Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects

Clement de Chaisemartin and Xavier D’Haultfoeuille(AER, 2020)

JSC Metrics Team 1: Wooyong Park, Joonwoo Shin, Wooin Park

Key Contributions

- shows that two-way fixed effects(TWFE) estimator is a biased estimator of average treatment effects(ATE) when the ATEs are heterogeneous across time and group

- decomposes the TWFE estimator to display the source of such bias - uneven weights.

- develops a framework to assess TWFE's robustness to heterogeneity in the staggered setting.

- constructs an alternative robust estimator($\mathrm{DID}_{\mathrm{M}}$)

- **RMK:** These results apply to any TWFE regressions, not only to those with staggered adoption. ▸ Detail
    - In the survey of the AER papers estimating TWFE regressions, less than 10 percent have a staggered adoption design.

DiD and TWFE revisited

Not all policies are carried out in a randomized fashion.
e.g. Did the unemployment rate really increase due to the minimum wage increase?

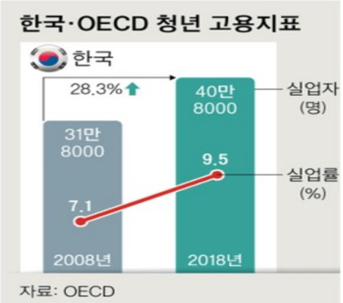


Figure 1: Example of First Difference in a Korean Article

DiD and TWFE revisited

Did the unemployment rate really increase due to the minimum wage increase?

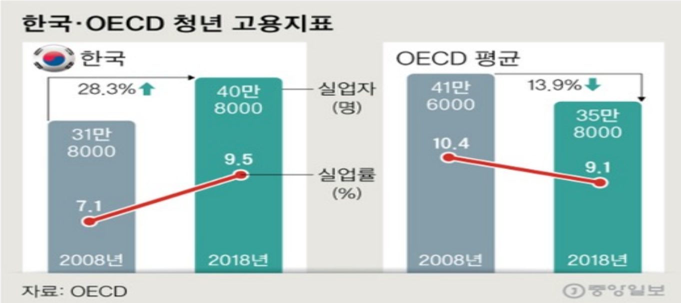


Figure 2: Example of DiD in a Korean Article

DiD and TWFE revisited

Key Assumptions

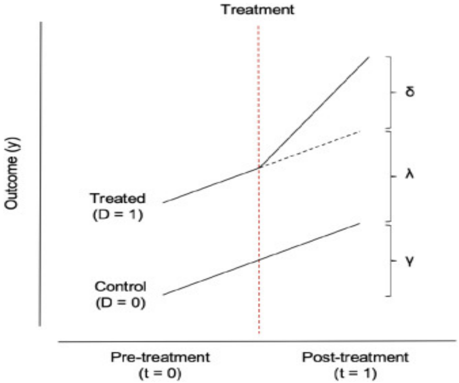
1. **Common Trends:** Without the policy, the potential first difference of the treated and control groups would be equal.

$$\begin{aligned} & E[Y_{i,2}(0)|D_i = 1] - [Y_{i,1}(0)|D_i = 1] \\ = \; & E[Y_{i,2}(0)|D_i = 0] - [Y_{i,1}(0)|D_i = 0] \end{aligned}$$

2. **No Anticipatory Effect:** Individuals do not modify their outcomes before the treatment by anticipating the policy.

$$E[Y_{i,t}(1)|D_i = 1] - [Y_{i,t}(0)|D_i = 1] \quad \text{for all } t < 2$$

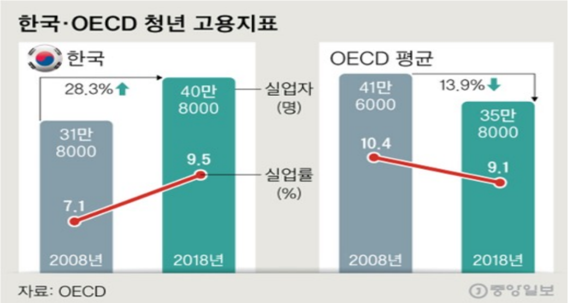
3. No compositional changes, No survival bias, etc.



Under the assumptions in the figure, $\delta = E[Y_{i,2}(1) - Y_{i,1}(0)|D_i = 1]_{=ATT}$

DiD and TWFE revisited

How can we transform such an illustration into a regression?



자료: OECD

$$
\begin{aligned}
\widehat{ATE} &= \text{Treated Difference} - \text{Compar. Difference} \\
&= (9.5 - 7.1) - (10.4 - 9.1)
\end{aligned}
$$

We want a regression setup with a coefficient β capturing the ATE.

DiD and TWFE revisited

Indicator variables would do!

	Pre	Post
Treated	x	o
Comparison	x	x

Table 1: Treatment Status

	Pre	Post
Treated	a	b
Comparison	c	d

Table 2: Y values

β has to capture (b-a) - (d-c),
and the following regression with indicator variables would do that.

$$Y_{i,g,t} = \alpha + \beta \cdot \mathbb{I}(t = \mathit{Post})\mathbb{I}(g = \mathit{Treated}) + \gamma \cdot \mathbb{I}(t = \mathit{Post}) + \delta \cdot \mathbb{I}(g = \mathit{Treated}) + \varepsilon_{i,g,t}$$

DiD and TWFE revisited

This is equivalent in terms of β to TWFE regression, a regression with both group and time fixed effects.

$$Y_{i,g,t} = \alpha + \beta \cdot \mathbb{I}(t = \mathit{Post})\mathbb{I}(g = \mathit{Treated}) + \gamma \cdot \mathbb{I}(t = \mathit{Post}) + \delta \cdot \mathbb{I}(g = \mathit{Treated}) + \varepsilon_{i,g,t}$$

is equivalent to

$$Y_{i,g,t} = \beta \cdot D_{i,g,t} + \gamma_t + \delta_g + \varepsilon_{i,g,t}$$

$D_{i,g,t}$: the treatment status of individual i in group g at time t

Staggered Diff-in-Diffs

Staggered treatment is a policy design where different groups have different implementation/dropout dates.

