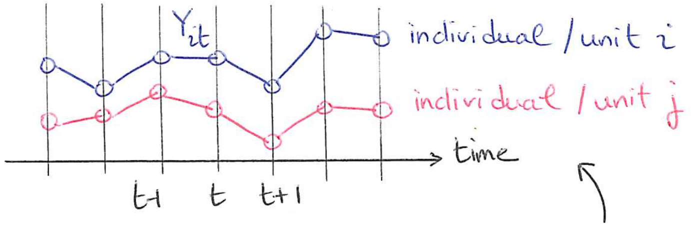
CI : PANEL DATA METHODS

Panel data (aka longitudinal data or cross-sectional time series) is data that include observations of the same (groups of) units (such as individuals, firms, households, products) over multiple time periods.



[Y_{it} = observation for individual i at time t.]

↘ We want to use panel data to estimate the effect of an intervention that affects some units in some time periods.

(generalisation from previous chapters where iid observations were considered)

This chapter covers :

- ↘ the difference estimator with panel data in an RCT
- ↘ the difference-in-difference estimator in an observational study & its application to RCTs.
- ↘ generalisation to stratified designs : block by block analysis & IPW
- ↘ generalisation to clustered experiments.

I_ THE DIFFERENCE ESTIMATOR

Consider n units $i=1,\dots,n$, each receiving a binary treatment assignment $W_i \in \{0,1\}$ completely at random. Let n_t denotes the number of treated units, and $n_c = n - n_t$ the number of control units. In an RCT, $W_i \perp \{Y_i(0), Y_i(1)\}$, where $Y_i(j)$ is the Potential Outcome of unit i receiving treatment j [$(j=1)$ = treated unit ; $(j=0)$ = control unit]. For each unit i, we observe the potential outcome corresponding to its treatment assignment $Y_i = W_i Y_i(1) + (1-W_i) Y_i(0)$. Throughout this chapter, we assume an infinite superpopulation model, so that $\{Y_i(0), Y_i(1)\} \sim \mathbb{P}$. We are interested in this section in the estimation of the ATE $= \Delta^\infty$

$$= \mathbb{E}\left(Y_i(1) - Y_i(0)\right).$$

$\mathbb{E}(.)$ under the joint probability $\mathbb{P}$.

No panel data yet : each unit is associated with a single observation Y_i [no time index]

In chapter CI : RANDOMISED CONTROL TRIALS, we introduced the unbiased & consistent difference estimator of Δ^∞,

$$\hat{\Delta} = \frac{1}{n_t}\sum_{i=1}^{n} W_i Y_i - \frac{1}{n_c}\sum_{i=1}^{n}(1-W_i)Y_i .$$

We re-express $\hat{\Delta}$ as the OLS estimate of Δ^∞ in the linear model

$$\boxed{Y_i = \beta_0 + \Delta^\infty W_i + \varepsilon_i .}$$

[DIFF | OLS]

↓

In matrix notation, $Y = X\beta + \varepsilon$ where $\beta = \begin{pmatrix} \beta_0 \\ \Delta^\infty \end{pmatrix}$,

$$X = \begin{pmatrix} 1 & W_1 \\ \vdots & \vdots \\ 1 & W_n \end{pmatrix}, \quad Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix},$$

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\Delta} \end{pmatrix} = (X^tX)^{-1}X^tY \quad ; \ \hat{\Delta} = \text{diff-estimator}$$

Assuming $\mathbb{E}\,\varepsilon = 0$ and writing $Cov(\varepsilon) = \mathbb{E}\,\varepsilon\varepsilon^t = \Sigma_\varepsilon$ the covariance matrix of $\hat{\beta}$ is

$$\Sigma_{\hat{\beta}} = (X^tX)^{-1}(X^t\Sigma_\varepsilon X)(X^tX) .$$

This linear representation together with the expression $\Sigma_{\hat{\beta}}$ is particularly useful for the construction of confidence intervals for $\hat{\Delta}$ under various assumptions on the correlation structure of the residual error (this appears to be crucial when later on extending the present results to panel data).

x Example 1: Homoskedastic errors : $\Sigma_\varepsilon = \sigma^2 I_n$, so that $\sigma^2 = Var\,\varepsilon_i$ for all i. In this case $\Sigma_{\hat{\beta}} = \sigma^2(X^tX)^{-1}$ (plug-in estimator for σ^2)

---

x Example 2: Heteroskedastic errors ; $Var\,\varepsilon_i = \sigma_i^2$

$$\Sigma_\varepsilon = \begin{pmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_n^2 \end{pmatrix} \leftarrow \text{independent errors.}$$

Then $\Sigma_{\hat{\beta}} = (X^tX)^{-1}\left(\sum_{i=1}^n \sigma_i^2 X_iX_i^t\right)(X^tX)^{-1}$,

where $X = \begin{pmatrix} — X_1^t — \\ \vdots \\ — X_n^t — \end{pmatrix}$.

A consistent estimator of $\Sigma_{\hat{\beta}}$ in this case is given by the Eicker-Huber-White (EHW) estimator:

$$\hat{\Sigma}_{EHW} = (X^tX)^{-1}\left(\sum_{i=1}^n \hat{\varepsilon}_i^2 X_iX_i^t\right)(X^tX)^{-1}$$

$$= (X^tX)^{-1}\left(X^t\begin{pmatrix} \hat{\varepsilon}_1^2 & & 0 \\ & \ddots & \\ 0 & & \hat{\varepsilon}_n^2 \end{pmatrix}X\right)(X^tX)^{-1},$$

where $\hat{\varepsilon}_i$ = i-th residual error = $Y_i - X_i^t\hat{\beta}$. Standard errors based on the EHW estimator are commonly called ROBUST.

↘ The consistency of $\hat{\Sigma}_{EHW}$ is derived in White (1980) under uniformly bounded assumptions of the error variances & covariance matrix of the regressors. Together with the asymptotic normality of $\sqrt{n}(\hat{\beta} - \beta)$, the EHW estimator can be used to construct appropriate confidence intervals with a desired nominal coverage.

× Example 3 = More generally, we may consider a block diagonal matrix

$$\underset{(n\times n)}{\Sigma_\varepsilon} = \left(\begin{array}{c|c|c|c} \Sigma_1 & 0 & & 0 \\ \hline 0 & \Sigma_2 & & \\ \hline & & \ddots & 0 \\ \hline 0 & 0 & & \Sigma_B \end{array}\right) \begin{array}{l} \updownarrow n_1 \\ \updownarrow n_2 \\ \\ \updownarrow n_B \end{array} \quad [n = n_1 + \dots + n_B]$$

$$= (\Sigma_{ij})_{\substack{i=1,\dots,n \\ j=1,\dots,n}}$$

↑ B blocs

↑ Units are clustered and assumed uncorrelated provided they belong to two different clusters.

To start with, we assume further the homoskedastic structure

$$\underset{(n_k\times n_k)}{\Sigma_k} = \begin{pmatrix} \sigma^2 & & \rho\sigma^2 \\ & \ddots & \\ \rho\sigma^2 & & \sigma^2 \end{pmatrix} \quad \forall k = 1, \dots, B$$

↑ constant correlation ρ for two units belonging to the same cluster.

In this case,

$$\Sigma_{\hat\beta} = (X^tX)^{-1}\left(\sum_{k=1}^{B} X_k^t \Sigma_k X_k\right)(X^tX)^{-1}$$

with

$$\underset{(n\times 2)}{X} = \begin{pmatrix} \boxed{X_1} \\ \vdots \\ \boxed{X_B} \end{pmatrix} \begin{array}{l} \updownarrow n_1 \\ \\ \updownarrow n_B \end{array}$$

$\longleftrightarrow 2$

In an RCT, units clustered together typically receive

the same treatment status : $\omega_i = \omega_j \quad \forall i,j \in C_k$, where C_k denotes the k-th cluster of size n_k.

The [DIFF IOLS] expression on page 3 becomes

$$\boxed{Y_{ik} = \beta_0 + \Delta^\omega \omega_k + \varepsilon_{ik}} \qquad \begin{array}{l} i = 1, \dots, n_k \\ k = 1, \dots, B \end{array}$$

In this case, $X = \begin{pmatrix} \boxed{X_1} \\ \vdots \\ \boxed{X_B} \end{pmatrix}$ with

$$\boxed{X_k} = \underbrace{\begin{pmatrix} 1 & \omega_k \\ \vdots & \vdots \\ 1 & \omega_k \end{pmatrix}}_{2} \updownarrow n_k = \underset{(n_k\times 1)}{\mathbb{1}_{n_k}} \underset{(1\times 2)}{\tilde X_k^t} \quad ; \left\{ \begin{array}{l} \mathbb{1}_{n_k} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \updownarrow n_k \\ \tilde X_k = \begin{pmatrix} 1 \\ \omega_k \end{pmatrix} \end{array} \right.$$

$$\Rightarrow X^tX = \sum_{k=1}^{B} n_k \tilde X_k \tilde X_k^t$$

$$\Rightarrow X^t \Sigma_\varepsilon X = \sum_{k=1}^{B} X_k^t \Sigma_k X_k$$

$$= \sum_{k=1}^{B} \tilde X_k \mathbb{1}_{n_k}^t \Sigma_k \mathbb{1}_{n_k} \tilde X_k^t$$

with
- $\Sigma_k = \sigma^2 \begin{pmatrix} 1 & & \rho \\ & \ddots & \\ \rho & & 1 \end{pmatrix}$
- $\Sigma_k \mathbb{1}_{n_k} = \begin{pmatrix} \vdots \\ 1 + (n_k - 1)\rho \\ \vdots \end{pmatrix} \sigma^2$
- $\mathbb{1}_{n_k}^t \Sigma_k \mathbb{1}_{n_k} = n_k \sigma^2 (1 + (n_k - 1)\rho)$

$$X^t \Sigma_\varepsilon X = \sum_{k=1}^{B} \sigma^2 n_k \left(1 + (n_k - 1)\rho\right) \tilde{X}_k \tilde{X}_k^t$$

defining $\hookrightarrow = \sum_{k=1}^{B} \sigma^2 n_k \tau_k \tilde{X}_k \tilde{X}_k^t$

$\tau_k = 1 + (n_k - 1)\rho$

$$\Rightarrow \Sigma_{\hat{\beta}} = \sigma^2 \Big(\sum_k n_k \tilde{X}_k \tilde{X}_k^t \Big)^{-1} \Big(\sum_k n_k \tau_k \tilde{X}_k \tilde{X}_k^t \Big) \Big(\sum_k n_k \tilde{X}_k \tilde{X}_k^t \Big)$$

In the special case where all groups have equal size $n_k = m \ \forall k = 1, .., B$, $\tau_k = 1 + (m-1)\rho \equiv \tau$, and

$$\boxed{\Sigma_{\hat{\beta}} = \underbrace{\sigma^2 \Big(\sum_{k=1}^{B} m \tilde{X}_k \tilde{X}_k^t \Big)^{-1}}_{\text{– Covariance of } \hat{\beta} \text{ ignoring the correlation structure –}} \underbrace{\tau}_{\text{MOULTON FACTOR}}}$$

$\Rightarrow$ The correlation structure of ε inflates the (co)variance by a factor τ. The Moulton Factor is maximal when $\rho = 1$, in which case $\tau = m$: each cluster carries a single unit of information. See Moulton (1990).

x <u>Remark</u>: The bloc-diagonal covariance structure can be justified using the representation

$$\varepsilon_{ik} = \underbrace{\nu_k + \eta_{ik}}_{\substack{\text{zero mean} \\ \text{\& variance } \sigma_\nu^2, \ \sigma_\eta^2, \ \text{iid}}} = \text{additive random effect model}$$

Then
- $\operatorname{cov}(\varepsilon_{ik}, \varepsilon_{jk}) = \mathbb{E}(\nu_k + \eta_{ik})(\nu_k + \eta_{jk})$

