NCRM National Centre for Research Methods

PEPA Programme Evaluation for Policy Analysis

Inference with difference-in-differences revisited

Robert Joyce (Institute for Fiscal Studies)

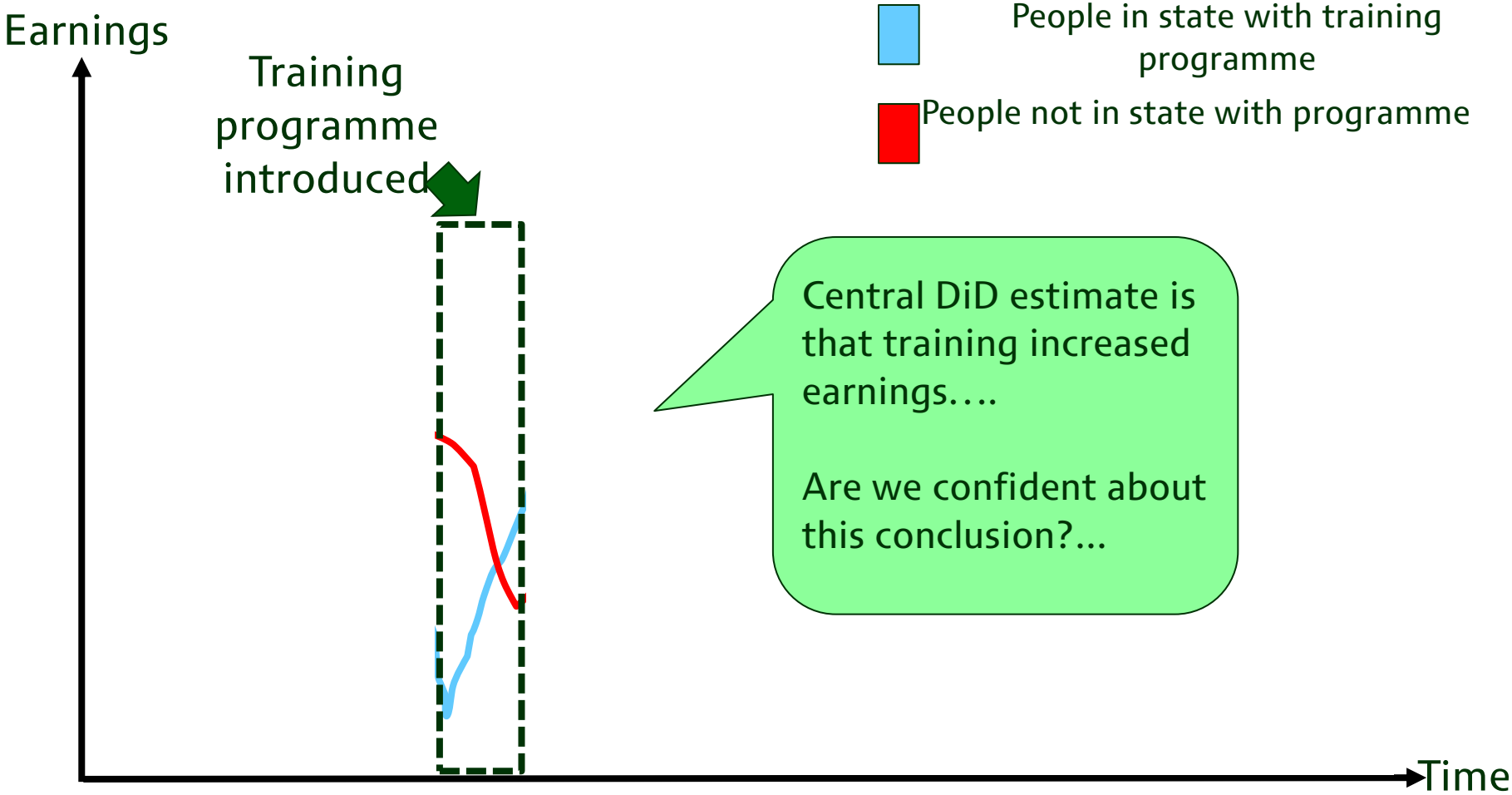
Joint work with Mike Brewer (Essex, IFS) and Thomas Crossley (Essex, IFS)

