

DiD for Big Data in R

Theoretical Background

Bradley Setzler, Penn State

This draft compiled on: October 28, 2022

Notation and Identification (1/4)

Notation and Definition of Treatment:

- i : unit of observation.
→ We will often say “individual”.
- t : calendar time. Sample time frame is \mathcal{T} .
→ We will often say “year”.
- $D_{i,t}$: indicator for currently receiving treatment.
→ Permanent treatment: $D_{i,t} = 1 \implies D_{i,t+1} = 1, t \in \mathcal{T}$.
- $G_i \equiv \min\{t : D_{i,t} = 1\}$: time that i first receives treatment.
→ We will often say “cohort” or “onset time”.
→ If i never receives treatment, we can write $G_i = \infty$.
→ Note: $D_{i,t} \equiv 1\{t \geq G_i\}$.
- $E_{i,t} \equiv (t - G_i)$: time since first treatment.
→ We will often say “event time”.
→ Sometimes we will consider fixing an event time e years after treatment versus b years before treatment.
- Potential outcomes: Let $Y_{i,t}(g)$ denote the outcome that is experienced if treated at g .
- Observed outcome:
$$Y_{i,t} = Y_{i,t}(\infty) + \sum_g 1\{G_i = g\}(Y_{i,t}(g) - Y_{i,t}(\infty)).$$

Notation and Identification (2/4)

Goal: Identify ATT at an event time. We assume throughout that the goal is to identify the average treatment effect on the treated (ATT) at event time e . Formally, we seek to identify,

$$ATT_e \equiv \mathbb{E}[Y_{i,t}(g) - Y_{i,t}(\infty) | E_{i,t} = e]$$

The identification challenge is that $\mathbb{E}[Y_{i,t}(\infty) | E_{i,t} = e]$ is a counterfactual object – it is the average outcome that *would have been experienced* by those receiving treatment for e years *if they had not received treatment*.

Decomposition into cohort-specific ATTs. Define,

$$ATT_{g,e} \equiv \mathbb{E}[Y_{i,g+e}(g) - Y_{i,g+e}(\infty) | G_i = g]$$

$$\omega_{g,e} \equiv \mathbb{E}[G_i = g | E_{i,t} = e]$$

where the cohort shares that have been treated at each event time ($\omega_{g,e}$) are observed. Rearranging terms,

$$ATT_e = \sum_g \omega_{g,e} ATT_{g,e}$$

Thus, given e , it is sufficient to identify $ATT_{g,e}$, $\forall g$ s.t. $\omega_{g,e} > 0$.

Notation and Identification (3/4)

Control Group: Define a control group membership indicator $C_{g,e}(G_i)$. At a minimum, $C_{g,e}(G_i)=1 \implies G_i > (g + e)$. We may further restrict C based on context, e.g., some consider the never-treated control group, $C_{g,e}(G_i)=1\{G_i=\infty\}$.

Assumption 1: Parallel trends.

$$\begin{aligned} \exists b < 0 \text{ s.t. } \mathbb{E}[Y_{i,g+e}(\infty) - Y_{i,g+b}(\infty) | C_{g,e}(G_i)=1] \\ = \mathbb{E}[Y_{i,g+e}(\infty) - Y_{i,g+b}(\infty) | G_i = g], \forall e \geq 0 \end{aligned}$$

This restricts the relationship between the treated group $G_i = g$ and the control group $C_{g,e}(G_i)=1$: the change in average outcome for the treated group *would have been the same in the absence of treatment* as that of the control group.

Assumption 2: No anticipation.

$$\exists b < 0 \text{ s.t. } \mathbb{E}[Y_{i,g+b}(g) | G_i = g] = \mathbb{E}[Y_{i,g+b}(\infty) | G_i = g], \forall g$$

This restricts *when* the treated cohorts respond to treatment.

Note: Both assumptions need only hold for the event time e and pre-period b chosen by the researcher.

Notation and Identification (4/4)

Difference-in-differences: Define the population estimator,

$$\text{DiD}_{g,e} \equiv \mathbb{E}[Y_{i,g+e} - Y_{i,g+b} | G_i = g] - \mathbb{E}[Y_{i,g+e} - Y_{i,g+b} | C_{g,e}(G_i) = 1]$$

It depends only on observed outcomes, not counterfactuals.

