Inference in difference-in-differences revisited

Robert Joyce (Institute for Fiscal Studies)

Joint work with Mike Brewer (Essex, IFS) and Thomas Crossley (Koc, IFS)

PEPA is based at the IFS and CEMMAP

Introduction

- Emerging literature on inference in common DiD designs

- Difficult to get test size right when
  - Treatment status varies at a group-time level
  - Grouped (clustered) error terms
  - Few groups
  - Serial correlation in group-time shocks

Main points

- With Monte Carlo simulations we make 3 points

1. Can get test size right with simple tweaks to standard methods, even with few groups

2. Problem is low power to detect real effects

3. FGLS combined with robust inference can help a lot

Outline

- Background/review
  - What is the problem?
  - What solutions have been proposed?
- Our simulation evidence
  - Methods
  - Results
- Summary and conclusions

Setup

- Model: $Y_{igt} = \alpha + \beta T_{gt} + \delta X_{igt} + \mu_g + \xi_t + u_{igt}$

$$E(u_{igt} \mid T_{gt}, X_{igt}, \mu_g, \xi_t) = 0$$

$$u_{igt} = \eta_{gt} + \varepsilon_{igt}$$

- Computation of $\widehat{\beta}_{OLS}$ equivalent to first running this regression…

$$Y_{igt} = \lambda_{gt} + \delta X_{igt} + u_{igt}$$

- …and then this, with error term $\omega_{gt} \equiv \eta_{gt} + \left(\hat{\lambda}_{gt} - \lambda_{gt}\right)$

$$\hat{\lambda}_{gt} = \alpha + \beta T_{gt} + \mu_g + \xi_t + \omega_{gt}$$

- True precision of $\widehat{\beta}_{OLS}$ depends almost entirely on # of group-time cells, not # of observations (if cell sizes are large)
  - Severe version of standard clustering problem (Moulton, 1990)

Accounting for variance of group-time shocks

1. Cluster-robust standard errors (Liang and Zeger, 1986)
   - Consistent, and t-stat ~ N(0,1), as # of clusters goes to infinity

2. Make assumptions about distribution of η_{gt}
   - E.G. something enabling finite sample inference with few clusters (Donald and Lang, 2007)

3. Bootstrap to estimate distribution of t-stat (Cameron et al, 2008)

"Cluster-robust" standard errors with few clusters (1)

- Bias-reducing adjustments proposed (see Bell and McCaffrey, 2002; Imbens and Kolesar, 2012)
  - Scale residuals in CRSE formula by sqrt(G/(G-1)). Stata does this (approx.)
  - BM propose more complex scaling (invalid in setup here)
- But t-stat ~N(0,1) also depends on # of clusters going to infinity
- With few clusters, CRSEs (inc. bias-adjusted ones) ***and standard normal critical values*** deliver double the correct test size (Bertrand et al, 2004; Cameron et al, 2008)

“Cluster-robust” standard errors with few clusters (2)

- But don’t have to use N(0,1) critical values
- Typical few-clusters approach uses t distribution: Stata uses t_{G-1}

- Bester et al (2011) showed that using t_{G-1} critical values and sqrt(G/(G-1))-scaled CRSEs (i.e. ~= Stata’s approach) can lead to tests of correct size ***with G fixed***

  - Asymptotics apply as group size tends to infinity
  - Requires homogeneity condition that won’t normally hold in DiD
  - But we find its performance in practice looks very promising…

Serial correlation

- Group-time shocks typically serially correlated too
    - Can lead to huge over-rejection of nulls if ignored (Bertrand et al, 2004)
- Cluster-robust SEs should therefore cluster at group level
- Hansen (2007) models process as AR(k) and uses FGLS estimation
    - Derives bias correction for AR(k) parameters, consistent as $G \to \infty$
- FGLS should be more efficient, but inference still tricky
    - FGLS SEs are wrong if AR(k) parameterisation is wrong
    - Can combine with cluster-robust SEs to control test size…
    - …but that doesn’t work well with few groups (…or does it?)

