Causal Inference using Difference-in-Differences

Lecture 7: Leveraging repeated cross-sectional data

Pedro H. C. Sant'Anna
Emory University

January 2025

Introduction

DiD procedures with Covariates

- We can include covariates into DiD to allow for covariate-specific trends.
  - Regression adjustments;
  - Inverse probability weighting;
  - Doubly Robust (augmented inverse probability weighting);
- All these are implemented in **DRDID** and **did** R packages, and **drdid** and **csdid** Stata packages.
- We can use them with **panel data** or **repeated cross-sectional data**.

Are there differences between these two cases?

For a given sample size, how much efficiency do we lose by not having balanced panel data?

We will focus on the 2x2 case

Let’s review our assumptions

Assumptions in 2x2 setup

Assumption (Conditional Parallel Trends Assumption)

$$\mathbb{E}\left[Y_{t=2}(\infty)|G=2,X\right]-\mathbb{E}\left[Y_{t=1}(\infty)|G=2,X\right]=\mathbb{E}\left[Y_{t=2}(\infty)|G=\infty,X\right]-\mathbb{E}\left[Y_{t=1}(\infty)|G=\infty,X\right] \quad a.s.$$

Assumption (No-Anticipation)

For all units i, $Y_{i,t}(g)=Y_{i,t}(\infty)$ for all groups in their pre-treatment periods, i.e., for all $t<g$.

Assumption (Strong Overlap Assumption)

The conditional probability of belonging to the treatment group, given observed characteristics X, is uniformly bounded away from 1. That is, for some $\epsilon>0$, $\mathbb{P}[G=2|X]<1-\epsilon$ almost surely.

Different Sampling Schemes

Panel data sampling scheme

Assumption (Panel Data Sampling Scheme)

The data $\{Y_{i,t=1}, Y_{i,t=2}, G_i, X_i\}_{i=i}^{n}$ *is a random sample of the population of interest.*

- Assumption states that we observe the same units in all time periods:
  No need to worry about compositional changes!
- Assumption does not restrict dependence between realized outcomes across periods.
- We observe covariates for all individuals.

Stationary Repeated cross-section data sampling scheme

Assumption (Stationary Repeated Cross-Section Data Sampling Scheme)

The pooled repeated cross-section data $\{Y_i, G_i, T_i, X_i\}_{i=1}^{n}$ *consist of iid draws from the mixture distribution*

$$\begin{aligned} P\left(Y \leq y, X \leq x, G = g, T = t\right) &= 1\{t = 2\} \cdot \lambda \cdot P\left(Y_{t=2} \leq y, X \leq x, G = g | T = 2\right) \\ &\quad + 1\{t = 1\} \cdot (1 - \lambda)\, P\left(Y_{t=1} \leq y, X \leq x, G = g | T = 1\right), \end{aligned}$$

where $(y, x, g, t) \in \mathbb{R} \times \mathbb{R}^k \times \{2, \infty\} \times \{1, 2\}$, $\lambda = \mathbb{P}\left(T = 2\right) \in (0, 1)$.

Furthermore, $\left(\mathrm{G}, \mathrm{X}\right) | \mathrm{T} = 1 \overset{d}{\sim} \left(\mathrm{G}, \mathrm{X}\right) | \mathrm{T} = 2.$

Stationary Repeated Cross-Section Data Sampling Schemes

- It accommodates the binomial sampling scheme where an observation i is randomly drawn from either $(Y_{t=2}, G, X)$ or $(Y_{t=1}, G, X)$ with fixed probability λ (here, T is a non-degenerated random variable).

- It also accommodates the "conditional" sampling scheme where $n_{t=2}$ observations are sampled from $(Y_{t=2}, G, X)$, $n_{t=1}$ observations are sampled from $(Y_{t=1}, G, X)$ and $\lambda = n_{t=2}/(n_{t=1} + n_{t=2})$ (here, T is treated as fixed).

- However, this assumption rules out compositional changes across periods: we are sampling from the same population of interest in both periods.

- RCS results of Abadie (2005) and Sant'Anna and Zhao (2020) really depends on this!

What if I want to allow for compositional changes?

