - Y_{i,t}(0))|X_i] d\Pr(X_i|D_i = 1) \\ &= \mathbb{E}\left[\frac{Y_{i,t}(1) - Y_{i,t}(0)}{\Pr(D_i = 1)} \cdot \frac{D_i - \Pr(D_i = 1|X_i)}{1 - \Pr(D_i = 1|X_i)}\right] \end{aligned} \quad (13)$$

Abadie does not specify how $\Pr(D_i = 1|X_i)$ should be estimated. Commonly, researchers use ML Logit. However, other methods that estimate the selection probability could also be used. The GME approach for binary discrete choice (Golan, Judge, and Perloff 1996) is known to have several advantages over traditional ML Logit: it is efficient for small samples, avoids strong parametric assumptions, handles multicollinearity and is resilient to ill-conditioned data. The

theory of GME Logit is recapped below (see Golan, Judge and Perloff 1996 for a more thorough development).

Assume that N units are observed where each unit is either selected into treatment or not. Let

$$(14) \quad p_i = \Pr(D_i = 1|X_i, \gamma) = F(X'_i \gamma) \quad \begin{matrix} \text{switched to } \gamma \text{ to} \\ \text{differentiate from} \\ \text{earlier } \beta. \end{matrix}$$

be the probability of observing unit $i \in [1, 2, \dots, N]$ in the treatment state, conditional on covariates X_i (of dimension $1 \times M$) and unknown parameters γ (of dimension $M \times 1$). If we have noisy data, we can write

$$(15) \quad y_i = F(\cdot) + e_i = p_i + e_i \quad \begin{matrix} \rightarrow \text{switched to } M \text{ so} \\ \text{as to not confuse w/ } X \\ \cancel{\text{and}} \text{ defined later} \end{matrix}$$

where $y_i \in [0, 1]$ is the observed outcome, p_i denotes the unknown and unobservable probability of selection and e_i is an error or noise component for each observation contained in $[-1, 1]$. We reparametrize the error component as:

$$e_i \stackrel{\text{def}}{=} \sum_h v_{ih} w_{ih}$$

(16)

where $\sum_h w_{ih} = 1$ are an H -dimensional vector of weights and v_i are the H -dimensional support space. (Golan, Judge, and Perloff 1996) suggest $v_i = \left[\frac{-1}{\sqrt{N}}, 0, \frac{1}{\sqrt{N}} \right]$ and thus $H = 3$.

We now maximize the entropy of the error augmented probability distribution:

$$\max_{p,w} \left(- \sum_i p_i \ln p_i - \sum_{ih} w_{ih} \ln w_{ih} \right)$$

(17)

subject to the M moment constraints imposed by the data, where the m -th constraint is:

$$\sum_i y_i X_{im} = \sum_i X_i p_i + \sum_{ih} X_i v_h w_{ih}$$

(18)

and the adding up constraint

$$\sum_h w_i = 1$$

(19)

(I omit the constraint the $\sum p_{i,j}=1$ present in a multinomial version of this model, since that requirement is naturally satisfied in the case of binary choice)  We can solve the above using the method of Lagrange multipliers. In matrix-form:

$$\mathcal{L} = -\mathbf{p} \ln \mathbf{p} - \mathbf{w}' \ln \mathbf{w} + \boldsymbol{\lambda}' (\mathbf{X}' \mathbf{p} + \mathbf{X}' \mathbf{V} \mathbf{w} - \mathbf{X}' \mathbf{y}) + \boldsymbol{\mu}' (1 - \mathbf{I}_T \mathbf{p}) + \boldsymbol{\rho}' (1 - \mathbf{1}' \mathbf{w})$$

(20)

where $\boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\rho}$ are Lagrange multipliers and $\boldsymbol{\gamma} = -\boldsymbol{\lambda}$. The first order conditions are:

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial p_i} &= -\ln p_i - 1 - \sum_m \lambda_m X_{im} - \mu_i = 0 \\ \frac{\partial \mathcal{L}}{\partial w_{ih}} &= -\ln w_{ih} - 1 - \sum_m \lambda_m X_{im} v_h - \rho_i = 0 \\ \frac{\partial \mathcal{L}}{\partial w_k} &= \sum_i y_i X_{im} - \sum_i X_{im} p_i - \sum_k X_{im} v_h w_{ih} = 0 \\ \frac{\partial \mathcal{L}}{\partial \rho_i} &= 1 - \sum_h w_{ih} = 0\end{aligned}$$

