Two-Way Fixed Effects, the Two-Way Mundlak Regression, and Difference-in-Differences Estimators

Jeff Wooldridge
Department of Economics
Michigan State University

September 23, 2021

1. Introduction
2. Equivalence of TWFE and Two-Way Mundlak
3. Interventions with Common Treatment Timing
4. Staggered Interventions
5. Comparison with Other Approaches
6. Testing and Relaxing Parallel Trends
7. Simulations
8. Concluding Remarks

1. Introduction

- For panel data, two-way fixed effects (TWFE) is a staple in empirical research.
  - Applied to "structural" models – say, production functions.
  - Applied to policy analysis – difference-in-differences.
- Used for all configurations of N and T.
- With small T, large N, time effects often absorbed into covariates.
  - Analyze as a one-way FE estimator.

- One-way Mundlak regression has proven useful for many purposes.
  - ▸ Leads to simple, robust, regression-based comparisons between FE and random effects estimation: Arellano (1993).
  - ▸ Produces insight into the pre-testing problem with Hausman tests.
  - ▸ Suggests how to allow heterogeneity to correlate with covariates in nonlinear models: Mundlak-Chamberlain device.

- Wooldridge (2019): The one-way Mundlak regression applies to unbalanced panels.
  - In the linear case, Mundlak still produces the complete-cases FE estimator.
  - Suggests correlated random effects for heterogeneous slopes and nonlinear models.

- Current paper: Shows the equivalence between the TWFE estimator and the obvious two-way Mundlak regression.
  - In latter case, focus is on pooled OLS, but results also hold for RE.
- Equivalence is simple but useful.
  - Further reveals the workings of TWFE.
  - Applications to staggered interventions and DiD.

- Advantages of TWFE for event studies:

1. We know properties of TWFE when the panel is unbalanced.
2. It is easy to test the null that treatment effects are homogeneous in a robust way.
3. Immediately extensions of the TWFE etimator to removing unit-specific trends can be applied with heterogenous treatment effects.

- Advantages of POLS for event studies:
  - Given equivalence in the linear case, POLS can be extended to nonlinear models.

2. Equivalence of TWFE and Two-Way Mundlak

- Motivation for TWFE estimation:

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + c_i + f_t + u_{it},\ \ t = 1, \ldots, T; i = 1, \ldots, N$$

  - $\mathbf{x}_{it}$ is $1 \times K$.
  - c_i are the unit-specific effects.
  - f_t are the time-specific effects.

- Equivalence results are algebraic.
- The two-way dummy variable regression:

$$y_{it} \text{ on } \mathbf{x}_{it}, 1, c2_i, ..., cN_i, f2_t, ..., fT_t, \ \ t = 1, \ldots, T; \ i = 1, \ldots, N.$$

  - Coefficients on $\mathbf{x}_{it}$ are $\hat{\boldsymbol{\beta}}_{FE}$ ($K \times 1$).
- $\mathbf{x}_{it}$ only includes variables that have some variation across i and t.

- Baltagi (2001): Two-way within transformation gives $\hat{\boldsymbol{\beta}}_{FE}$.

$$\bar{\mathbf{x}}_{i\cdot} = T^{-1} \sum_{t=1}^{T} \mathbf{x}_{it}$$

$$\bar{\mathbf{x}}_{\cdot t} = N^{-1} \sum_{i=1}^{N} \mathbf{x}_{it}$$

$$\bar{\mathbf{x}} = (NT)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} \mathbf{x}_{it} = N^{-1} \sum_{i=1}^{N} \bar{\mathbf{x}}_{i\cdot} = T^{-1} \sum_{t=1}^{T} \bar{\mathbf{x}}_{\cdot t}$$

$$\ddot{\mathbf{x}}_{it} = (\mathbf{x}_{it} - \bar{\mathbf{x}}_{i\cdot}) - N^{-1}\sum_{i=1}^{N}(\mathbf{x}_{it} - \bar{\mathbf{x}}_{i\cdot}) = \mathbf{x}_{it} - \bar{\mathbf{x}}_{i\cdot} - \bar{\mathbf{x}}_{\cdot t} + \bar{\mathbf{x}}$$

$$\ddot{y}_{it} = y_{it} - \bar{y}_i - \bar{y}_t + \bar{y}$$

- $\hat{\boldsymbol{\beta}}_{FE}$ is also the pooled OLS estimator from

$$\ddot{y}_{it} \text{ on } \ddot{\mathbf{x}}_{it},\ t = 1,\ldots,T;\ i = 1,\ldots,N.$$

- Alternatively, consider the *two-way Mundlak regression*.
- Pooled OLS of

$$y_{it} \text{ on } 1, \mathbf{x}_{it}, \bar{\mathbf{x}}_{i\cdot}, \bar{\mathbf{x}}_{\cdot t},\ t = 1,\ldots,T; i = 1,\ldots,N.$$

  - ▸ Let $\hat{\boldsymbol{\beta}}_M$ be the coefficients $\mathbf{x}_{it}$.

