EC708 Discussion 6
DiD and Clustered SE

Yan Liu

Department of Economics
Boston University

March 17, 2023

Outline

1. Difference-in-Difference (DiD)
2. Clustered Standard Errors

Table of Contents

1. Difference-in-Difference (DiD)
2. Clustered Standard Errors

Difference-in-Difference (DiD)

Setup

- Outcome Y_{it} and treatment D_{it} are observed for $i = 1, \ldots, N$ and $t = 1, 2$.
- Observed outcome is based on potential outcomes $(Y_{it}(1), Y_{it}(0))$ through $Y_{it} = Y_{it}(1)D_{it} + Y_{it}(0)(1 - D_{it})$.
- Let $G_i \in \{0, 1\}$ indicate groups:
  - $G_i = 1$ (treated group): $D_{i1} = 0, D_{i2} = 1$
  - $G_i = 0$ (control group): $D_{i1} = D_{i2} = 0$
  - $N_1 = \sum_{i=1}^{N} G_i, N_0 = N - N_1$

Difference-in-Difference (DiD)

Identification of ATT

Average treatment effect for the treated (ATT):

$$\tau = E[Y_{i2}(1) - Y_{i2}(0) | G_i = 1].$$

Define the selection bias as

$$\mathrm{SB}_t = E[Y_{it}(0)|G_i = 1] - E[Y_{it}(0)|G_i = 0], \quad t = 1, 2.$$

If $\mathrm{SB}_2 = 0$ (randomized experiments), the ATT is identified as

$$\begin{aligned} \tau &= E[Y_{i2}(1)|G_i = 1] - E[Y_{i2}(0)|G_i = 0] \\ &= E[Y_{i2}|G_i = 1] - E[Y_{i2}|G_i = 0]. \end{aligned}$$

However, we may not have a properly designed experiment ...

Difference-in-Difference (DiD)

Identification of ATT

- Instead of $\mathrm{SB}_2 = 0$, we assume the selection bias is stable: $\mathrm{SB}_2 = \mathrm{SB}_1$.
- This is equivalent to the **parallel trend** assumption:

$$E[Y_{i2}(0) - Y_{i1}(0)|G_i = 1] = E[Y_{i2}(0) - Y_{i1}(0)|G_i = 0].$$

- Then, the ATT is identified as

$$\begin{aligned}\tau &= E[Y_{i2}(1) - Y_{i2}(0)|G_i = 1] \\ &= E[Y_{i2}(1) - Y_{i1}(0)|G_i = 1] - E[Y_{i2}(0) - Y_{i1}(0)|G_i = 0] \\ &= E[Y_{i2} - Y_{i1}|G_i = 1] - E[Y_{i2} - Y_{i1}|G_i = 0].\end{aligned}$$

Under the parallel trend assumption, a consistent estimator for the ATT is

$$\hat{\tau} = \Big(\frac{1}{N_1}\sum_{i:G_i=1} Y_{i2} - \frac{1}{N_1}\sum_{i:G_i=1} Y_{i1}\Big) - \Big(\frac{1}{N_0}\sum_{i:G_i=0} Y_{i2} - \frac{1}{N_0}\sum_{i:G_i=0} Y_{i1}\Big).$$

This is called the **difference-in-difference (DiD)** estimator. Consider the linear panel data model with **two-way fixed effects (TWFE)**:

$$Y_{it} = D_{it}\beta + A_i + F_t + U_{it},\ i = 1, \ldots, N,\ t = 1, 2.$$

It turns out that the OLS estimator $\hat{\beta}$ is numerically equivalent to $\hat{\tau}$.

- Unit mean: $\bar{D}_i = \frac{1}{2}\sum_{t=1}^{2} D_{it}$
- Time mean: $\tilde{D}_t = \frac{1}{N}\sum_{i=1}^{N} D_{it}$
- Full-sample mean: $\bar{\tilde{D}} = \frac{1}{2N}\sum_{i=1}^{N}\sum_{t=1}^{2} D_{it}$
- Fixed-effects adjusted treatment: $\ddot{D}_{it} = D_{it} - \bar{D}_i - \tilde{D}_t + \bar{\tilde{D}}$

