DiD for Big Data in R

Theoretical Background

Bradley Setzler, Penn State

This draft compiled on: November 21, 2022

Notation and Parameters of Interest (1/2)

Notation and Definition of Treatment:

- i: unit of observation.
  $\rightarrow$ We will often say "individual".
- t: calendar time. Sample time frame is $\mathcal{T}$.
  $\rightarrow$ We will often say "year".
- $D_{i,t}$: indicator for currently receiving treatment.
  $\rightarrow$ 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.
  $\rightarrow$ We will often say "cohort" or "onset time".
  $\rightarrow$ If i never receives treatment, we can write $G_i = \infty$.
  $\rightarrow$ Note: $D_{i,t} \equiv 1\{t \geq G_i\}$.
- $E_{i,t} \equiv (t - G_i)$: time since first treatment.
  $\rightarrow$ We will often say "event time".
  $\rightarrow$ 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 Parameters of Interest (2/2)

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,

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

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,

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

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

$$\mathrm{ATT}_e = \sum_g \omega_{g,e} \mathrm{ATT}_{g,e}$$

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

Identification (1/2)

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

Assumption 1: Parallel trends.

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

This restricts the relationship between the treated group $G_i = g$ and the control group $\mathcal{C}_{g,e}(G_i)=1$: 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.

Identification (2/2)

Difference-in-differences: Define the sample DiD estimator s.t.,

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

It is defined in terms of observable outcomes, not counterfactuals.

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

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

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

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

Identification: Thus, $\text{DiD}_{g,e} \to \text{ATT}_{g,e}$ if the parallel-trends and no-anticipation assumptions hold for the pair (g, e).

Identification for the event-time average: If parallel-trends and no-anticipation hold $\forall g$ s.t. $\omega_{g,e} > 0$, the above results imply,

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

Estimation and Inference (1/3)

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

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

where we require that parallel-trends and no-anticipation holds for one of these 3 possible control groups:

$$\begin{array}{rl} \text{“all”} & \mathcal{C}_{g,e}(G_i) = 1\{G_i > (g+e)\} \\ \text{“future-treated”} & \mathcal{C}_{g,e}(G_i) = 1\{G_i > (g+e) \;\&\; G_i < \infty\} \\ \text{“never-treated”} & \mathcal{C}_{g,e}(G_i) = 1\{G_i = \infty\} \end{array}$$

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: $e = -5, ..., 5$.
2. Which pre-period should be the base? Default: $b = -1$.
3. Which of the 3 control selections $\mathcal{C}$ to use? Default: “all”.

Estimation and Inference (2/3)

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

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

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

Hypothesis Testing: Since $\text{ATT}_{g,e} \equiv \mu_{A,g} - \mu_{B,g}$, consider,

$$\text{Test statistic: } \text{DiD}_{g,e} = \overline{A}_g - \overline{B}_g, \quad \text{Null } H_0 : \mu_{A,g} - \mu_{B,g} = 0$$

Central Limit Theorem: Denote the population variances by $\sigma^2_{A,g} \equiv \text{Var}[A_i | i \in \mathcal{T}_g]$ and $\sigma^2_{B,g} \equiv \text{Var}[B_i | i \in \mathcal{C}_g]$. By the CLT under the null, with samples drawn independently across i,

$$\overline{A}_g \sim_d \mathcal{N}\left(\mu_{A,g},\ \sigma^2_{A,g}/N_{A,g}\right), \quad \overline{B}_g \sim_d \mathcal{N}\left(\mu_{B,g},\ \sigma^2_{B,g}/N_{B,g}\right)$$

$$\implies \text{DiD}_{g,e} = \left(\overline{A}_g - \overline{B}_g\right) \sim_d \mathcal{N}\left(0,\ \sigma^2_{A,g}/N_{A,g} + \sigma^2_{B,g}/N_{B,g}\right)$$

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

Estimation and Inference (3/3)