MONTE CARLO SIMULATIONS

Monte Carlo simulations (1)

- Use women's log-earnings from CPS (N ~=750k), as in Bertrand et al (2004), Cameron et al (2008), Hansen (2007)

- Collapse to state-year level using covariate-adjusted means
  - As in other papers, we find test size can't be controlled in micro-data

- Repeat the following 5000 times, varying G from 6 to 50:
  - Sample G states at random with replacement
  - Randomly choose G/2 states to be 'treated'
  - Randomly choose a year from which treated states will be treated
  - Estimate treatment 'effect'
  - Test (true) null of no effect using nominal 5%-level test

Monte Carlo simulations (2)

- Model: $Y_{ict} = \alpha + \beta T_{ct} + \delta X_{ict} + \mu_c + \xi_t + \eta_{ct} + \varepsilon_{ict}$

1. Collapse to state-time level by estimating λ_{ct}

$$Y_{ict} = \lambda_{ct} + \delta X_{ict} + \varepsilon_{ict}$$

2. Monte Carlos look at inference based on following regression

$$\hat{\lambda}_{ct} = \alpha + \mu_c + \xi_t + \beta T_{ct} + \omega_{ct}$$

- 1 accounts for grouping of errors at state-time level
- ***Issue then is dealing with finite number of states, and serial correlation in state-time shocks***

Rejection rates with tests of nominal 5% size, for 'placebo treatments' with 30 years of CPS earnings data

	Number of groups (US states), half of which are treated			
Inference method	**50**	**20**	**10**	**6**
Assume iid	0.422*	0.420*	0.404*	0.412*

Notes:

* Indicates that rejection rate from 5000 Monte Carlo replications is statistically significantly different from 0.05.

Uses sample of CPS data defined and aggregated to state-year level in same way as in Bertrand, Duflo and Mullainathan, except we use data from 1979 to 2009 (rather than 1999). Monte Carlos work in same way as in row 4 of Table 2 of that paper.

Rejection rates with tests of nominal 5% size, for 'placebo treatments' with 30 years of CPS earnings data

	Number of groups (US states), half of which are treated			
Inference method	**50**	**20**	**10**	**6**
Assume iid	0.422*	0.420*	0.404*	0.412*
CRSE, N(0,1) critical vals	0.048	0.061*	0.079*	0.107*

Notes:

* Indicates that rejection rate from 5000 Monte Carlo replications is statistically significantly different from 0.05.

Uses sample of CPS data defined and aggregated to state-year level in same way as in Bertrand, Duflo and Mullainathan, except we use data from 1979 to 2009 (rather than 1999). Monte Carlos work in same way as in row 4 of Table 2 of that paper.

Rejection rates with tests of nominal 5% size, for 'placebo treatments' with 30 years of CPS earnings data

	Number of groups (US states), half of which are treated			
Inference method	**50**	**20**	**10**	**6**
Assume iid	0.422*	0.420*	0.404*	0.412*
CRSE, N(0,1) critical vals	0.048	0.061*	0.079*	0.107*
CRSE*sqrt(G/(G-1)), N(0,1)	0.044	0.056	0.075*	0.104*

Notes:

* Indicates that rejection rate from 5000 Monte Carlo replications is statistically significantly different from 0.05.

Uses sample of CPS data defined and aggregated to state-year level in same way as in Bertrand, Duflo and Mullainathan, except we use data from 1979 to 2009 (rather than 1999). Monte Carlos work in same way as in row 4 of Table 2 of that paper.

PEPA Programme Evaluation for Policy Analysis

NCRM National Centre for Research Methods

Rejection rates with tests of nominal 5% size, for ‘placebo treatments’ with 30 years of CPS earnings data

	Number of groups (US states), half of which are treated			
Inference method	**50**	**20**	**10**	**6**
Assume iid	0.422*	0.420*	0.404*	0.412*
CRSE, N(0,1) critical vals	0.048	0.061*	0.079*	0.107*
CRSE*sqrt(G/(G-1)), N(0,1)	0.044	0.056	0.075*	0.104*
CRSE*sqrt(G/G-1), t_{G-1}	0.042*	0.046	0.050	0.049

Notes:

* Indicates that rejection rate from 5000 Monte Carlo replications is statistically significantly different from 0.05.