  $= \sigma_\nu^2 \quad i \neq j$
- $\operatorname{var}(\varepsilon_{ik}) = \sigma_\nu^2 + \sigma_\eta^2 \equiv \sigma_\varepsilon^2$
- $\operatorname{corr}(\varepsilon_{ik}, \varepsilon_{jk}) = \dfrac{\sigma_\nu^2}{\sigma_\nu^2 + \sigma_\eta^2} \equiv \rho$

& thus $\operatorname{cov}(\varepsilon_{ik}, \varepsilon_{jk}) = \rho(\sigma_\nu^2 + \sigma_\eta^2) = \rho\, \sigma_\varepsilon^2$ ∎

Liang & Zeger (1986) relax Moulton's model of constant correlation and consider general covariance matrices Σ_k: $\Sigma_{\hat{\beta}} = (X^t X)^{-1} \Big(\sum_{k=1}^{B} X_k^t \Sigma_k X_k \Big) (X^t X)^{-1}$

↑ Estimated using

$$\hat{\Sigma}_{LZ} = (X^t X)^{-1} \sum_{k=1}^{B} \Big(\sum_{i=1}^{n_k} (Y_{ik} - \hat{\beta}^t X_{ik}) X_{ik} \Big) \Big(\sum_{i=1}^{n_k} (Y_{ik} - \hat{\beta}^t X_{ik}) X_{ik} \Big)^t \times (X^t X)^{-1}$$

– plugged-in residuals –

since

$$\sum_{k=1}^{B} \underset{(2\times n_k)}{X_k^t} \underset{(n_k \times n_k)}{\Sigma_k} \underset{(n_k \times 2)}{X_k} = \mathbb{E}\Big\{ \sum_{k=1}^{B} \underset{(2\times n_k)}{X_k^t} \underset{(n_k\times 1)}{\varepsilon_k} \underset{(1\times n_k)}{\varepsilon_k^t} \underset{(n_k\times 2)}{X_k} \Big\}$$

$$= \mathbb{E}\Big\{ \sum_{k=1}^{B} \Big(\sum_{i=1}^{n_k} \varepsilon_i \underset{(2\times 1)}{X_{ik}} \Big) \Big(\sum_{i=1}^{n_k} \varepsilon_i \underset{(1\times 2)}{X_{ik}^t} \Big) \Big\}$$

$$\underset{(n_k\times 2)}{X_k} = \overset{\xleftrightarrow{\;2\;}}{\boxed{\;- X_{ik}^t -\;}} \ \updownarrow n_k \qquad \underset{(n\times 2)}{X} = \begin{pmatrix} \overset{\leftarrow 2 \rightarrow}{\boxed{X_1}} \\ \vdots \\ \boxed{X_B} \end{pmatrix} \begin{matrix} \updownarrow n_1 \\ \\ \updownarrow n_B \end{matrix}$$

Summary : $\hat{\Delta}$ is the OLS estimate of Δ^∞ in the linear model $Y_i = \beta_0 + \Delta^\infty W_i + \varepsilon_i$.
We established that under (i) homoskedastic errors, (ii) heteroskedastic errors and (iii) general bloc-diagonal covariance structure for ε, we can consistently estimate the covariance matrix of $\begin{pmatrix} \hat{\beta}_0 \\ \hat{\Delta} \end{pmatrix}$, and thus construct (asymptotically) valid confidence intervals for the difference in means estimator of the ATE.

× Remark : Relative Scale

In the linear representation $Y_i = \beta_0 + \Delta^\infty W_i + \varepsilon_i$,

$$\beta_0 = \mathbb{E}(Y_i \mid W_i = 0) = \mathbb{E}(Y_i(0) \mid \cancel{W_i = 0})$$

$$\begin{aligned}\Delta^\infty &= \mathbb{E}(Y_i \mid W_i = 1) - \mathbb{E}(Y_i \mid W_i = 0) \\ &= \mathbb{E}(Y_i(1) \mid \cancel{W_i = 1}) - \mathbb{E}(Y_i(0) \mid \cancel{W_i = 0}) \\ &= \mathbb{E}(Y_i(1) - Y_i(0))\end{aligned}$$

$$\Rightarrow \delta^\infty := \frac{\mathbb{E}(Y_i(1) - Y_i(0))}{\mathbb{E}(Y_i(0))} = \frac{\Delta^\infty}{\beta_0} = \text{relative "lift"}$$

A natural estimator for δ^∞ is $\hat{\delta}^\infty := \dfrac{\hat{\Delta}^\infty}{\hat{\beta}_0}$; whose variance can be easily derived from $\Sigma_{\hat{\beta}}$ using the delta method

$$\text{var}\left(\frac{X}{Y}\right) \approx \frac{\mu_X^2}{\mu_Y^2}\left(\frac{\sigma_X^2}{\sigma_Y^2} - 2\frac{\text{cov}(X,Y)}{\mu_X \mu_Y} + \frac{\sigma_Y^2}{\mu_Y^2}\right)$$

∎

- Panel Data : Units $i = 1, \dots, n$ are observed over some period of time $t = 1, \dots, T$. Each unit i receives a treatment status $W_i \in \{0, 1\}$ that they keep throughout the experiment.
  ↘ completely at random

Assume a common baseline γ_0 for all P.O. $Y_{it}(0)$, so that $Y_{it}(0) = \gamma_0 + \eta_{it}$; $\mathbb{E}\,\eta_{it} = 0$. In addition, let Δ_{it} be the effect of the treatment on unit i at time t, $Y_{it}(1) = Y_{it}(0) + \Delta_{it}$.

↘ may vary across units and time.
↘ assume a general bivariate distribution for Δ_{it} ; over units & time.

× Estimation Method : $Y_{it} = \beta_0 + \Delta W_i + \varepsilon_{it}$, $\mathbb{E}\,\varepsilon_{it} = 0$
where $Y_{it} = W_i Y_{it}(0) + (1 - W_i) Y_{it}(1)$.

↘ Errors are clustered over time

$$\underset{(nT \times nT)}{\Sigma_\varepsilon} = \begin{array}{c} \left(\begin{array}{c|c|c|c} \Sigma_1 & 0 & & \\ \hline 0 & \Sigma_2 & & \\ \hline & & \ddots & 0 \\ \hline & & 0 & \Sigma_n \end{array}\right) \\ \overset{\longleftrightarrow}{T} \quad \overset{\longleftrightarrow}{T} \quad \quad \quad \overset{\longleftrightarrow}{T} \end{array} \begin{array}{c} \updownarrow T \\ \updownarrow T \\ \\ \updownarrow T \end{array}$$

The OLS estimator of Δ is

$$\hat{\Delta} = \frac{1}{n_t T} \sum_{i \mid W_i = 1} \sum_{t=1}^{T} Y_{it} - \frac{1}{n_c T} \sum_{i \mid W_i = 0} \sum_{t=1}^{T} Y_{it}$$

$$\xrightarrow[n_t, n_c \to \infty]{} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}(Y_{it}(1) \mid \cancel{W_i = 1}) - \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}(Y_{it}(0) \mid \cancel{W_i = 0})$$

$$= \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}(Y_{it}(1) - Y_{it}(0)) = \text{averaged treatment effect over the test period.}$$

II - DIFFERENCE - IN - DIFFERENCES.

We relax the common baseline assumption of page 10 and put

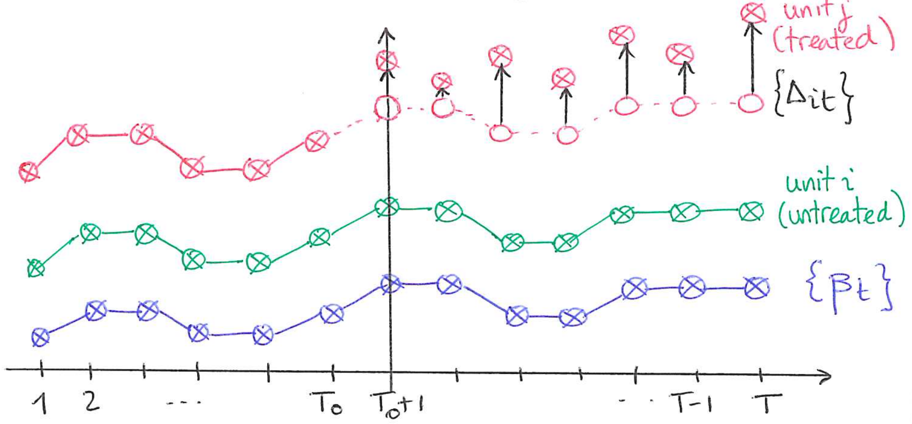
$$Y_{it}(0) = \alpha_i + \beta_t + \varepsilon_{it} .$$

unit-specific effect time-specific effect

⇒ Units evolve in parallel following a baseline time series $\{\beta_t\}$, shifted by unit-specific quantities α_i.

The diff-in-diff model is $\boxed{Y_{it} = \alpha_i + \beta_t + \Delta_{it} W_{it} + \varepsilon_{it}}$

[DID]



× **Estimation Method**: TWFE

$$Y_{it} = \alpha_i + \beta_t + \Delta W_{it} + \varepsilon_{it}$$

↘ Let $\hat{\Delta}$ denote the OLS estimator of Δ in the TWFE model

Remark: Important to dissociate the TWFE model with the [DID] assumption. In section II.3 page 18 we introduce another linear model for estimation of a causal parameter within the [DID] framework.

↘ $t = 1, .., T_0, T_0+1, .., T$ where indexes 1 to T_0 denote observations before the treatment starts. $t = T_0+1, .., T$ denotes the test period.

↘ For treated units, we assume that the treatment starts at time $t = T_0+1$ for all of them. In a subsequent section, we discuss the case of a staggered rollout.

[For a treated unit i, $W_{it} = 1 \quad \forall t = T_0+1, .., T$]

[For all units, $W_{it} = 0 \quad \forall t = 1, .., T_0$]

↘ The treatment is not necessarily randomized here (compared with section I & previous chapters). The TWFE model is usually considered in an observational study, where some units are more likely to receive the treatment than others.

In Appendix A, we show that the OLS estimator of Δ is

$$\boxed{\begin{aligned}\hat{\Delta} = &\left\{ \frac{1}{n_t(T-T_0)} \sum_{i \in trt} \sum_{t \geq T_0+1} Y_{it} - \frac{1}{n_t T_0} \sum_{i \in trt} \sum_{t \leq T_0} Y_{it} \right\} \\ &\qquad\qquad \text{mean diff in the trt group} \\ &- \left\{ \frac{1}{n_c(T-T_0)} \sum_{i \in ctl} \sum_{t \geq T_0+1} Y_{it} - \frac{1}{n_c T_0} \sum_{i \in ctl} \sum_{t \leq T_0} Y_{it} \right\} \\ &\qquad\qquad \text{mean diff in the ctl group}\end{aligned}}$$

↑ The difference-in-differences estimator

Q: Which quantity does $\hat{\Delta}$ identify?

II.1. Identification with two time periods

In this section, we assume that $T_0 = 1$, $T = 2$: for each unit, we observe a single value Y_{i1} in pre-test, and Y_{i2} in test period. Then

$$\hat{\Delta} = \frac{1}{n_t} \sum_{i \in trt} (Y_{i2} - Y_{i1}) - \frac{1}{n_c} \sum_{i \in ctl} (Y_{i2} - Y_{i1})$$

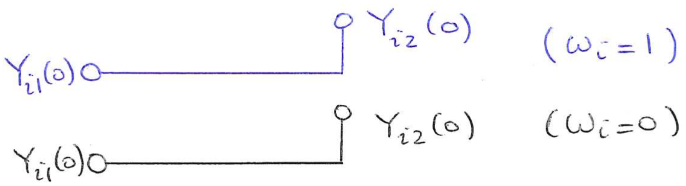
$$\xrightarrow[n_t, n_c \to \infty]{} \mathbb{E}(Y_{i2} - Y_{i1} \mid W_i = 1) - \mathbb{E}(Y_{i2} - Y_{i1} \mid W_i = 0)$$

Indentification involves P.O. only. To get rid of observed values Y_{it}, we make the following two assumptions.