◂ Key Contributions

e.g. State/County-level laws

	t_1	t_2	t_3	t_4	t_5
g_1	X	O	O	O	X
g_2	X	X	O	O	O
g_3	X	X	X	X	O
g_4	X	X	X	X	X
g_5	X	X	X	X	O

Table 3: Generic Stagg. Treatment

	t_1	t_2	t_3	t_4	t_5
g_1	O	O	O	O	O
g_2	X	O	O	O	O
g_3	X	X	O	O	O
g_4	X	X	X	O	O
g_5	X	X	X	X	O
g_∞	X	X	X	X	X

Table 4: Staggered **Adoption**

Staggered Diff-in-Diffs

Would the TWFE still survive in staggered treatment designs?

	t_1	t_2	t_3	t_4	t_5
g_1	x	o	o	o	x
g_2	x	x	o	o	o
g_3	x	x	x	x	o
g_4	x	x	x	x	x
g_5	x	x	x	x	o

Table 3: Generic Stagg. Treatment

	t_1	t_2	t_3	t_4	t_5
g_1	o	o	o	o	o
g_2	x	o	o	o	o
g_3	x	x	o	o	o
g_4	x	x	x	o	o
g_5	x	x	x	x	o
g_∞	x	x	x	x	x

Table 4: Staggered **Adoption**

Staggered Diff-in-Diffs

Would the TWFE still survive in staggered treatment designs?

Not if ATEs are heterogenous across group and time!

	t_1	t_2	t_3	t_4	t_5
g_1	x	o	o	o	x
g_2	x	x	o	o	o
g_3	x	x	x	x	o
g_4	x	x	x	x	x
g_5	x	x	x	x	o

Table 3: Generic Stagg. Treatment

	t_1	t_2	t_3	t_4	t_5
g_1	o	o	o	o	o
g_2	x	o	o	o	o
g_3	x	x	o	o	o
g_4	x	x	x	o	o
g_5	x	x	x	x	o
g_∞	x	x	x	x	x

Table 4: Staggered **Adoption**

Introduction

- A popular method to estimate ATE is DID, and is implemented by estimating regressions that control for group and time fixed effects(TWFE).
    - 19% of all empirical articles in AER between 2010 and 2012 have used TWFE to estimate the effect of a treatment on an outcome.

- When the treatment effect is constant across groups and over time, such regressions estimate the ATE under the “common trends” assumption.

Introduction

TWFE Specification:

$$Y_{i,g,t} = \beta_{fe} \cdot D_{i,g,t} + \gamma_t + \delta_g + \varepsilon_{i,g,t}$$

Under the common trends,

$$\beta_{fe} = E\left(\sum_{(g,t):D_{g,t}=1} W_{g,t} \Delta_{g,t} \right)$$

$\Delta_{g,t}$: the ATE for each treated cells (g,t)
$W_{g,t} = \frac{N_{g,t}}{N_1} w_{g,t}, \quad \sum_{D_{g,t}=1} W_{g,t} = 1$
Ideally, $w_{g,t} = 1$.

Setup

Notation

- Group and Time: $(g, t) \in \{1, ..., G\} \times \{1, ..., T\})$
- Number of obs.: $N = \sum_{g,t} N_{g,t}$
- Treatment status: $D_{igt} (=^{\text{sharp}} D_{g,t})$
- Potential outcome: $Y_{igt}(D_{igt}) \in \{Y_{igt}(0), Y_{igt}(1)\}$

Setup

Notation for (g,t) Aggregation

- $D_{g,t} = \frac{1}{N_{g,t}} \sum_{i=1}^{N_{g,t}} D_{i,g,t} \in \{0,1\}$ if sharply designed (i.e., all individuals in the group shares the same treatment status)
- $Y_{g,t}(1) = \frac{1}{N_{g,t}} \sum_{i=1}^{N_{g,t}} Y_{i,g,t}(1)$ (Similar with $Y_{g,t}(0)$)

Notation for Further Aggregation

For any variable $Z_{i,g,t}$,

- $Z_{g,.} = \frac{1}{N_{g,.}} \sum_{t=1}^{T} N_{g,t} Z_{g,t}$
- $Z_{.,t} = \frac{1}{N_{.,t}} \sum_{g=1}^{G} N_{g,t} Z_{g,t}$
- $Z_{.,.} = \frac{1}{N} \sum_{g,t} N_{g,t} Z_{g,t}$

Setup

Assumption 1 (Balanced Panel of Groups)

For all $(g, t) \in \{1, \ldots, G\} \times \{1, \ldots, T\}, N_{g,t} > 0$

Assumption 2 (Sharp Design)

For all $(g, t) \in \{1, \ldots, G\} \times \{1, \ldots, T\}$ and $i \in \{1, 2, ... \ldots N_{g,t}\}$, $\mathrm{D}_{i,g,t} = D_{g,t}$

Setup

Assumption 3 (Independent Groups) → i.i.d. condition

The vectors $(Y_{g,t}(0), Y_{g,t}(1), D_{g,t})$ are mutually independent.

Assumption 4 (Strong Exogeneity)

For all $(g, t) \in \{1, \ldots, G\} \times \{1, \ldots, T\}$,

$$E(Y_{g,t}(0) - Y_{g,t-1}(0) | D_{g,1}, D_{g,2}, D_{g,3}, \ldots, D_{g,T})$$

$$= E(Y_{g,t}(0) - Y_{g,t-1}(0))$$

Assumption 5 (Common Trends)

For $t \geq 2$, $E(Y_{g,t}(0) - Y_{g,t-1}(0))$ does not vary across groups.

Two-way Fixed Effects Regressions

1. The Estimand: ATT

How do we define the ATT (average treatment effect of the treated group)?

- We define Δ^{TR}, which is the weighted average of ATE from the treated sample.
- ATT is (δ^{TR}), the expected value of Δ^{TR}.

$$\Delta^{TR} = \frac{1}{N_1} \sum_{(i,g,t):D_{g,t}=1} [Y_{i,g,t}(1) - Y_{i,g,t}(0)]$$

$$\delta^{TR} = E(\Delta^{TR})$$

Two-way Fixed Effects Regressions

Define the ATE of a cell (g, t) to be

$$\Delta_{g,t} = \frac{1}{N_{g,t}} \sum_{i=1}^{N_{g,t}} [Y_{i,g,t}(1) - Y_{i,g,t}(0)]$$

Then δ^{TR} is equal to the expectation of a weighted average of the treated cells' $\Delta_{g,t}$.

$$\delta^{TR} = E\left[\sum_{g,t:D_{g,t}=1} \frac{N_{g,t}}{N_1} \Delta_{g,t}\right] \tag{2}$$

