Unpooled-DiD
Difference-in-Differences with unpoolable data

Sunny Karim, Matthew Webb, Nicole Austin and Erin Strumpf

August 1, 2023

Table of Content

Canonical DiD Setup: 2X2

- Two groups: Treatment ($G_i = 1$) and Control ($G_i = 0$)
- Two time periods: Pre ($t = 0$) and Post ($t = 1$)

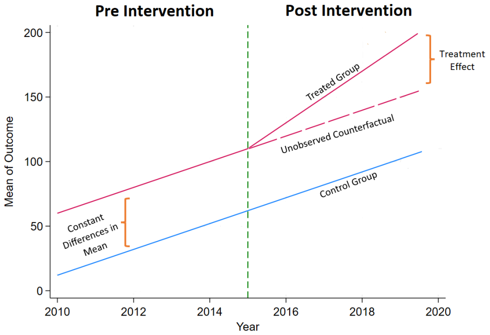


Figure 1: Canonical DiD setup

Key Identifying Assumptions

1. **Parallel trends/Conditional Parallel trends assumption** (Roth et al., 2022)
2. **No anticipation/strong exogeneity** (De Chaisemartin and d'Haultfoeuille, 2020a; Abadie, 2005)
3. **Homogeneous Treatment Effect** across both time and units (Roth et al., 2022)
4. **No staggered adoption** (De Chaisemartin and d'Haultfoeuille, 2020a; Callaway and Sant'Anna, 2021)
5. **Single isolationed treatment** (de Chaisemartin and D'Haultfœuille, 2020b)
6. The data is poolable

Conventional Estimate

- **Conventional Estimation** with repeated cross sectional data:

$$Y_{i,t} = \beta_0 + \beta_1 treat_i + \beta_2 post_t + \beta_3 treat_i * post_t + \beta_4 X_i + \epsilon_{i,t} \quad (1)$$

- Estimate of the ATT:

$$\begin{aligned} \widehat{ATT} = & \Big[E[Y_1 | G_i = 1, X_i] - E[Y_0 | G_i = 1, X_i] \Big] \\ & - \Big[E[Y_1 | G_i = 0, X_i] - E[Y_0 | G_i = 0, X_i] \Big] \end{aligned} \quad (2)$$

What is $\hat{\beta}_3$?

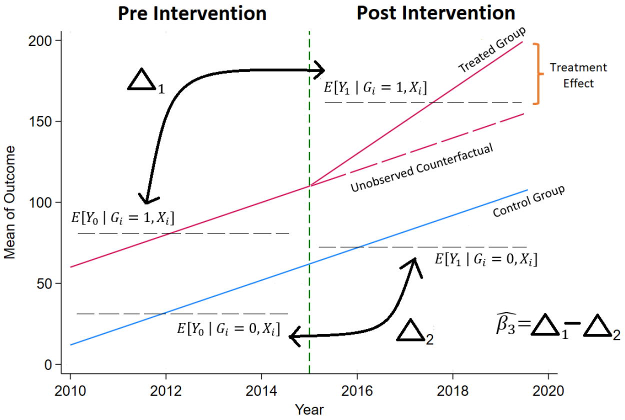


Figure 2: $\hat{\beta}_3$ from the conventional estimate

Why is data unpoolable?

- Prevalent problem in **Health Economics**
- Administrate Health Data is unpoolable
  - legal restrictions in data sharing (siloed data)
  - In Canada, separate provincial health insurers
- Data cannot be combined together to do DiD analysis using traditional methods
- Missed opportunity for research
  - CIHR's Institute of Health Services and Policy Research labelled Canada as a "policy laboratory"

Unpooled Regressions with covariates

- For j = {T, C}

$$Y_{i,t}^{T} = \lambda_1^j pre_t + \lambda_2^j post_t + \lambda_3 X_{i,t}^j + \nu_{i,t}^j \quad (3)$$

Or alternatively:

$$Y_{i,t}^{j} = \lambda_0 + \lambda_1^j post_t^j + \lambda_2^j X_{i,t}^j + \eta_{i,t}^j \quad (4)$$

Proof

- $\widehat{ATT} = (\widehat{\lambda_2^T} - \widehat{\lambda_1^T}) - (\widehat{\lambda_2^C} - \widehat{\lambda_1^C}) = \hat{\lambda_1^T} - \hat{\lambda_1^C}$

What is the ATT?