(I) **Parallel trends**

$$\mathbb{E}(Y_{i2}(0) - Y_{i1}(0) \mid W_i = 1) = \mathbb{E}(Y_{i2}(0) - Y_{i1}(0) \mid W_i = 0)$$

[In absence of treatment, the two groups evolve by the same amount (on average)]



(II) **No anticipation** : $Y_{i1}(0) = Y_{i1}(1)$, i treated

Individuals in the treatment group do not anticipate the upcoming treatment in pre-test period

---

Result : Under (I) and (II), the OLS estimator $\hat{\Delta}$ of Δ in the TWFE model $Y_{it} = \alpha_i + \beta_t + \Delta W_{it} + \varepsilon_{it}$ identifies the ATT $= \Delta = \mathbb{E}(Y_{i2}(1) - Y_{i2}(0) \mid W_i = 1)$

test period trt group

Indeed,

$$ATT = \mathbb{E}(Y_{i2}(1) - Y_{i2}(0) \mid W_i = 1)$$

$$= \mathbb{E}(Y_{i2}(1) \mid W_i = 1) - \underbrace{\mathbb{E}(Y_{i2}(0) \mid W_i = 1)}_{\begin{array}{c} \parallel \\ \mathbb{E}(Y_{i1}(0) \mid W_i = 1) \\ + \mathbb{E}(Y_{i2}(0) \mid W_i = 0) \\ - \mathbb{E}(Y_{i1}(0) \mid W_i = 0) \end{array}}$$

[Under (I)]

$$= \left\{ \mathbb{E}(Y_{i2}(1) \mid W_i = 1) - \boxed{\mathbb{E}(Y_{i1}(0) \mid W_i = 1)} \right\}$$

$$- \left\{ \mathbb{E}(Y_{i2}(0) \mid W_i = 0) - \mathbb{E}(Y_{i1}(0) \mid W_i = 0) \right\}$$

[Under (II)] $= \mathbb{E}(Y_{i1}(1) \mid W_i = 1)$

$$= \left\{ \mathbb{E}(Y_{i2}\cancel{(1)} \mid W_i = 1) - \mathbb{E}(Y_{i1}\cancel{(1)} \mid W_i = 1) \right\}$$

$$- \left\{ \mathbb{E}(Y_{i2}\cancel{(0)} \mid W_i = 0) - \mathbb{E}(Y_{i1}\cancel{(0)} \mid W_i = 0) \right\}$$

$$= \mathbb{E}(Y_{i2} - Y_{i1} \mid W_i = 1) - \mathbb{E}(Y_{i2} - Y_{i1} \mid W_i = 0)$$

$$= \Delta$$

$= \text{limit of } \hat{\Delta}$ (p.13) as $n_t, n_c \to \infty$

× **Remark** : Abuse of notation : $\begin{pmatrix} Y_{it}(0) = Y_{it}(0,0) \\ Y_{it}(1) = Y_{it}(0,1) \end{pmatrix}$ pre-period test period "full trajectories" x **Remark** : The ATNT := $\mathbb{E}(Y_{i2}(1)-Y_{i2}(0)\mid W_i=0)$ is identified under

(1') $$\mathbb{E}(Y_{i2}(1)-Y_{i1}(1)\mid W_i=1) = \mathbb{E}(Y_{i2}(1)-Y_{i1}(1)\mid W_i=0)$$

(II') $$\mathbb{E}(Y_{i1}(0)-Y_{i1}(1)\mid W_i=0)=0$$

i.e. under (1'), (II'), Δ = ATNT.

When W_{it} are drawn at random, (1), (1'), (II), (II') are satisfied and we identify the ATE:

$$\begin{aligned}\text{ATE} &= \text{ATT}\times\mathbb{P}(W_{i2}=1)+\text{ATNT}\times\mathbb{P}(W_{i2}=0)\\ &= \Delta\times\mathbb{P}(W_{i2}=1)+\Delta\times\mathbb{P}(W_{i2}=0)\\ &= \Delta\end{aligned}$$

∎

II. 2 Identification in the general case.

Assumptions (1) and (II) generalize nicely in the multi-period case. Put

$$\boxed{\text{ATT}(T_0,T)=\frac{1}{T-T_0}\sum_{t=T_0+1}^{T}\mathbb{E}(Y_{it}(1)-Y_{it}(0)\mid W_i=1)}$$

= average ATT over $t=T_0+1,\dots,T$

(A) Parallel trends

$$\frac{1}{T-T_0}\sum_{t=T_0+1}^{T}\mathbb{E}(Y_{it}(0)\mid W_i=1)-\frac{1}{T_0}\sum_{t=1}^{T_0}\mathbb{E}(Y_{it}(0)\mid W_i=1)$$

test period = pre-test

$$\frac{1}{T-T_0}\sum_{t=T_0+1}^{T}\mathbb{E}(Y_{it}(0)\mid W_i=0)-\frac{1}{T_0}\sum_{t=1}^{T_0}\mathbb{E}(Y_{it}(0)\mid W_i=0)$$

(B) No Anticipation

here again, we abused notation, as these must be understood with a path of treatment (p. 14)

$$\forall t=1,\dots,T_0:\ Y_{it}(0)=Y_{it}(1)\quad,\ \forall i \text{ with } W_i=1$$

Result: Under (A) and (B), the OLS estimator $\hat{\Delta}$ (bottom of page 12) converges to ATT(T_0,T) as $n_t, n_c\to\infty$.

Indeed,

$$\begin{aligned}\text{ATT}(T_0,T) &= \frac{1}{T-T_0}\sum_{t=T_0+1}^{T}\mathbb{E}(Y_{it}(1)-Y_{it}(0)\mid W_i=1)\\ &= \frac{1}{T-T_0}\sum_{t=T_0+1}^{T}\mathbb{E}(Y_{it}(1)\mid W_i=1)\\ &\quad -\frac{1}{T-T_0}\sum_{t=T_0+1}^{T}\mathbb{E}(Y_{it}(0)\mid W_i=1)\end{aligned}$$

[Under (A)]

$$\begin{aligned}&= \frac{1}{T_0}\sum_{t=1}^{T_0}\mathbb{E}(Y_{it}(0)\mid W_i=1)\\ &\quad +\frac{1}{T-T_0}\sum_{t=T_0+1}^{T}\mathbb{E}(Y_{it}(0)\mid W_i=0)\\ &\quad -\frac{1}{T_0}\sum_{t=1}^{T_0}\mathbb{E}(Y_{it}(0)\mid W_i=0)\end{aligned}$$

[under (B)]

$$= \frac{1}{T_0}\sum_{t=1}^{T_0}\mathbb{E}(Y_{it}(1)\mid W_i=1)$$

$$\begin{aligned}&= \left\{\frac{1}{T-T_0}\sum_{t=T_0+1}^{T}\mathbb{E}(Y_{it}(\cancel{1})\mid W_i=1)-\frac{1}{T_0}\sum_{t=1}^{T_0}\mathbb{E}(Y_{it}(\cancel{1})\mid W_i=1)\right\}\\ &\quad -\left\{\frac{1}{T-T_0}\sum_{t=T_0+1}^{T}\mathbb{E}(Y_{it}(\cancel{0})\mid W_i=0)-\frac{1}{T_0}\sum_{t=1}^{T_0}\mathbb{E}(Y_{it}(\cancel{0})\mid W_i=0)\right\}\end{aligned}$$

∎

x **Remark** = Instead, we may collapse the multi-period data to a single pre-test and test : consider the time aggregated outcomes

$$\left\{\begin{aligned} Y_{i,pre} &= \sum_{t=1}^{T_0} Y_{it} \\ Y_{i,test} &= \sum_{t=T_0+1}^{T} Y_{it} \end{aligned}\right.$$

and compute the diff-in-diff estimator

$$\tilde{\Delta} = \frac{1}{n_t}\sum_{i\in trt}(Y_{i,test}-Y_{i,pre}) - \frac{1}{n_c}\sum_{i\in ctl}(Y_{i,test}-Y_{i,pre})$$

The difference with $\hat{\Delta}$ lies in the extra averaging $\frac{1}{T_0}$ and $\frac{1}{T-T_0}$ & therefore in the interpretation of the quantity $\hat{\Delta}$ and $\tilde{\Delta}$ identify.

$\hat{\Delta} \rightarrow$ per unit, per epoch t

$\tilde{\Delta} \rightarrow$ per unit, over a whole time period.

Because of this, $\tilde{\Delta}$ tends to have a higher variance than $\hat{\Delta}$.

↳ To account for temporal correlation when computing the variance of $\hat{\Delta}$, one may cluster errors over time, see p.10. Note that this is implicitely suggested in the two-periods case, where

$$\hat{\Delta} = \frac{1}{n_t}\sum_{i\in trt}(Y_{i2}-Y_{i1}) - \frac{1}{n_c}\sum_{i\in ctl}(Y_{i2}-Y_{i1})$$

$$\downarrow$$

$$\text{var}\,\hat{\Delta} = \frac{1}{n_t}\text{var}(Y_{i2}-Y_{i1}\,|W_i=1) + \frac{1}{n_c}\text{var}(Y_{i2}-Y_{i1}\,|W_i=0)$$

Temporal correlation is accounted for here since the variance of the difference $(Y_{i2}-Y_{i1})$ is computed in each group.

$$\text{var}_1(Y_{i2}-Y_{i1}) = \text{var}\,Y_{i2} + \underbrace{\text{var}\,Y_{i1} - 2\,\text{cov}(Y_{i2},Y_{i1})}$$

↑ Shorthand for $\text{var}(\ldots|W_i=1)$

If ≤ 0, the resulting diff-in-diff estimator has smaller variance than the difference estimator. To get this, (Y_{i1}, Y_{i2}) must be sufficiently correlated to counter balance $\text{var}\,Y_{i1}$. ▦

II.3. Alternative Representation.

Instead of the TWFE model $Y_{it} = \alpha_i + \beta_t + \Delta W_{it} + \varepsilon_{it}$, the difference-in-differences estimator $\hat{\Delta}$ (page 12) can be seen to be the OLS estimate of Δ in the alternative linear representation

[DID|OLS]

$$\boxed{\begin{aligned} Y_{it} = a_0 &+ a_1\,\mathbb{1}(i \text{ is in the treatment group}) \\ &+ a_2\,\mathbb{1}(t \geq T_0+1) \\ &+ \Delta\,\mathbb{1}(t \geq T_0+1 \;\&\; i\in trt) + \varepsilon_{it}. \end{aligned}}$$

This representation is particularly suited to compute relative effects and confidence intervals. Note that =

- $\hat{a}_0 \rightarrow \frac{1}{T_0} \sum_{t=1}^{T_0} \mathbb{E}(Y_{it} \mid W_i = 0)$ (as $n_t, n_c \rightarrow \infty$)
- $\hat{a}_1 \rightarrow \frac{1}{T-T_0} \sum_{t=T_0+1}^{T} \mathbb{E}(Y_{it} \mid W_i = 0) - \frac{1}{T_0} \sum_{t=1}^{T_0} \mathbb{E}(Y_{it} \mid W_i = 0)$
- $\hat{a}_2 \rightarrow \frac{1}{T_0} \sum_{t=1}^{T_0} \mathbb{E}(Y_{it} \mid W_i = 1) - \frac{1}{T_0} \sum_{t=1}^{T_0} \mathbb{E}(Y_{it} \mid W_i = 0)$

$$\Rightarrow \hat{a}_0 + \hat{a}_1 + \hat{a}_2 \longrightarrow \frac{1}{T-T_0} \sum_{t=T_0+1}^{T} \mathbb{E}(Y_{it}(0) \mid W_i = 0) - \frac{1}{T_0} \sum_{t=1}^{T_0} \mathbb{E}(Y_{it}(0) \mid W_i = 0) + \boxed{\frac{1}{T_0} \sum_{t=1}^{T_0} \mathbb{E}(Y_{it}(1) \mid W_i = 1)}$$

No anticipation

$$= \frac{1}{T_0} \sum_{t=1}^{T_0} \mathbb{E}(Y_{it}(0) \mid W_i = 1)$$

// trends

$$= \frac{1}{T-T_0} \sum_{t=T_0+1}^{T} \mathbb{E}(Y_{it}(0) \mid W_i = 1)$$

= reference value for the treatment group in the test period.

$$\Rightarrow \frac{\hat{\Delta}}{\hat{a}_0 + \hat{a}_1 + \hat{a}_2} = \text{relative lift}$$

↑ variance is computed using the delta method.