Two-way Fixed Effects Regressions

2. The Estimator: TWFE

Main Regression:

$$Y_{i,g,t} = \beta \cdot D_{g,t} + \gamma_g + \lambda_t + \upsilon_{i,g,t}$$

Regression of the treatment status on TWFEs:

$$D_{g,t} = \alpha + \gamma_g + \lambda_t + \varepsilon_{g,t}$$

Construct the weights.

$$w_{g,t} = \frac{\varepsilon_{g,t}}{\sum_{(g,t):D_{g,t}=1} \frac{N_{g,t}}{N_1} \varepsilon_{g,t}}$$

Two-way Fixed Effects Regressions

2. The Estimator: TWFE

Main Regression:

$$Y_{i,g,t} = \beta \cdot D_{g,t} + \gamma_g + \lambda_t + \upsilon_{i,g,t}$$

Regression of the treatment status on TWFEs:

$$D_{g,t} = \alpha + \gamma_g + \lambda_t + \varepsilon_{g,t}$$

Construct the weights.

$$w_{g,t} = \frac{\varepsilon_{g,t}}{\sum_{(g,t):D_{g,t}=1} \frac{N_{g,t}}{N_1} \varepsilon_{g,t}}$$

Two-way Fixed Effects Regressions

2. The Estimator: TWFE

Main Regression:

$$Y_{i,g,t} = \beta \cdot D_{g,t} + \gamma_g + \lambda_t + \upsilon_{i,g,t}$$

Regression of the treatment status on TWFEs:

$$D_{g,t} = \alpha + \gamma_g + \lambda_t + \varepsilon_{g,t}$$

Construct the weights.

$$w_{g,t} = \frac{\varepsilon_{g,t}}{\sum_{(g,t):D_{g,t}=1} \frac{N_{g,t}}{N_1} \varepsilon_{g,t}}$$

Two-way Fixed Effects Regressions

Theorem 1

▸ proof

Suppose that Assumptions 1∼5 hold. Then, by the Frisch-Waugh Theorem,

$$\beta_{fe} = E\left[\sum_{g,t:D_{g,t}=1} \frac{N_{g,t}}{N_1} w_{g,t}\Delta_{g,t}\right]$$

Compare this to the estimand(ATT):

$$\delta^{TR} = E\left[\sum_{g,t:D_{g,t}=1} \frac{N_{g,t}}{N_1} \Delta_{g,t}\right]$$

Two-way Fixed Effects Regressions

Theorem 1 ▸ proof

Suppose that Assumptions 1∼5 hold. Then, by the Frisch-Waugh Theorem,

$$\beta_{fe} = E\left[\sum_{g,t:D_{g,t}=1} \frac{N_{g,t}}{N_1} w_{g,t} \Delta_{g,t}\right]$$

Compare this to the estimand(ATT):

$$\delta^{TR} = E\left[\sum_{g,t:D_{g,t}=1} \frac{N_{g,t}}{N_1} \Delta_{g,t}\right]$$

Robustness to Heterogeneous Effects

By Theorem 1,

$$\beta_{fe} = E\left[\sum_{(g,t):D_{g,t}=1} \frac{N_{g,t}}{N_1} w_{g,t}\Delta_{g,t}\right]$$

But we have

$$\delta^{TR} = E\left[\sum_{(g,t):D_{g,t}=1} \frac{N_{g,t}}{N_1} \Delta_{g,t}\right]$$

Hence it is clear the bias of β_{fe} depends on the weights $w_{g,t}$

Example

Simple Staggered Adoption Design

- Two groups : g=1,2, and three periods : t= 1, 2, 3.

- Treatment according to following table:

	t_1	t_2	t_3
g_1	x	x	o
g_2	x	o	o

Table 5: Staggered Adoption Design

- $N_{g,t}$ is constant on t.

Example

In this simple example we get

$$\beta_{fe} = \frac{1}{2}E[\Delta_{1,3}] + E[\Delta_{2,2}] - \frac{1}{2}E[\Delta_{2,3}].$$

Suppose, for instance,

$$E[\Delta_{1,3}] = E[\Delta_{2,2}] = 1, \; E[\Delta_{2,3}] = 4.$$

Then we have $\beta_{fe} = -\frac{1}{2} < 0$, even though the treatment effect in each groups are positive!

Example

Two main observations from this simple example is

- (i) How weights are distributed affect bias of TWFE estimator

- (ii) If weights are structurally correlated to ATE's they are being attached to, we may worry about bias.

Sufficient Condition

Corollary 2

Let

$$\widetilde{\Delta_{g,t}} = E(\Delta_{g,t}|\mathbf{D}),\ \widetilde{\Delta^{TR}} = E(\Delta^{TR}|\mathbf{D}),\ \widetilde{\beta_{fe}} = E(\widehat{\beta_{fe}}|\mathbf{D})$$

If assumptions 1 to 5 hold and

$$E\left[\sum_{(g,t):D_{g,t}=1} \frac{N_{g,t}}{N_1}(w_{g,t}-1)(\widetilde{\Delta_{g,t}} - \widetilde{\Delta^{TR}})\right] = 0$$

then $\beta_{fe} = \delta^{TR}$

Robust Measure

We denote

$$\sigma(\tilde{\Delta}) = \left(\sum_{(g,t):D_{g,t}=1} \frac{N_{g,t}}{N_1} (\widetilde{\Delta_{g,t}} - \widetilde{\Delta^{TR}})^2 \right)^{\frac{1}{2}}$$

$$\sigma(\mathbf{w}) = \left(\sum_{(g,t):D_{g,t}=1} \frac{N_{g,t}}{N_1} (w_{g,t} - 1)^2 \right)^{\frac{1}{2}}$$

standard deviation of conditional ATE’s, and **w**-weights, respectively.