Uses sample of CPS data defined and aggregated to state-year level in same way as in Bertrand, Duflo and Mullainathan, except we use data from 1979 to 2009 (rather than 1999). Monte Carlos work in same way as in row 4 of Table 2 of that paper.

PEPA Programme Evaluation for Policy Analysis

NCRM National Centre for Research Methods

Rejection rates with tests of nominal 5% size, for ‘placebo treatments’ with 30 years of CPS earnings data

	Number of groups (US states), half of which are treated			
Inference method	**50**	**20**	**10**	**6**
Assume iid	0.422*	0.420*	0.404*	0.412*
CRSE, N(0,1) critical vals	0.048	0.061*	0.079*	0.107*
CRSE*sqrt(G/(G-1)), N(0,1)	0.044	0.056	0.075*	0.104*
CRSE*sqrt(G/G-1), t_{G-1}	0.042*	0.046	0.050	0.049
Wild cluster bootstrap-t	0.054	0.055	0.059*	0.062*

Notes:

* Indicates that rejection rate from 5000 Monte Carlo replications is statistically significantly different from 0.05.

Uses sample of CPS data defined and aggregated to state-year level in same way as in Bertrand, Duflo and Mullainathan, except we use data from 1979 to 2009 (rather than 1999). Monte Carlos work in same way as in row 4 of Table 2 of that paper.

Rejection rates with tests of nominal 5% size, for ‘placebo treatments’ with 30 years of CPS earnings data

	Number of groups (US states), half of which are treated			
Inference method	**50**	**20**	**10**	**6**
Assume iid	0.422*	0.420*	0.404*	0.412*
CRSE, N(0,1) critical vals	0.048	0.061*	0.079*	0.107*
CRSE*sqrt(G/(G-1)), N(0,1)	0.044	0.056	0.075*	0.104*
CRSE*sqrt(G/(G-1)), t_{G-1}	**0.042***	**0.046**	**0.050**	**0.049**
Wild cluster bootstrap-t	0.054	0.055	0.059*	0.062*

Notes:

* Indicates that rejection rate from 5000 Monte Carlo replications is statistically significantly different from 0.05.

Uses sample of CPS data defined and aggregated to state-year level in same way as in Bertrand, Duflo and Mullainathan, except we use data from 1979 to 2009 (rather than 1999). Monte Carlos work in same way as in row 4 of Table 2 of that paper.

Checking robustness to the data generating process

- CPS provided one dgp to test methods on - perhaps we got lucky

- To check robustness we simulate our own state-time shocks

$$\lambda_{ct}^{sim} = \hat{\mu}_c + \hat{\xi}_t + \omega_{ct}^{sim}$$

$$\omega_{ct}^{sim} = \rho \omega_{c,t-1}^{sim} + \upsilon_{ct} \quad t = 2, \ldots, 30$$