Impose parallel trends: By the parallel-trends assumption, we can replace the second expectation as follows:

$$\text{DiD}_{g,e} \equiv \mathbb{E}[Y_{i,g+e} - Y_{i,g+b} | G_i = g] - \mathbb{E}[Y_{i,g+e}(\infty) - Y_{i,g+b}(\infty) | G_i = g]$$

Impose no anticipation: By the no-anticipation assumption, we can cancel out the two terms involving $Y_{i,g+b}$:

$$\text{DiD}_{g,e} \equiv \mathbb{E}[Y_{i,g+e} | G_i = g] - \mathbb{E}[Y_{i,g+e}(\infty) | G_i = g] = \text{ATT}_{g,e}$$

Thus, we have proven that $\text{DiD}_{g,e} = \text{ATT}_{g,e}$ if the parallel-trends and no-anticipation assumptions hold for the pair (g, e) .

Result: If parallel-trends and no-anticipation hold $\forall g$ s.t. $\omega_{g,e} > 0$,

$$\text{ATT}_e = \sum_{g: \omega_{g,e} > 0} \omega_{g,e} \text{DiD}_{g,e}$$

Estimator used by DiD for Big Data (1/4)

Estimator based on averages. Replacing population means with sample means, the package implements the following DiD:

$$\text{DiD}_{g,e} = \underbrace{\mathbb{E}[Y_{i,g+e} - Y_{i,g+b} | G_i = g]}_{\text{Difference for treated group}} - \underbrace{(\mathbb{E}[Y_{i,g+e} - Y_{i,g+b} | C_{g,e}(G_i) = 1])}_{\text{Difference for control group}}$$

where, following Callaway and Sant'Anna (2021), we require that parallel-trends and no-anticipation holds for one of these 3 possible control groups:

“all”	$C_{g,e}(G_i) = 1\{G_i > (g + e)\}$
“future-treated”	$C_{g,e}(G_i) = 1\{G_i > (g + e) \ \& \ G_i < \infty\}$
“never-treated”	$C_{g,e}(G_i) = 1\{G_i = \infty\}$

Researcher choices. The DiD researcher must make 3 choices:

1. What is the range of event times e for which you would like ATT_e estimates? Default: $e = -5, \dots, 5$.
2. Which pre-period should be the base? Default: $b = -1$.
3. Which of the 3 control selections C to use? Default: “all”.

Estimator used by DiD for Big Data (2/4)

Fix some (e, b, g) . Consider testing $\text{ATT}_{g,e} = 0$.

Notation: Consider treatment group $\mathcal{T} \equiv 1\{G_i = g\}$ and control group \mathcal{C} , which could be any of the three options above.

Define within- i differences $A_i \equiv Y_{i,g+e} - Y_{i,g+b}$, $i \in \mathcal{T}$ with mean $\mu_A \equiv \mathbb{E}[A_i | i \in \mathcal{T}]$, and $B_i \equiv Y_{i,g+e} - Y_{i,g+b}$, $i \in \mathcal{C}$ with mean $\mu_B \equiv \mathbb{E}[B_i | i \in \mathcal{C}]$, where subscripts are dropped if unambiguous.

Hypothesis Testing: Since $\text{ATT}_{g,e} \equiv \mu_A - \mu_B$, consider,

Test statistic: $\text{DiD}_{g,e} = \bar{A} - \bar{B}$, Null $H_0 : \mu_A - \mu_B = 0$

Central Limit Theorem: Denote the population variances by $\sigma_A^2 \equiv \text{Var}[A_i | i \in \mathcal{T}]$ and $\sigma_B^2 \equiv \text{Var}[B_i | i \in \mathcal{C}]$. By the CLT under the null, and that the samples are drawn independently across i ,

$$\begin{aligned}\bar{A} &\sim_d \mathcal{N}(\mu_A, \sigma_A^2/N_A), \quad \bar{B} \sim_d \mathcal{N}(\mu_B, \sigma_B^2/N_B) \\ \implies \text{DiD}_{g,e} = (\bar{A} - \bar{B}) &\sim_d \mathcal{N}(0, \sigma_A^2/N_A + \sigma_B^2/N_B)\end{aligned}$$

Thus, $\text{SE}(\text{DiD}_{g,e}) = \sqrt{\sigma_A^2/N_A + \sigma_B^2/N_B}$. The empirical counterpart is trivial to compute (e.g. no matrix inversion needed).