Repeated cross-section data sampling scheme

Assumption (Repeated Cross-Section Data Sampling Scheme)

The pooled data $\{Y_i, G_i, T_i, X_i\}_{i=1}^{n}$ *consist of iid draws from*

$$\begin{aligned} P\left(Y \leq y, X \leq x, G = g, T = t\right) \quad = \quad & 1\{t=2\} \cdot \lambda \cdot P\left(Y_{t=2} \leq y, X \leq x, G = g | T = 2\right) \\ & + 1\{t=1\} \cdot (1-\lambda)\, P\left(Y_{t=1} \leq y, X \leq x, G = g | T = 1\right), \end{aligned}$$

where $(y, x, g, t) \in \mathbb{R} \times \mathbb{R}^{k} \times \{2, \infty\} \times \{1, 2\}$, $\lambda = \mathbb{P}\left(T = 2\right) \in (0, 1)$.

- Not many results are available for this case in the literature.
- In Sant'Anna and Xu (2023), we have worked out the details on how to allow compositional changes while doing DiD.

Repeated cross-section data sampling scheme

- Not many results are available for this case in the literature.
- In Sant'Anna and Xu (2023), we have worked out the details on how to allow compositional changes while doing DiD.
    - Derive the semiparametric efficiency bound for this case;
    - Propose nonparametric, data-driven estimators that achieve the semiparametric efficiency bound;
    - Propose DML estimators that can leverage modern ML methods;
    - Propose Hausman-type tests for compositional changes.
    - Derive the semiparametric efficiency bound for cases where part of the data is a balanced panel and another part is repeated cross-sectional (like in CPS).
- We won't have time to dig into these details today, as these results are very recent—I may update these slides in the future.

Difference 1:

Most DiD estimators with RCS data impose a no-compositional changes assumption.

This is not the case when panel data is available.

Doubly Robust estimators

Doubly Robust DiD procedure with Panel

- Sant'Anna and Zhao (2020) considered the following doubly robust estimand when panel data are available:

$$ATT^{dr,p} = \mathbb{E}\left[\left(\frac{D}{\mathbb{E}[D]} - \frac{\frac{p(X)(1-D)}{1-p(X)}}{\mathbb{E}\left[\frac{p(X)(1-D)}{1-p(X)}\right]}\right)\left(\Delta Y - m_{\Delta}^{G=\infty}(X)\right)\right],$$

where $D = 1\{G = 2\}$.

- Note that we only need to model the evolution of Y given X for the untreated units.
- No need to model the behavior of the treated units! Pretty neat!
- Estimator can achieve the semiparametric efficiency bound (even without modelling the out. evol. among treated).

Doubly Robust DiD procedure with stationary repeated cross-section

- Sant'Anna and Zhao (2020) **two** different DR estimands with RCS data.
- First one mimics the panel data one:

$$ATT_1^{dr,rc} =$$

$$\mathbb{E}\left[\left(\frac{D\cdot 1\{T=2\}}{\mathbb{E}\left[D\cdot 1\{T=2\}\right]} - \frac{D\cdot 1\{T=1\}}{\mathbb{E}\left[D\cdot 1\{T=1\}\right]}\right)\cdot\left(Y - \left(m^{rc}_{G=\infty,t=2}(X) - m^{rc}_{G=\infty,t=1}(X)\right)\right)\right]$$

$$-\mathbb{E}\left[\left(\frac{\frac{p(X)(1-D)\cdot 1\{T=2\}}{1-p(X)}}{\mathbb{E}\left[\frac{p(X)(1-D)\cdot 1\{T=2\}}{1-p(X)}\right]} - \frac{\frac{p(X)(1-D)\cdot 1\{T=1\}}{1-p(X)}}{\mathbb{E}\left[\frac{p(X)(1-D)\cdot 1\{T=1\}}{1-p(X)}\right]}\right)\cdot\left(Y - \left(m^{rc}_{G=\infty,t=2}(X) - m^{rc}_{G=\infty,t=1}(X)\right)\right)\right]$$

- Model the evolution of Y given X for the untreated units:
  need two models because we do not observe the same units over time.

However, this DR estimator is not efficient!

We can do better!

Doubly Robust DiD procedure with repeated cross-section

Sant'Anna and Zhao (2020) second DR DiD estimand also relies on outcome regression models for the treated unit:

$$ATT_2^{dr,rc} = ATT_1^{dr,rc}$$

$$+ \left(\mathbb{E}\left[m^{rc}_{G=2,t=2}(X) - m^{rc}_{G=\infty,t=2}(X)\middle| D=1\right] - \mathbb{E}\left[m^{rc}_{G=2,t=2}(X) - m^{rc}_{G=\infty,t=2}(X)\middle| D=1, T=2\right]\right)$$

$$- \left(\mathbb{E}\left[m^{rc}_{G=2,t=1}(X) - m^{rc}_{G=\infty,t=1}(X)\middle| D=1\right] - \mathbb{E}\left[m^{rc}_{G=2,t=1}(X) - m^{rc}_{G=\infty,t=1}(X)\middle| D=1, T=1\right]\right),$$

- Need to model the behavior of the treated units.
- Estimator can now achieve the semiparametric efficiency bound!