PEPA is based at the IFS and CEMMAP

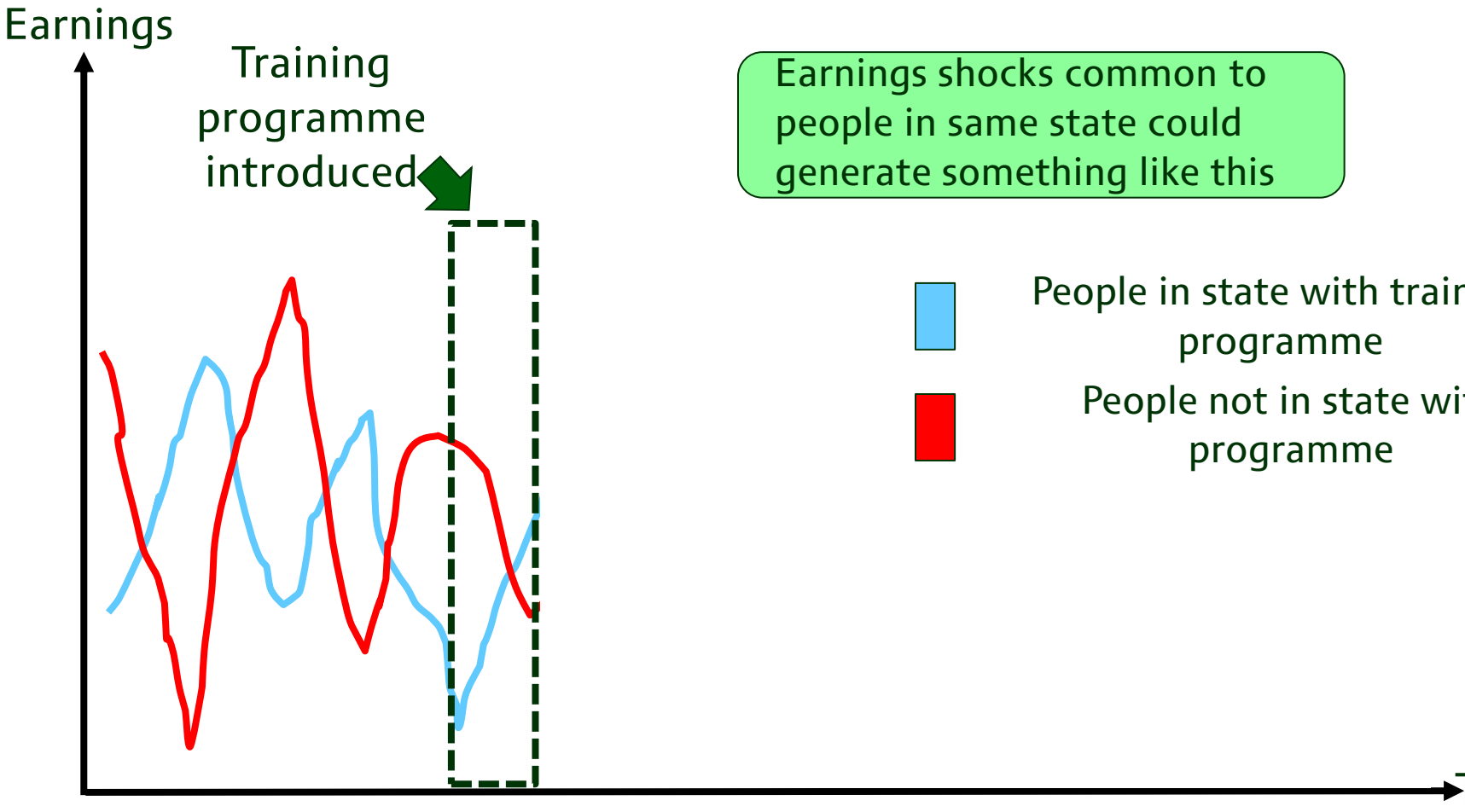
Institute for Fiscal Studies

cemmap centre for microdata methods and practice

Inference in DiD



One reason not to be so confident

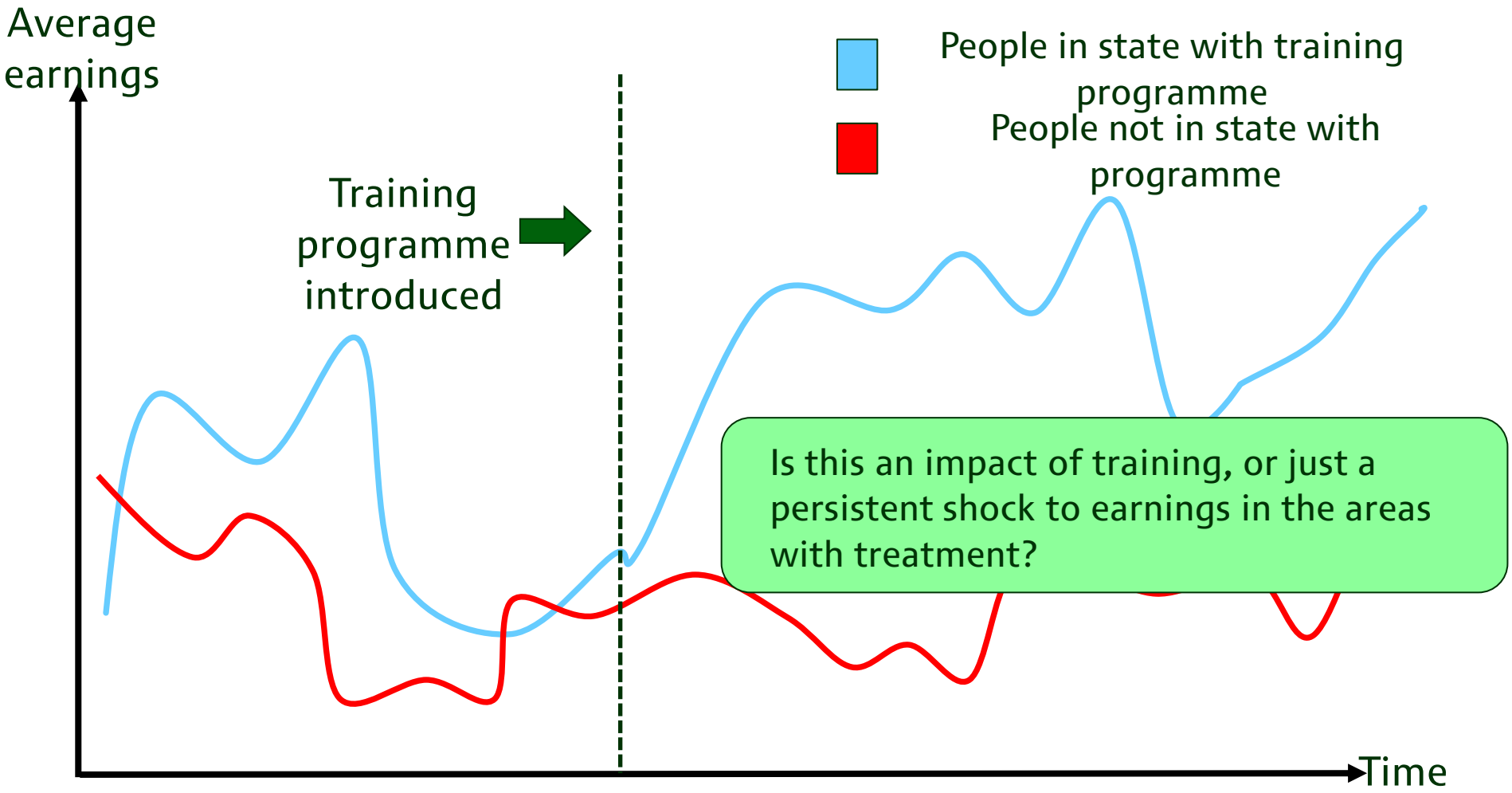


Earnings shocks common to people in same state could generate something like this

People in state with training programme

People not in state with programme

Serial correlation (another reason not to be so confident)