Estimator used by DiD for Big Data (3/4)

We usually are not interested in the ATT for a specific cohort g . Instead, we are usually interested in the ATT that is e years after treatment, averaging across cohorts.

Average Effects by Event Time. Let ω_g denote the share of treated units in cohort g . Since treated cohorts are independent of one another at a given e , CLT implies,

$$\text{DiD}_e = \sum_g \omega_g \text{DiD}_{g,e}, \quad \text{SE}(\text{DiD}_e) = \sqrt{\sum_{g \in \mathcal{G}} \omega_g^2 \text{SE}(\text{DiD}_{g,e})^2}$$

Average across Event Times: Similarly, we are often interested in an average of DiD_e across event times. Letting \mathcal{E} denote a set of event times (e.g. $\mathcal{E} = \{1, 2, 3\}$), we are typically interested in the equally-weighted average, which has the following SE:

$$\text{DiD}_{\mathcal{E}} = \frac{1}{|\mathcal{E}|} \sum_{\mathcal{E}} \text{DiD}_e, \quad \text{SE}(\text{DiD}_{\mathcal{E}}) = \frac{1}{|\mathcal{E}|} \sqrt{\sum_{e \in \mathcal{E}} \text{SE}(\text{DiD}_e)^2}$$

The R package provides all possible DiD_e with their SEs by default, and user-friendly options to estimate various $\text{DiD}_{\mathcal{E}}$ and their SEs.

Estimator used by DiD for Big Data (4/4)

Event study parameters: For $\mathcal{H} \in \{\mathcal{T}, \mathcal{C}\}$, one may wish to plot averages across event times:

$$\bar{Y}_{g,e}^{\mathcal{H}} = \sum_{i \in \mathcal{H}_g} Y_{i,g+e}, \quad \text{SE}(\bar{Y}_{g,e}^{\mathcal{H}}) = \sqrt{\frac{\sigma_{\mathcal{H},g,e}^2}{|\mathcal{H}_g|}}, \quad \sigma_{\mathcal{H},g,e}^2 = \text{Var}[Y_{i,g+e} | i \in \mathcal{H}_g]$$

$$\bar{Y}_e^{\mathcal{H}} = \sum_{g \in \mathcal{G}} \omega_g \bar{Y}_{g,e}^{\mathcal{H}}, \quad \text{SE}(\bar{Y}_e^{\mathcal{H}}) = \sqrt{\sum_g \omega_g^2 \text{Var}(\bar{Y}_{g,e}^{\mathcal{H}})},$$

The package provides these means and SEs by default.

Plots: The package also provides automated plots, both for presenting the event study parameters and for presenting the ATT estimates. These can be plotted by (g, e) or by e (average over g).

Accounting for Covariates (1/4)

Recall: Definition of parallel trends.

$$\begin{aligned}\exists b < 0 \text{ s.t. } \mathbb{E}[Y_{i,g+e}(\infty) - Y_{i,g+b}(\infty) | \mathcal{C}_{g,e}(G_i)=1] \\ = \mathbb{E}[Y_{i,g+e}(\infty) - Y_{i,g+b}(\infty) | G_i = g], \forall e \geq 0\end{aligned}$$

Time-invariant Covariate: Suppose potential outcomes in the case with no-treatment are given by the model,

$$Y_{i,t}(\infty) = \alpha_i + \mu_t + X_i \beta_t + \epsilon_{i,t}$$

Is the parallel-trends condition violated? Note that,

$$Y_{i,g+e}(\infty) - Y_{i,g+b}(\infty) = (\mu_{g+e} - \mu_{g+b}) + X_i(\beta_{g+e} - \beta_{g+b})$$

Plugging this into the expectations above, we see that:

1. If $\beta_{g+e} = \beta_{g+b}$, then parallel trends holds, as X_i cancels out in $Y_{i,g+e}(\infty) - Y_{i,g+b}(\infty)$.
2. If $\beta_{g+e} \neq \beta_{g+b}$, then parallel trends holds only if,

$$\mathbb{E}[X_i | \mathcal{C}_{g,e}(G_i)=1] = \mathbb{E}[X_i | G_i = g]$$

Accounting for Covariates (2/4)