Doubly Robust DiD procedure with repeated cross-section

- Both DR DiD estimators for RCS data are consistent for the ATT under the exact same conditions:

- Even if the regression model for the outcome evolution for the treated group is misspecified, $ATT_2^{dr,rc}$ is consistent for the ATT (provided that either the pscore or the regression models for out. evol. among untreated units are correctly specified).

- However, in general, $ATT_2^{dr,rc}$ is more efficient than $ATT_1^{dr,rc}$.

- In fact, Sant'Anna and Zhao (2020) shows that $ATT_2^{dr,rc}$ is (locally) semiparametrically efficient.

Difference 2:

We need to model the outcome evol. among treated units in RCS if we want to achieve semiparametric efficiency bound!

Comparing Semiparametric efficiency bounds

Semiparametric efficiency: Panel vs. Repeated cross-section data

Corollary

Assume that T is independent of (Y_1, Y_0, D, X), and the other regularity conditions stated in Sant'Anna and Zhao (2020) hold. Then,

$$
\begin{aligned}
& V_{eff}^{rc,2} - V_{eff}^{p,2} \\
& = \frac{1}{\mathbb{E}[D]^2} \mathbb{E}\left[D \left(\sqrt{\frac{1-\lambda}{\lambda}} \left(Y_{t=2} - m_{G=2,t=2}(X) \right) + \sqrt{\frac{\lambda}{1-\lambda}} \left(Y_{t=1} - m_{G=2,t=1}(X) \right) \right)^2 \right. \\
& \left. + \frac{(1-D)\, p(X)^2}{(1-p(X))^2} \left(\sqrt{\frac{1-\lambda}{\lambda}} \left(Y_{t=2} - m_{G=\infty,t=2}(X) \right) + \sqrt{\frac{\lambda}{1-\lambda}} \left(Y_{t=1} - m_{G=\infty,t=1}(X) \right) \right)^2 \right] \geq 0,
\end{aligned}
$$

where $\lambda = \mathbb{P}(T = 2)$

Semiparametric efficiency: Panel vs. Repeated cross-section data

- Efficiency loss is convex in λ:
  loss of efficiency is bigger when the pre and post-treatment sample sizes are more imbalanced.

- Optimal λ depends on the data: $\lambda = \tilde{\sigma}_2 / \left(\tilde{\sigma}_1 + \tilde{\sigma}_2\right)$, where, for $t = 1, 2$

$$\tilde{\sigma}_t^2 = \mathbb{E}\left[D\left(Y_t - m_{G=2,t}\left(X\right)\right)^2 + \frac{\left(1-D\right)p\left(X\right)^2}{\left(1-p\left(X\right)\right)^2}\left(Y_t - m_{g=\infty,t}\left(X\right)\right)^2\right]$$

- In principle, one may benefit from "oversampling" from either the pre or post-treatment period.

- However, it is, in general, not feasible to know the optimal λ during the design stage: $\tilde{\sigma}_2^2$ depends on post-treatment data!

- $\lambda = 0.5$ is a "reasonable" choice, in practice.

Difference 3:

The best RCS DiD estimator is always less efficient than the best Panel DiD estimator for a given sample size.

Take-way messages

Take-way message

- When dealing with RCS data, we must consider compositional changes.
  - **did** R package and **csdid** Stata package assumes stationarity.
  - Check Sant'Anna and Xu (2023) for how to make your model more flexible and test for compositional changes.
- With RCS, there are important benefits of modeling the outcome evolution of treated units when doing DiD.
- Overall, for a given sample size, RCS is less efficient than Panel data.

  Loss of efficiency is bigger when the pre and post-treatment sample sizes are more imbalanced.
- See Sant'Anna and Zhao (2020) and Sant'Anna and Xu (2023) for more details!

References

Abadie, Alberto, "Semiparametric Difference-in-Differences Estimators," *The Review of Economic Studies*, 2005, *72* (1), 1–19.

Sant'Anna, Pedro H. C. and Qi Xu, "Difference-in-Differences with Compositional Changes," *arXiv:2304.13925*, 2023.

Sant'Anna, Pedro H. C. and Jun Zhao, "Doubly robust difference-in-differences estimators," *Journal of Econometrics*, November 2020, *219* (1), 101–122.