THEOREM: Provided the $K \times K$ matrix

$$\sum_{i=1}^{N} \sum_{t=1}^{T} \ddot{\mathbf{x}}_{it}' \ddot{\mathbf{x}}_{it}$$

is nonsingular,

$$\hat{\boldsymbol{\beta}}_M = \hat{\boldsymbol{\beta}}_{FE}$$

Moreover, in the extended regression

$$y_{it} \text{ on } 1, \mathbf{x}_{it}, \bar{\mathbf{x}}_{i\cdot}, \bar{\mathbf{x}}_{\cdot t}, \mathbf{z}_i, \mathbf{m}_t, t = 1, \ldots, T; i = 1, \ldots, N$$

for time-constant variables $\mathbf{z}_i$ and unit-constant variables $\mathbf{m}_t$, the coefficients on $\mathbf{x}_{it}$ are still $\hat{\boldsymbol{\beta}}_{FE}$. □

- Proof uses Frisch-Waugh partialling out.
- Coefficients on $\bar{\mathbf{x}}_{i\cdot}$ and $\bar{\mathbf{x}}_{\cdot t}$ *do* change with the inclusion of $\mathbf{z}_i$, $\mathbf{m}_t$.
  - Basis of robust, regression-based Hausman tests.
- Suppose a regressor is an interaction of the form

$$x_{itj} = z_{ij} \cdot m_{tj}$$

  - Then

$$\bar{x}_{i\cdot j} = z_{ij} \cdot \bar{m}_j, \ \bar{x}_{\cdot tj} = \bar{z}_j m_{tj}$$

  - Mundlak regression includes z_{ij}, m_{tj} as controls.

3. Interventions with Common Treatment Timing

- T time periods.
  - $t = 1, \ldots, q - 1$ are control periods.
  - Intervention happens at $t = q$, remains in place.
- Treatment indicator:

$$w_{it} = d_i \cdot p_t$$

$$d_i = 1 \text{ if (eventually) treated}$$

$$p_t = fq_t + \cdots + fT_t = 1 \text{ if a post treatment period}$$

- Homogeneous treatment effect.
- Equation that motivates TWFE:

$$y_{it} = \beta w_{it} + c_i + g_t + u_{it}, t = 1, \ldots, T; i = 1, 2, \ldots, N$$

$$\begin{aligned} \bar{w}_{i\bullet} &= d_i \bar{p} \\ \bar{w}_{\bullet t} &= \bar{d} p_t \end{aligned}$$

- TWM regression is equivalent to the DID regression

$$y_{it} \text{ on } 1, w_{it}, d_i, p_t, t = 1, \ldots, T; i = 1, \ldots, N$$

  - $\hat{\beta}_{DD} = \hat{\beta}_{FE}$. Enough to control for d_i, p_t.

- The TWFE estimator has the familiar form

$$\hat{\beta}_{FE} = \hat{\beta}_{DD} = \left(\bar{y}_1^{post} - \bar{y}_0^{post}\right) - (\bar{y}_1^{pre} - \bar{y}_0^{pre})$$

- Using separate time dummies $f2_t, \ldots, fT_t$ in place of p_t has no effect on $\hat{\beta}_{DD}$.
- Adding time-constant controls, $\mathbf{x}_i$ or $d_i \cdot \mathbf{x}_i$, has no effect on $\hat{\beta}_{DD}$.

- Allow TEs to change over treatment period:

$$y_{it} = \beta_q(w_{it} \cdot fq_t) + \cdots + \beta_T(w_{it} \cdot fT_t) + c_i + g_t + u_{it}$$

- Can use TWFE to estimate the β_r.

$$w_{it} \cdot fr_t = d_i \cdot p_t \cdot fr_t = d_i(fq_t + \cdots + fT_t)fr_t = d_i fr_t$$

- Time averages are proportional to d_i.
- Cross-sectional averages proportional to fr_t.
- TWM equation:

$$y_{it} = \alpha + \beta_q(w_{it} \cdot fq_t) + \cdots + \beta_T(w_{it} \cdot fT_t) + \zeta d_i + \theta_q fq_t + \cdots + \theta_T fT_t + e_{it}$$

- POLS and RE give identical estimates.

- Allow time-constant covariates, $\mathbf{x}_i$.
  - Adding $\mathbf{x}_i$ or $d_i \cdot \mathbf{x}_i$ to the regression does not change the $\hat{\beta}_r$.
- Instead, also include interactions with time dummies and treatment:

$$\begin{aligned} y_{it} = {} & \beta_q(w_{it} \cdot fq_t) + \cdots + \beta_T(w_{it} \cdot fT_t) + [w_{it} \cdot fq_t \cdot (\mathbf{x}_i - \boldsymbol{\mu}_1)]\boldsymbol{\gamma}_q \\ & + \cdots + [w_{it} \cdot fT_t \cdot (\mathbf{x}_i - \boldsymbol{\mu}_1)]\boldsymbol{\gamma}_T \\ & + (fq_t \cdot \mathbf{x}_i)\boldsymbol{\delta}_q + \cdots + (fT_t \cdot \mathbf{x}_i)\boldsymbol{\delta}_T + c_i + g_t + u_{it} \end{aligned}$$

$$\boldsymbol{\mu}_1 \equiv E(\mathbf{x}_i | d_i = 1)$$

- Can estimate by TWFE or TWM.