(21)

From which we obtain the estimated probability distributions:

$$\hat{p}_i = \frac{\exp(-\sum_m \widehat{\lambda}_m X_{im})}{1 + \exp(-\sum_m \widehat{\lambda}_m X_{im})}$$

(22)

and

$$w_{ih} = \frac{\exp(-\sum_m X_{im} \widehat{\lambda}_m v_h)}{\sum_h \exp(-\sum_m X_{im} \widehat{\lambda}_m v_h)}$$

(23)

Note that the above can be computed using a “dual unconstrained” method that is usually less computationally intensive but not discussed here for brevity (see (Golan 2017) for details).

IV. Simulation

To evaluate the performance of alternative estimators of the ATT, I conduct Monte Carlo simulations based on a stylized data generating process (DGP) designed to reflect common features of observational studies, including selection into treatment on observables and covariate-dependent potential outcomes.

I begin by describing the baseline simulation. Since I am describing the data generating process for each unit i , I drop the i subscripts in this section. The simulated dataset contains $N = 500$ observations and the simulation is replicated across 5,000 Monte Carlo runs. Each observation is characterized by two baseline covariates, X_1 and X_2 , independently drawn from a standard normal distribution (note that X_1 and X_2 just represent two vectors of covariates, rather than unit or time subscripts):

$$\begin{aligned} X_1 &\sim \mathcal{N}(0,1) \\ X_2 &\sim \mathcal{N}(0,1) \end{aligned} \tag{24}$$

The baseline outcome at $t = 0$, denoted Y_0 , is modeled as a linear function of the covariates, with additive normal noise:

$$\begin{aligned} Y_0 &= 5 + 0.5X_1 + 0.3X_2 + \epsilon_0, \\ \epsilon_0 &\sim \mathcal{N}(0,1) \end{aligned} \tag{25}$$

This structure ensures that baseline outcomes are systematically related to observed characteristics. In addition to level differences, the untreated trend also depends on the covariates:

$$\Delta_0 = 0.4X_1 - 0.1X_2 \tag{26}$$

This trend is added to all individuals, regardless of treatment status, mimicking scenarios in which outcome trajectories differ across subgroups even in the absence of intervention. Treatment is assigned based on a logistic selection model that depends on the covariates:

$$\Pr(D = 1|X_1, X_2) = \text{invlogit}(0.5X_1 - 0.8X_2) \tag{27}$$

where invlogit is $\Pr(D = 1|X_1, X_2) = \frac{\exp(0.5X_1 - 0.8X_2)}{\exp(0.5X_1 - 0.8X_2) + 1}$. Each unit is assigned to treatment with a probability equal to this score. This introduces selection on observables, such that the probability of receiving treatment is systematically related to covariates that also influence potential outcomes. Equations (25) and (26) are the crucial deviations from the unconditional parallel trends assumption that makes TWFE a biased estimator of the ATT. Since treatment into selection and the trend are both dependent on covariates X_1 and X_2 , the treated and untreated groups no longer follow parallel trends.

The outcome in the post-treatment period ($t = 1$), denoted Y_1 , is constructed as the untreated outcome plus the untreated trend, a constant treatment effect for the treated, and additional noise:

$$\begin{aligned} Y_1 &= Y_0 + \Delta_0 + 2 \cdot D + \epsilon_1, \\ \epsilon_1 &\sim \mathcal{N}(0, 1) \end{aligned} \tag{28}$$

Thus, the ATT is set to a constant value of 2 across all treated units, independent of covariates. However, since both the baseline outcome and the untreated trend depend on X_1 and X_2 , failure to account for these covariates can lead to biased estimates of the treatment effect.

Under the baseline assumptions (large samples, well-behaved data), GME Logit and ML logit models of selection into treatment should behave similarly. To test whether GME Logit outperforms ML Logit under certain circumstances, I test several cases where GME is known to outperform. Firstly, GME is known to outperform ML logit for all finite samples. These differences are likely to be highlighted when sample sizes are very small. I modify the above (baseline) simulation, so that the dataset is of size $N = 25$ observations. Next, I test how the two estimators perform when the probability of receiving treatment is very small (this is done by shifting the selection equation to:

$$\Pr(D = 1|X_1, X_2) = \text{invlogit}(-3 + 0.4X_1 - 0.1X_2) \tag{29}$$

which causes only ~5% of observations to be selected into treatment). GME logit is also known to outperform ML logit when covariates are highly collinear. I simulate this by setting $X_2 = X_1 + u$, $u \sim \mathcal{N}(0, 0.01)$. The tables below summarize the simulation settings.