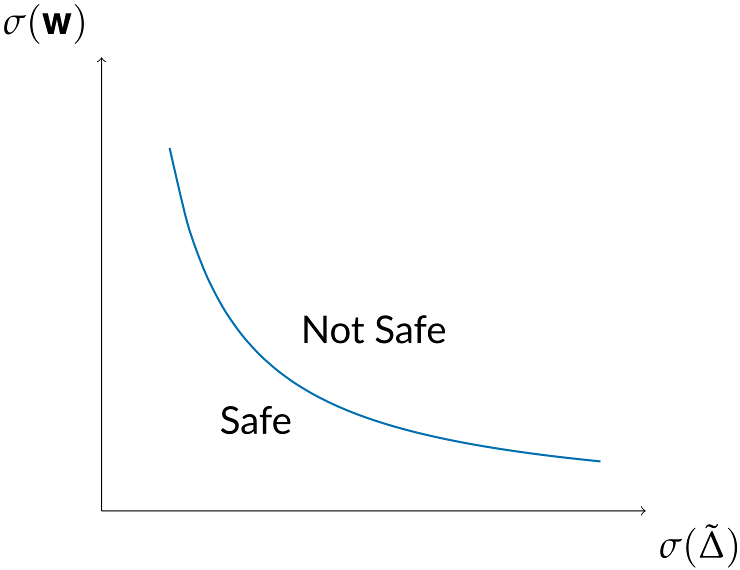
Robust Measure

Corollary 1

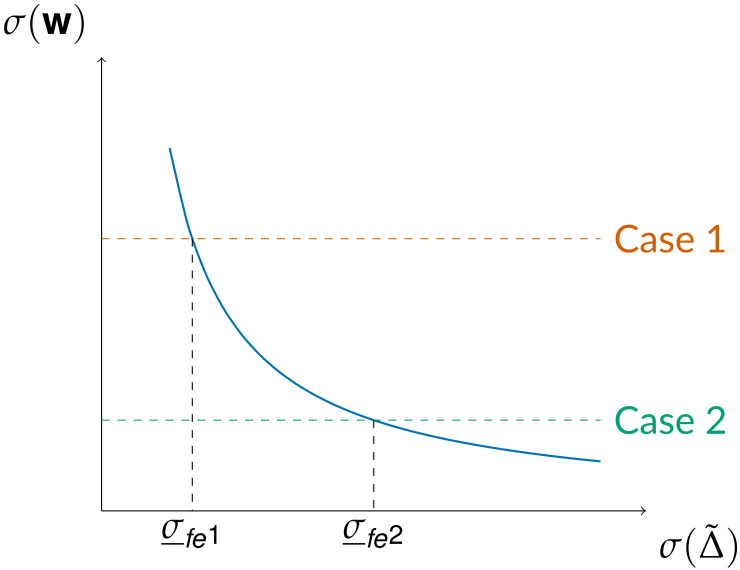
(i) If $\sigma(\mathbf{w}) > 0$, the minimal value of $\sigma(\widetilde{\Delta})$ compatible with $\widetilde{\beta_{fe}}$ and $\widetilde{\Delta^{TR}} = 0$ is

$$\underline{\sigma_{fe}} = \frac{|\widetilde{\beta_{fe}}|}{\sigma(\mathbf{w})}$$

Robust Measure



Robust Measure



An Alternative Estimand

- Under the situation where treatment effects are heterogeneous across groups or over time, Let

$$\delta^s = E\left[\frac{1}{N_s}\sum_{(i,g,t):t\geq 2, D_{g,t}\neq D_{g,t-1}}[Y_{i,g,t}(1) - Y_{i,g,t}(0)]\right]$$

  - with $N_s = \sum_{(g,t):t\geq 2, D_{g,t}\neq D_{g,t-1}} N_{g,t}$
  - The term δ^s is the ATE of all switching cells.

- We now show that δ^s can be unbiasedly estimated by a weighted average of DID estimators under the following assumptions.

Assumption 9, 10

- Assumption 9 (Strong Exogeneity for $Y(1)$)

$$^{\forall}(g, t) \in \{1, \dots, G\} \times \{2, \dots, T\},$$

$$\mathrm{E}\left(Y_{g,t}(1) - Y_{g,t-1}(1) \middle| D_{g,1}, \dots, D_{g,T}\right) = E\left(Y_{g,t}(1) - Y_{g,t-1}(1)\right)$$

- Assumption 10 (Common Trends for Y(1))
  $^{\forall}t \geq 2, E(Y_{g,t}(1) - Y_{g,t-1}(1))$ does not vary across g.

Assumption 11

- Assumption 11 (Existence of “Stable” Groups)

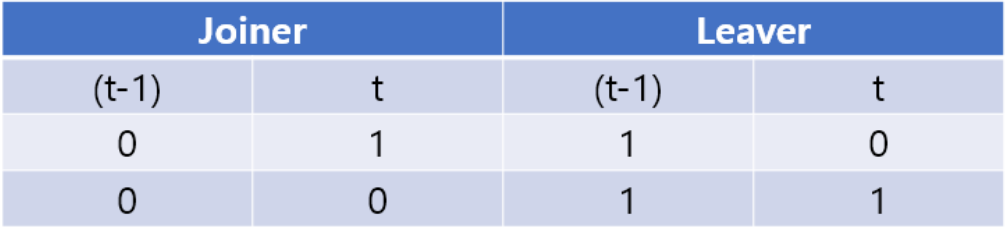
  $^{\forall}t \in \{2, \ldots, T\}$,

  (1) **Joiner**: If there is at least one $g \in \{1, \ldots, G\}$ such that $D_{g,t-1} = 0$, $D_{g,t} = 1$, then there exists at least one $g^{'} \neq g, g^{'} \in \{1, \ldots, G\}$ such that $D_{g^{'},t-1} = D_{g^{'},t} = 0$

  (2) **Leaver**: If there is at least one $g \in \{1, \ldots, G\}$ such that $D_{g,t-1} = 1$, $D_{g,t} = 0$, then there exists at least one $g^{'} \neq g, g^{'} \in \{1, \ldots, G\}$ such that $D_{g^{'},t-1} = D_{g^{'},t} = 1$

Assumption 11

- Assumption 11 (Existence of “Stable” Groups)



Joiner		Leaver	
(t-1)	t	(t-1)	t
0	1	1	0
0	0	1	1

Assumption 12

- Assumption 12 (Mean Independence between a Group's Outcome and Other Groups Treatments)

  $^{\forall}g$ and t ,

  $$E(Y_{g,t}(0) \mid \mathbf{D}) = E(Y_{g,t}(0) \mid \mathbf{D}_g) \quad \text{and} \quad E(Y_{g,t}(1) \mid \mathbf{D}) = E(Y_{g,t}(1) \mid \mathbf{D}_g)$$