- TWM includes the time-constant variables d_i, $\mathbf{x}_i$, $d_i \cdot \mathbf{x}_i$.
- Need time dummies for fq_t, ..., fT_t.

$$
\begin{aligned}
y_{it} = {} & \alpha + \beta_q(w_{it} \cdot fq_t) + \cdots + \beta_T(w_{it} \cdot fT_t) + [w_{it} \cdot fq_t \cdot (\mathbf{x}_i - \boldsymbol{\mu}_1)]\boldsymbol{\gamma}_q \\
& + \cdots + [w_{it} \cdot fT_t \cdot (\mathbf{x}_i - \boldsymbol{\mu}_1)]\boldsymbol{\gamma}_T \\
& + (fq_t \cdot \mathbf{x}_i)\boldsymbol{\delta}_q + \cdots + (fT_t \cdot \mathbf{x}_i)\boldsymbol{\delta}_T + \zeta d_i + \mathbf{x}_i\boldsymbol{\xi} + (d_i \cdot \mathbf{x}_i)\boldsymbol{\lambda} \\
& + \theta_q fq_t + \cdots + \theta_T fT_t + e_{it}
\end{aligned}
$$

  - Harmless to include fs_t for $s = 2, \ldots, q - 1$.
- Replace $\boldsymbol{\mu}_1$ with $\bar{\mathbf{x}}_1 = N_1^{-1} \sum_{i=1}^{N} d_i \cdot \mathbf{x}_i$.

• Pooled OLS regression

$$y_{it} \text{ on } 1, d_i, \mathbf{x}_i, d_i \cdot \mathbf{x}_i, f2_t, \ldots, fT_t, f2_t \cdot \mathbf{x}_i, \ldots, fT_t \cdot \mathbf{x}_i,$$
$$w_{it} \cdot d_i \cdot fq_t, \ldots, w_{it} \cdot d_i \cdot fT_t$$
$$w_{it} \cdot d_i \cdot fq_t \cdot \dot{\mathbf{x}}_i, \ldots, w_{it} \cdot d_i \cdot fT_t \cdot \dot{\mathbf{x}}_i$$

$$\dot{\mathbf{x}}_i = \mathbf{x}_i - \bar{\mathbf{x}}_1$$

▸ Estimation is same without w_{it}.

▸ Introducing w_{it} is convenient for obtaining standard errors that account for sampling error in $\bar{\mathbf{x}}_1$.

- To use Stata's `margins` option, do not center the covariates:

$$
\begin{aligned}
y_{it} \text{ on } & 1, d_i, \mathbf{x}_i, d_i \cdot \mathbf{x}_i, f2_t, \ldots, fT_t, f2_t \cdot \mathbf{x}_i, \ldots, fT_t \cdot \mathbf{x}_i, \\
& w_{it} \cdot d_i \cdot fq_t, \ldots, w_{it} \cdot d_i \cdot fT_t, \\
& w_{it} \cdot d_i \cdot fq_t \cdot \mathbf{x}_i, \ldots, w_{it} \cdot d_i \cdot fT_t \cdot \mathbf{x}_i
\end{aligned}
$$

```
reg y d x1 ... xK c.d#c.x1 ... c.d#c.xK
   i.year i.year#c.x1 ... i.year#c.xK
   c.w#c.d#c.fq ... c.w#c.d#c.fT
   c.w#c.d#c.fq#c.x1 ... c.w#c.d#c.fq#c.xK
⋮
   c.w#c.d#c.fT#c.x1 ... c.w#c.d#c.fT#c.xK
   vce(cluster id)
```

```
margins, dydx(w) at(d = 1 fq = 1 fqp1 = 0
   ... fT = 0), subpop(if d == 1) vce(uncon)
margins, dydx(w) at(d = 1 fq = 0 fqp1 = 1
   ... fT = 0), subpop(if d == 1) vce(uncon)
⋮
margins, dydx(w) at(d = 1 fq = 0 fqp1 = 0
   ... fT = 1), subpop(if d == 1) vce(uncon)
```

- Same with TWFE.
- See did_4.do.

What is Being Estimated?

- Potential outcomes, $y_t(0)$ and $y_t(1)$.
- Treatment effect for a generic unit:

$$te_t = y_t(1) - y_t(0)$$

- ATT in each treated period:

$$\tau_t \equiv E[y_t(1) - y_t(0)|d = 1],\ t = q, q+1, \ldots, T$$

Assumption NA (**No Anticipation**): For $t < q$,

$$E[y_t(1) - y_t(0)|d = 1] = 0. \ \square$$

- The ATTs prior to the intervention are zero.

Assumption CT (**Common Trend**): With the (eventually) treated indicator d,

$$E[y_t(0) - y_1(0)|d] = E[y_t(0) - y_1(0)] \equiv \theta_t, \ t = 2, \ldots, T. \ \square$$

- Does not depend on treatment status.