Remark: Both this linear model and the TWFE allow identification of the ATT under the parallel trend assumption.

II. 4. Checking for pre-trends

We cannot check for parallel trends in the test period. However, we can check how plausible the assumption holds in pre-test. Consider the following TWFE model

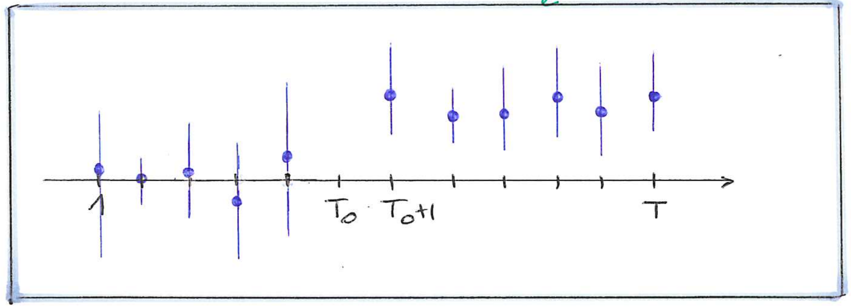
$$\boxed{Y_{it} = \alpha_i + \beta_t + \sum_{\substack{s=1 \\ s \neq T_0}}^{T} \gamma_s W_i \mathbb{1}(t = s) + \varepsilon_{it}}$$

One coefficient is arbitrarily removed to avoid overspecifying the linear model (here, the last day before intervention)

[see the Appendix for a formal proof] — p.47

& generate an "event study plot"

Coeffs $\hat{\gamma}_s$ & confidence bounds



Limitations

- No guarantees that // trends hold in test period
- Typically low power (often fail to reject the null)
- Conditions the analysis on "passing pre-trends" → selection bias

II_5_ Limitations of the TWFE approach.

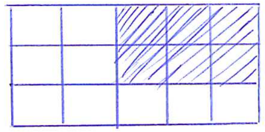
Consider the TWFE model (page 11) with general $\{W_{it}\}$:

$$Y_{it} = \alpha_i + \beta_t + \Delta W_{it} + \varepsilon_{it} .$$

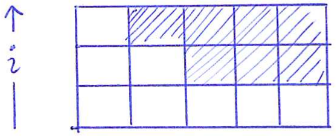
Let $\Delta_{it} = Y_{it}(1) - Y_{it}(0)$ = unit i treatment effect.

When there is no heterogeneity between units & across time, $\mathbb{E}\Delta_{it} = \Delta_0 \ \forall (i,t)$, the OLS estimate $\hat{\Delta}$ of Δ in the TWFE model recovers the correct value $\mathbb{E}\hat{\Delta} = \Delta_0$.

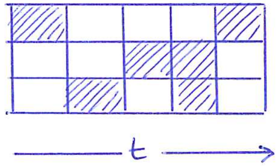
↘ The discussion in this section holds true with general treatment patterns, such as



= Classical design (Sections II.1-3)



= Staggered design



= General pattern

Issues arise when there is heterogeneity in the treatment effect either across units or time. We state next a simplified version of a result in de Chaisemartin & d'Haultfoeuille (2020), proved under the following assumptions:

(I) Balanced Panel : Observe Y_{it} $\forall i, \forall t$

(II) Independent individuals

$(Y_{i1}(0), Y_{i1}(1), W_{i1}, \dots, Y_{iT}(0), Y_{iT}(1), W_{iT})$

$i = 1, \dots, n$ are independent (time correlation is allowed)

(III) Common Trends

$\forall t \geq 2 \quad \mathbb{E}(Y_{it}(0) - Y_{i,t-1}(0))$ independent of i

(IV) Strong Exogeneity

Shocks are independent of the past, present & future treatments

$$\mathbb{E}(Y_{it}(0) - Y_{i,t-1}(0) \mid W_{i1}, \dots, W_{iT}) = \mathbb{E}(Y_{it}(0) - Y_{i,t-1}(0)) .$$

Under (I),(II),(III),(IV), the OLS estimator $\hat{\Delta}$ of Δ in $Y_{it} = \alpha_i + \beta_t + \Delta W_{it} + \varepsilon_{it}$ satisfies

$$\mathbb{E}\hat{\Delta} = \mathbb{E}\left[\sum_{(i,t)|W_{it}=1} r_{it} \Delta_{it} \right]$$

where $\rightarrow$ $r_{it} = \varepsilon_{it} \Big/ \sum_{(i,t)|W_{it}=1} \varepsilon_{it}$

$\rightarrow$ ε_{it} = residual in $W_{it} = \gamma + \overbrace{\gamma_i + \delta_t}^{\text{FE}} + \varepsilon_{it}$

↘ Weights sum to one; but can be negative

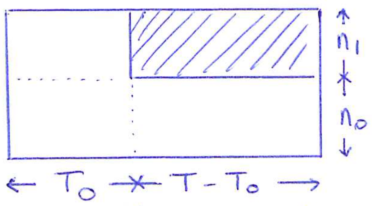
↘ When the treatment effect does not vary across units & time, the OLS estimator $\hat{\Delta}$ is unbiased for the ATT.

- Example 1 : Classical design.

T_0 = # epochs in pre-test ; $T - T_0$ = # epochs in test

$n_1 = \#$ treated units , $n_0 = \#$ control units

Then one can show that $\varepsilon_{it} = W_{it} - \overline{W}_{i\cdot} - \overline{W}_{\cdot t} + \overline{W}_{\cdot\cdot}$



$\Big\}\ \overline{D}_{i\cdot} = \dfrac{T - T_0}{T}$

$\overline{D}_{\cdot t} = \dfrac{n_1}{n}$ $\qquad \overline{D}_{\cdot\cdot} = \dfrac{n_1(T - T_0)}{nT}$

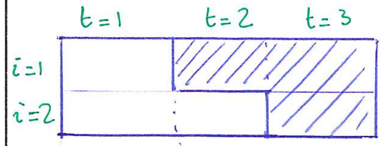
↘ $\varepsilon_{it} = \dfrac{n_0 T_0}{nT} \quad \forall (i,t) \mid W_{it} = 1$

↘ $r_{it} = \dfrac{1}{n_1 (T - T_0)}$ — " —

$$\Rightarrow \mathbb{E}\hat{\Delta} = \frac{1}{n_1(T - T_0)} \sum_{(i,t)|W_{it}=1} \mathbb{E}\Delta_{it} \text{ , as required .}$$

Heterogeneity of treatment effects is not an issue here. ▣

- Example 2 = Staggered design



$\overline{D}_{i\cdot} = 2/3$

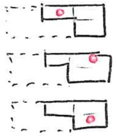
$\overline{D}_{i\cdot} = 1/3$

↙ 2 units , 3 time periods

$\overline{D}_{\cdot t} = 1/2 \quad \overline{D}_{\cdot t} = 1$

$(i,t) \in$

$$\varepsilon_{it} = \begin{cases} 1 - 2/3 - 1/2 + 1/2 = 1/3 \\ 1 - 2/3 - 1 + 1/2 = -1/6 \\ 1 - 1/3 - 1 + 1/2 = 1/6 \end{cases}$$



$$r_{it} = \begin{cases} 1 \\ -1/2 \\ 1/2 \end{cases}$$

$$\Rightarrow \mathbb{E}\hat{\Delta} = 1 \times \mathbb{E}\Delta_{1,2} - \frac{1}{2}\mathbb{E}\Delta_{1,3} + \frac{1}{2}\mathbb{E}\Delta_{2,3}$$

$$\neq \frac{1}{3}\left(\mathbb{E}\Delta_{1,2} + \mathbb{E}\Delta_{1,3} + \mathbb{E}\Delta_{2,3} \right).$$

Worse, $\mathbb{E}\hat{\Delta}$ may be negative when all $\mathbb{E}\Delta_{it}$ are > 0.

Take e.g. $\mathbb{E}\Delta_{1,3} = 4$ & $\mathbb{E}\Delta_{1,2} = \mathbb{E}\Delta_{2,3} = 1$.

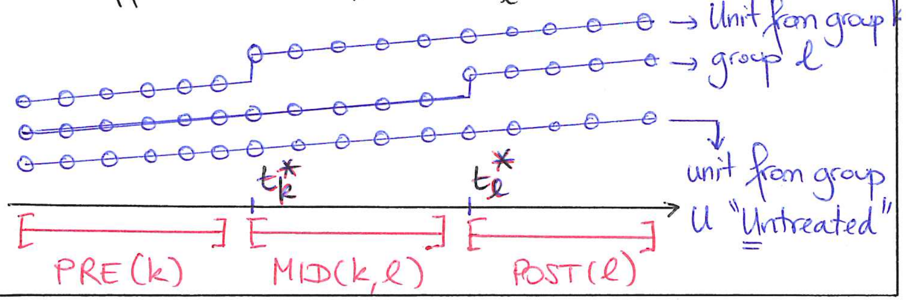
Then $\mathbb{E}\hat{\Delta} = -1/2$.

↳ Negative weights are of concern only under heterogeneous effects. For example, with $\mathbb{E}\Delta_{1,3} = \mathbb{E}\Delta_{1,2} = \mathbb{E}\Delta_{2,3} = 1$, $\mathbb{E}\hat{\Delta} = 1$.

x Remark : Since $\varepsilon_{it} = D_{it} \underbrace{- \overline{D}_{i\cdot} - \overline{D}_{\cdot t}}_{} + \overline{D}_{\cdot\cdot}$,

negative weights in a staggered design are more likely on early adopters in late time periods. ▣

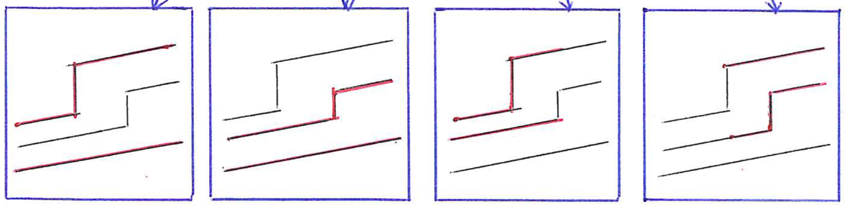
x Remark : Goodman-Bacon (2018) provides helpful intuition why this is happening in a staggered design. He shows that $\hat{\Delta}$ can be decomposed as a weighted sum of canonical diff-in-diff estimates, where early adopters are mistakenly taken as control units "forbidden comparisons". Specifically, consider n units whose treatment status turns on at two different dates t_k^* and t_ℓ^*



Then

sum to one

$$\hat{\Delta} = s_{ku}\hat{\Delta}_{ku} + s_{\ell u}\hat{\Delta}_{\ell u} + s^{k}_{k\ell}\hat{\Delta}^{k}_{k\ell} + s^{\ell}_{k\ell}\hat{\Delta}^{\ell}_{k\ell}$$



Here, the untreated group is correctly used as a control group

Here, the treated group is used as control

$$\hat{\Delta}_{ku} = \left(\bar{Y}_k^{POST(k)} - \bar{Y}_k^{PRE(k)}\right) - \left(\bar{Y}_u^{POST(k)} - \bar{Y}_u^{PRE(k)}\right)$$

$$\hat{\Delta}_{\ell u} = \left(\bar{Y}_\ell^{POST(\ell)} - \bar{Y}_\ell^{PRE(\ell)}\right) - \left(\bar{Y}_u^{POST(\ell)} - \bar{Y}_u^{PRE(\ell)}\right)$$

↳ $\hat{\Delta}_{ku}$; $\hat{\Delta}_{\ell u}$ are legitimate did estimators

$$\hat{\Delta}^{k}_{k\ell} = \left(\bar{Y}_k^{MID(k,\ell)} - \bar{Y}_k^{PRE(k)}\right) - \left(\bar{Y}_\ell^{MID(k,\ell)} - \bar{Y}_\ell^{PRE(k)}\right)$$

$$\hat{\Delta}^{\ell}_{k\ell} = \left(\bar{Y}_\ell^{POST(\ell)} - \bar{Y}_\ell^{MID(k,\ell)}\right) - \left(\bar{Y}_k^{POST(\ell)} - \bar{Y}_k^{MID(k,\ell)}\right)$$

acting as treatment group ↑ — acting as control group ↑

⇒ TWFE $\hat{\Delta}$ is inappropriate in a staggered design.