Table 1: Baseline simulation settings. All simulations have 5,000 Monte Carlo runs.

Observations (N)	500
Covariate Distribution (X_1, X_2)	$X_1 \sim \mathcal{N}(0,1)$ $X_2 \sim \mathcal{N}(0,1)$
Heterogenous Trend (Δ_0)	$\Delta_0 = 0.4X_1 - 0.1X_2$
Selection equation	$\Pr(D = 1 X_1, X_2) = \text{invlogit}(0.5X_1 - 0.8X_2)$
Errors	$\epsilon_0 \sim \mathcal{N}(0,1)$ $\epsilon_1 \sim \mathcal{N}(0,1)$

Table 2: Modifications to the baseline case for specific scenarios. All simulations have 5,000 Monte Carlo runs

Case: Small sample size	$N = 25$
Case: Rare selection	$\Pr(D = 1 X_1, X_2) = \text{invlogit}(-3 + 0.5X_1 - 0.8X_2)$
Case: Highly collinear X_1, X_2	$X_2 = X_1 + u, u \sim \mathcal{N}(0,0.01)$

For the estimation of the GME Logit model, I use a support space of 3 with $v_{ih} = \left[\frac{-1}{\sqrt{n}}, \frac{0}{\sqrt{n}} \right]$ as suggested by Golan, Judge and Perloff (1996). I do not use any prior distributions for GME Logit. Note that unlike the GME Linear Model (discussed later), GME Logit does not require us to reparametrize the coefficients ($\hat{\beta}$) and the error is naturally bounded between 0 and 1.

First Stage

I begin my analysis of the simulation results by first examining how successful ML Logit and GME Logit are at recovering the selection mechanism. I examine two aspects of the methods' performance: (1) how well each predicts treatment assignment (treated vs. untreated) and (2) how accurately each recovers the true model parameters. To assess (1), I report the number of misclassifications (following (Golan, Judge, and Perloff 1996)) as well as the Area Under the Receiver Operating Curve (AUC-ROC). The AUC-ROC is a widely used metric that summarizes a classifier's ability to rank positive cases above negative ones across all possible threshold values. A value of 0.5 corresponds to guessing at random, while a value of 1 indicates perfect classification. For reporting misses, I set the classifier threshold to 0.5 (that is, for each method, when the predicted probability is > 0.5 , the observation is set to be in the treated group). For assessing (2), I report the Mean Squared Error (MSE) for each method as in Golan Judge and Perloff (1996).

Table 3: M_{gme} Performance for Stage 1 over 5,000 Monte Carlo simulation runs. Each run has 500 observations, except for the small sample case, which has 25 observations. The worst, average and best misclassification rates are reported (misclassification is defined as predicted incorrectly to receive treatment when the unit did not receive treatment or predicted incorrectly to not receive treatment, when the unit did in-fact receive treatment). The average AUC and the MSE (calculated as bias² + variance) are also reported.

		Misclassification Rate				
		Worst case	Average	Best Case	AUC	MSE
Baseline	GME	40.8%	33.1%	26.2%	0.73	0.02
	ML Logit	40.8%	33.1%	26.2%	0.73	0.02
Small Sample	GME	60.0%	29.1%	0%	0.76	0.24
	ML Logit	60.0%	29.1%	0%	0.76	625.48
Rare Selection	GME	11.2%	6.65%	3.0%	0.75	0.06
	ML Logit	11.2%	6.65%	3.0%	0.75	0.06
Highly Collinear	GME	50.6%	43.6%	35.6%	0.59	117.1
	ML Logit	50.6%	43.6%	35.6%	0.59	118.5

The results for stage 1 are similar to those obtained by Golan, Judge and Perloff (1996) in their exploration of this method. In the baseline case, GME Logit performs as well as ML Logit at classification, and has similar MSE. In addition to the metrics reported by them, I also include the AUC-ROC. These are also comparable between GME Logit and ML Logit.

Comparing GME Logit's performance with ML Logit for deviations from the baseline case, I note that both methods perform equally well at classification (whether looking at misclassification rates or AUC-ROC). Although the GME Logit has much lower MSE than ML Logit for small samples, these might not translate over to gains in the second stage of estimation. Only the estimate of the probability of selection into treatment (i.e. the propensity score) enters the second stage of estimation, and here both methods give similar results.