• With time-constant covariates:

Assumption CCT (Conditional Common Trends): For treatment indicator d and covariates $\mathbf{x}$,

$$E[y_t(0) - y_1(0)|d, \mathbf{x}] = E[y_t(0) - y_1(0)|\mathbf{x}],\ t = 2, \ldots, T. \ \square$$

• Abadie (2005) uses this with $T = 2$.

• Similar to Callaway and Sant'Anna (2021); Sun and Abraham (2021); and others.

• Add a linearity (in $\mathbf{x}$) assumption for the conditional means

- Under NA, CCT, and linearity, I show

$$
\begin{aligned}
E(y_t|d,\mathbf{x}) = {} & \eta + \lambda d + \dot{\mathbf{x}}\boldsymbol{\kappa} + (d \cdot \dot{\mathbf{x}})\boldsymbol{\varphi} + \theta_2 f2_t + \cdots + \theta_T fT_t \\
& + (f2_t \cdot \dot{\mathbf{x}})\boldsymbol{\pi}_q + \cdots + (fT_t \cdot \dot{\mathbf{x}})\boldsymbol{\pi}_T \\
& + \tau_q(d \cdot fq_t) + \cdots + \tau_T(d \cdot fT_t) \\
& + (d \cdot fq_t \cdot \dot{\mathbf{x}})\boldsymbol{\rho}_q + \cdots + (d \cdot fT_t \cdot \dot{\mathbf{x}})\boldsymbol{\rho}_T
\end{aligned}
$$

$$
\dot{\mathbf{x}} = \mathbf{x} - E(\mathbf{x}|d = 1)
$$

4. Staggered Interventions

- Wooldridge (2005): Usual TWFE with heterogeneous slopes.
- TWFE under recent scrutiny for staggered (and more general) interventions.
- de Chaisemartin and D'Haultfœuille (2020), Goodman-Bacon (2021), Callaway and Sant'Anna (2021), Sun and Abraham (2021).
- Just showed TWFE is fine in common timing case.
  - Allow TEs to change across t and with $\mathbf{x}$.
- Can use TWFE or TWM in staggered case and allow lots of heterogeneity.

- First intervention period is $t = q$.
- Subsequent treatment in each period after q, up to T.
  - Might have gaps.
- No reversibility.
- Initially, a never treated group.

- Define potential outcomes:

$$y_t(\infty)\text{: never treated state}$$

$$y_t(r),\, r \in \{q, q+1, \ldots, T\}\text{: first exposure in } r$$

- Define treatment cohorts by dummies: d_q, ..., d_T.
  - $d_r = 1$ if unit first enters treatment in period r.

- Define ATTs relative to the never treated state:

$$\tau_{rt} \equiv E[y_t(r) - y_t(\infty)|d_r = 1],\ r = q, \ldots, T; t = r, \ldots, T$$

  - For cohort r, can estimate ATTs for $t = r, r+1, \ldots, T$.

Assumption NA (**No Anticipation**, **Staggered**): For treatment cohorts $r = q,\ q+1,\ \ldots, T$,

$$E[y_t(r) - y_t(\infty)|\mathbf{d}] = 0,\ t < r.\ \square$$

Assumption CTS (**Common Trend**, **Staggered**): With the exposure dummies d_q, ..., d_T,

$$E[y_t(\infty) - y_1(\infty)|d_q,\ldots,d_T] = E[y_t(\infty) - y_1(\infty)] \equiv \theta_t,\ t = 2,\ldots,T. \ \square$$

- Similar to Callaway and Sant'Anna; Sun and Abraham; others.
- Under Assumptions NA and CTS for a random draw i:

$$\begin{aligned} E(y_{it}|\mathbf{d}_i) &= \eta + \lambda_q d_{iq} + \cdots + \lambda_T d_{iT} + \sum_{s=2}^{T} \theta_s fs_t \\ &+ \sum_{r=q}^{T}\sum_{s=r}^{T} \tau_{rs}(w_{it} \cdot d_{ir} \cdot fs_t),\ t = 1,\ldots,T \end{aligned}$$

- This is the Mundlak equation.
  - Time dummies for $t < q$ are redundant.
- Mundlak regression:

$$
\begin{aligned}
&y_{it} \text{ on } 1, d_{iq}, \ldots, d_{iT}, fq_t, \ldots, fT_t, \\
&\qquad w_{it} \cdot d_{iq} \cdot fq_t, \ldots, w_{it} \cdot d_{iq} \cdot fT_t, \ldots, w_{it} \cdot d_{iT} \cdot fT_t
\end{aligned}
$$

  - Include every interaction that makes sense as a treatment indicator.
  - The cohort and year dummies are controls.
  - If there is no cohort r, drop all terms with d_{ir}.

- Equivalently, can use TWFE:

$$y_{it} = \sum_{r=q}^{T}\sum_{s=r}^{T} \tau_{rs}(w_{it} \cdot d_{ir} \cdot fs_t) + c_i + f_t + u_{it},\ t = 1,\ldots,T;\ i = 1,\ldots,N$$

- This "extended" TWFE allows more heterogeneity than imposing

$$\tau_{rs} = \tau,\ r = q,\ldots,T;\ s = r,\ldots,T$$

- Can aggregate the estimates or impose restrictions.
- Note: New Stata command `xtdidregress` estimates constant effect model.