$$\omega_{c1}^{sim} = (1 - \rho^2)^{-\frac{1}{2}} \upsilon_{c1}$$

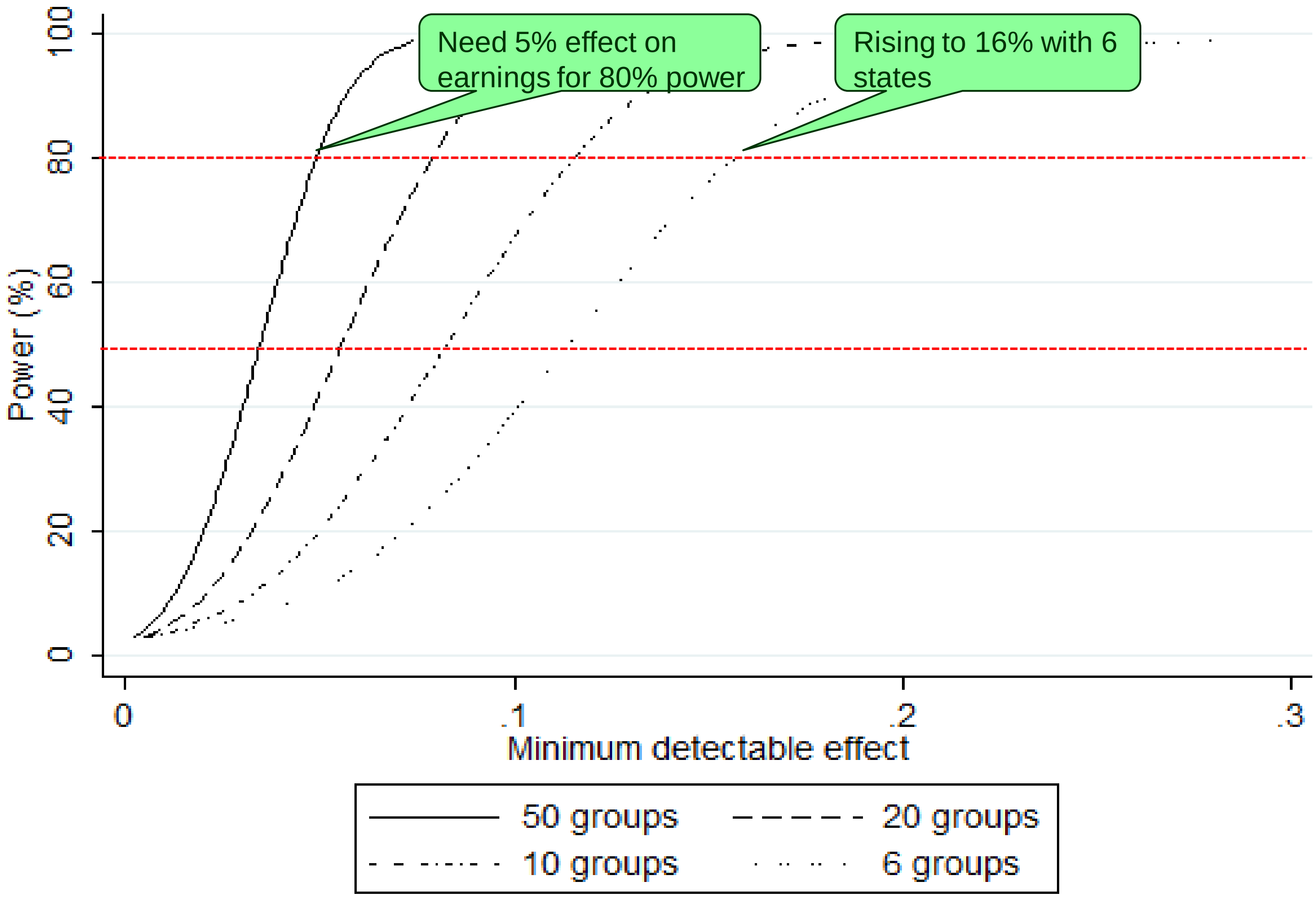
- White noise drawn from t distribution with d degrees of freedom

- In paper we also run simulations using CPS employment outcomes, and all conclusions carry over to that case

Rejection rates under various error processes with 6 groups, using CRSE*sqrt(G/G-1) and t_{G-1} critical values

		AR(1) parameter				
d (controls non-normality in white noise)	0	0.2	0.4	0.6	0.8	Varies by group
2	0.055*	0.058*	0.058*	0.058*	0.052	0.051
4	0.055*	0.058*	0.056*	0.056*	0.051	0.054*
20	0.053	0.059*	0.057*	0.057*	0.051	0.054*
60	0.056*	0.061*	0.058*	0.057*	0.053	0.053
120	0.056*	0.060*	0.057*	0.057*	0.052	0.052