Second Stage

Recall that in my data generating process, the true treatment effect is “hard-coded” to a value of 2.0 (that is, a unit that receives treatment will have a 2.0 higher value for the outcome $Y_{i,t=1}$ if it receives treatment than it would otherwise). I now examine the ability of the second stage

estimation (which simply involves replacing the weighted population expectations in equation 13 with their sample counterparts) to recover this treatment effect. In addition to the results obtained with weights derived from GME Logit and ML Logit first steps, I also show the results for the unweighted (or equal weighted) estimator. This estimator is biased and not consistent with the true ATT, but is shown to give researchers an idea of what to expect when unconditional parallel trends assumption is violated in the underlying data but the analysis does not account for the violation.

Table 4: Second Stage regression results. See above for specification of each case (leftmost column). The true ATT is 2.0. Column headers indicate the model used for first stage propensity score estimation.

	MSE			Bias		
	GME	ML	Unweighted	GME	ML	Unweighted
Baseline	0.114	0.113	0.244	0.039	0.039	0.076
Small Sample	0.976	1.485	0.380	-0.739	-0.908	0.150
Rare Selection	0.073	0.074	0.149	-0.019	-0.027	0.280
Highly Collinear	0.032	0.032	0.018	0.122	0.123	0.038

In general, ML Logit and GME Logit perform equally well (when measured by bias or MSE) in recovering the true treatment effect. This is even the case when the first stage GME estimator has lower MSE (e.g. for small sample sizes). In the second stage, the parameter estimates of the selection equation are not explicitly used. Only the propensity (likelihood of selection into treatment conditional on the covariates) enters the second stage weighting scheme. Golan, Judge and Perloff (1996) show that the GME estimator performs equally well with ML Logit when comparing misclassification rates but has lower variance and lower MSE. Thus, the results of the second stage are not unexpected. However, researchers might still prefer to use the GME Logit for the first stage if estimating the parameters of the selection mechanism is important, since GME Logit has lower variance than ML Logit.

Extension: Sub-population treatment effects

So far, our treatment effect has been constant (set to 2.0 in the previous examples) regardless of the value of the covariates. Thus, although the baseline values ($Y_0 = 5 + 0.5X_1 + 0.3X_2 + \epsilon_0$), trend ($\Delta_0 = 0.4X_1 - 0.1X_2$) and selection into treatment ($\Pr(D = 1|X_1, X_2) = \text{invlogit}(0.5X_1 - 0.8X_2)$) all depend on the covariates, the treatment effect for the entire population remains

constant regardless of the values of X_1 and X_2 . This might not be an accurate way to model how treatment works. Going back to the job training example, we might imagine that younger and more skilled workers could benefit more from training. If X_1 and X_2 are workers' ages and a proxy for skill, then the true treatment effect would be $\tau = g(X_1, X_2)$. Abadie's weighting method above would still recover the population-wide ATT, even if the treatment effect is dependent on covariates. However, researchers might be interested in studying the treatment effect for a substratum of the population (for example, for younger workers, for males vs. females, etc.). Abadie (2005) also suggests a way to estimate these subpopulation treatment effects, which is outlined below.

Consider X_k to be a deterministic function of the underlying covariates X (for example, X_k maybe a subset of X , or may contain indicator variables for X). Let $\mathcal{G} = \{g(X_k; \theta) | \theta \in \Theta \subset \mathbb{R}^k\}$ be a class of approximating functions, square integrable with respect to $\Pr(X_k | D = 1)$. Then a least squares approximation from \mathcal{G} to $\mathbb{E}[Y_1(1) - Y_0(1) | X_k]$ is given by $g(X_k; \theta_0)$, where:

$$\theta_0 = \operatorname{argmin}_{\theta \in \Theta} \mathbb{E}[\{\mathbb{E}[Y_1(1) - Y_0(1) | X_k, D = 1] - g(X_k; \theta)\}^2 | D = 1] \quad (30)$$

For example, if $\mathcal{G} = \{X_k' \theta | \theta \in \Theta \subset \mathbb{R}^k\}$ then θ_0 defines a linear LS approximation to $\mathbb{E}[Y_1(1) - Y_0(1) | X_k]$. Abadie shows that the θ_0 can be recovered as:

$$\theta_0 = \operatorname{argmin}_{\theta \in \Theta} \mathbb{E}[\Pr(D = 1 | X) \cdot \{\rho_0 \cdot (Y_1 - Y_0) - g(X_k; \theta)\}^2] \quad (31)$$

In the case when $\mathcal{G} = \{X_k' \theta | \theta \in \Theta \subset \mathbb{R}^k\}$, this simply becomes weighted LS (with weights obtained from the first stage).