- Add covariates.

Assumption CCTS (**Conditional Common Trends**, **Staggered**):

For exposure indicators d_r and covariates $\mathbf{x}$,

$$E[y_t(0) - y_1(0)|d_q, \ldots, d_T, \mathbf{x}] = E[y_t(0) - y_1(0)|\mathbf{x}],\ t = 2, \ldots, T. \ \square$$

- Assume all conditional expectations are linear in $\mathbf{x}$.
- This means linearity conditional on each $d_r = 1, r = q, \ldots, T$.

$$\dot{\mathbf{x}}_r \equiv \mathbf{x} - E(\mathbf{x}|d_r = 1), r = q, \ldots, T$$

• Under Assumptions NA, CCTS, and linearity:

$$
\begin{aligned}
E(y_t|d_q,\ldots,d_T,\mathbf{x}) = {} & \eta + \sum_{r=q}^{T} \lambda_r d_r + \mathbf{x}\boldsymbol{\kappa} + \sum_{r=q}^{T} (d_r \cdot \mathbf{x})\boldsymbol{\zeta}_r \\
& + \sum_{s=2}^{T} \theta_s fs_t + \sum_{s=2}^{T} (fs_t \cdot \mathbf{x})\boldsymbol{\pi}_t \\
& + \sum_{r=q}^{T}\sum_{s=r}^{T} \tau_{rs}(d_r \cdot fs_t) + \sum_{r=q}^{T}\sum_{s=r}^{T} (d_r \cdot fs_t \cdot \dot{\mathbf{x}}_r)\boldsymbol{\rho}_{rs}.
\end{aligned}
$$

- The regression is

$$
\begin{aligned}
y_{it} \text{ on } & 1, d_{iq}, \ldots, d_{iT}, \mathbf{x}_i, d_{iq} \cdot \mathbf{x}_i, \ldots, d_{iT} \cdot \mathbf{x}_i, \\
& f2_t, \ldots, fT_t, f2_t \cdot \mathbf{x}_i, \ldots, fT_t \cdot \mathbf{x}_i, \\
& w_{it} \cdot d_{iq} \cdot fq_t, \ldots, w_{it} \cdot d_{iq} \cdot fT_t, \ldots, \\
& w_{it} \cdot d_{i,q+1} \cdot f(q+1)_t, \ldots, w_{it} \cdot d_{i,q+1} \cdot fT_t, \ldots, w_{it} \cdot d_{iT} \cdot fT_t, \\
& w_{it} \cdot d_{iq} \cdot fq_t \cdot \dot{\mathbf{x}}_{iq}, \ldots, w_{it} \cdot d_{iq} \cdot fT_t \cdot \dot{\mathbf{x}}_{iT}, \\
& w_{it} \cdot d_{i,q+1} \cdot f(q+1)_t \cdot \dot{\mathbf{x}}_{iq}, \ldots, w_{it} \cdot d_{i,q+1} \cdot fT_t \cdot \dot{\mathbf{x}}_{iT}, \ldots, \\
& w_{it} \cdot d_{iT} \cdot fT_t \cdot \dot{\mathbf{x}}_{iT}
\end{aligned}
$$

$$
\dot{\mathbf{x}}_{ir} = \mathbf{x}_i - \bar{\mathbf{x}}_r = \mathbf{x}_i - N_r^{-1} \sum_{h=1}^{N} d_{hr} \mathbf{x}_h .
$$

- RE gives identical estimates.
  - Improving over POLS requires allowing more general patterns of serial correlation and maybe time-varying variances.
- Equivalently, drop everything in the first two lines except $f2_t \cdot \mathbf{x}_i$, ..., $fT_t \cdot \mathbf{x}_i$ and use TWFE.
- Can use Stata and `margins` to account for sampling variation in $\bar{\mathbf{x}}_r$.
  - See staggered_6.do.

- Often want to aggregate the effects.
  - Can average all ATTs for a single effect.
  - Average by cohort.
- Or, impose restrictions before estimation.
  - Treatment effect only differs by intensity, not calendar time.

Efficiency of POLS

THEOREM 6.2: Write the conditional mean equation with a composite error as

$$y_{it} = \eta + \sum_{r=q}^{T} \lambda_r d_{ir} + \mathbf{x}_i \boldsymbol{\kappa} + \sum_{r=q}^{T} (d_{ir} \cdot \mathbf{x}_i) \boldsymbol{\zeta}_r + \sum_{s=2}^{T} \theta_s fs_t + \sum_{s=2}^{T} (fs_t \cdot \mathbf{x}_i) \boldsymbol{\pi}_s$$

$$+ \sum_{r=q}^{T} \sum_{s=r}^{T} \tau_{rs} (w_{it} \cdot d_{ir} \cdot fs_t) + \sum_{r=q}^{T} \sum_{s=r}^{T} (w_{it} \cdot d_{ir} \cdot fs_t \cdot \dot{\mathbf{x}}_{ir}) \boldsymbol{\rho}_{rs} + a_i + u_{it}$$