Introduction

- Emerging literature on inference in common DiD designs
- Main focus has been on test size when
  - Grouped errors and no variation in treatment status within group-time cells (Moulton 1990; Donald and Lang 2007)
  - Serially correlated errors and treatments within groups (Bertrand et al 2004; Hansen 2007)
- Standard solution would be 'cluster-robust' standard errors (CRSEs) but asymptotics apply as $G \to \infty$
  - With small G, wild cluster bootstrap-t the leading method (Cameron et al 2008)
  - Alternatives include small-G adjustments to CRSE-based inference (e.g. scaling SEs; using t reference dist.), but literature hasn't found they work well

With Monte Carlo simulations we make 3 points

1. Test size need not be a concern

2. Problem is low power to detect real effects

3. FGLS combined with robust inference can help a lot

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 $\hat{\beta}_{OLS}$ equivalent to first running this regression…

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

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

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

- If cell sizes are large, true precision of $\hat{\beta}_{OLS}$ depends almost entirely on # of group-time cells (not observations)

Monte Carlo experiments

- Use women's log-earnings from CPS (1979-2008), as in Bertrand et al (2004), Cameron et al (2008), Hansen (2007)

- Collapse to state-year level using covariate-adjusted means

- Repeat the following 5000 times, varying G from 6 to 50:
  - Sample G states at random with replacement
  - Randomly choose some (initially 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
- Then count how often null was rejected (out of 5000)

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**

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	0.429*	0.424*	0.422*	0.413*

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	0.429*	0.424*	0.422*	0.413*
CRSE, N(0,1) critical vals	0.059*	0.073*	0.110*	0.175*

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	0.429*	0.424*	0.422*	0.413*
CRSE, N(0,1) critical vals	0.059*	0.073*	0.110*	0.175*
CRSE*sqrt(G/(G-1)), t_{G-1}	0.045	0.041*	0.042*	0.052

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	0.429*	0.424*	0.422*	0.413*
CRSE, N(0,1) critical vals	0.059*	0.073*	0.110*	0.175*
CRSE*sqrt(G/(G-1)), t_{G-1}	0.045	0.041*	0.042*	0.052
Wild cluster bootstrap-t	0.044	0.041*	0.048	0.059*

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	0.429*	0.424*	0.422*	0.413*
CRSE, N(0,1) critical vals	0.059*	0.073*	0.110*	0.175*
CRSE*sqrt(G/(G-1)), t_{G-1}	**0.045**	**0.041***	**0.042***	**0.052**
Wild cluster bootstrap-t	**0.044**	**0.041***	**0.048**	**0.059***

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.

Alternative data generating processes

- To check robustness we simulate our own state-time shocks
- In doing so we vary degree of serial correlation and non-normality

$$\omega_{ct}^{sim} = \rho\omega_{c,t-1}^{sim} + f^1(d)\upsilon_{ct} \quad t = 2,\ldots,30$$
$$\omega_{c1}^{sim} = f^2(d)\upsilon_{c1}$$

- Error term generated by AR(1) process with parameter ρ
- White noise υ_{ct} drawn from t distribution with d degrees of freedom
- f^1 and f^2 scale white noise so that, with $\rho = 0.4$, variance of state-time shocks in each period equals that in the CPS

Rejection rates under various error processes with 10 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
4	0.054	0.052	0.049	0.051	0.056*	0.053
20	0.051	0.050	0.050	0.049	0.049	0.051
60	0.053	0.050	0.050	0.047	0.053	0.053
120	0.052	0.051	0.053	0.053	0.054	0.055*

Unbalanced designs (back to the CPS data)

	Number of treated states out of 10			
	5	4	3	2
CRSE*sqrt(G/(G-1)), t_{G-1}	0.042*	0.051	0.074*	0.150*
Wild cluster bootstrap-t	0.048	0.054	0.052	0.018*

Unbalanced designs (back to the CPS data)

	Number of treated states out of 10			
	5	4	3	2
CRSE*sqrt(G/(G-1)), t_{G-1}	0.042*	0.051	0.074*	0.150*
Wild cluster bootstrap-t	0.048	0.054	0.052	0.018*
	Number of treated states out of 50			
	25	15	10	5
CRSE*sqrt(G/(G-1)), t_{G-1}	0.045	0.052	0.060*	0.119*
Wild cluster bootstrap-t	0.044	0.051	0.046	0.060*