An Alternative Estimator

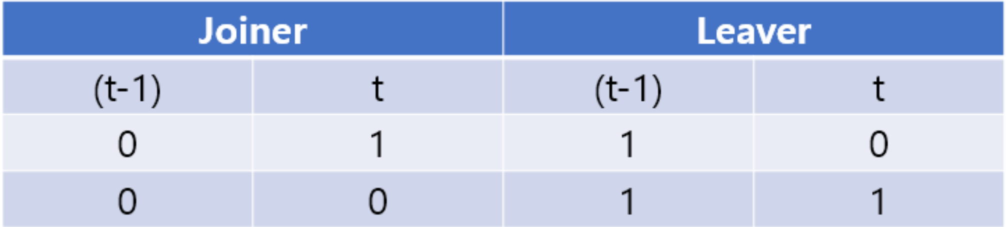
- For all $t \in \{2, \ldots, T\}$ and for all $(d, d^{'}) \in \{0, 1\}^2$,
- Let

$$N_{d,d',t} = \sum_{g:D_{g,t}=d, D_{g,t-1}=d'} N_{g,t}$$

[joiner] $\text{DiD}_{+,t} = \sum_{g:D_{g,t}=1,D_{g,t-1}=0} \frac{N_{g,t}}{N_{1,0,t}}(Y_{g,t} - Y_{g,t-1}) - \sum_{g:D_{g,t}=D_{g,t-1}=0} \frac{N_{g,t}}{N_{0,0,t}}(Y_{g,t} - Y_{g,t-1})$

[Leaver] $\text{DiD}_{-,t} = \sum_{g:D_{g,t}=1,D_{g,t-1}=0} \frac{N_{g,t}}{N_{1,1,t}}(Y_{g,t} - Y_{g,t-1}) - \sum_{g:D_{g,t}=0,D_{g,t-1}=1} \frac{N_{g,t}}{N_{0,1,t}}(Y_{g,t} - Y_{g,t-1})$

An Alternative Estimator



Joiner		Leaver	
(t-1)	t	(t-1)	t
0	1	1	0
0	0	1	1

[joiner] $\mathrm{DiD}_{+,t} = \sum_{g:D_{g,t}=1,D_{g,t-1}=0} \frac{N_{g,t}}{N_{1,0,t}}(Y_{g,t} - Y_{g,t-1}) - \sum_{g:D_{g,t}=D_{g,t-1}=0} \frac{N_{g,t}}{N_{0,0,t}}(Y_{g,t} - Y_{g,t-1})$

[Leaver] $\mathrm{DiD}_{-,t} = \sum_{g:D_{g,t}=1,D_{g,t-1}=0} \frac{N_{g,t}}{N_{1,1,t}}(Y_{g,t} - Y_{g,t-1}) - \sum_{g:D_{g,t}=0,D_{g,t-1}=1} \frac{N_{g,t}}{N_{0,1,t}}(Y_{g,t} - Y_{g,t-1})$

$E[\text{DiD}_M] = \delta^s$

- If Assumptions 1, 2, 3, 4, 5, and 9-12 hold, then $E[\text{DiD}_M] = \delta^s$

$$\text{DiD}_M = \sum_{t=2}^{T} \left(\frac{N_{1,0,t}}{N_s} \text{DiD}_{+,t} + \frac{N_{0,1,t}}{N_s} \text{DiD}_{-,t} \right)$$

- weighted sum of **Joiners'** treatment effect & **Leavers'** treatment effect
- computed by the following Stata packages: *fuzzydid*, *did multiplegt*

Limitation of Our Alternative Estimator

- (1) Homogeneous treatment effect:

$$\text{Var}(\hat{\beta_{fe}}) << Var(DiD_M)$$

- (2) Heterogeneous treatment effect:

$$\text{Var}(\hat{\beta_{fe}}) < Var(DiD_M)$$

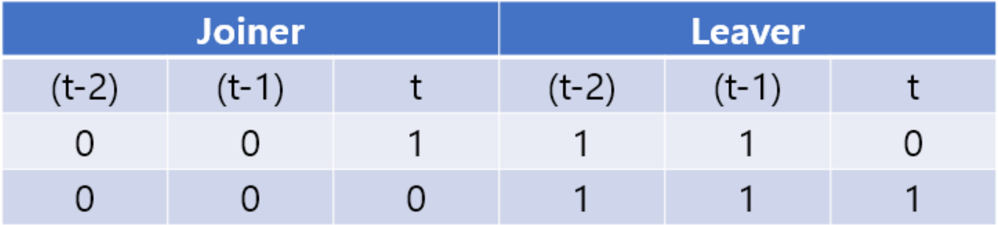
Assumption 13 (Existence of “Stable” Groups for the Placebo Test)

- $\forall t \in \{3, \ldots, T\}$,

  (1) **Joiner**: If there is at least one $g \in \{1, \ldots, G\}$ such that $D_{g,t-2} = D_{g,t-1} = 0$ and $D_{g,t} = 1$, then there exists at least one $g^{'} \neq g, g^{'} \in \{1, \ldots, G\}$ such that $D_{g^{'},t-2} = D_{g^{'},t-1} = D_{g^{'},t} = 0$

  (2) **Leaver**: If there is at least one $g \in \{1, \ldots, G\}$ such that $D_{g,t-2} = D_{g,t-1} = 1$, $D_{g,t} = 0$, then there exists at least one $g^{'} \neq g, g^{'} \in \{1, \ldots, G\}$ such that $D_{g^{'},t-2} = D_{g^{'},t-1} = D_{g^{'},t} = 1$

Assumption 13 (Existence of “Stable” Groups for the Placebo Test)



Joiner			Leaver		
(t-2)	(t-1)	t	(t-2)	(t-1)	t
0	0	1	1	1	0
0	0	0	1	1	1

Assumption 13 - Placebo Test

- $^\forall t \in \{2, \ldots, T\}$ and $\forall (d, d', d'') \in \{0,1\}^3$, let

$$N_{d,d',d'',t} = \sum_{g: D_{g,t}=d,\, D_{g,t-1}=d',\, D_{g,t-2}=d''} N_{g,t}$$

- *d'': the number of obs with treatment status at period $t-2$, d' at period $t-1$, and d at period t.*

- Let

$$N_s^{pl} = \sum_{(g,t): t \geq 3,\, D_{g,t} \neq D_{g,t-1} = D_{g,t-2}} N_{g,t},$$