$$E(a_i | \mathbf{d}_i, \mathbf{x}_i) = 0, \; E(\mathbf{u}_i | a_i, \mathbf{d}_i, \mathbf{x}_i) = \mathbf{0}$$

$$\mathbf{u}_i' \equiv (u_{i1}, u_{i2}, \ldots, u_{iT})$$

• Assume in addition that

$$
\begin{aligned}
Var(a_i|\mathbf{d}_i, \mathbf{x}_i) &= \sigma_a^2 \\
Var(\mathbf{u}_i|a_i, \mathbf{d}_i, \mathbf{x}_i) &= \sigma_u^2 \mathbf{I}_T
\end{aligned}
$$

• The POLS estimator $\hat{\boldsymbol{\tau}} = (\hat{\tau}_{rs})$ has the following properties:

(i) $\hat{\boldsymbol{\tau}}$ is the BLUE of $\hat{\boldsymbol{\tau}}$ conditional on $(\mathbf{D}, \mathbf{X})$ for any realization where the rank condition holds.

(ii) $\hat{\boldsymbol{\tau}}$ is asymptotically efficient in the class of estimators consistent under NA, CCTS, linearity. $\square$

- POLS is BLUE because it equals RE using the true variance-covariance matrix.
  - POLS = RE follows from Wooldridge (2019).
- Under assumptions similar in spirit, Borusyak, Jaravel, and Spiess (2021) show their imputation estimator is BLUE.
- POLS is not efficient under serial correlation in $\{u_{it} : t = 1, \ldots, T\}$ or heteroskedasticity in (c_i, u_{it}).
  - Could use, say, an unrestricted feasible GLS estimator.
- How to improve efficiency of imputation?

All Units Eventually Treated

- Regression approach (POLS/ETWFE) extends immediately if all units are treated by period T.
- Generally, the TEs are

$$y_t(r) - y_t(T),\ r = q,\ldots,T-1;\ t = r,\ldots,T$$

  - The gain in period t from first being treated in the earlier period r rather than the last period.

- The identified parameters are

$$\tau_{(r:T),t} \equiv E[y_t(r) - y_t(T)|d_r = 1],\ r = q, \ldots, T-1;\ t = r, \ldots, T$$

- The NA and CT assumptions are stated for the potential outcome $y_t(T)$.
- If there *could* have been a never treated group, under NA

$$y_t(T) = y_t(\infty),\ t < T$$

5. Comparision with Other Methods

Long Differencing with Regression Adjustment/IPW

- Callaway and Sant'Anna (2021) extend Abadie (2005) to multiple periods, staggered interventions.
  - Also combine regression with inverse probability weighting for "doubly robust" estimation.
- Long differencing is inefficient: Does not use all control units available.
  - Can be more resilient to violations of parallel trends.

- Consider $T = 3$ with staggered entry in $t = 2$ and $t = 3$.

$$\tau_{22} = E[y_2(2) - y_2(\infty)|d_2 = 1]$$

- POLS/ETWFE will use the $d_\infty = 1$ and $d_3 = 1$ cohorts as control groups to estimate τ_{22}.
  - Neither group has been treated at $t = 2$.
- Callaway and Sant'Anna use the never treated group.
  - Can see this in the output of the Stata user-written command `csdid`.

Imputation Estimators

• Recall two ways to estimate τ_{att} in the cross-sectional treatment effect setting assuming unconfoundedness.

1. Pooled OLS:

$$y_i \text{ on } 1, d_i, \mathbf{x}_i, d_i \cdot (\mathbf{x}_i - \bar{\mathbf{x}}_1), i = 1, \ldots, N$$

▸ $\hat{\tau}_{att}$ is the coefficient on d_i.

2. Imputation. Using only the N_0 controls,

$$y_i \text{ on } 1, \mathbf{x}_i \text{ if } d_i = 0$$

▸ For each of the N_1 treated units, impute an estimate of $y_i(0)$:

$$\hat{y}_i(0) = \hat{\alpha}_0 + \hat{\boldsymbol{\beta}}_0 \mathbf{x}_i \text{ if } d_i = 1$$

$$\widehat{te}_i \equiv y_i - \hat{y}_i(0) = y_i - \hat{\alpha}_0 - \hat{\boldsymbol{\beta}}_0 \mathbf{x}_i,$$

$$\tilde{\tau}_{att} \equiv N_1^{-1} \sum_{i=1}^{N} d_i \cdot \widehat{te}_i = \bar{y}_1 - N_1^{-1} \sum_{i=1}^{N} d_i \cdot \hat{y}_i(0) = \bar{y}_1 - \left(\hat{\alpha}_0 + \bar{\mathbf{x}}_1 \hat{\boldsymbol{\beta}}_0 \right)$$

- Well known that

$$\tilde{\tau}_{att} = \hat{\tau}_{att}$$

- Same is true in the staggered DiD setting.