But what about power? Minimum detectable effects on log(earnings) using 5% level hypothesis tests



PEPA Programme Evaluation for Policy Analysis

NCRM National Centre for Research Methods

Increasing power, whilst controlling test size using CRSE*sqrt(G/G-1) and t_{G-1} critical values

	G=50		G=20		G=6	
	No effect	Effect of +0.02 log-points	No effect	Effect of +0.02 log-points	No effect	Effect of +0.02 log-points
OLS, robust	0.042	0.220	0.046	0.118	0.049	0.073

Note: FGLS is implemented assuming an AR(2) process for the state-time shocks. For the BC-FGLS procedure, see Hansen (2007).

Increasing power, whilst controlling test size using CRSE*sqrt(G/G-1) and t_{G-1} critical values

	G=50		G=20		G=6	
	No effect	Effect of +0.02 log-points	No effect	Effect of +0.02 log-points	No effect	Effect of +0.02 log-points
OLS, robust	0.042	0.220	0.046	0.118	0.049	0.073
FGLS	0.100	0.460	0.106	0.275	0.126	0.191

Note: FGLS is implemented assuming an AR(2) process for the state-time shocks. For the BC-FGLS procedure, see Hansen (2007).

Increasing power, whilst controlling test size using CRSE*sqrt(G/G-1) and t_{G-1} critical values

	G=50		G=20		G=6	
	No effect	Effect of +0.02 log-points	No effect	Effect of +0.02 log-points	No effect	Effect of +0.02 log-points
OLS, robust	0.042	0.220	0.046	0.118	0.049	0.073
FGLS	0.100	0.460	0.106	0.275	0.126	0.191
FGLS, robust	0.047	0.348	0.053	0.175	0.061	0.096

Note: FGLS is implemented assuming an AR(2) process for the state-time shocks. For the BC-FGLS procedure, see Hansen (2007).

Increasing power, whilst controlling test size using CRSE*sqrt(G/G-1) and t_{G-1} critical values

	G=50		G=20		G=6	
	No effect	Effect of +0.02 log-points	No effect	Effect of +0.02 log-points	No effect	Effect of +0.02 log-points
OLS, robust	0.042	0.220	0.046	0.118	0.049	0.073
FGLS	0.100	0.460	0.106	0.275	0.126	0.191
FGLS, robust	0.047	0.348	0.053	0.175	0.061	0.096
BC-FGLS	0.068	0.395	0.077	0.224	0.099	0.150