It is known (see (Golan, Judge, and Miller 1996; Golan 2007) that the generalized maximum entropy linear model (GME Linear) is superior to LS for finite samples and ill-behaved data. Thus, it is natural to extend the weighted LS (WLS or weighted LS) proposed of Abadie to a weighted GME approach, to see how it performs. To test this, I replace the constant treatment effect $\tau = 2.0$ with

$$\tau = 2.0 + 0.5X_1 - 0.3X_2 \quad (32)$$

Both weighted LS regression and weighted GME regression should be consistent estimators of the conditional average treatment effect (32). For the weights, I use the predicted selection probabilities obtained by performing a GME Logit in the first stage $w_i = \Pr(D_i = 1|X_i)$. Since the GME Logit function in Stata does not accept weights, I obtain GME results by multiplying each observation by $\sqrt{w_i}$.

For the GME Linear estimation, we also need to specify a support space for the parameters and a support space for the errors. Since the model is now in a “weighted space” (each observation weighted by $\sqrt{w_i}$), the empirical $\pm 3s$ rule is no longer appropriate. I also use a wide support space for the parameters, to capture extreme parameter values. Equation (33) specifies the support spaces:

$$(33) \quad w_{ih} = [-100, -91.67, \dots, 0, \dots, 91.67, 100]; \{h: [1, 25]\}$$

$$Z = \begin{bmatrix} -100 & 0 & 100 \\ -100 & 0 & 100 \\ -100 & 0 & 100 \end{bmatrix}$$

The error support space for each observation w_i is an equally spaced vector of size 25, ranging from -100 to 100. The parameter support space is a 3×3 matrix with each parameter having the support space $[-100, 0, 100]$.

Table 5 shows weighted LS and OLS results in head to head comparison. In general, both methods behave similarly and are successful at obtaining the conditional average treatment effect. However, GME Logit has significantly lower variance in the case of highly collinear variables. The weighted LS regression has over 300x  MSE as the GME estimator. Thus, researchers should prefer the GME based estimator since it performs as well as WLS in all circumstances and strictly better when covariates are highly collinear.

Table 5: Estimate of the conditional average treatment effect. The true conditional treatment effect is $\tau(X_1, X_2) = 2.0 + 0.5X_1 - 0.3X_2$

	Bias						MSE					
	β_0		β_1		β_2		β_0		β_1		β_2	
	WLS	GME	WLS	GME	WLS	GME	WLS	GME	WLS	GME	WLS	GME
Baseline	0.001	-0.003	0.009	-0.001	-0.008	0.003	0.020	0.023	0.047	0.080	0.049	0.111
Small Sample	-0.005	-0.093	0.107	0.074	-0.133	-0.122	0.435	0.389	0.714	0.628	0.657	0.574
Rare Selection	-0.180	-0.195	0.003	0.009	-0.116	-0.119	0.176	0.176	0.155	0.151	0.146	0.143
Collinear	-0.001	-0.005	0.074	-0.634	-0.074	0.633	0.017	0.017	194.47	0.505	194.42	0.504

In general, researchers should prefer the info-metrics based methods for estimating selection probabilities (stage 1) and for estimating the conditional treatment effects. The info-metrics method outperforms ML Logit for small sample sizes and the info-metrics based conditional average treatment effect outperforms weighted LS when covariates are highly collinear.

Application to Real World Data

I now apply this method to a real-world dataset. I use the data provided by LaLonde (1986) for his evaluation of the impact of the National Supported Work (NSW) demonstration. The data is readily available in many software packages as well as through the NBER website. As described by LaLonde, the NSW program randomly assigned individuals to a treatment or control group. Those in the treatment group received “supported work” where participants were employed at construction, service or similar industries in a supportive but performance-oriented work environment. Those in the control group were left to fend for themselves. LaLonde examines the data for male and female participants separately. For brevity, I focus on the analysis for male participants only. I also only use the info-metrics based estimators (due to the reasons discussed earlier). For the male group, the outcome of interest is the 1978 real earnings due to participation in the program (or earnings growth from baseline 1975 wages in a difference-in-differences context). Since we have data from a control group, the difference in mean 1978 earnings between the experimental and control group is an unbiased estimator of the treatment effect. LaLonde asks, what if researchers used an alternative control group to analyze the data instead of the actual