- Weights are non-negative & depend on the subsamples sizes squared & subsample variance. Weights are larger when the two groups are similar in size & when the treatment occurs in the middle of the time window.

III _ STRATIFICATION

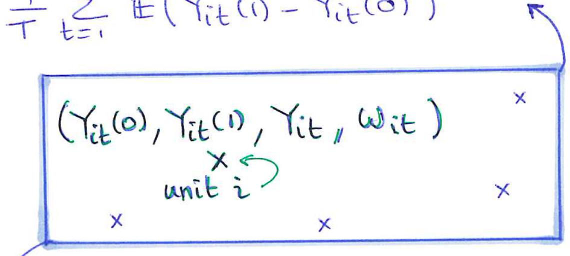
In Sections I and II, we considered a collection of units (randomized or not), and derived estimators of the ATE/ATT when units are observed over some period of time.

No pre-test + RCT : $W_{it} = W_i \perp \{Y_{it}(0), Y_{it}(1)\}$ $\forall i, \forall t$

The OLS estimator $\hat{\Delta}$ of Δ in

$Y_{it} = \beta_0 + \Delta W_i + \varepsilon_{it}$ is a consistent estimator of

$$ATE = \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left(Y_{it}(1) - Y_{it}(0)\right)$$ (p.10)



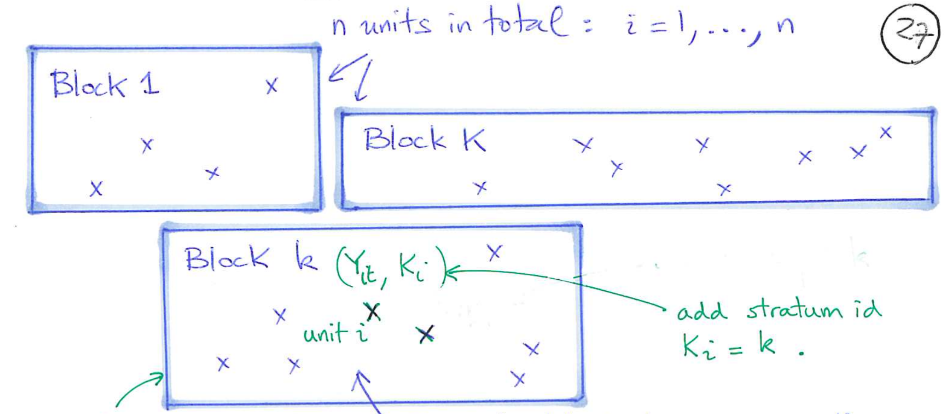
Pre-test + Observational data : W_{it} are not randomized. The OLS estimator $\hat{\Delta}$ of Δ in

$Y_{it} = \alpha_i + \beta_t + \Delta W_{it} + \varepsilon_{it}$ is a consistent estimator of

$$ATT(T_0, T) = \frac{1}{T-T_0}\sum_{t=T_0+1}^{T} \mathbb{E}\left(Y_{it}(1) - Y_{it}(0) \mid W_i = 1\right)$$

under assumptions (A) Parallel trends & (B) No anticipation (p.18)

Instead of a single collection of units where these results hold, we may consider K of them (aka strata).



n_k units in the k-th block

In each block k, assume either an RCT, or that parallel trends + no anticipation holds.

To analyse such data and recover ATE, ATT(T_0, T), we may

(i) analyse block by block and aggregate the point estimates

(ii) pull all the data together and solve for Δ in $Y_{it} = \beta_0 + \Delta W_i + \varepsilon_{it}$ (RCT) or $Y_{it} = \alpha_i + \beta_t + \Delta W_{it} + \varepsilon_{it}$ (DiD) using a proper weighting of the units, to account for potential imbalance across blocks → IPW approach.

III.1. Analysis block by block

An analysis block by block requires a substantial amount of data (treated / control units), so that asymptotic results are well approximated in each block.

In the k-th stratum, let $\hat{\Delta}_k$ denote the OLS estimator of Δ_k in

$$Y_{it} = \beta_{0k} + \Delta_k W_i + \varepsilon_{it} \quad , \quad \{i \mid K_i = k\}, \; t = 1, \dots, T$$

(difference est / RCT)

(n_k units)

or

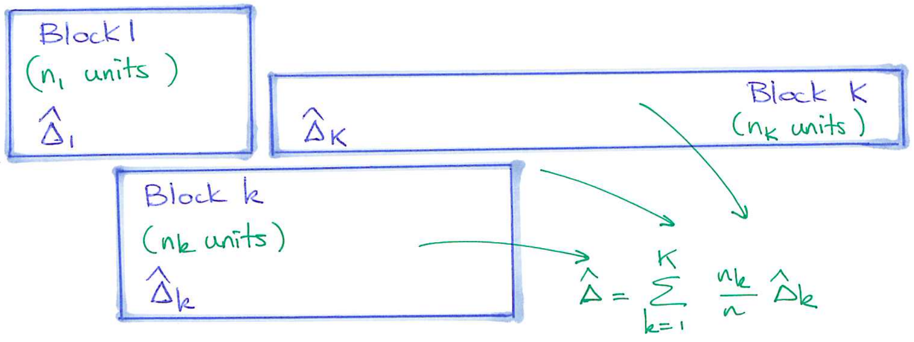
$$Y_{it} = \alpha_{ik} + \beta_{tk} + \Delta_k W_{it} + \varepsilon_{it} \quad , \quad \{i \mid K_i = k\}, \; t = 1, \dots, T$$

(diff-in-diff / no anticipation + // trends)

Let $n_1 + \dots + n_K = n$. The aggregate estimator $\hat{\Delta}_{AGG}$ is defined by

$$\boxed{\hat{\Delta}_{AGG} = \sum_{k=1}^{K} \frac{n_k}{n} \hat{\Delta}_k}$$

The variance of $\hat{\Delta}$ can be easily computed using $\operatorname{var} \hat{\Delta} = \sum_{k=1}^{K} \left(\frac{n_k}{n}\right)^2 \operatorname{var} \hat{\Delta}_k$.



Put

- $\text{ATE}(k) := \frac{1}{T} \sum_{t=1}^{T} E(Y_{it}(1) - Y_{it}(0) \mid K_i = k)$
- $\text{ATT}(k; T_0, T) := \frac{1}{T - T_0} \sum_{t=T_0+1}^{T} E(Y_{it}(1) - Y_{it}(0) \mid K_i = k, W_i = 1)$.

As n_{1k} (= # treated units in the k-th stratum) and $n_{0k} = n_k - n_{1k} \to \infty$, the aggregate estimator

$$\hat{\Delta}_{AGG} \to \sum_{k=1}^{K} \pi_k \underbrace{ATE(k)}_{= ATE} \quad (\text{difference est ; RCT})$$

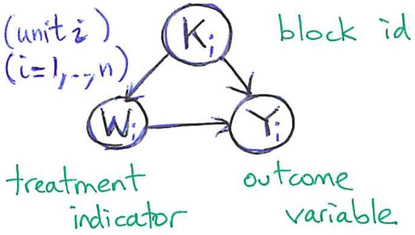
$$\to \sum_{k=1}^{K} \pi_k \underbrace{ATT(k ; T_0, T)}_{= ATT(T_0, T)} , \quad (\text{did ; no anticipation } + \text{ // trends})$$

where $\frac{n_k}{n} \to \pi_k$ as $n_k \to \infty$.

III. 2. IPW

We discuss the difference and diff - in - diff estimators separately.

- The difference estimator: in each block $k = 1, .., K$, we perform an RCT. In other words, when pulling the data together, the block / stratum id confounds the treatment effect :



⇒ Consider an IPW approach, weighting unit i in stratum k by $\frac{1}{\mathbb{P}(W_i = 1 \mid K_i = k)}$ if i is in the treatment group, and by $\frac{1}{1 - \mathbb{P}(W_i = 1 \mid K_i = k)}$ if i is in the control group.

Let $\hat{\Delta}_{IPW}$ denote the WLS estimate of Δ in

$$Y_{it} = \beta_0 + \Delta W_i + \varepsilon_{it} , \quad \left(\begin{matrix} i = 1, ..., n \\ t = 1, .., T \end{matrix} \right)$$

with weight

$$z_i = z(W_i, K_i) = \begin{cases} \dfrac{1}{\mathbb{P}(W_i = 1 \mid K_i = k)} & \text{if } i \in \text{trt}, K_i = k \\ \dfrac{1}{1 - \mathbb{P}(W_i = 1 \mid K_i = k)} & \text{if } i \in \text{ctl}, K_i = k \end{cases}$$

- Result : The IPW estimator of Δ is the weighted difference in means

$$\hat{\Delta}_{IPW} = \underbrace{\frac{\sum_{i | W_i = 1, t} z(W_i, K_i) Y_{it}}{\sum_{i | W_i = 1, t} z(W_i, K_i)}}_{\text{[treated units]}} - \underbrace{\frac{\sum_{i | W_i = 0, t} z(W_i, K_i) Y_{it}}{\sum_{i | W_i = 0, t} z(W_i, K_i)}}_{\text{[control units]}}$$

(see proof on the next page)

Then $$\hat{\Delta}_{IPW} \xrightarrow[(n \to \infty)]{} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}(Y_{it}(1) - Y_{it}(0)) = ATE$$

Lets consider the treated units in $\hat{\Delta}_{IPW}$. The control units are treated similarly. Multiplying the numerator and denominator by $\frac{1}{nT}$, we see that

- $\frac{1}{nT} \sum_{i | W_i = 1, t} z(W_i, K_i) \xrightarrow[(n \to \infty)]{} \mathbb{E}\{W z(W, K)\}$

$$= \sum_k z(1, k) \, \mathbb{P}(W = 1, K = k)$$

$$= \sum_k \mathbb{P}(K = k) = 1$$

- Similarly, $\frac{1}{nT} \sum_{i | W_i = 1, t} z(W_i, K_i) Y_{it}(1)$

$$\xrightarrow[(n \to \infty)]{} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}(Y_{it}(1))$$

proof: $(\hat\beta_0, \hat\Delta_{IPW}) = \underset{\beta_0, \Delta}{\arg\min} \left\{ \sum_{i=1}^{nT} z(W_i, K_i)(Y_i - \beta_0 - \Delta W_i)^2 \right\}$

Put $\hat\beta = \begin{pmatrix} \hat\beta_0 \\ \hat\Delta_{IPW} \end{pmatrix}$.