By the Frisch–Waugh–Lovell theorem,

$$\hat{\beta} = \frac{\frac{1}{2N}\sum_{i=1}^{N}\sum_{t=1}^{2} Y_{it}\ddot{D}_{it}}{\frac{1}{2N}\sum_{i=1}^{N}\sum_{t=1}^{2} \ddot{D}_{it}^{2}}. \tag{1}$$

We can calculate

$$\bar{D}_i = \begin{cases} \frac{1}{2} & \text{if } G_i = 1 \\ 0 & \text{if } G_i = 0 \end{cases}, \quad \tilde{D}_t = \begin{cases} 0 & \text{if } t = 1 \\ \frac{N_1}{N} & \text{if } t = 2 \end{cases}, \quad \tilde{\bar{D}} = \frac{N_1}{2N}.$$

Hence,

$$\ddot{D}_{it} = \begin{cases} 0 - \frac{1}{2} - 0 + \frac{N_1}{2N} = -\frac{N_0}{2N} & \text{if } G_i = 1, t = 1 \\ 1 - \frac{1}{2} - \frac{N_1}{N} + \frac{N_1}{2N} = \frac{N_0}{2N} & \text{if } G_i = 1, t = 2 \\ 0 - 0 - 0 + \frac{N_1}{2N} = \frac{N_1}{2N} & \text{if } G_i = 0, t = 1 \\ 0 - 0 - \frac{N_1}{N} + \frac{N_1}{2N} = -\frac{N_1}{2N} & \text{if } G_i = 0, t = 2 \end{cases}.$$

Numerator for $\hat{\beta}$:

$$\frac{1}{2N}\sum_{i=1}^{N}\sum_{t=1}^{2} Y_{it}\ddot{D}_{it} = \frac{1}{2N}\Big[\Big(\frac{N_0}{2N}\sum_{i:G_i=1} Y_{i2} - \frac{N_0}{2N}\sum_{i:G_i=1} Y_{i1}\Big) - \Big(\frac{N_1}{2N}\sum_{i:G_i=0} Y_{i2} - \frac{N_1}{2N}\sum_{i:G_i=0} Y_{i1}\Big)\Big].$$

Denominator for $\hat{\beta}$:

$$\frac{1}{2N}\sum_{i=1}^{N}\sum_{t=1}^{2} \ddot{D}_{it}^2 = \frac{1}{2N}\Big[2N_1\Big(\frac{N_0}{2N}\Big)^2 + 2N_0\Big(\frac{N_1}{2N}\Big)^2\Big] = \frac{N_0 N_1}{4N^2}.$$

Put together,

$$\hat{\beta} = \Big(\frac{1}{N_1}\sum_{i:G_i=1} Y_{i2} - \frac{1}{N_1}\sum_{i:G_i=1} Y_{i1}\Big) - \Big(\frac{1}{N_0}\sum_{i:G_i=0} Y_{i2} - \frac{1}{N_0}\sum_{i:G_i=0} Y_{i1}\Big).$$

Suppose there are three groups: $G_i \in \{U, E, L\}$

- U: untreated group
- E: early treatment group, which receives treatment at time t_1
- L: late treatment group, which receives treatment at time $t_2 > t_1$

There are three types of time windows to consider

- $\text{PRE}(E) : t < t_1$ and $\text{PRE}(L) : t < t_2$
- $\text{POST}(E) : t \geq t_1$ and $\text{POST}(L) : t \geq t_2$
- $\text{MID} : t_1 \leq t < t_2$

It turns out that the TWFE estimator is an average of 2×2 DiD estimators (Goodman-Bacon, 2021):