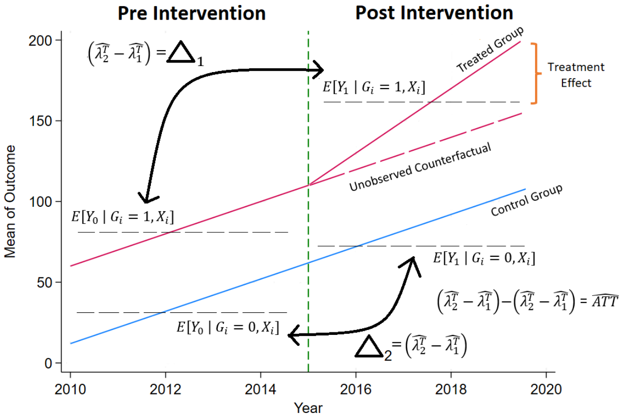


Figure 3: $\hat{ATT}$ from the unpooled regressions

Standard Errors and p values

- Standard Error $= \sqrt{(SE_{\hat{\lambda_1^T}})^2 + (SE_{\hat{\lambda_2^T}} + (SE_{\hat{\lambda_1^C}})^2 + (SE_{\hat{\lambda_2^C}})^2}$
- This will be equivalent to:
- Standard Error $= \sqrt{(SE_{\lambda_1})^2 + (SE_{\gamma_1})^2}$
- t-stats for inference $= \dfrac{\widehat{ATT}}{\text{Standard Error}}$

Data Generating Process (DGP)

- **3 Cases**
  1. No covariates
  2. Single time invariant covariate and homogeneous effect of X
  3. Single time invariant covariate with heterogeneous effect of X
- Done with both equal and unequal sample sizes
- True ATT $= 0.1$

Monte Carlo Simulations

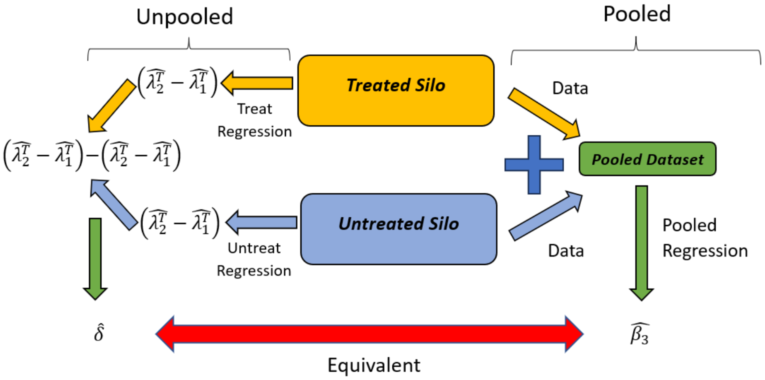
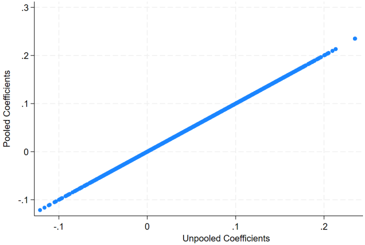
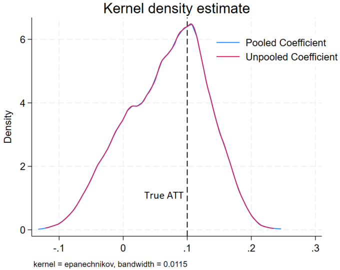