control group. He constructs several sets of control groups: one based on the Current Population Survey (CPS-1), one based on the Panel Study of Income Dynamics (PSID-1), and several others that are obtained by sub-setting CPS-1 or PSID-1 samples to match pre-treatment earnings and demographic characteristics of program participants. LaLonde's results with respect to this dataset (as well as using from the actual control group, which serves as the "gold standard" estimate of the treatment effect) is reproduced below.

Table 6 Treatment effects for male NWS participants as reported in LaLonde (1986) using the actual control group, and CPS-1 group drawn from the current population survey with similar characteristics as the treatment group. Standard Errors in parenthesis

Comparison Group	Treatment Less Comparison Group Earnings		Diff-in-Diff: Difference in Earnings Growth 1975 to 1978 Less Comparison Group	
	Unadjusted	Adjusted ¹	Without Age	With Age
Actual Controls	\$886 (476)	\$798 (472)	\$847 (560)	\$856 (558)
CPS-1	-\$8,870 (562)	-\$4,416 (577)	\$1,714 (452)	\$195 (441)

The gold standard point estimate of the ATT (as shown in row 1) is \$886 (or \$798 if we adjust for difference in age, schooling, etc. between treatment and comparison groups). LaLonde also performs a diff-in-diff analyses (this is strictly not needed when we are using actual RCT control group as our comparison group, since the simple difference in means between the control and treatment groups in the treatment period is an unbiased estimator of the ATT. However, a diff-in-diff is also a consistent and unbiased estimator of the ATT). LaLonde performs both a simple diff-in-diff and one that adjusts for the difference in age between the treatment and control group (other characteristics such as education and ethnicity remain constant for each individual across periods, so are not used in the diff-in-diff). Point estimates from each method range from \$798 (for the demographics adjusted difference in mean earnings) to \$886 (for unadjusted differences between the treatment and control groups), with the diff-in-diff estimators falling in between. At the $\alpha = 0.05$ level of significance, we cannot reject the null hypothesis that the treatment had no effect on

¹ The exogenous variables used in the regression adjusted equations are age, age squared, years of schooling, high school dropout status and race

earnings. This is in line with other analysis of the NWS program which finds no effect on earnings for male participants (e.g. (A. Smith and E. Todd 2005; Dehejia and Wahba 1999)).

I now focus on row 2, which answers the question ‘if researchers did not have a RCT control group but used a comparison group such as the current population survey, how successful would they be at matching the “gold standard” ATT estimate’. The CPS-1 group consists basically of all male headed-households in the CPS who were less than age 55 in 1975 with matched social security numbers (Westat’s matching criteria). The naïve difference in 1978 earnings between the treatment group and CPS comparison group (column 1) recovers a large negative impact of participating in the program. Using the experimental control group as our reference, we know participation in the program is associated with a ~\$886 increase in earnings. A naïve comparison between the CPS-1 sample’s wages and the treatment group’s wages is not successful at recovering the ATT. The large negative impact obtained when using the CPS sample as the comparison group is likely due to differences in demographic characteristics between the treatment group and the CPS sample. Column 2 attempts to adjust for some demographic characteristics such as educational attainment, race and age through a regression. However, the estimate of the program’s effect continues to be negative, likely due to uncontrolled for (and unobservable) differences between the CPS-1 group and NSW treatment group. Columns 3 and 4 show the results of a diff-in-diff type estimation (using 1975 earnings as a baseline). Results in column 4 controls for age (and age squared) in the DiD estimate. The CPS group shows wildly different impact of program participation when age is controlled for, and none are close to our point estimate of the ATT.

I now apply Abadie’s weighting method, using the GME based first stage described earlier. For estimating the selection equation, I use the age, age², years of education, an indicator for whether the individual has a high school degree, an indicator for the individual’s race or ethnicity (white is baseline), and an indicator for whether the individual is married. I obtain standard errors through the bootstrap. I use the same error support weight vector as specified earlier: $\left[\frac{-1}{\sqrt{N}}, 0, \frac{+1}{\sqrt{N}} \right]$. Table 7 shows the results. The estimate of the program’s effect is now much closer to one obtained from using the experimental control group. The standard error of the estimate is also similar. If researchers used the CPS sample but weight the observations so that treatment and control groups

look similar, they would obtain estimates and reach similar conclusions as they would using an RCT.