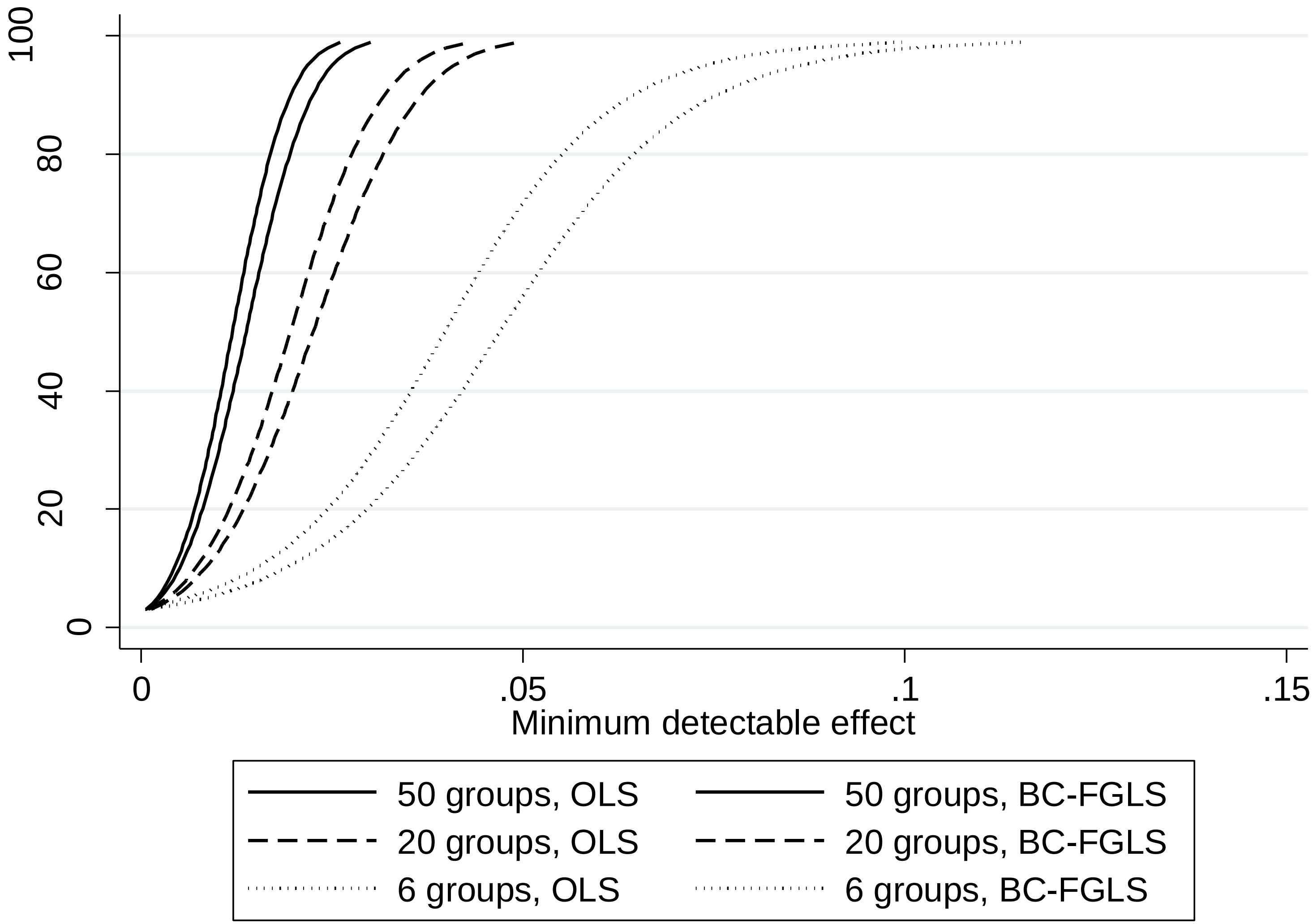
Note: FGLS is implemented assuming an AR(2) process for the state-time shocks. For the BC-FGLS procedure, see Hansen (2007).

Increasing power, whilst controlling test size using CRSE*sqrt(G/G-1) and t_{G-1} critical values

	G=50		G=20		G=6	
	No effect	Effect of +0.02 log-points	No effect	Effect of +0.02 log-points	No effect	Effect of +0.02 log-points
OLS, robust	0.042	0.220	0.046	0.118	0.049	0.073
FGLS	0.100	0.460	0.106	0.275	0.126	0.191
FGLS, robust	0.047	0.348	0.053	0.175	0.061	0.096
BC-FGLS	0.068	0.395	0.077	0.224	0.099	0.150
BC-FGLS, robust	0.049	0.365	0.057	0.187	0.064	0.103

Note: FGLS is implemented assuming an AR(2) process for the state-time shocks. For the BC-FGLS procedure, see Hansen (2007).

Minimum detectable effects on log(earnings) using 5% level hypothesis tests: OLS vs BC-FGLS estimation



Summary and conclusions

- Literature is right that DiD designs can pose problems for inference
- But we find that correct test size can be achieved, even with few groups, using very straightforward methods
- Key problem is low power
- We therefore recommend that researchers think seriously about the efficiency of DiD estimation (not just consistency and test size)
- We have shown how FGLS combined with robust inference can help significantly, *without* compromising test size, even with *few groups*