Figure 4: $A\hat{T}T$ from the unpooled regressions

Results



(a) Coefficients match for both methods



(b) Coefficients centered at true value

Unpooldid Program

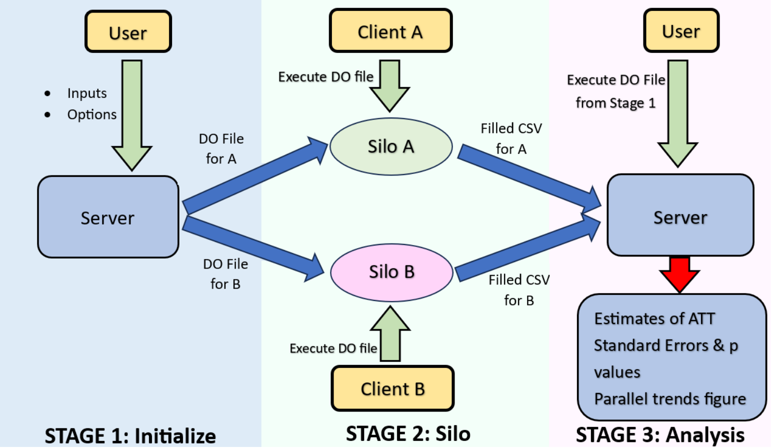


Figure 5: Unpooldid STATA program

Unpooldid commands

1. **Stage 1:** Initializes the program, creates scripts for remaining calls (Server)

   Command: unpooldid *depvars* [*indepvars*] [if] [in] [weight] , siloinfo() stage(1) names() [*options*]

2. **Stage 2:** Called once for each silo, output necessary statistics (Silo specific client)

   Command: unpooldid *depvars* [*indepvars*] [if] [in] [weight] , stage(2) names() siloinfo() [*options*]

3. **Stage 3:** Uses the output from Stage 2 and produces the analysis (Server)

   Command: unpooldid *depvars* [*indepvars*] [if] [in] [weight] , stage(3) names() siloinfo() [*options*]

Unpooldid Inputs and options

- *siloinfo* -
    - will be entered as a string
    - (begin period, period first treated, end period)
    - period first treated will be 0 for the control silo
- *names* - labels for the silos
- *stage* - 1 (init), 2 (silo), 3 (analysis)
- *sample* - to restrict sample for analysis, i.e. by age, gender, etc
- *nograph* - do not produce parallel trends figures
- *cluster*

Unpooldid example

1. **Stage 1:**

   unpooldid y x w, siloinfo(2001,2003,2006\2001,0,2006) stage(1) names(Ontario Quebec) nograph cluster(group) sample(w=0)

2. **Stage 2:**

   unpooldid y x w , stage(2) names(Ontario) siloinfo(2001, 2003, 2006) nograph cluster(group) sample(w=0)

   unpooldid y x w , stage(2) names(Quebec) siloinfo(2001, 0, 2006) nograph cluster(group) sample(w=0)

3. **Stage 3:**

   unpooldid y x w , stage(3) siloinfo(2001,2003,2006\2001,0,2006) names(Ontario Quebec) nograph cluster(group) sample(w=0)

Pre-test for Parallel trends

- In Stage 2, both unconditional and conditional means for each period will be collected in the csv file
- In Stage 3, the server will use these means to plot two figures: an unconditional figure, and a conditional figure for the evolution of outcome
- In the output, the server will display both a unconditional and a conditional parallel trends figure

Questions?

Feedback welcome
Please Email:
SunnyKarim@cmail.carleton.ca
for any suggestions

Appendix

Proof: We know that $pre_t^j = (1 - post_t^j)$. Substituting this into Equation (4):

$$Y_{i,t}^j = \lambda_0 + \lambda_1^j 1post_t^j + \lambda_2^j X_{i,t}^j + \eta_{i,t}^j$$

$$\Rightarrow Y_{i,t}^j = \lambda_1^j(1 - post_t^j) + \lambda_2^j post_t^j + \nu_{i,t}^j$$

$$\therefore Y_{i,t}^j = \lambda_1^j + (\lambda_2^j - \lambda_1^j)post_t^j + \nu_{i,t}^j \tag{5}$$

back

References I

Abadie, Alberto (2005). "Semiparametric difference-in-differences estimators". In: *The review of economic studies* 72.1, pp. 1–19.

Callaway, Brantly and Pedro HC Sant'Anna (2021). "Difference-in-differences with multiple time periods". In: *Journal of Econometrics* 225.2, pp. 200–230.

De Chaisemartin, Clément and Xavier D'Haultfœuille (2020b). "Two-way fixed effects regressions with several treatments". In: *arXiv preprint arXiv:2012.10077*.

De Chaisemartin, Clément and Xavier d'Haultfoeuille (2020a). "Two-way fixed effects estimators with heterogeneous treatment effects". In: *American Economic Review* 110.9, pp. 2964–2996.

References II

Roth, Jonathan et al. (2022). "What's trending in difference-in-differences? A synthesis of the recent econometrics literature". In: *arXiv preprint arXiv:2201.01194*.