Average Effects by Event Time. Let $\omega_g \equiv \mathbb{E}[G_i = g | G_i < \infty]$ denote the share of treated units in cohort g. We can define,

$$\text{DiD}_e \equiv \sum_{g \in \mathcal{G}} \omega_g \text{DiD}_{g,e} = \sum_{g \in \mathcal{G}} \omega_g \left(\overline{A}_g - \overline{B}_g \right), \ \text{ATT}_e \equiv \sum_{g \in \mathcal{G}} \omega_g (\mu_{A,g} - \mu_{B,g})$$

Recall: For a given (g, e), the treated and control group are mutually exclusive. Thus, independence across i ensures that $\text{Cov}(\overline{A}_g, \overline{B}_g) = 0$. We used this result on the previous slide to obtain simple SEs with no covariance terms for $\text{DiD}_{g,e} = \overline{A}_g - \overline{B}_g$.

Repeated i in the event time average. Unlike $\text{DiD}_{g,e}$, DiD_e depends on both $\overline{B}_g$ and $\overline{B}_{g'}$ for different cohorts g, g'. The same individual i often appears in multiple control groups (e.g. never-treated units), implying $\text{Cov}(\overline{B}_g, \overline{B}_{g'}) \neq 0$, so we cannot ignore covariance terms when calculating the SE for DiD_e. Similarly, if $g < g'$, we often have $\text{Cov}(\overline{B}_g, \overline{A}_{g'}) \neq 0$, since some members i of the control group at g later enter the treated g'.

Delta Method: Stacking the various $\overline{A}_g$ and $\overline{B}_g$ terms used by DiD_e, we apply the delta-method to obtain a simple matrix algebra representation of the standard error w.r.t. their covariance matrix.

Regression Representation (1/4)

Recall: Hypothesis Testing for a single cohort-event pair:
Above, we showed inference on $\text{DiD}_{g,e}$ using,

Test statistic: $\text{DiD}_{g,e} = \overline{A}_g - \overline{B}_g$, Null $H_0 : \mu_{A,g} - \mu_{B,g} = 0$

$$\implies \text{DiD}_{g,e} = \left(\overline{A}_g - \overline{B}_g\right) \sim_d \mathcal{N}\left(0,\ \sigma^2_{A,g}/N_{A,g} + \sigma^2_{B,g}/N_{B,g}\right)$$

Vector notation: Define the following:

- Treatment and control group corresponding to (g, e) is $\mathcal{H}_{g,e} \equiv \{i \ :\ G_i = g \text{ or } \mathcal{C}_{g,e}(G_i) = 1\}$.
- Difference relative to b: $Y^{\Delta(b)}_{i,g+e} \equiv (Y_{i,g+e} - Y_{i,g+b})$
- Vector of outcomes: $\tilde{Y}^{\Delta(b)}_{g+e} = (Y^{\Delta(b)}_{i,g+e})_{i \in \mathcal{H}_{g,e}}$.
- Vector of treatments: $\tilde{D}_{g,e} = (1\{G_i = g\})_{i \in \mathcal{H}_{g,e}}$
- Vector of errors: $\tilde{\epsilon}^{\Delta(b)}_{g+e} = (\tilde{\epsilon}^{\Delta(b)}_{i,g+e})_{i \in \mathcal{H}_{g,e}}$

Vector regression representation: We can write,

$$\tilde{Y}^{\Delta(b)}_{g+e} = \tilde{\mu}_{g,e} + \tilde{D}_{g,e}\delta_{g,e} + \tilde{\epsilon}^{\Delta(b)}_{g+e}$$

Applying OLS to the regression, $\delta^{\text{OLS}}_{g,e} = \text{DiD}_{g,e}$ holds numerically, and the standard error estimate is also the same. Note: Robust (Huber) SEs are required in OLS to replicate the SEs above.

Regression Representation (2/4)

Comment 1: Avoids unit fixed-effects.

- Implementations of DiD using OLS often control for unit fixed-effects to control for composition differences.
- My approach sidesteps this issue by working with the differences $(Y_{i,g+e} - Y_{i,g+b})$ and a single (g, e) pair at a time, which mechanically removes fixed-effects.

Comment 2: Avoids clustering on unit.