⚠ For convenience we denote the nT observations using a single subscript

Then $\hat\beta = (X^t Z X)^{-1} X^t Z Y$,

where $\underset{(nT \times nT)}{Z} = \mathrm{diag}\{ z(W_i, K_i) \}$

$$\underset{(nT\times 2)}{X} = \begin{pmatrix} 1 & W_1 \\ \vdots & \vdots \\ 1 & W_{nT} \end{pmatrix}, \quad \underset{(nT\times 1)}{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_{nT} \end{pmatrix}$$

After calculations, $X^t X = \begin{pmatrix} \sum z_i & \sum z_i W_i \\ \sum z_i W_i & \sum z_i W_i \end{pmatrix}$

$z_i = z(W_i, K_i)$ $i = 1, \dots, nT$

$$X^t Z Y = \begin{pmatrix} \sum z_i Y_i \\ \sum z_i W_i Y_i \end{pmatrix}$$

$$\hat\beta = \begin{pmatrix} \dfrac{\sum z_i Y_i (1 - W_i)}{\sum z_i (1 - W_i)} \\ \\ \dfrac{(\sum z_i)(\sum z_i W_i Y_i) - (\sum z_i W_i)(\sum z_i Y_i)}{(\sum z_i W_i)(\sum z_i (1 - W_i))} \end{pmatrix}$$

$$\Rightarrow \hat\beta_0 = \frac{\sum_{i|W_i=0} z_i Y_i}{\sum_{i|W_i=0} z_i} \; ; \; \hat\beta_0 + \hat\Delta_{IPW} = \frac{\sum_{i|W_i=1} z_i Y_i}{\sum_{i|W_i=1} z_i}$$

■

× Remark :

$$\hat\Delta_{AGG} = \sum_{k=1}^{K} \left(\frac{n_k}{n} \right) \hat\Delta_k \quad \text{(page 22)}$$

$$= \underbrace{\sum_{k=1}^{K} \frac{n_k}{n} \left(\frac{1}{n_{1k} T} \sum_t \sum_{i|W_i=1} Y_{it} \right.}_{\text{focusing on the first term,}} \left. - \frac{1}{n_{0k} T} \sum_t \sum_{i|W_i=0} Y_{it} \right)$$

$$= \frac{1}{nT} \sum_{t=1}^{T} \sum_{k=1}^{K} \sum_{\substack{i|W_i=1 \\ K_i=k}} \frac{n_k}{n_{1k}} Y_{it}$$

$$\frac{n_k}{n_{1k}} = \frac{1}{\mathbb{P}(W_i = 1 | K_i = k)} = z(W_i, K_i)$$

$$= \frac{1}{nT} \sum_{t=1}^{T} \sum_{i|W_i=1} z(W_i, K_i) Y_{it}$$

In addition, note that

$$\frac{1}{nT} \sum_{t=1}^{T} \sum_{i|W_i=1} z(W_i, K_i) = \frac{1}{nT} \sum_{t=1}^{T} \underbrace{\sum_{i|W_i=1} \frac{n_k}{n_{1k}}} = 1$$

$$= \sum_{k=1}^{K} \sum_{\substack{i|W_i=1 \\ K_i=k}} \frac{n_k}{n_{1k}} = \sum_{k=1}^{K} n_k = n$$

Thus

$$\hat\Delta_{AGG} = \frac{\frac{1}{nT} \sum_{t=1}^{T} \sum_{i|W_i=1} z(W_i, K_i) Y_{it}}{\frac{1}{nT} \sum_{t=1}^{T} \sum_{i|W_i=1} z(W_i, K_i)} - \frac{\text{similar term for control}}{\text{similar term for control}}$$

$= \hat\Delta_{IPW}$. $\Rightarrow$ Point estimates $\hat\Delta_{AGG}$ and $\hat\Delta_{IPW}$ coincide.

■

× Summary for the difference estimator

- Set-up =. $k = 1, \dots, K$ blocks
    - In each block, randomize units completely at random : $\{Y_{it}(0), Y_{it}(1)\} \perp W_i \mid K_i = k$

- Analysis block per block (page 22)

$\hat{\Delta}_k$ = OLS estimator of Δ_k in $Y_{it} = \beta_{0k} + \Delta_k W_i + \varepsilon_{it}$ for units i in the k-th stratum

$$\hat{\Delta}_{AGG} = \sum_{k=1}^{K} \frac{n_k}{n} \hat{\Delta}_k$$ ← Need large n_k to estimate the variance of $\hat{\Delta}_k$ accurately

- IPW estimator (page 24)

All data pulled together

$\hat{\Delta}_{IPW}$ = WLS estimator of Δ in $Y_{it} = \beta_0 + \Delta W_i + \varepsilon_{it}$, where observation Y_{it} has weight $z(W_i, K_i)$

$$= \begin{cases} 1/\mathbb{P}(W_i = 1 \mid K_i = k) & \text{if } i \text{ is treated} \\ 1/\mathbb{P}(W_i = 0 \mid K_i = k) & \text{if } i \text{ is in control.} \end{cases}$$

Then . $$\hat{\Delta}_{IPW} = \frac{1}{nT} \sum_{t=1}^{T} \sum_{i \mid W_i = 1} z(W_i, K_i) Y_{it} - \frac{1}{nT} \sum_{t=1}^{T} \sum_{i \mid W_i = 0} z(W_i, K_i) Y_{it}$$

$$= \hat{\Delta}_{AGG}$$ (page 26)

- Also,

$$\hat{\Delta}_{AGG} = \hat{\Delta}_{IPW} \xrightarrow{n \to \infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}(Y_{it}(1) - Y_{it}(0))$$

$= ATE$

* n_k = # units in k-th block

n_{1k} = # treated units in the k-th block

In the summary page 27, we see that the analysis block per block requires the number n_k of units to be large in each block to estimate the variance of $\hat{\Delta}_k$ (and thus of $\hat{\Delta}_{AGG}$) accurately.

This condition can be relaxed with the IPW estimator : units (Y_{it}, K_i) are assumed to be drawn at random, and we only require to observe many of them to get an accurate estimate of the variance of $\hat{\Delta}_{IPW}$ (even when clustering errors over $t = 1, \dots, T$).

⇒ With small blocks, the IPW estimator is preferred over the aggregated estimator.

- The difference-in-differences estimator

We begin by stating a general result on the WLS solution of the diff-in-diff linear model, with general weights:

* **Result** : The WLS estimate $\hat{\beta}_3$ of β_3 in

$$Y_{it} = \beta_0 + \beta_1 \mathbb{1}(i \in trt) + \beta_2 \mathbb{1}(t > T_0) + \beta_3 \mathbb{1}(i \in trt, T > T_0) + \varepsilon_{it}$$

with weight z_i (constant weight $\forall t$) is

$$\hat{\beta}_3 = \underbrace{\left\{ \frac{\sum_{\substack{i \in trt \\ t \geq T_0+1}} z_i Y_{it}}{\sum_{\substack{i \in trt \\ t \geq T_0+1}} z_i} - \frac{\sum_{\substack{i \in trt \\ t \leq T_0}} z_i Y_{it}}{\sum_{\substack{i \in trt \\ t \leq T_0}} z_i} \right\}}_{\text{treated units}} - \underbrace{\left\{ \frac{\sum_{\substack{i \in ctl \\ t \geq T_0+1}} z_i Y_{it}}{\sum_{\substack{i \in ctl \\ t \geq T_0+1}} z_i} - \frac{\sum_{\substack{i \in ctl \\ t \leq T_0}} z_i Y_{it}}{\sum_{\substack{i \in ctl \\ t \leq T_0}} z_i} \right\}}_{\text{control units}}$$

↑ test ↑ pre-test

× Case I = Randomized treatment

Units are randomized (completely at random) in each block. We use IPW weights to recover the ATE.

$$z_i = \begin{cases} 1 / \mathbb{P}(W_i = 1 \mid K_i = k) & \text{if } i \text{ is treated} \\ 1 / \mathbb{P}(W_i = 0 \mid K_i = k) & \text{if } i \text{ is in control.} \end{cases}$$

Proceeding as before,

$$\hat{\beta}_3 \xrightarrow{(n\to\infty)} \left\{ \frac{1}{T-T_0} \sum_{t=T_0+1}^{T} \mathbb{E} Y_{it}(1) - \cancel{\frac{1}{T_0} \sum_{t=1}^{T_0} \mathbb{E} Y_{it}(1)} \right\}$$

$$- \left\{ \frac{1}{T-T_0} \sum_{t=T_0+1}^{T} \mathbb{E} Y_{it}(0) - \cancel{\frac{1}{T_0} \sum_{t=1}^{T_0} \mathbb{E} Y_{it}(0)} \right\}$$

$$= \frac{1}{T-T_0} \left\{ \sum_{t=T_0+1}^{T} \mathbb{E}\left(Y_{it}(1) - Y_{it}(0) \right) \right\} .$$

since $\forall t = 1, .., T_0$, $\mathbb{E}\, Y_{it}(1) = \mathbb{E}\, Y_{it}(0)$ due to randomization (assuming no anticipation)

↘ Note that when checking balance of the KPI on the treatment & control groups, one must use weighted averages (using IPW weights) since

$$\mathbb{E}\, Y_{it}(1) = \mathbb{E}\, \mathbb{E}\, Y_{it}(1) \mid K_i$$

$$\approx \sum_{k=1}^{K} \frac{n_k}{n} \frac{1}{n_{1k}} \sum_{\substack{i \mid W_i = 1 \\ K_i = k}} Y_{it}$$

$$= \frac{1}{n} \sum_{i \mid W_i = 1} \left(\frac{n_{1k}}{n_k} \right)^{-1} Y_{it} \qquad \blacksquare$$

× Case II = Observation Study

Assume // trend & no anticipation in each block:

[cond // trend] [two time periods]

$$\mathbb{E}\left(Y_{i1}(0) - Y_{i0}(0) \mid W_i = 0, \; K_i = k \right)$$

$$= \mathbb{E}\left(Y_{i1}(0) - Y_{i0}(0) \mid W_i = 1, \; K_i = k \right)$$

[no anticipation] $Y_{i0}(0) = Y_{i0}(1)$ [in pre-test]

↘ Again, abusing notation $\left(\begin{array}{l} Y_{it}(0) = Y_{it}(0,0) \\ Y_{it}(1) = Y_{it}(0,1) \end{array} \right.$

(pre-test has subscript 0 & test subscript 1)

– full trajectory –

Theorem Abadie (2005) [two time periods]

$$\text{ATT} := \mathbb{E}\left(Y_{i1}(1) - Y_{i1}(0) \mid W_i = 1 \right)$$

$$= \mathbb{E}\left(Y_{i1} - Y_{i0} \mid W_i = 1 \right)$$

$$- \mathbb{E}\left\{ \alpha(K_i) (Y_{i1} - Y_{i0}) \mid W_i = 0 \right)$$

observed quantities → the ATT is identified.

with $$\alpha(k) = \frac{\mathbb{P}(W_i = 1 \mid K_i = k)}{\mathbb{P}(W_i = 1)} \times \frac{\mathbb{P}(W_i = 0)}{\mathbb{P}(W_i = 0 \mid K_i = k)}$$

↘ Holds with general vector / real-valued covariates K_i.

proof =

- First term = $\mathbb{E}\left(Y_{i1} - Y_{i0} \mid W_i = 1 \right) = \mathbb{E}\left(Y_{i1}(1) - Y_{i0}(1) \mid W_i = 1 \right)$

$$= \mathbb{E}\left(Y_{i1}(1) - Y_{i0}(0) \mid W_i = 1\right)$$

(no anticipation)