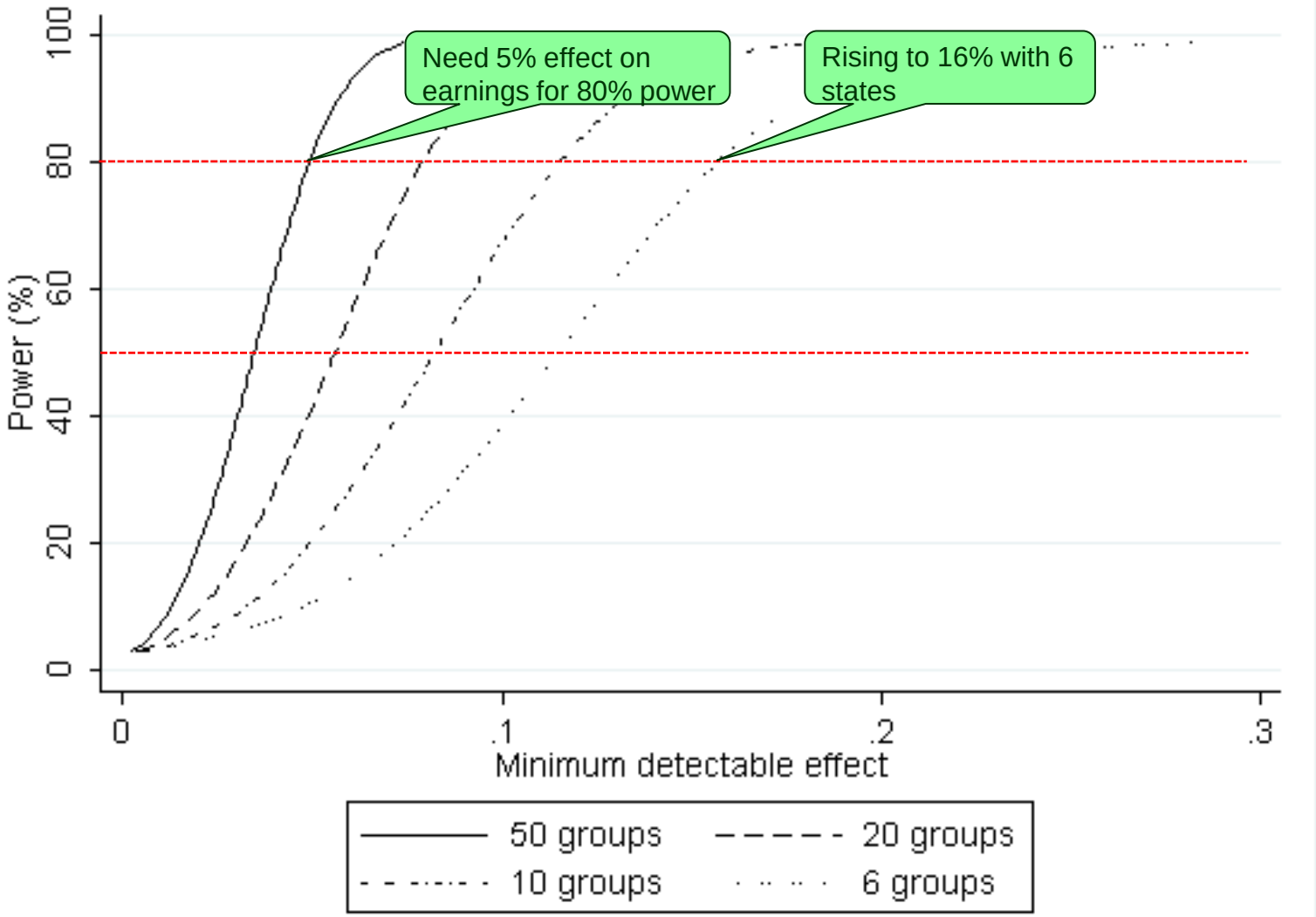
BUT WHAT ABOUT POWER?

But what about power?