- Implementations of DiD using OLS typically have to cluster on i because the same individual i appears on multiple rows of the regression (a pre-period and a post-period $Y_{i,t}$).
- My approach sidesteps this issue by working with the differences $(Y_{i,g+e} - Y_{i,g+b})$ and a single (g, e) pair at a time, which implies that each individual appears on at most one row – there is no repeated i to cluster on.

Comment 3: Avoids issues with unbalanced panels.

- Some DiD implementations include i that are missing in the base period, or require i to be non-missing in all periods.
- My approach avoids these two extremes: $(Y_{i,g+e} - Y_{i,g+b})$ is exactly the contribution of unit i to $\text{DiD}_{g,e}$, so i should be included if and only if $Y_{i,g+e}$ and $Y_{i,g+b}$ are observed.

Regression Representation (3/4)

Stacked Notation: For a given e, we can stack the vectors defined on the previous slide across cohorts g:

- Define $\mathcal{G}_e \equiv \{g \ : \ \omega_{g,e} > 0\}$, which are all the treatment cohorts g for which we observe the outcome at event time e.
- $\overline{Y}_e^{\Delta(b)} \equiv (\tilde{Y}_{g+e}^{\Delta(b)})_{g \in \mathcal{G}_e}$, $\overline{\epsilon}_e^{\Delta(b)} \equiv (\tilde{\epsilon}_{g+e}^{\Delta(b)})_{g \in \mathcal{G}_e}$, $\overline{D}_e \equiv (\tilde{D}_{g,e})_{g \in \mathcal{G}_e}$. Note that these vectors can re-use the same individual i on multiple rows (but at different points in time for i).
- Let $H(g)$ be an indicator that is 1 for the set of rows in $\overline{D}_e$ that correspond to the treatment and control groups for treated cohort g, and zero otherwise.

Stacked Regression for Multiple DiD: Given this notation,

$$\overline{Y}_e^{\Delta(b)} = \sum_{g \in \mathcal{G}_e} H(g)\overline{\mu}_e + \sum_{g \in \mathcal{G}_e} H(g)\overline{D}_e\overline{\delta}_e + \overline{\epsilon}_e^{\Delta(b)}$$

where $\overline{\delta}_e \equiv (\delta_{g,e})_{g \in \mathcal{G}_e}$, $\overline{\mu}_e \equiv (\tilde{\mu}_{g,e})_{g \in \mathcal{G}_e}$. Applying OLS to the regression, $\overline{\delta}_e^{\text{OLS}} = (\delta_{g,e}^{\text{OLS}})_{g \in \mathcal{G}_e} = (\text{DiD}_{g,e})_{g \in \mathcal{G}_e}$ holds numerically.

Regression Representation (4/4)

Recall: Stacked Regression for Multiple DiD:

$$\overline{Y}_e^{\Delta(b)} = \sum_{g \in \mathcal{G}_e} H(g)\overline{\mu}_e + \sum_{g \in \mathcal{G}_e} H(g)\overline{D}_e\overline{\delta}_e + \overline{\epsilon}_e^{\Delta(b)}$$

Applying OLS, $\overline{\delta}_e^{\text{OLS}} = (\delta_{g,e}^{\text{OLS}})_{g \in \mathcal{G}_e} = (\text{DiD}_{g,e})_{g \in \mathcal{G}_e}$.

Variance-covariance matrix: Let $\Omega_e \equiv \text{Var}(\overline{\delta}_e^{\text{OLS}})$ denote the variance-covariance matrix for the vector of estimates $\overline{\delta}_e$.

Since each observation used in the stacked regression is sampled independently across i, the only rows that can be correlated are the rows that correspond to the same unit i.

This satisfies the standard Liang-Zeger assumptions, so we should generally use clustered standard errors on the unit identifier i.

Test Statistic and Inference: Let $\omega_e = (\omega_{g,e})_{g \in \mathcal{G}_e}$, which satisfies $\omega_e'\mathbf{1} = 1$. The test statistic of interest and its variance are,

$$\text{DiD}_e = \omega_e'\overline{\delta}_e, \quad \text{Var}(\text{DiD}_e) = \omega_e'\Omega_e\omega_e$$

where we replace Ω_e with its (clustered) OLS estimate in practice.