- Second term $= \mathbb{E}\left\{(Y_{i1} - Y_{i0})\,\alpha(K_i) \mid W_i = 0\right\}$

$$= \mathbb{E}\left\{(Y_{i1}(0) - Y_{i0}(0))\,\alpha(K_i) \mid W_i = 0\right\}$$

$$= \mathbb{E}\left[\mathbb{E}\left\{ \text{---}''\text{---} \mid W_i = 0,\ K_i\right\} \mid W_i = 0\right]$$

$$= \mathbb{E}\left[\alpha(K_i)\,\underset{\Psi(K_i)}{\underset{\equiv}{\boxed{\mathbb{E}\left\{(Y_{i1}(0) - Y_{i0}(0)) \mid W_i = 0, K_i\right\}}}} \mid W_i = 0\right]$$

$$= \mathbb{E}\left[\alpha(K_i)\,\Psi(K_i) \mid W_i = 0\right]$$

$$= \sum_k \alpha(k)\,\Psi(k)\,\underset{\dfrac{\mathbb{P}(W_i=0 \mid K_i=k)}{\mathbb{P}(W_i=0)}\mathbb{P}(K_i=k)}{\underset{\parallel}{\boxed{\mathbb{P}(K_i = k \mid W_i = 0)}}}$$

definition of $\alpha(k)$ $\downarrow$

$$= \sum_k \Psi(k)\,\frac{\mathbb{P}(W_i = 1 \mid K_i = k)}{\mathbb{P}(W_i = 1)}\,\mathbb{P}(K_i = k)$$

$$= \sum_k \Psi(k)\,\mathbb{P}(K_i = k \mid W_i = 1)$$

Conditional // trend assumption $\downarrow$

$$= \sum_k \mathbb{E}\left\{(Y_{i1}(0) - Y_{i0}(0)) \mid W_i = 1, K_i = k\right\} \times \mathbb{P}(K_i = k \mid W_i = 1)$$

$$= \mathbb{E}\left[\mathbb{E}\left(Y_{i1}(0) - Y_{i0}(0) \mid W_i = 1, K_i\right) \mid W_i = 1\right]$$

$$= \mathbb{E}\left[Y_{i1}(0) - Y_{i0}(0) \mid W_i = 1\right]$$

In addition, note that

$$\mathbb{E}\left(\alpha(K_i) \mid W_i = 0\right) = \sum_k \alpha(k)\,\mathbb{P}(K_i = k \mid W_i = 0)$$

$$= \sum_k \frac{\mathbb{P}(W_i=0)}{\mathbb{P}(W_i=1)}\,\frac{\mathbb{P}(W_i = 1 \mid K_i = k)}{\mathbb{P}(W_i = 0 \mid K_i = k)} \times \mathbb{P}(K_i = k \mid W_i = 0)$$

$$= \frac{\mathbb{P}(K_i = k)}{\mathbb{P}(W_i = 0)}$$

$$= \sum_k \frac{\mathbb{P}(\cancel{W_i=0})}{\mathbb{P}(W_i=1)}\,\mathbb{P}(W_i = 1 \mid K_i = k)\,\frac{\mathbb{P}(K_i = k)}{\mathbb{P}(\cancel{W_i=0})}$$

$$= \sum_k \mathbb{P}(K_i = k \mid W_i = 1)$$

$$= 1$$

and likewise in the treatment group.

<u>Consequence</u>: The ATT can be estimated using WLS:

$$Y_{it} = \beta_0 + \beta_1 \mathbb{1}(i \text{ is trt}) + \beta_2 \mathbb{1}(t = 1) + \beta_3 \mathbb{1}(i \text{ is trt},\ t = 1) + \varepsilon_{it} \qquad \left(\begin{matrix} t = 0, 1 \\ i = 1, \dots, n \end{matrix}\right.$$

$$z_i = \begin{cases} \alpha(k) & \text{if } i \text{ in block } k \text{ is in control} \\ 1 & \text{if } i \text{ is treated.} \end{cases}$$

The WLS estimate of $\hat{\beta}_3$ identifies the ATT.

The result generalizes to multi-periods assuming

(A) Conditional Parallel Trends

$$\frac{1}{T-T_0}\sum_{t=T_0+1}^{T}\mathbb{E}(Y_{it}(0)\mid W_i=1, K_i=k)-\frac{1}{T_0}\sum_{t=1}^{T_0}\mathbb{E}(Y_{it}(0)\mid W_i=1,K_i=k)$$

$$=$$

$$\frac{1}{T-T_0}\sum_{t=T_0+1}^{T}\mathbb{E}(Y_{it}(0)\mid W_i=0, K_i=k)-\frac{1}{T_0}\sum_{t=1}^{T_0}\mathbb{E}(Y_{it}(0)\mid W_i=0,K_i=k)$$

(B) No anticipation

$$Y_{it}(\underbrace{0,\dots,0}_{T_0},\underbrace{0,\dots,0}_{T-T_0}) = Y_{it}(\overbrace{0,\dots,0}^{T_0},\overbrace{1,\dots,1}^{T-T_0}) \quad \forall t=1,\dots,T_0$$

$$(\Leftrightarrow\ Y_{it}(0)=Y_{it}(1)\ ,\ t\le T_0)\ \forall i \text{ with } W_i=1$$

Then

$$ATT(T_0,T)=\frac{1}{T-T_0}\sum_{t=T_0+1}^{T}\mathbb{E}(Y_{it}(1)-Y_{it}(0)\mid W_i=1)$$

P.O.

$$=\left\{\frac{1}{T-T_0}\sum_{t=T_0+1}^{T}\mathbb{E}(Y_{it}\mid W_i=1)-\frac{1}{T_0}\sum_{t=1}^{T_0}\mathbb{E}(Y_{it}\mid W_i=1)\right\}$$

Observed quantities

$$-\left\{\frac{1}{T-T_0}\sum_{t=T_0+1}^{T}\mathbb{E}\left[\alpha(K_i)Y_{it}\mid W_i=0\right]-\frac{1}{T_0}\sum_{t=1}^{T_0}\mathbb{E}\left[\alpha(K_i)Y_{it}\mid W_i=0\right]\right\}$$

& proceed as before to estimate the ATT using weighted least squares: $Y_{it}=\beta_0+\beta_1\mathbb{1}(trt)+\beta_2\mathbb{1}(test)+\beta_3\mathbb{1}(trt,test)+\varepsilon_{it}$.

weight z_i

× Summary for the did estimator

- Set-up:
  - $k=1,\dots,K$ blocks
  - In each block, assume parallel trends & no anticipatio

[p. 30 → two time periods
p. 33 → multi-periods]

- Analysis block per block (p.22)

$\hat{\Delta}_k$ = OLS estimator of Δ_k in $Y_{it}=\alpha_{ik}+\beta_{tk}+\Delta W_{it}+\varepsilon_{it}$

$(\forall i \mid K_i=k)\ ;\ t=1,\dots,T$

$$\hat{\Delta}_{AGG}=\sum_{k=1}^{K}\frac{n_k}{n}\hat{\Delta}_k$$

- IPW estimator

$\hat{\Delta}_{IPW}$ = WLS estimator of Δ in (p.28)

$$Y_{it}=\beta_0+\beta_1\mathbb{1}(W_i=1)+\beta_2\mathbb{1}(t\geqslant T_0+1)+\Delta\mathbb{1}(W_i=1,\ t\geqslant T_0+1)+\varepsilon_{it}$$

with weight z_i.

(A) If treatment is randomized, (p.29)

$$z_i=\begin{cases}1/\,\mathbb{P}(W_i=1\mid K_i=k) & \text{if } i \text{ is treated (in } k\text{-th stratum)}\\ 1/\,\mathbb{P}(W_i=0\mid K_i=k) & \text{if } i \text{ is in control}\end{cases}$$

(B) If treatment is not randomized (under // trends & no anticipation)

$$z_i=z(W_i,K_i)=\begin{cases}1 & \text{if } W_i=1,\ K_i=k\\ \alpha(k) & \text{if } W_i=0,\ K_i=k\end{cases}$$

(p.30)

$$\hat{\Delta}_{AGG},\ \hat{\Delta}_{IPW}\xrightarrow{n\to\infty}ATT(T_0,T)\ \left[=ATE \text{ if trt is rdmized}\right]$$

x Appendix A.1. We show that the OLS estimator of Δ in $Y_{it} = \alpha_i + \beta_t + \Delta W_{it} + \varepsilon_{it}$ is the difference-in-differences $\hat{\Delta}$ on page 12. We proceed in three steps

- ↘ Step I = [toolbox] the Frisch-Waugh Theorem
- ↘ Step II = One-way Fixed Effects
- ↘ Step III = Two-way Fixed Effects.

- Step I : The FW theorem

[Reg 1] $$\underset{(n\times1)}{Y} = \underset{(n\times k_1)}{X_1}\beta_1 + \underset{(n\times k_2)}{X_2}\beta_2 + \underset{(n\times1)}{\varepsilon}$$

[Reg 2] $$\underset{(n\times n)(n\times1)}{M_1 Y} = M_1 \underset{(n\times k_2)}{X_2}\beta_2 + \underset{(n\times1)}{u}$$

where $M_1 = I - X_1(X_1^t X_1)^{-1}X_1^t$
= projection matrix onto the $\perp$ column space of X_1

Then

(i) OLS estimates of β_2 in [Reg 1] and in [Reg 2] are identical

(ii) OLS residuals from [Reg 1] and [Reg 2] are identical.

↘ Why is this interesting? When k_1 is large, we can solve numerically a simpler problem if one is interested in estimating β_2.

---

proof =

[Reg 1] Let $\hat{\beta}_1, \hat{\beta}_2$ = OLS estimates of β_1, β_2

[Reg 2] Put $\tilde{\beta}_2 = (X_2^t M_1 X_2)^{-1} X_2^t M_1 Y$ = OLS estimate of β_2.
↖ idempotent

Note that

$$\boxed{Y = \boxed{P_X Y} + (I-P_X)Y = \boxed{X_1\hat{\beta}_1 + X_2\hat{\beta}_2} + M_X Y} \quad (*)$$

↑ where $M_X = I - P_X$; $\underset{n\times(k_1+k_2)}{X = [X_1 \;\; X_2]}$

$M_X = I - X(X^tX)^{-1}X^t$

↓ pre-multiplying by $X_2^t M_1$:

$$\boxed{X_2^t M_1 Y = \boxed{X_2^t M_1 X_1 \hat{\beta}_1} + X_2^t M_1 X_2 \hat{\beta}_2 + \boxed{X_2^t M_1 M_X Y}}$$

= 0 since M_1 wipes off X_1

taking the transpose:
$\boxed{M_X M_1} X_2 = \boxed{M_X} X_2 = 0$

since $P_1 P_X = P_X P_1 = P_1$
$\Rightarrow M_X M_1 = (I-P_X)(I-P_1)$
$= I - P_X - P_1 + P_X P_1$
$= I - P_X = M_X$

since M_X wipes off all columns in X

$\Rightarrow$ It follows that
$\hat{\beta}_2 = (X_2^t M_1 X_2)^{-1} X_2^t M_1 Y = \tilde{\beta}_2$; which concludes the first part of the theorem.

Pre-multiplying (*) by M_1 yields

$$\boxed{M_1 Y} = \boxed{M_1 X_2 \hat{\beta}_2} + \boxed{M_x Y} \longleftarrow$$

It follows that $M_x Y$ is the vector of residuals in [Reg 2]

But it is immediate to see that $M_x Y$ is also the vector of residuals in [Reg 1]. This concludes the second part of the thm.