Table 7: Treatment effects using Abadie's weighting scheme (with GME first step)

Comparison Group	Difference in Earnings
Controls	\$847 (560)
CPS-1 with Abadie's method (GME first stage)	\$575 (562)

Conclusion

This paper advances the conditional difference-in-differences (DiD) framework by integrating an information-theoretic approach, specifically generalized maximum entropy (GME) methods, to enhance the estimation of treatment effects under violations of the parallel trends assumption. Through Monte Carlo simulations, I demonstrate that GME Logit matches or outperforms maximum likelihood logistic regression (ML Logit) in estimating selection probabilities, particularly in small samples and rare selection scenarios, due to its lower mean squared error and robustness to ill-conditioned data. While both methods yield comparable ATT estimates in the second stage, GME Logit's advantages in parameter estimation make it a valuable tool when modeling selection mechanisms is a priority. Furthermore, extending Abadie's (2005) weighted least squares approach to a weighted GME Linear estimator for subpopulation treatment effects shows promise in handling heterogeneous treatment effects, especially in finite samples or with complex data structures.

Applying the method to LaLonde's (1986) National Supported Work dataset, I find that the GME-based DiD estimator, using Abadie's weighting scheme, produces an ATT estimate (\$575) closer to the randomized controlled trial benchmark (\$886) than traditional DiD methods when using a non-experimental comparison group like CPS-1. This underscores the method's ability to mitigate biases from covariate imbalances, offering a robust alternative when conditional parallel trends hold.

These findings have implications for empirical research, particularly in observational studies where data limitations challenge standard assumptions. The GME-based approach provides a flexible and efficient framework for causal inference, accommodating small samples, rare events, and collinear covariates. Future research could explore its performance in multi-period or

staggered treatment settings, integrate doubly robust methods, or develop computational optimizations for large-scale applications. By leveraging information-theoretic principles, this work contributes to the ongoing evolution of causal inference methods.

References

- Abadie, Alberto. 2005. "Semiparametric Difference-in-Differences Estimators." *The Review of Economic Studies* 72 (1): 1–19.
- Abadie, Alberto, and Guido W. Imbens. 2006. "Large Sample Properties of Matching Estimators for Average Treatment Effects." *Econometrica* 74 (1): 235–67. <https://doi.org/10.1111/j.1468-0262.2006.00655.x>.
- Angrist, Joshua D. 1990. "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records." *The American Economic Review* 80 (3): 313–36.
- . 1991. "The Draft Lottery and Voluntary Enlistment in the Vietnam Era." *Journal of the American Statistical Association* 86 (415): 584–95. <https://doi.org/10.1080/01621459.1991.10475083>.
- Angrist, Joshua D., and Guido W. and Imbens. 1995. "Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity." *Journal of the American Statistical Association* 90 (430): 431–42. <https://doi.org/10.1080/01621459.1995.10476535>.
- Angrist, Joshua D., and Alan B. Krueger. 1991. "Does Compulsory School Attendance Affect Schooling and Earnings?*." *The Quarterly Journal of Economics* 106 (4): 979–1014. <https://doi.org/10.2307/2937954>.
- . 1999. "Chapter 23 - Empirical Strategies in Labor Economics." In *Handbook of Labor Economics*, edited by Orley C. Ashenfelter and David Card, 3:1277–1366. Elsevier. [https://doi.org/10.1016/S1573-4463\(99\)03004-7](https://doi.org/10.1016/S1573-4463(99)03004-7).
- Angrist, Joshua D., and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton: Princeton University Press.
- . 2010. "The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con out of Econometrics." *Journal of Economic Perspectives* 24 (2): 3–30. <https://doi.org/10.1257/jep.24.2.3>.
- Ashenfelter, Orley. 1978. "Estimating the Effect of Training Programs on Earnings." *The Review of Economics and Statistics* 60 (1): 47–57. <https://doi.org/10.2307/1924332>.
- Ashenfelter, Orley, and David Card. 1984. "Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs." Working Paper. Working Paper Series. National Bureau of Economic Research. <https://doi.org/10.3386/w1489>.
- Athey, Susan, and Guido W. Imbens. 2006. "Identification and Inference in Nonlinear Difference-in-Differences Models." *Econometrica* 74 (2): 431–97. <https://doi.org/10.1111/j.1468-0262.2006.00668.x>.
- Baker, Andrew, Brantly Callaway, Scott Cunningham, Andrew Goodman-Bacon, and Pedro H. C. Sant'Anna. 2025. "Difference-in-Differences Designs: A Practitioner's Guide." arXiv. <https://doi.org/10.48550/arXiv.2503.13323>.
- Callaway, Brantly, Andrew Goodman-Bacon, and Pedro H. C. Sant'Anna. 2024. "Difference-in-Differences with a Continuous Treatment." Working Paper. Working Paper Series. National Bureau of Economic Research. <https://doi.org/10.3386/w32117>.
- Callaway, Brantly, and Pedro H. C. Sant'Anna. 2021. "Difference-in-Differences with Multiple Time Periods." *Journal of Econometrics*, Themed Issue: Treatment Effect 1, 225 (2): 200–230. <https://doi.org/10.1016/j.jeconom.2020.12.001>.