Assumption 13 - Placebo Test

$$\begin{aligned}\text{DiD}^{pl}_{+,t} = & \sum_{g:D_{g,t}=1,D_{g,t-1}=D_{g,t-2}=0} \frac{N_{g,t}}{N_{1,0,0,t}}(Y_{g,t-1} - Y_{g,t-2}) \\ & - \sum_{g:D_{g,t}=D_{g,t-1}=D_{g,t-2}=0} \frac{N_{g,t}}{N_{0,0,0,t}}(Y_{g,t-1} - Y_{g,t-2})\end{aligned}$$

$$\begin{aligned}\text{DiD}^{pl}_{-,t} = & \sum_{g:D_{g,t}=D_{g,t-1}=D_{g,t-2}=1} \frac{N_{g,t}}{N_{1,1,1,t}}(Y_{g,t-1} - Y_{g,t-2}) \\ & - \sum_{g:D_{g,t}=0,D_{g,t-1}=D_{g,t-2}=1} \frac{N_{g,t}}{N_{0,1,1,t}}(Y_{g,t-1} - Y_{g,t-2})\end{aligned}$$

$E[\text{DiD}_M^{pl}] = 0$

- If Assumptions 1, 2, 4, 5, 9, 10, 12 and 13 hold then $E[\text{DiD}_M^{pl}] = 0$

$$\text{DiD}_M^{pl} = \sum_{t=3}^{T} \left(\frac{N_{1,0,0,t}}{N_s^{pl}} \text{DiD}_{+,t}^{pl} + \frac{N_{0,1,1,t}}{N_s^{pl}} \text{DiD}_{-,t}^{pl} \right)$$

- $E[\text{DiD}_M^{pl}] = 0$ is a testable implication of Assumptions 4,5,9,10.
- Finding DiD_M^{pl} significantly different from 0 = Those assumptions are violated (experience different trends).
- Another placebo test: Callaway and Sant'Anna (2018) in staggered adoption designs.

Example

C : Union Membership Premium

Data: National Longitudinal Survey(Youth Sample)

Model:

$$Y_{i,g,t} = u_g + v_t + \beta_{fe} D_{i,g,t} + \delta \mathbf{X}_{i,g,t} + \epsilon_{i,g,t}$$

$$Y_{i,g,t} : \text{log(wage)},\ D_{i,g,t} : \text{Union Membership status}$$

Result:

$$\hat{\beta_{fe}} = 0.107(0.030^{***})$$

which is consistent with literature : *Vella and Verbeek(1998), Jakubson(1991).*

Example

- 820 and 196 weights attached to β_{fe} are estimated to be strictly positive and negative, respectively.

- $\underline{\hat{\sigma}_{fe}} = 0.097$
    - The estimate of ATT may have a nonsignificant value regardless of β_{fe}'s significance when the treatment effect across (group)$\times$(time) is equal or larger than 0.097, which is possible to happen.

Example

The data shows that stable groups assumption holds; hence we can calculate DID_M and it is given by

$$DID_M = 0.041(0.034),$$

which is significantly different from $\hat{\beta_{fe}} = 0.107$ (with t-stat = 2.60).

Conclusion

- Regardless of the TWFE’s popularity in the estimation of ATE(20% of AER empirical articles(2010-2012)), there is no reason to assume it will always capture the desired estimand.

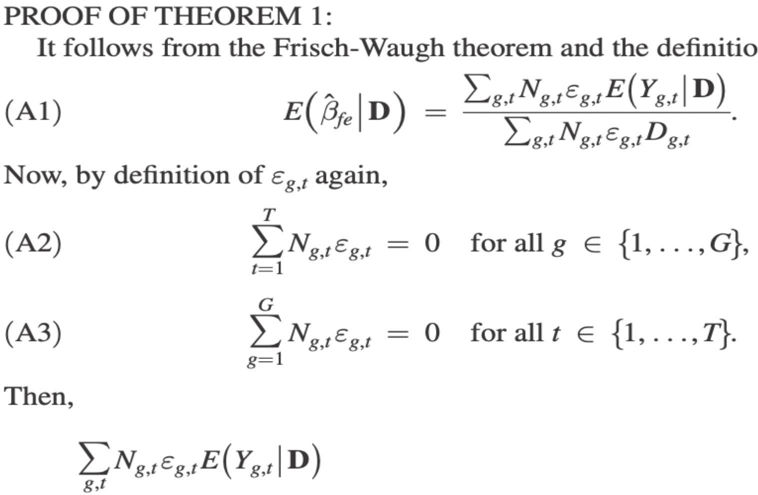
- Under common trends, TWFE estimates the weighted sum of the treatment effect of each group and time, and it could even be negative.

- Such negativity and bias are problematic when the treatment effects are heterogeneous.

- In this paper, we studied (i) why it is the case, (ii) how to check its credibility, and (iii) an alternative estimator whose use is not limited to staggered adoption designs.

Appendix

Proof of Theorem 1



PROOF OF THEOREM 1:

It follows from the Frisch-Waugh theorem and the definition of $\varepsilon_{g,t}$ that

$$E\left(\hat{\beta}_{fe}\middle|\mathbf{D}\right) = \frac{\sum_{g,t} N_{g,t}\varepsilon_{g,t}E\left(Y_{g,t}\middle|\mathbf{D}\right)}{\sum_{g,t} N_{g,t}\varepsilon_{g,t}D_{g,t}}. \tag{A1}$$

Now, by definition of $\varepsilon_{g,t}$ again,

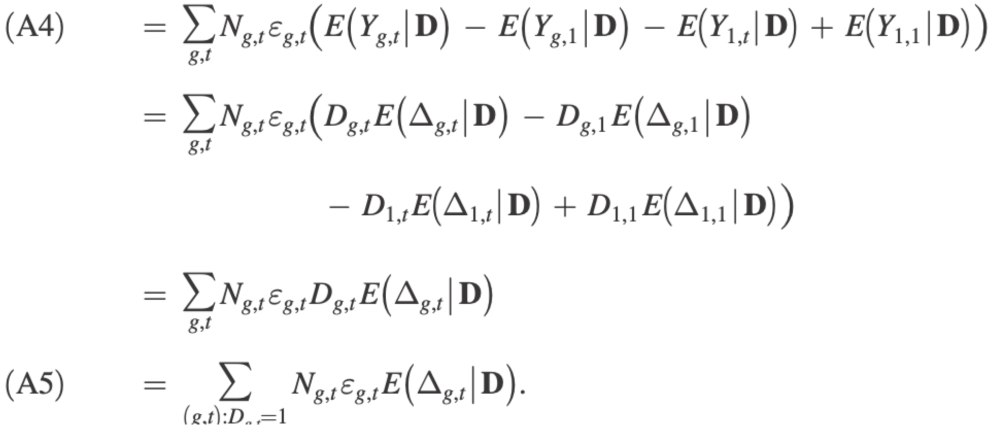
$$\sum_{t=1}^{T} N_{g,t}\varepsilon_{g,t} = 0 \quad \text{for all } g \in \{1,\ldots,G\}, \tag{A2}$$

$$\sum_{g=1}^{G} N_{g,t}\varepsilon_{g,t} = 0 \quad \text{for all } t \in \{1,\ldots,T\}. \tag{A3}$$