$\uparrow$ Regressand in [Reg 2]

$\nwarrow$ $= M_1 X_2 \tilde{\beta}_2$ since we just proved that $\hat{\beta}_2 = \tilde{\beta}_2$

- Step II = One-Way Fixed Effects. Baltagi

$$\boxed{Y_{it} = \alpha_i + \Delta W_{it} + \varepsilon_{it}}$$

vector notation

$$\boxed{\underset{(nT\times 1)}{Y} = \underset{(nT\times n)}{Z_\alpha}\, \underset{(n\times 1)}{\alpha} + \underset{(nT\times 1)}{W}\, \underset{(1\times 1)}{\Delta} + \underset{(nT\times 1)}{\varepsilon}}$$

$$Y = \begin{pmatrix} Y_{11} \\ \vdots \\ Y_{1T} \\ \hline \vdots \\ \hline \vdots \end{pmatrix} \begin{array}{l} \updownarrow \text{unit } 1 \\ \\ \\ \\ \updownarrow \text{unit } n \end{array} \qquad \alpha = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix} = FE \qquad W = \begin{pmatrix} W_{11} \\ \vdots \\ W_{1T} \\ \hline \vdots \\ \hline \vdots \end{pmatrix}$$

Put $\mathbb{1}_T = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \in \mathbb{R}^T$. Then $Z_\alpha = I_n \underset{\text{Kronecker product}}{\otimes} \mathbb{1}_T$

$$Z_\alpha = \begin{array}{c} \longleftarrow n \longrightarrow \\ \begin{array}{|cc|} \hline \begin{matrix}1\\ 1\\ 1\end{matrix} & 0 \\ \hline 0 \; \begin{matrix}1\\ 1\\ 1\end{matrix} & 0 \\ \hline \multicolumn{2}{|c|}{\cdots/\cdots} \\ \hline 0 & \begin{matrix}1\\ 1\\ 1\end{matrix} \\ \hline \end{array} \end{array} \begin{array}{l} \updownarrow T \\ \updownarrow T \\ \\ \updownarrow T \end{array}$$

We are interested in estimating Δ using FW.

$$\Rightarrow P_\alpha := Z_\alpha (Z_\alpha^t Z_\alpha)^{-1} Z_\alpha^t = I_n \otimes \bar{J}_T$$

where $\bar{\bar{J}}_T = \frac{1}{T} \mathbb{1}_T \mathbb{1}_T^t$

$$= \left[\begin{array}{c|c|c} \bar{J}_T & & \\ \hline & & \\ \hline & & \bar{\bar{J}}_T \end{array}\right] = \text{time-averages for each unit}$$

$P_\alpha Y$ has elements $\frac{1}{T} \sum_{t=1}^{T} Y_{it} = \bar{Y}_{i\cdot}$

$\Rightarrow$ $Q_\alpha = I - P_\alpha$ plays the same role as M_1 in FW (Step I)
= deviations from the mean

$Q_\alpha Y$ has elements $Y_{it} - \bar{Y}_{i\cdot}$

FW theorem with Q_α implies that $\boxed{\hat{\Delta} = (\underbrace{W^t Q_\alpha W}_{})^{-1} W^t Y}$

= a scalar; much smaller dimension than inverting an $(n+1)\times(n+1)$ matrix

In addition, we can recover the FE using:

$$Y_{it} = \alpha_i + \Delta W_{it} + \varepsilon_{it}$$

$$\bar{Y}_{i\cdot} = \alpha_i + \Delta \bar{W}_{i\cdot} + \bar{\varepsilon}_{i\cdot}$$

Once $\hat{\Delta}$ is computed, we get $\hat{\alpha}_i = \bar{Y}_{i\cdot} - \hat{\Delta} \bar{W}_{i\cdot}$

x Remark: Equivalently, we may use the representation $Y_{it} = \mu + \alpha_i + \Delta W_{it} + \varepsilon_{it}$; imposing $\sum \alpha_i = 0$ for identifiability.

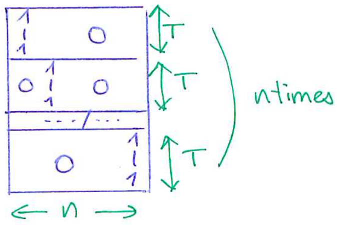
$$\Rightarrow \begin{array}{l} \bar{Y}_{i\cdot} = \mu + \alpha_i + \Delta \bar{W}_{i\cdot} + \bar{\varepsilon}_{i\cdot} \\ \bar{Y}_{\cdot\cdot} = \mu \qquad\;\; + \Delta \bar{W}_{\cdot\cdot} + \bar{\varepsilon}_{\cdot\cdot} \end{array} \quad \Rightarrow \quad \begin{array}{l} \hat{\mu} = \bar{Y}_{\cdot\cdot} - \hat{\Delta} \bar{W}_{\cdot\cdot} \\ \hat{\alpha}_i = \bar{Y}_{i\cdot} - \hat{\mu} - \hat{\Delta} \bar{W}_{i\cdot} \end{array}$$

• Step III = Two-Way FE

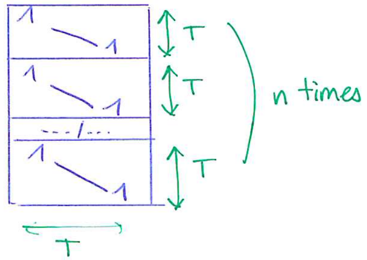
$$Y_{it} = \alpha_i + \beta_t + \Delta W_{it} + \varepsilon_{it}$$

$$\underset{nT\times 1}{Y} = \underset{nT\times n}{Z_\alpha}\,\underset{n\times 1}{\alpha} + \underset{nT\times T}{Z_\beta}\,\underset{T\times 1}{\beta} + \underset{nT\times 1}{W}\,\underset{1\times 1}{\Delta} + \underset{nT\times 1}{\varepsilon}$$

$\underset{nT\times n}{Z_\alpha} = I_n \otimes \mathbb{1}_T =$



$\underset{nT\times T}{Z_\beta} = \mathbb{1}_n \otimes I_T =$



Apply FW with

$$\underset{nT\times nT}{Q} := I_n \otimes I_T - I_n \otimes \bar{J}_T - \bar{J}_n \otimes I_T + \bar{J}_n \otimes \bar{J}_T$$

$$= E_n \otimes E_T \quad \text{where} \quad E_n := I_n - \bar{J}_n$$

$$E_T := I_T - \bar{J}_T$$

One can check that Q wipes out the unit & time FE (and the intercept if there is one). Moreover, $QY = \tilde{Y}$ where $\tilde{Y}_{it} := Y_{it} - \bar{Y}_{i.} - \bar{Y}_{.t} + \bar{Y}_{..}$

= double de-meaned

$$\bar{Y}_{..} = \frac{1}{nT}\sum_{i,t} Y_{it}$$

We get $\hat{\Delta} = (W^t Q W)^{-1} W^t Q Y$

A scalar to invert instead of an $(n+T+1)\times(n+T+1)$ matrix!

We recover the FE using

$$Y_{it} = \mu + \alpha_i + \beta_t + \Delta W_{it} + \varepsilon_{it}$$

$$\bar{Y}_{i.} = \mu + \alpha_i + \Delta \bar{W}_{i.} + \bar{\varepsilon}_{i.}$$

$$\bar{Y}_{.t} = \mu + \beta_t + \Delta \bar{W}_{.t} + \bar{\varepsilon}_{.t}$$

$$\bar{Y}_{..} = \mu + \Delta \bar{W}_{..} + \bar{\varepsilon}_{..}$$

$$\Rightarrow \begin{cases} \hat{\mu} = \bar{Y}_{..} - \hat{\Delta}\bar{W}_{..} \\ \hat{\alpha}_i = (\bar{Y}_{i.} - \bar{Y}_{..}) - (\bar{W}_{i.} - \bar{W}_{..})\hat{\Delta} \\ \hat{\beta}_t = (\bar{Y}_{.t} - \bar{Y}_{..}) - (\bar{W}_{.t} - \bar{W}_{..})\hat{\Delta} \end{cases}$$

× **Remark** : We can re-express $\hat{\Delta}$ in a simplified way:

$\tilde{Y} := QY$

$\tilde{W} := QW$

Then $\hat{\Delta} = (W^t Q W)^{-1} W^t Q Y$ $\quad Q^t = Q$

$$= (W^t Q^t W)^{-1} W^t Q^t Y$$

$$= (\tilde{W}^t W)^{-1} \tilde{W}^t Y$$

$$= \frac{\sum_{i,t} \tilde{W}_{it} Y_{it}}{\sum_{i,t} \tilde{W}_{it} W_{it}}$$

$n_1 = \#$ treated units

$$\tilde{W}_{it} = W_{it} - \bar{W}_{i.} - \bar{W}_{.t} + \bar{W}_{..}$$

$\bar{W}_{i.} = \frac{T-T_0}{T}\mathbb{1}(i \in trt)$

$\bar{W}_{.t} = \frac{n_1}{n}\mathbb{1}(t \geqslant T_0 + 1)$

$\bar{W}_{..} = \frac{n_1(T-T_0)}{nT}$

$$\sum_{i,t} \tilde{w}_{it} Y_{it}$$

$$= \boxed{\sum_{i,t} \omega_{i,t} Y_{it}} = \sum_{t \geq T_0+1} \sum_{i \in trt} Y_{it} =: S_{trt,\, test}$$

$$- \boxed{\sum_{i,t} \bar{\omega}_{i.} Y_{it}} = \frac{T-T_0}{T} \sum_{t=1}^{T} \sum_{i \in trt} Y_{it} = \frac{T-T_0}{T} \left(S_{trt,pre} + S_{trt,test} \right)$$

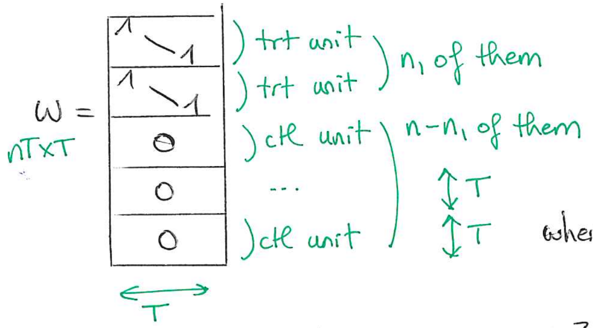
$$- \boxed{\sum_{i,t} \bar{\omega}_{.t} Y_{it}} = \frac{n_1}{n} \sum_{t \geq T_0+1} \sum_{i=1}^{n} Y_{it} = \frac{n_1}{n} \left(S_{trt,test} + S_{ctl,test} \right)$$

$$+ \boxed{\sum_{i,t} \bar{\omega}_{..} Y_{it}} = \frac{n_1 (T-T_0)}{nT} \Big(S_{trt,pre} + S_{trt,test} + S_{ctl,pre} + S_{ctl,test} \Big)$$

Likewise, one can show that $\sum_{i,t} w_{it} \tilde{w}_{it} = \frac{T_0 (T-T_0) n_1 (n-n_1)}{nT}$

It follows that $\hat{\Delta} = did$, as required. ■

× Appendix A.2 To see why the model with all coefs γs is overspecified, note that $Y = Z_\alpha \alpha + Z_\beta \beta + W \gamma + \varepsilon$, where



$$= \begin{pmatrix} | & & | \\ \underline{\omega}_1 & \cdots & \underline{\omega}_T \\ | & & | \end{pmatrix}$$

where $\underline{\omega}_1 + \cdots + \underline{\omega}_T = \begin{pmatrix} | \\ \vdots \\ | \\ 0 \\ \vdots \\ 0 \end{pmatrix}$

$= z_{\alpha_1} + \cdots + z_{\alpha_{n_1}}$

= colums of the n_1 treated units in Z_α

$$\underset{(nT \times n)}{Z_\alpha} = \begin{pmatrix} | & & | \\ \underline{z}_{\alpha_1} & \cdots & \underline{z}_{\alpha_n} \\ | & & | \end{pmatrix}$$

$Q(\omega_1 + \cdots + \omega_T) = Q(z_{\alpha_1} + \cdots + z_{\alpha_{n_1}}) = 0 \Rightarrow W^t Q W$ is not invertible.