Accounting for Time-varying Covariates (1/2)

Recall: Definition of parallel trends.

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

Time-varying Covariates Model: Suppose that there are covariates X that vary over time and affect the potential outcome through a time-invariant coefficient β:

$$Y_{it}(\infty) = \alpha_i + \mu_t + X_{it}\beta + \epsilon_{it}$$

Is parallel trends violated? Plugging this model into the definition above, parallel trends requires,

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

Thus, we only need to control for time-varying covariates that evolve differentially for the treated and control groups.

For example, age changes the same for all i, so it doesn't make sense to control for age.

Accounting for Time-varying Covariates (2/2)

Recall: Parallel trends violation:

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

Regression Correction for a Single DiD: Define the vector $\tilde{X}^{\Delta(b)}_{g,e} \equiv (X_{i,g+e} - X_{i,g+b})_{i \in \mathcal{H}_{g,e}}$. To control for these time-varying covariates, we can modify the vector regression as follows:

$$\tilde{Y}^{\Delta(b)}_{g+e} = \tilde{\mu}_{g,e} + \tilde{D}_{g,e}\delta_{g,e} + \tilde{X}^{\Delta(b)}_{g,e}\beta_{g,e} + \tilde{\epsilon}^{\Delta(b)}_{g+e}$$

This controls directly for the parallel-trends violation. We allow for $\beta_{g,e}$ to be cohort-event-specific, which is very flexible.

Regression Correction for a Multiple DiD: Define the stacked terms $\overline{X}^{\Delta(b)}_{e} \equiv (\tilde{X}^{\Delta(b)}_{g+e})_{g \in \mathcal{G}_e}$, $\overline{\beta}_e = (\beta_{g+e})_{g \in \mathcal{G}_e}$. The correction is,

$$\overline{Y}^{\Delta(b)}_{e} = \sum_{g \in \mathcal{G}_e} H(g)\overline{\mu}_e + \sum_{g \in \mathcal{G}_e} H(g)\overline{D}_e\overline{\delta}_e + \sum_{g \in \mathcal{G}_e} H(g)\overline{X}^{\Delta(b)}_{e}\overline{\beta}_e + \overline{\epsilon}^{\Delta(b)}_{e}$$

Accounting for Common Group Shocks (1/2)

Recall: Definition of parallel trends.

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

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

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

where X_i is a $N \times K$ matrix of indicators for membership in group $k = 1, ..., K$, and ξ is a K-length vector of group-specific shocks.

Group Shocks: Is the parallel-trends assumption violated?

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

Plugging this into the expectations, parallel trends requires,

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

Group Shock Bias: If X_i and $(\xi_{g+e} - \xi_{g+b})$ are independent conditional on treatment/control status, parallel trends still holds.

However, if certain X_i are selected for treatment status based on their unobserved growth $(\xi_{g+e} - \xi_{g+b})$, parallel trends fails.

Accounting for Common Group Shocks (2/2)

Recap: If $Y_{i,t}(\infty) = \alpha_i + \mu_t + X_i \xi_t + \epsilon_{i,t}$, parallel trends requires $\mathbb{E}[X_i(\xi_{g+e} - \xi_{g+b}) | \mathcal{C}_{g,e}(G_i) = 1] = \mathbb{E}[X_i(\xi_{g+e} - \xi_{g+b}) | G_i = g]$.

Clustered SEs for group shocks: If parallel-trends holds, then, DiD is consistent, but the i.i.d. assumption fails for the errors due to the common group shock.

This is easy to correct by using clustered (Liang-Zeger) standard errors based on the group membership X.

Controlling for group shocks: If parallel-trends fails due to the common group shocks, we need to control for the group shocks.

This can be done by forming the high-dimensional matrix that fully interacts X with event-time, then controlling for this in the stacked regression with time-varying covariates defined above.

Because this is a high-dimension fixed-effects regression, it deviates from our no-fixed-effect philosophy. However, it seems unavoidable in the presence of treatment-correlated group-specific shocks.