Recap: If $Y_{i,t}(\infty) = \alpha_i + \mu_t + X_i\beta_t + \epsilon_{i,t}$ and $\beta_{g+e} \neq \beta_{g+b}$, then parallel trends holds only if,

$$\mathbb{E}[X_i | C_{g,e}(G_i)=1] = \mathbb{E}[X_i | G_i = g]$$

Discrete correction: If $X_i \in \mathcal{X}$ is discrete, parallel trends holds if we condition on treatment and control groups with the same X_i :

$$\mathbb{E}[X_i | C_{g,e}(G_i)=1 \ \& \ X_i = x] - \mathbb{E}[X_i | G_i = g \ \& \ X_i = x] = x - x = 0$$

I.e. parallel trends must hold mechanically when conditioning on treatment and control groups within the same $X_i = x$ bin.

Testing: Define $\text{DiD}_{g,e}(x)$ as follows:

$$\mathbb{E}[Y_{i,g+e} - Y_{i,g+b} | G_i = g \ \& \ X_i = x] - \mathbb{E}[Y_{i,g+e} - Y_{i,g+b} | C_{g,e}(G_i)=1 \ \& \ X_i = x]$$

Then, we can test that $\text{ATT}_{g,e}$ is zero using,

$$\text{DiD}_{g,e} = \sum_{\mathcal{X}} \omega_g(x) \text{DiD}_{g,e}(x), \quad \omega_g(x) = \mathbb{E}[X_i = x | G_i = g]$$

$$\text{SE}(\text{DiD}_{g,e}) = \sqrt{\sum_{x \in \mathcal{X}} (\omega_g(x))^2 \text{SE}(\text{DiD}_{g,e}(x))^2}$$

Accounting for Covariates (3/4)

Time-varying Covariates: Suppose that it is the covariates X rather than the coefficient β that varies over time:

$$Y_{i,t}(\infty) = \alpha_i + \mu_t + X_{i,t}\beta + \epsilon_{i,t}$$

Is parallel trends violated? Note that,

$$Y_{i,g+e}(\infty) - Y_{i,g+b}(\infty) = (\mu_{g+e} - \mu_{g+b}) + (X_{i,g+e} - X_{i,g+b})\beta$$

Plugging this into the expectations above, parallel trends requires

$$\mathbb{E}[X_{i,g+e} - X_{i,g+b} | \mathcal{C}_{g,e}(G_i)=1] = \mathbb{E}[X_{i,g+e} - X_{i,g+b} | G_i = g]$$

Though this will not hold in general, it holds if X is age or time, as $X_{i,g+e} - X_{i,g+b} = (e - b)$ is the same for $G_i = g$ and \mathcal{C} .

Discrete correction for time-varying covariates: Suppose $X \in \mathcal{X}$ is discrete. Then, the difference $\tilde{X}_i \equiv X_{i,g+e} - X_{i,g+b}$ is also discrete. We can condition on $\tilde{X}_i = x$ and use the same approach we used for time-invariant covariates above.

Accounting for Covariates (4/4)

Recall: We consider the case in which X varies over time:

$$Y_{i,t}(\infty) = \alpha_i + \mu_t + X_{i,t}\beta + \epsilon_{i,t}$$

Regression representation: Let $\mathcal{H}_{g,e}$ denote the union of the treated and control group for a particular (g, e) pair. Then,

$$\begin{aligned}(Y_{i,g+e} - Y_{i,g+b}) &= \tilde{\mu} + 1\{G_i = g\}(Y_{i,g+e}(g) - Y_{i,g+e}(\infty)) \\ &\quad + (X_{i,g+e} - X_{i,g+b})\beta + \tilde{\epsilon}, \quad \forall i \in \mathcal{H}_{g,e}\end{aligned}$$

which is a regression equation *with no fixed effects* that can be estimated separately for each g, e pair. In particular, the coefficient on $1\{G_i = g\}$ recovers $\mathbb{E}[Y_{i,g+e}(g) - Y_{i,g+e}(\infty) | G_i = g]$.

OLS: Thus, we can apply OLS to regress $Y_{i,g+e} - Y_{i,g+b}$ on $1\{G_i = g\}$ and $X_{i,g+e} - X_{i,g+b}$, using the standard OLS point estimate and SE for the coefficient on $1\{G_i = g\}$. Note that this OLS formulation accommodates both discrete and continuous $X_{i,t}$.