Then,

$$\sum_{g,t} N_{g,t}\varepsilon_{g,t}E\left(Y_{g,t}\middle|\mathbf{D}\right)$$

Proof of Theorem 1



$$(A4) \quad = \sum_{g,t} N_{g,t}\varepsilon_{g,t}\Big(E\big(Y_{g,t}\,|\,\mathbf{D}\big) - E\big(Y_{g,1}\,|\,\mathbf{D}\big) - E\big(Y_{1,t}\,|\,\mathbf{D}\big) + E\big(Y_{1,1}\,|\,\mathbf{D}\big)\Big)$$

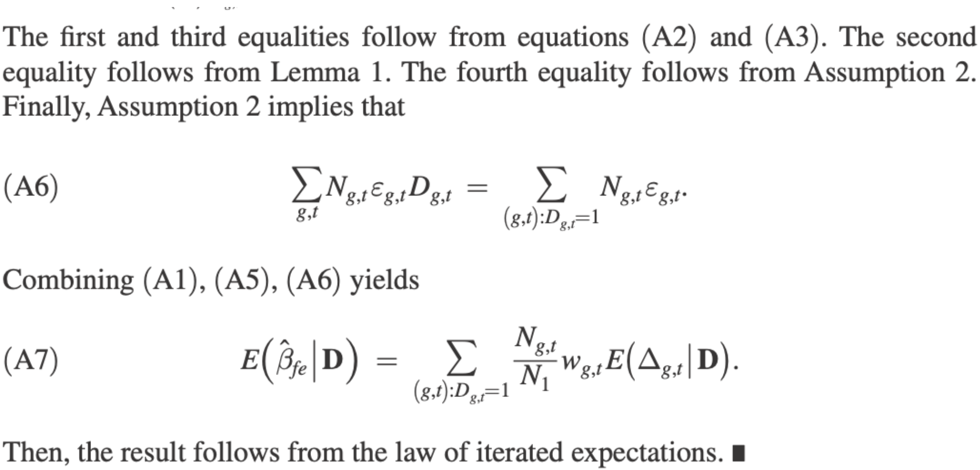
$$= \sum_{g,t} N_{g,t}\varepsilon_{g,t}\Big(D_{g,t}E\big(\Delta_{g,t}\,|\,\mathbf{D}\big) - D_{g,1}E\big(\Delta_{g,1}\,|\,\mathbf{D}\big)$$

$$- D_{1,t}E\big(\Delta_{1,t}\,|\,\mathbf{D}\big) + D_{1,1}E\big(\Delta_{1,1}\,|\,\mathbf{D}\big)\Big)$$

$$= \sum_{g,t} N_{g,t}\varepsilon_{g,t}D_{g,t}E\big(\Delta_{g,t}\,|\,\mathbf{D}\big)$$

$$(A5) \quad = \sum_{(g,t):D_{g,t}=1} N_{g,t}\varepsilon_{g,t}E\big(\Delta_{g,t}\,|\,\mathbf{D}\big).$$

Proof of Theorem 1



The first and third equalities follow from equations (A2) and (A3). The second equality follows from Lemma 1. The fourth equality follows from Assumption 2. Finally, Assumption 2 implies that

$$\sum_{g,t} N_{g,t}\varepsilon_{g,t}D_{g,t} = \sum_{(g,t):D_{g,t}=1} N_{g,t}\varepsilon_{g,t}. \tag{A6}$$

Combining (A1), (A5), (A6) yields

$$E\left(\hat{\beta}_{fe}\middle|\mathbf{D}\right) = \sum_{(g,t):D_{g,t}=1}\frac{N_{g,t}}{N_1}w_{g,t}E\left(\Delta_{g,t}\middle|\mathbf{D}\right). \tag{A7}$$

Then, the result follows from the law of iterated expectations. ■

◂ Theorem 1

Proof of Corollary 1

first point.-If the assumptions hold and $\Delta^{TR} = 0$, then

$$\begin{cases} \widetilde{\beta_{fe}} = \sum_{(g,t):D_{gt}=1} \frac{N_{g,t}}{N_1} w_{g,t} \widetilde{\Delta_{g,t}} \\ 0 = \sum_{(g,t):D_{g,t}=1} \frac{N_{g,t}}{N_1} \widetilde{\Delta_{g,t}} \end{cases} \tag{1}$$

Then Cauchy-Schwartz inequality yields the result :

$$|\widetilde{\beta_{fe}}| = \left| \sum_{(g,t):D_{g,t}=1} \frac{N_{g,t}}{N_1} (w_{g,t} - 1) \left(\widetilde{\Delta_{g,t}} - \widetilde{\Delta^{TR}} \right) \right| \leq \sigma(\mathbf{w}) \sigma(\tilde{\Delta})$$

Proof of Corollary 1

second point.-First we assume $\widetilde{\beta_{fe}} > 0$. We solve following problem.

$$\min_{\Delta \in \mathbb{R}^n} \sum_{i=1}^{n} \frac{N_{(i)}}{N_1} \left(\Delta_{(i)} - \widetilde{\Delta^{TR}} \right)^2 \tag{2}$$

with constraints:

$$\widetilde{\beta_{fe}} = \sum_{i=1}^{n} \frac{N_{(i)}}{N_1} w_{(i)} \Delta_{(i)}, \qquad \Delta_{(i)} \leq 0, \quad \forall i = 1, 2, \cdots, n$$

Proof of Corollary 1

This is quadratic programming with symmetric, positive semi-definite matrix. For the linear term in the quadratic problem is 0, the solution exists if and only if the feasible set is nonempty(*Frank and Wolf*, *1956*). Note that

$$\sum_{i=1}^{n} \frac{N_{(i)}}{N_1} \left(\Delta_{(i)} - \sum_{i=1}^{n} \frac{N_{(i)}}{N_1} \Delta_{(i)} \right)^2 = \sum_{i=1}^{n} \frac{N_{(i)}}{N_1} \Delta_{(i)}^2 - \left(\sum_{i=1}^{n} \frac{N_{(i)}}{N_1} \Delta_{(i)} \right)^2 .$$

