Two-stage differences in differences

John Gardner
University of Mississippi
jrgardne@olemiss.edu
jrgcmu.github.io

WEAI International Conference
March, 2021

Introduction: The problem

- In the 2$\times$2 case, difference-in-differences regression identifies the ATT (i.e., the ATE for the treated group)
- Recent literature: With multiple groups/periods, this does not hold if ATTs vary by group/treatment duration:
  - DD regression identifies a weighted average of group$\times$period-specific ATTs, where the weights may actually be negative (Borusyak and Jaravel, 2017; de Chaisemartin and D'Haltfoeuille, 2020; Sun and Abraham, 2020)
  - Equivalently, DD regression represents a (positive, variance) weighted average of all 2$\times$2 DDs, so identifies a weighted average of ATTs plus *changes* in ATTs (Goodman-Bacon, 2018)

Introduction: Existing solutions

- Stacked DD (Gormley and Matsa, 2014; Cengiz et al. 2019, Deshpandi and Li, 2019, e.g.)
  - Stack treated/controls for each adoption into a “tall” dataset, using relative time instead of calendar time
  - IDs weighted average of treatment effects
- Aggregation: Estimate each group$\times$period effects, then aggregate them somehow
  - Callaway and Sant’anna, 2020: Use individual 2×2 DD regressions, IPW, or a doubly robust combination
  - Sun and Abraham, 2020: Use one regression with interactions between treatment-status, group and period

Introduction: This paper

- Provides simple insight into why DD fails to identify a reasonable average treatment effect with multiple groups/periods
- Based on this approach, develops a simple and intuitive new approach to estimation that works with multiple groups/periods

Motivation: Setup

- Index groups by g and periods by p, group 0 is never treated, group 1 adopts treatment in period 1, group 2 adopts in period 2, etc.
- Groups may consist of individuals i, periods may consist of shorter time units t
- Think of g as groups of states that are treated at the same time and p as groups of years during which they become treated

Motivation: Causal model

- The ATT for group g in period p:

$$\beta_{gp} = E(Y_{1gpit} - Y_{0gpit} | g, p)$$

  where (Y_{0gpit}, Y_{1gpit}) are underlying counterfactual outcomes

- Parallel trends:

$$E(Y_{gpit} | g, p, D_{gp}) = \lambda_g + \gamma_p + \beta_{gp} D_{gp},$$

  where D_{gp} is an indicator for whether group g is treated in period p

Motivation: The 2×2 case

- In the 2×2 case, the DD regression

$$E(Y_{gpit}|g, p, D_{gp}) = \lambda_g + \gamma_p + \beta_{gp} D_{gp}$$

  is the same as the "manual" DD

$$(\mu_{11} - \mu_{10}) - (\mu_{01} - \mu_{00}) = \beta_{11}$$

- Can think of this as the difference in outcomes between the treated and control groups, *after removing group and time effects* (λ_g and γ_t)

Motivation: Understanding the problem

- We now know that this doesn't always extend to the case of multiple groups/periods
- DD has been around forever. Why did it take so long to realize this?
- What's wrong with this logic?

  > *Mean outcomes are linear in group effects, period effects, and treatment status, so regression DD identifies the overall average ATT*

Motivation: The general case

- Rewrite parallel trends as

$$E(Y_{gpit}|g,p,D_{gp}) = \lambda_g + \gamma_p + E(\beta_{gp}|D_{gp}=1)D_{gp} \\ + [\beta_{gp} - E(\beta_{gp}|D_{gp}=1)]D_{gp}$$

  where $E(\beta_{gp}|D_{gp}=1)$ is the "overall average" ATT
- The "error term" $[\beta_{gp} - E(\beta_{gp}|D_{gp}=1)]D_{gp}$ is *not necessarily* mean-zero conditional on g, p and D_{gp}
- $\Rightarrow E(Y_{gpit}|g,p,D_{gp})$ is *not necessarily* a linear function of those variables, so regression DD *may not* identify it
- It *is* linear when there is only one treated group or when all of the group-specific ATTs are the same (so sometimes regression DD works, sometimes it doesn't)
- Can say more about what regression DD does identify (DD estimand)

Solution: Two-stage differences in differences

- In the 2×2 case, regression DD is the same as regressing outcomes on treatment status, *after removing group and period effects*
- This suggests a simple extension to the multiple groups/periods case:
  1. Estimate the model

     $$Y_{gpit} = \lambda_g + \gamma_p + \varepsilon_{gpit}$$

     on the sample of untreated observations (those with $D_{gp} = 0$)
  2. Regress adjusted outcomes

     $$\tilde{Y}_{gpit} = Y_{gpit} - \hat{\lambda}_g - \hat{\gamma}_p$$

     on treatment status D_{gp}

Solution: Why it works

- Parallel trends implies that

$$E(Y_{gpit}|g, p, D_{gpit}) - \lambda_g - \gamma_p = \beta_{gp} D_{gp}$$
$$= E(\beta_{gp}|D_{gp} = 1) D_{gp} + [\beta_{gp} - E(\beta_{gp}|D_{gp} = 1)] D_{gp}$$

- But the "error term" $[\beta_{gp} - E(\beta_{gp}|D_{gp} = 1)] D_{gp}$ in this regression *is* mean zero conditional on D_{gp}
- $\Rightarrow$ A regression of $\tilde{Y}_{gpit}$ on D_{gp} *does* identify $E(\beta_{gp}|D_{gp} = 1)$
- Consistent as number of observations per group grows (from continuous mapping theorem)