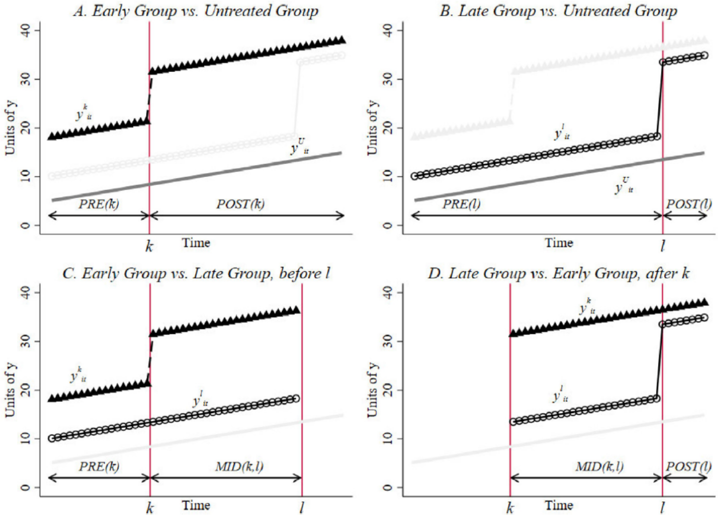
$$\hat{\beta} = \sum_{g \in \{E,L\}} s_{gU} \hat{\beta}_{gU}^{2\times 2} + s_{EL}^{E} \hat{\beta}_{EL}^{2\times 2,E} + s_{EL}^{L} \hat{\beta}_{EL}^{2\times 2,L},$$

where

$$\begin{aligned}
\hat{\beta}_{gU}^{2\times 2} &= (\bar{Y}_g^{\text{POST}(g)} - \bar{Y}_g^{\text{PRE}(g)}) - (\bar{Y}_U^{\text{POST}(g)} - \bar{Y}_U^{\text{PRE}(g)}), \\
\hat{\beta}_{EL}^{2\times 2,E} &= (\bar{Y}_E^{\text{MID}} - \bar{Y}_E^{\text{PRE}(E)}) - (\bar{Y}_L^{\text{MID}} - \bar{Y}_L^{\text{PRE}(E)}), \\
\hat{\beta}_{EL}^{2\times 2,L} &= (\bar{Y}_L^{\text{POST}(L)} - \bar{Y}_L^{\text{MID}}) - (\bar{Y}_E^{\text{POST}(L)} - \bar{Y}_E^{\text{MID}}).
\end{aligned}$$

Difference-in-Difference (DiD)

DiD with Variation in Treatment Timing



What does each 2×2 DiD estimator capture? Define

- $Y_{it}(E)/Y_{it}(L)$: potential outcome if treated early/late
- $Y_{it}(0)$: untreated potential outcome

Two parameters for causal interpretation:

- For $g \in \{E, L\}$ and a date range W, define the group-time ATT as

$$\text{ATT}_g(W) = \frac{1}{T_W} \sum_{t \in W} E[Y_{it}(g) - Y_{it}(0) | G_i = g].$$

- For $g \in \{U, E, L\}$ and two date ranges W_1, W_0, define the difference over time in average untreated potential outcomes as

$$\Delta Y_g^0(W_1, W_0) = \frac{1}{T_{W_1}} \sum_{t \in W_1} E[Y_{it}(0) | G_i = g] - \frac{1}{T_{W_0}} \sum_{t \in W_0} E[Y_{it}(0) | G_i = g].$$

We can show that

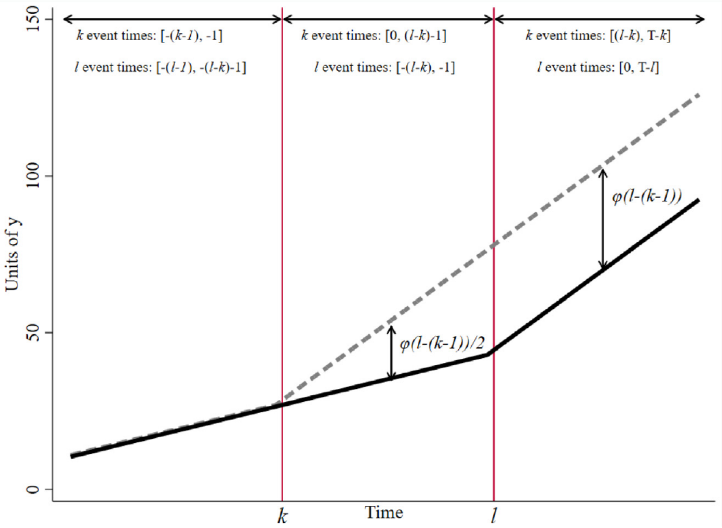
$$
\begin{aligned}
\text{plim}_{N\to\infty}\hat{\beta}^{2\times 2}_{gU} &= \text{ATT}_g(\text{POST}(g)) \\
&\quad + [\Delta Y^0_g(\text{POST}(g), \text{PRE}(g)) - \Delta Y^0_U(\text{POST}(g), \text{PRE}(g))], \\
\text{plim}_{N\to\infty}\hat{\beta}^{2\times 2,E}_{EL} &= \text{ATT}_E(\text{MID}) \\
&\quad + [\Delta Y^0_E(\text{MID}, \text{PRE}(E)) - \Delta Y^0_L(\text{MID}, \text{PRE}(E))], \\
\text{plim}_{N\to\infty}\hat{\beta}^{2\times 2,L}_{EL} &= \text{ATT}_L(\text{POST}(L)) \\
&\quad + [\Delta Y^0_L(\text{POST}(L), \text{MID}) - \Delta Y^0_E(\text{POST}(L), \text{MID})] \\
&\quad - [\text{ATT}_E(\text{POST}(L)) - \text{ATT}_E(\text{MID})].
\end{aligned}
$$