$$
\begin{aligned}
E(y_{it}|d_{iq},\ldots,d_{iT},\mathbf{x}_i) = \eta &+ \sum_{r=q}^{T}\lambda_r d_{ir} + \mathbf{x}_i\boldsymbol{\kappa} + \sum_{r=q}^{T}(d_{ir}\cdot\mathbf{x}_i)\boldsymbol{\zeta}_r \\
&+ \sum_{s=2}^{T}\theta_s fs_t + \sum_{s=2}^{T}(fs_t\cdot\mathbf{x}_i)\boldsymbol{\pi}_s \\
&+ \sum_{r=q}^{T}\sum_{s=r}^{T}\tau_{rs}(w_{it}\cdot d_{ir}\cdot fs_t) \\
&+ \sum_{r=q}^{T}\sum_{s=r}^{T}(w_{it}\cdot d_{ir}\cdot fs_t\cdot\dot{\mathbf{x}}_{ir})\boldsymbol{\rho}_{rs}.
\end{aligned}
$$

(i) Using the $w_{it} = 0$ observations, run the pooled regression and obtain the $\hat{\eta}$, $\hat{\lambda}_r$, $\hat{\boldsymbol{\zeta}}_r$, $\hat{\theta}_s$, $\hat{\boldsymbol{\pi}}_s$.

(ii) For the $w_{it} = 1$ subsample, obtain

$$\widehat{te}_{it} = y_{it} - \left[\hat{\eta} + \sum_{r=q}^{T} \hat{\lambda}_r d_{ir} + \mathbf{x}_i \hat{\boldsymbol{\kappa}} + \sum_{r=q}^{T} (d_{ir} \cdot \mathbf{x}_i) \hat{\boldsymbol{\zeta}}_r \right.$$
$$\left. + \sum_{s=2}^{T} \hat{\theta}_s fs_t + \sum_{s=2}^{T} (fs_t \cdot \mathbf{x}_i) \hat{\boldsymbol{\pi}}_s \right]$$

$$\tilde{\tau}_{rt} = N_{rt}^{-1} \sum_{i=1}^{N} d_{ir} \widehat{te}_{it}$$

• Can show that

$$\tilde{\tau}_{rt} = \hat{\tau}_{rt}, r = q, \ldots, T, t = r, \ldots, T$$

• Also, the estimates $\hat{\eta}$, $\hat{\lambda}_r$, $\hat{\boldsymbol{\zeta}}_r$, $\hat{\theta}_s$, $\hat{\boldsymbol{\pi}}_s$ from the imputation method are the same as the POLS estimates.

• Not quite the same as BJS (2021): they use fixed effects in the first step.

6. Testing and Relaxing Parallel Trends

- Need at least two pre-treatment periods.
- Suppose $T = 3$, intervention at $t = 3$.
- Without covariates, run the regression

$$\Delta y_{i2} \text{ on } 1, d_i, i = 1, \ldots, N$$

  - Heteroskedasticity-robust t statistic on d_i.

- Two pooled OLS approaches yield the same statistic:

$$y_{it} \text{ on } 1, d_i, f2_t, d_i \cdot f2_t, f3_t, d_i \cdot f3_t, t = 1,2,3; i = 1,\ldots,N$$

  - Cluster-robust t statistic on $d_i \cdot f2_t$.

- Or use a heterogeneous linear time trend:

$$y_{it} \text{ on } 1, d_i, f2_t, d_i \cdot t, f3_t, d_i \cdot f3_t, t = 1,2,3; i = 1,\ldots,N$$

  - Cluster-robust t statistic on $d_i \cdot t$.

- Statistics are identical.
  - Coefficients on $d_i \cdot f3_t$ can be very different.

• Using $d_i \cdot t$, the coefficient on $d_i \cdot f3_t$ is a DiDiD estimator:

$$
\begin{aligned}
\hat{\tau}_3 &= N_1^{-1} \sum_{i=1}^{N} d_i \cdot \Delta^2 y_{i3} - N_0^{-1} \sum_{i=1}^{N} (1 - d_i) \cdot \Delta^2 y_{i3} \\
&= [(\bar{y}_{3,treat} - \bar{y}_{2,treat}) - (\bar{y}_{2,treat} - \bar{y}_{1,treat})] \\
&\quad - [(\bar{y}_{3,control} - \bar{y}_{2,control}) - (\bar{y}_{2,control} - \bar{y}_{1,control})] \\
&= (\overline{\Delta y}_{3,treat} - \overline{\Delta y}_{3,control}) - (\overline{\Delta y}_{2,treat} - \overline{\Delta y}_{2,control})
\end{aligned}
$$

• Testing strategies in the general case:

1. In the full POLS regression, add interactions $d_{ir} fs_t$ for $s < r$, do joint test.

2. In the full POLS regression, add heterogenous linear trends

$$d_{iq} \cdot t, \ldots, d_{iT} \cdot t$$

and use a joint test.

▸ This works as a correction, too, if the differences in trends are linear in t.