Solution: Advantages

- Intuitive: Difference between treatment and control group after removing group/period effects
- Easy to implement:
  - Don't have to reshape data
  - Don't need to estimate and manually aggregate individual group/period effects
  - Don't need any special software
- Can use standard two-step GMM results to correct SEs for first-stage estimation of $\hat{\lambda}_g$ and $\hat{\gamma}_p$ (Newey and McFadden, 1994)

Solution: Implementation

Can be implemented in one (long) line of Stata code:

```
gmm (eq1: (y-{xb: i.year}-{xg: ibn.id})*(1-d)) ///
  (eq2: y-{xb:} - {xg:} - {delta}*d), ///
  instruments(eq1: i.year ibn.id) ///
  instruments(eq2: d) winitial(identity) ///
  onestep quickderivatives vce(cluster id)
```

(Estimates both regressions simultaneously as a joint GMM estimator)

Extensions

- Easy to include covariates
- Can be adapted to identify other average treatment effect measures (e.g., average effect of being treated for $\bar{P}$ periods instead of average over all groups and periods)
- Sun and Abraham (2020) show that a similar problem applies to event-study regressions of the form

$$Y_{gpit} = \lambda_g + \gamma_p + \sum_{r=-R}^{P} \beta_r D_{rgp} + \varepsilon_{gpit},$$

  where D_{rgp} is an indicator for the treatment being adopted for $r \in \{-R, \ldots, 0, \ldots P\}$ periods
- The 2SDD approach extends readily to this case

Simulations: DGP

- 250 datasets, 50 units, 10 periods
- DGP:

$$Y_{gpit} = \lambda_i + \gamma_t + \beta_{gp} D_{gp} + \varepsilon_{gpit},$$

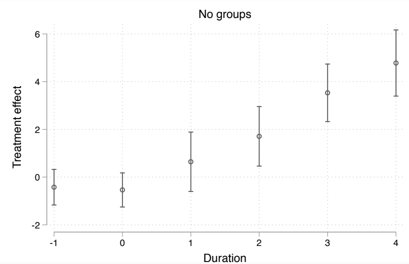
  $\lambda_i, \varepsilon_{gpit} \sim N$
- Three treatment groups adopt (one in period four, one in five, one in six)
- Equal/unequal group sizes
- ATT varies differently by treatment duration for each group

Simulations: Results

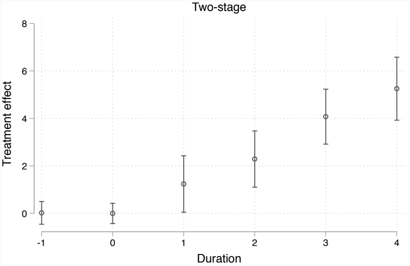
	Simulation 1	Simulation 2
True	4.08	3.46
Diff-in-diff	3.51	2.71
	(1.06)	(0.24)
Aggregated	4.12	3.48
	(1.02)	(0.23)
Two-stage	4.12	3.48
	(0.28)	(0.23)

Group sizes equal in sim 1 and unequal in sim 2

Simulations: Results



Regression approach suggests parallel trends violated (it's not)



2S approach identifies correct (duration-specific) average effects

Application: Autor (2003)

- Autor (2003), effects of limiting employment at will on employment in temporary help services sector (THS)
- 12 states adopt between 1997 and 1996 for 177 possible group$\times$period-specific ATTs

Diff-in-diff	0.108
	(0.105)
Aggregated	0.096
	(0.183)
Two-stage	0.099
	(0.176)

- Event-study results (not shown) are similar
- Can also examine the DD weights (DD weights)

Conclusion

- The two-stage approach is
  - Intuitive
  - Quick and easy to implement
  - Effective
- Simulation evidence (and an empirical application) illustrate these characteristics

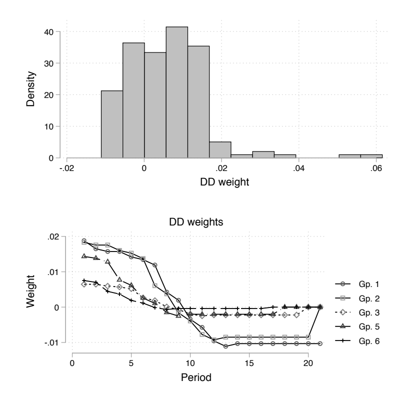
Regression DD estimand General case

- What *does* regression DD identify?
- It can be shown that $\beta^* = \sum_{g=1}^{G} \sum_{p=g}^{P} \omega_{gp} \beta_{gp}$, where

$$\beta_{gp} = \frac{[(1 - P_g) - (P_p - P)]\pi_{gp}}{\sum_{g=1}^{G} \sum_{p=1}^{P} [(1 - P_g) - (P_p - P)]\pi_{gp}},$$

  $P_g = P(D_{gp} = 1|g)$, $P_p = P(D_{gp} = 1|p)$, $P = P(D_{gp} = 1)$ and $\pi = P(g, p)$
- Intuition: Longer treated, more of TE attributed to group effects; more units treated, more of TE attributed to time effects
- Weights sum to one, but can be negative (also, if the β_{gp}'s are all the same, they don't matter)

Application: DD weights Application



- Weights are negative for some group-periods
- Weights decrease as groups treated for more periods and in periods where more groups are treated (this is only for the first 5 groups)