Two sources of bias:

- Timing groups' differential trends
- Changes in ATT over time $\Rightarrow$ negative weights

Difference-in-Difference (DiD)

DiD with Variation in Treatment Timing



New estimators for staggered timing:

- Consider as building blocks the group-time ATT, $\text{ATT}_{g,t}$:
  De Chaisemartin and d'Haultfoeuille (2020); Callaway and Sant'Anna (2021); Sun and Abraham (2021)
- Run a stacked regression (match each treated unit to "clean" controls):
  Cengiz et al. (2019); Deshpande and Li (2019)

Table of Contents

1 Difference-in-Difference (DiD)

2 Clustered Standard Errors

Clustered Standard Errors

Variance Inflation for OLS

Consider a setting in which each unit ℓ belongs to a cluster $C_\ell \subset \{1, \ldots, N\}$.

- each household belongs to some geographical area (e.g., state)
- each individual can be viewed as a cluster in panel data

For simplicity, begin with OLS

$$Y_\ell = X_\ell' \beta + U_\ell.$$

Suppose U_ℓ are homoskedastic and equicorrelated within the cluster:

$$E[U_\ell U_m] = \begin{cases} 0 & C_\ell \neq C_m \\ \rho_u \sigma^2 & C_\ell = C_m, \ell \neq m \\ \sigma^2 & \ell = m \end{cases}.$$

Clustered Standard Errors

Variance Inflation for OLS

We can calculate

$$\text{Var}\Big(\sum_{\ell} X_\ell U_\ell \Big| X_1, \ldots, X_N\Big) = \sigma^2 \sum_{\ell} X_\ell X_\ell' + \rho_u \sigma^2 \sum_{\substack{\ell \neq m: \\ C_\ell = C_m}} X_\ell X_m'.$$

If we further assume

- constant cluster size L,
- $X_\ell = X_m$ if $C_\ell = C_m$ (within-group-constant explanatory variable),

the variance of the OLS estimator is

$$\text{Var}(\hat{\beta} | X_1, \ldots, X_N) = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1} [1 + \rho_u (L - 1)].$$

More generally, for the kth regressor, the default OLS variance estimate based on $s^2(\mathbf{X}'\mathbf{X})^{-1}$ should be inflated by

$$\tau_k \simeq 1 + \rho_{x_k}\rho_u(\bar{N}_g - 1),$$

where

- ρ_{x_k}: measure of within-cluster correlation of the kth regressor
- ρ_u: within-cluster error correlation
- $\bar{N}_g$: average cluster size

Clustered Standard Errors

DiD: Setup

Consider estimating the average effect of a binary policy d_{it} on outcome y_{it}:

$$y_{it} = \gamma d_{it} + w_{it}'\beta + \alpha_i + \delta_t + u_{it}.$$

- d_{it} varies by state and over time
  - For "treated states", $d_{it} = 0$ for $t \leq t^*$ and $d_{it} = 1$ for $t > t^*$
  - For "control states", $d_{it} = 0$ for all t
- w_{it}: vector of additional controls
- α_i, δ_t: state and year fixed effects

Clustered Standard Errors

DiD: Robust Inference

Bertrand et al. (2004) demonstrated the importance of using cluster-robust standard errors in DiD settings.