- The imputation result holds for adding heterogeneous trends.
  - ▸ So the test is identical to using only the $w_{it} = 0$ observations and doing a joint test on

$$d_{iq} \cdot t, ..., d_{iT} \cdot t$$

  - ▸ The test for pre-trends is not contaminated by using the long regression and all observations provided a full set of heterogeneous treatment effects is allowed.
  - ▸ Same property as the BJS (2021) test for pre-trends.

7. Simulations

- $N = 500$, $T = 6$, staggered entry at $q = 4$.
- One covariate. CT imposed conditional on x.
- $R^2 = 0.127$.
- Cohort shares: $\rho_\infty = 0.241$, $\rho_4 = 0.358$, $\rho_5 = 0.291$,

$\rho_6 = 0.225$.

- 1,000 replications.

	ATT	No Control		POLS		CS		Het. Trend	
$N = 500$	Mean	Mean	SD	Mean	SD	Mean	SD	Mean	SD
τ_{44}	3.99	3.99	0.287	3.99	0.288	3.99	0.362	3.99	0.396
τ_{45}	4.19	4.19	0.288	4.19	0.289	4.20	0.367	4.20	0.513
τ_{46}	4.59	4.60	0.307	4.60	0.316	4.60	0.372	4.61	0.662
τ_{55}	3.03	3.02	0.322	3.03	0.326	3.03	0.446	3.02	0.423
τ_{56}	3.62	3.62	0.326	3.63	0.358	3.63	0.430	3.62	0.521
τ_{66}	2.05	2.05	0.410	2.04	0.474	2.04	0.644	2.05	0.546

- Rejection rate of common trends test (3 df, 5% level): 0.045

- Generate outcomes with different linear trends for d_4, d_5, and d_6.

	ATT	POLS		CS		Het. Trend	
$N = 500$	Mean	Mean	SD	Mean	SD	Mean	SD
τ_{44}	3.99	2.42	0.288	2.99	0.362	3.99	0.396
τ_{45}	4.19	1.52	0.291	2.20	0.367	4.20	0.513
τ_{46}	4.59	0.75	0.317	1.60	0.372	4.61	0.662
τ_{55}	3.03	1.99	0.329	2.53	0.446	3.02	0.423
τ_{56}	3.62	1.91	0.358	2.63	0.430	3.62	0.521
τ_{66}	2.05	1.05	0.474	1.70	0.644	2.05	0.546

8. Concluding Remarks

- Equivalence between TWFE and TWM has applications to DiD estimators with common and staggered entry.
  - "Extended" TWFE allows for flexible treatment effects.
  - TWFE some resilience to unbalanced panels.

- The FE approach extends to exponential mean functions:

$$E(y_{it}|d_{iq},\ldots,d_{iT},\mathbf{x}_i,c_i) = c_i \exp\Bigg[\sum_{s=2}^{T}\theta_s fs_t + \sum_{s=2}^{T}(fs_t \cdot \mathbf{x}_i)\boldsymbol{\pi}_s$$

$$+ \sum_{r=q}^{T}\sum_{s=r}^{T}\tau_{rs}(w_{it} \cdot d_{ir} \cdot fs_t)$$

$$+ \sum_{r=q}^{T}\sum_{s=r}^{T}(w_{it} \cdot d_{ir} \cdot fs_t \cdot \dot{\mathbf{x}}_{ir})\boldsymbol{\rho}_{rs}\Bigg]$$

  - Use FE Poisson estimator with cluster-robust inference.

- Pooled methods can be used with any nonlinear model.

$$
\begin{aligned}
E(y_{it}|d_{iq},\ldots,d_{iT},\mathbf{x}_i) = G\Bigg[& \eta + \sum_{r=q}^{T} \beta_r d_{ir} + \mathbf{x}_i\boldsymbol{\kappa} + \sum_{r=q}^{T} (d_{ir} \cdot \mathbf{x}_i)\boldsymbol{\eta}_r \\
& + \sum_{s=2}^{T} \gamma_s fs_t + \sum_{s=2}^{T} (fs_t \cdot \mathbf{x}_i)\boldsymbol{\pi}_s \\
& + \sum_{r=q}^{T} \sum_{s=r}^{T} \delta_{rs}(w_{it} \cdot d_{ir} \cdot fs_t) \\
& + \sum_{r=q}^{T} \sum_{s=r}^{T} (w_{it} \cdot d_{ir} \cdot fs_t \cdot \mathbf{x}_i)\boldsymbol{\xi}_{rs} \Bigg]
\end{aligned}
$$

- $G(\cdot) = \exp(\cdot)$ for $y_{it} \geq 0$.
- $G(\cdot) = \Lambda(\cdot) = \exp(\cdot)/[1 + \exp(\cdot)]$ for $0 \leq y_{it} \leq 1$ (binary or fractional)
  - ▸ Use pooled quasi-MLE in the linear exponential family.
  - ▸ Benefit to using the canonical link: pooled and imputation methods are identical, as in the linear case.

- Can combine insights from regression – which uses all information in the assumptions – with IPW for efficient doubly robust estimation.
  - ▸ Details to be worked out.
  - ▸ Have to be explicit about overlap assumptions.
- Currently thinking about staggered exit.
  - ▸ Cohorts are now indexed by entry and exit date.