	Number of groups (US states), half of which are treated			
	50	20	10	6
Effect on log-earn = 0.02				
CRSE*sqrt(G/(G-1)), t_{G-1}	0.238	0.134	0.088	0.074
Wild cluster bootstrap-t	0.225	0.125	0.093	0.074
Effect on log-earn = 0.05				
CRSE*sqrt(G/(G-1)), t_{G-1}	0.822	0.513	0.273	0.168
Wild cluster bootstrap-t	0.799	0.490	0.283	0.167
Effect on log-earn = 0.10				
CRSE*sqrt(G/(G-1)), t_{G-1}	1.000	0.919	0.718	0.448
Wild cluster bootstrap-t	0.999	0.898	0.712	0.429
Effect on log-earn = 0.15				
CRSE*sqrt(G/(G-1)), t_{G-1}	1.000	0.995	0.904	0.755
Wild cluster bootstrap-t	1.000	0.992	0.896	0.700

Note:
Following Davidson and Mackinnon (1998), the nominal significance level used to determine whether to reject the null hypothesis is that which gives a test of true size 0.05. This nominal significance level is obtained from the 5th percentile of the empirical distribution of p-values from Monte Carlo simulations under a true null.

Minimum detectable effects on log(earnings) using CRSE*sqrt(G/G-1) and t_{G-1} critical values, 5% level tests



PEPA Programme Evaluation for Policy Analysis

NCRM National Centre for Research Methods

Increasing power using feasible GLS

	G=50		G=20		G=6	
	No effect	Effect of +0.05 log-points	No effect	Effect of +0.05 log-points	No effect	Effect of +0.05 log-points
OLS, robust	0.045	0.810	0.041	0.467	0.052	0.168

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

Increasing power using feasible GLS

	G=50		G=20		G=6	
	No effect	Effect of +0.05 log-points	No effect	Effect of +0.05 log-points	No effect	Effect of +0.05 log-points
OLS, robust	0.045	0.810	0.041	0.467	0.052	0.168
FGLS	0.106	0.985	0.101	0.799	0.124	0.434

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

Increasing power using feasible GLS

	G=50		G=20		G=6	
	No effect	Effect of +0.05 log-points	No effect	Effect of +0.05 log-points	No effect	Effect of +0.05 log-points
OLS, robust	0.045	0.810	0.041	0.467	0.052	0.168
FGLS	0.106	0.985	0.101	0.799	0.124	0.434
FGLS, robust	0.049	0.957	0.045	0.670	0.061	0.255

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

Increasing power using feasible GLS

	G=50		G=20		G=6	
	No effect	Effect of +0.05 log-points	No effect	Effect of +0.05 log-points	No effect	Effect of +0.05 log-points
OLS, robust	0.045	0.810	0.041	0.467	0.052	0.168
FGLS	0.106	0.985	0.101	0.799	0.124	0.434
FGLS, robust	0.049	0.957	0.045	0.670	0.061	0.255
BC-FGLS	0.073	0.978	0.070	0.763	0.096	0.384