- d_{it} is highly serially correlated within each cluster by construction
- u_{it} may also be correlated within the cluster.
- Clustering should be on state, assuming error independence across states.

Clustered Standard Errors

DiD: Robust Inference

Let units be ordered by cluster. Consider the covariance matrix of the relevant error vector

$$V = \begin{bmatrix} \Omega_1 & 0 & 0 & \dots & 0 \\ 0 & \Omega_2 & 0 & \dots & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & \dots & \dots & 0 & \Omega_N \end{bmatrix}.$$

For conducting robust inference, one does not need to know the form of Ω_i.

Clustered Standard Errors

- Let T_i denote the number of units that belong to cluster i.
- Let X_i be a $T_i \times k$ matrix stacking $1 \times k$ vectors x'_ℓ for all ℓ such that $C_\ell = i$.
- Consider asymptotics under which $N \to \infty$ with T_i being finite for all i

Many estimators $\sqrt{N}(\hat{\beta} - \beta)$ are asymptotically normal with asymptotic variance

$$\begin{aligned}\mathrm{AsyVar}(\hat{\beta}) &= Q^{-1} \lim_{N\to\infty} \Big(\frac{1}{N} \sum_{i=1}^{N} E[X'_i \Omega_i X_i] \Big) Q^{-1} \\ &= Q^{-1} \lim_{N\to\infty} \Big(\frac{1}{N} \sum_{i=1}^{N} \sum_{\ell : C_\ell = i} \sum_{m : C_m = i} E[v_{\ell,m} x_\ell x'_m] \Big) Q^{-1}\end{aligned}$$

for some $k \times k$ matrix Q, where $v_{\ell,m}$ is the (ℓ, m) component of V.

Clustered Standard Errors

Estimated version:

$$\widehat{\text{AsyVar}}(\hat{\beta}) = \hat{Q}^{-1}\Big(\frac{1}{N}\sum_{i=1}^{N}\sum_{\ell:C_\ell=i}\sum_{m:C_m=i}\hat{\epsilon}_\ell\hat{\epsilon}_m x_\ell x_m']\Big)\hat{Q}^{-1}.$$

- Balanced clusters (T_i is the same for all i): consistency shown by White (1984, p.134–142)
- Unbalanced clusters: consistency shown by Liang and Zeger (1986)
- Performance of the approximation relies on N and T_i

Consider the case with no within-group varying explanatory variables:

$$Y_{ig} = a + X_g\beta + \alpha_g + \varepsilon_{ig}.$$

Donald and Lang (2007) propose a two-step estimator:

1. Take group means: $\hat{d}_g = \frac{1}{N_g}\sum_{i=1}^{N_g} Y_{ig}$
2. Calculate the “between-groups” estimator of β by regressing $\hat{d}_g$ on X_g

The second-stage becomes

$$\hat{d}_g = \bar{Y}_g = a + X_g\beta + \underbrace{\alpha_g + \bar{\varepsilon}_g}_{=\eta_g}.$$

Clustered Standard Errors

DiD: Small Number of Clusters

Rewrite the second-stage as

$$\tilde{Y}_g = \tilde{X}_g \beta + \tilde{\eta}_g,$$

where $\sim$ denotes a deviation from the mean. The t-statistic is

$$t_\beta = \frac{\hat{\beta} - \beta}{\hat{\sigma}_\eta (\sum_g \tilde{X}_g^2)^{1/2}}, \quad \hat{\sigma}_\eta^2 = \frac{1}{G-2} \sum_{g=1}^{G} (\tilde{Y}_g - \tilde{X}_g \hat{\beta})^2.$$

For t_β to (approximately) have a $t(G-2)$ distribution, it is sufficient that $\tilde{\eta}_g \sim N(0, \sigma_\eta^2)$. Some possibilities:

- Finite N_g: $\alpha_g \sim N(0, \sigma_\alpha^2)$, $\varepsilon_{ig} \sim N(0, \sigma_\varepsilon^2)$ and $N_g \equiv N$
- Large N_g:
  - $\alpha_g \sim N(0, \sigma_\alpha^2)$, ε_{ig} satisfy LLN
  - $\alpha_g \sim N(0, \sigma_\alpha^2 / N_g)$, ε_{ig} satisfy CLT, $N_{g'} / N_g \to 1$ for $g' \neq g$