Proof of Corollary 1

Karush-Kuhn-Tucker Necessary Conditions are given by

$$\Delta_{(i)} = \Delta^{\tilde{TR}} + \lambda w_{(i)} - \gamma_{(i)},$$

$$\sum_{i=1}^{n} \frac{N_{(i)}}{N_1} w_{(i)} \Delta_{(i)} = \widetilde{\beta_{fe}},$$

$$\gamma_{(i)} \geq 0,$$

$$\gamma_{(i)} \Delta_{(i)} = 0.$$

Proof of Corollary 1

Observe that $\Delta_{(i)} = 0$ if and only if $\Delta^{\tilde{TR}} + \lambda w_{(i)} \geq 0$. Hence if $\Delta^{\tilde{TR}} + \lambda w_{(i)} < 0$, $\Delta_{(i)}$ would be nonzero, which further implies $\gamma_{(i)} = 0$, and $\Delta_{(i)} = \Delta^{\tilde{TR}} + \lambda w_{(i)}$. Therefore we established

$$\Delta_{(i)} = \min\{\tilde{\Delta}^{TR} + \lambda w_{(i)},\ 0\} \tag{*}$$

Above equation implies $\Delta_{(i)} \leq \tilde{\Delta}^{TR} + \lambda w_{(i)}$, whence $\tilde{\Delta} TR \leq \tilde{\Delta}^{TR} + \lambda$. Therefore λ is non-negative. As a consequence, $\tilde{\Delta}^{TR} + \lambda_{(i)}$ is decreasing in i, so is $\Delta_{(i)}$. Then it must be the case $\Delta_{(n)} < 0$, for otherwise $\Delta_{(i)} = 0$ for all i, so $\widetilde{\beta_{fe}} = 0$.

Proof of Corollary 1

Put $s = \min\{i \in \{1, \cdots, n\} | \Delta_{(i)} < 0\}$. Using (*), we get:

$$\tilde{\Delta}^{TR} \sum_i \frac{N_{(i)}}{N_1} \Delta_{(i)} = P_s \tilde{\Delta}^{TR} + \lambda S_s,$$

whence

$$\tilde{\Delta}^{TR} = \frac{\lambda S_s}{1 - P_s}. \tag{**}$$

Proof of Corollary 1

Using (*) we get,

$$\Delta_{(i)} = \lambda\{\frac{S_s}{1-P_s} + w_{(i)}\}.$$

Again by (*),

$$\widetilde{\beta_{fe}} = \sum_{i \geq s} \frac{N_{(i)}}{N_1} w_{(i)} \Delta_{(i)} = \lambda\{\frac{{S_s}^2}{1-P_s} + w_{(i)}\}$$

thus

$$\lambda = \frac{\widetilde{\beta_{fe}}}{T_s + S_s^2/(1-P_s)}.$$

Proof of Corollary 1

Therefore we have :

$$\begin{aligned}
\underline{\underline{\sigma_{fe}^2}} &= \sum_{i \geq s} \frac{N_{(i)}}{N_1} \left(\lambda w_{(i)}\right)^2 + \sum_{i < s} \frac{N_{(i)}}{N_1} \left(\Delta^{\tilde{T}R}\right)^2 \\
&= \lambda^2 T_s + (1 - P_s) \left(\frac{\lambda S_s}{1 - P_s}\right)^2 \\
&= \lambda^2 \left[T_s + \frac{S_s^2}{1 - P_s}\right] \\
&= \frac{\widetilde{\beta_{fe}^2}}{T_s + S_s^2 / (1 - P_s)}.
\end{aligned}$$

Proof of Corollary 1

The result falls out immediately, for (*) and (**) imply that $s = \min\{i \in \{1, \ldots, n\} : w_{(i)} < -S_{(i)}/(1 - P_{(i)})\}$.

For the case $\widetilde{\beta_{fe}} < 0$, put $\Delta'_{(i)} = -\Delta_{(i)}$ and $\widetilde{\beta'_{fe}} = -\widetilde{\beta_{fe}}$. Then we get

$$\underline{\underline{\sigma^2_{fe}}} = \min_{\Delta'_{(1)} \leq 0, \ldots, \Delta'_{(i)} \leq 0} \sum_{i=1}^{n} \frac{N_{(i)}}{N_1} {\Delta'_{(i)}}^2 - \left(\sum_{i=1}^{n} \frac{N_{(i)}}{N_1} \Delta'_{(i)} \right)^2 .$$

subject to

$$\widetilde{\beta'_{fe}} = \sum_{i=1}^{n} \frac{N_{(i)}}{N_1} w_{(i)} \Delta'_{(i)}$$

Proof of Corollary 1

This is nothing but what we have done so far. Therefore we obtain

$$\underline{\underline{\sigma_{fe}^2}} = \frac{\widetilde{\beta_{fe}'^2}}{T_s + S_s^2/(1 - P_s)} = \frac{\widetilde{\beta_{fe}^2}}{T_s + S_s^2/(1 - P_s)}.$$

This completes the proof. □

Proof of Corollary 2

We have

$$\begin{aligned}
\beta_{fe} &= E\left(\sum_{(g,t):D_{g,t}=1} \frac{N_{g,t}}{N_1} w_{g,t} \tilde{\Delta_{g,t}}\right) \\
&= E\left(\left(\sum_{(g,t):D_{g,t}=1} \frac{N_{g,t}}{N_1} w_{g,t}\right) \widetilde{\Delta^{TR}}\right) \\
&= E(\widetilde{\Delta^{TR}}) \\
&= \delta^{TR}.
\end{aligned}$$

Proof of Corollary 2

First equality is a consequence of law of iterated expectations and the fact

$$E(\hat{\beta_{fe}}|\mathbf{D}) = \sum_{(g,t):D_{g,t}=1} \frac{N_{g,t}}{N_1} w_{g,t} E(\Delta_{g,t}|\mathbf{D}),$$

which was demonstrated in the proof of Theorem 1. The second equality follows from Assumption 7. By definition of $w_{g,t}$, we have $\sum_{(g,t):D_{g,t}=1}(N_{g,t}/N_1)w_{g,t} = 1$, which implies the third equality. Last step is then obtained by law of iterated expectations. $\square$