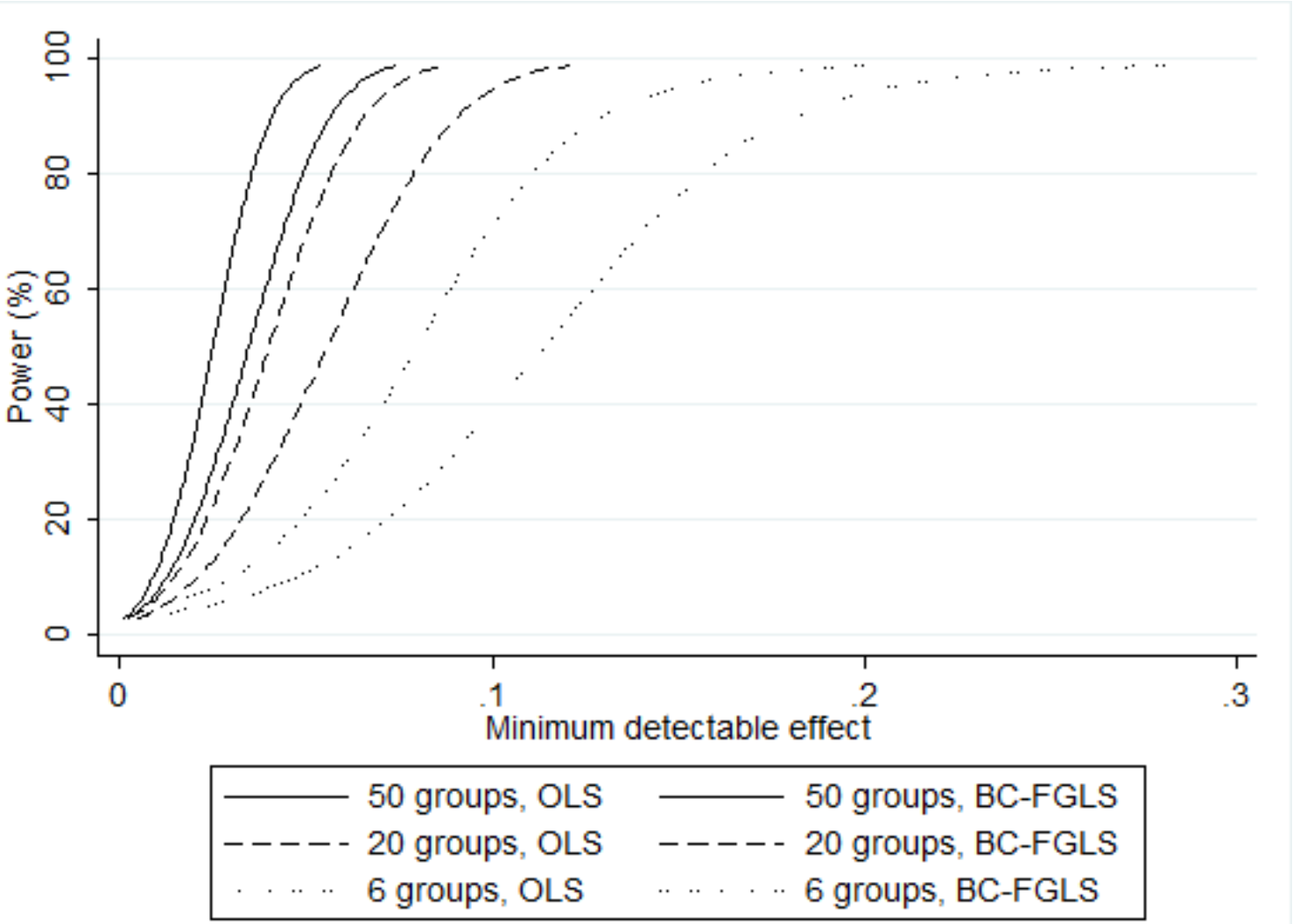
Note: FGLS is implemented assuming an AR(2) process for the state-time shocks. For the BC-FGLS procedure, see Hansen (2007).

Increasing power using feasible GLS

	G=50		G=20		G=6	
	No effect	Effect of +0.05 log-points	No effect	Effect of +0.05 log-points	No effect	Effect of +0.05 log-points
OLS, robust	0.045	0.810	0.041	0.467	0.052	0.168
FGLS	0.106	0.985	0.101	0.799	0.124	0.434
FGLS, robust	0.049	0.957	0.045	0.670	0.061	0.255
BC-FGLS	0.073	0.978	0.070	0.763	0.096	0.384
BC-FGLS, robust	0.049	0.955	0.045	0.696	0.065	0.286

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



PEPA Programme Evaluation for Policy Analysis

NCRM National Centre for Research Methods

More on FGLS

- In paper we look at power gains and size properties under mis-specified error processes and varying time dimension

- If process severely misspecified (if it's really MA(1)), there's no power gain, but size can still be controlled (even with few groups)
- If less severely misspecified (if it's really heterogeneous AR(2)), still big power gains
- Power improvement noticeable as long as T>=10

- Punchline: BC-FGLS may have big benefits (i.e. higher power); and it won't hurt you (i.e. you can control size), even if parametric assumptions about error process are wrong and G is small

Summary and conclusions

- Literature is right that DiD designs can pose problems for inference
- But controlling test size need not be big problem
- Key problem is low power
- We therefore recommend that researchers think seriously about the efficiency of DiD estimation (not just consistency and test size)
- BC-FGLS combined with robust inference can help significantly, *without* compromising test size, even with *few groups*