- Cunningham, Scott. 2021. *Causal Inference: The Mixtape*. New Haven London: Yale University Press.
- . 2023. “The First and Second Waves of Difference-in-Differences.” Substack newsletter. *Scott’s Mixtape Substack* (blog). November 9, 2023. <https://causalinf.substack.com/p/the-first-and-second-waves-of-difference>.
- Golan, Amos. 2007. “Information and Entropy Econometrics — A Review and Synthesis.” *Foundations and Trends® in Econometrics* 2 (1–2): 1–145. <https://doi.org/10.1561/0800000004>.
- . 2017. *Foundations of Info-Metrics: Modeling and Inference with Imperfect Information*. New York, NY: Oxford University Press.
- Golan, Amos, George G. Judge, and Douglas Miller. 1996. *Maximum Entropy Econometrics: Robust Estimation with Limited Data*. Series in Financial Economics and Quantitative Analysis. Chichester [England] ; New York: Wiley.
- Golan, Amos, George Judge, and Jeffrey M. Perloff. 1996. “A Maximum Entropy Approach to Recovering Information From Multinomial Response Data.” *Journal of the American Statistical Association* 91 (434): 841–53. <https://doi.org/10.2307/2291679>.
- Goodman-Bacon, Andrew. 2021. “Difference-in-Differences with Variation in Treatment Timing.” *Journal of Econometrics*, Themed Issue: Treatment Effect 1, 225 (2): 254–77. <https://doi.org/10.1016/j.jeconom.2021.03.014>.
- Halkiewicz, Stanisław. 2024. “What Are Difference-in-Difference Models? An Overview of History, Theory and Applications.” In , 219–31.
- Heckman, James, Hidehiko Ichimura, Jeffrey Smith, and Petra Todd. 1998. “Characterizing Selection Bias Using Experimental Data.” *Econometrica* 66 (5): 1017–98. <https://doi.org/10.2307/2999630>.
- Heckman, James J., Hidehiko Ichimura, and Petra E. Todd. 1997. “Matching As An Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme.” *The Review of Economic Studies* 64 (4): 605–54. <https://doi.org/10.2307/2971733>.
- LaLonde, Robert J. 1986. “Evaluating the Econometric Evaluations of Training Programs with Experimental Data.” *The American Economic Review* 76 (4): 604–20.
- Lester, Richard A. 1946. “Shortcomings of Marginal Analysis for Wage-Employment Problems.” *The American Economic Review* 36 (1): 63–82.
- Rambachan, Ashesh, and Jonathan Roth. 2023. “A More Credible Approach to Parallel Trends.” *The Review of Economic Studies* 90 (5): 2555–91. <https://doi.org/10.1093/restud/rdad018>.
- Roth, Jonathan. 2022. “Pretest with Caution: Event-Study Estimates after Testing for Parallel Trends.” *American Economic Review: Insights* 4 (3): 305–22.
- Sant’Anna, Pedro H. C., and Jun Zhao. 2020. “Doubly Robust Difference-in-Differences Estimators.” *Journal of Econometrics* 219 (1): 101–22. <https://doi.org/10.1016/j.jeconom.2020.06.003>.
- Sun, Liyang, and Sarah Abraham. 2021. “Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects.” *Journal of Econometrics*, Themed Issue: Treatment Effect 1, 225 (2): 175–99. <https://doi.org/10.1016/j.jeconom.2020.09.006>.