Inference (and power) with difference-in-differences

Mike Brewer (Essex and IFS/PEPA)
Thomas Crossley (Essex and IFS/PEPA)
Robert Joyce (IFS/PEPA)

Overview

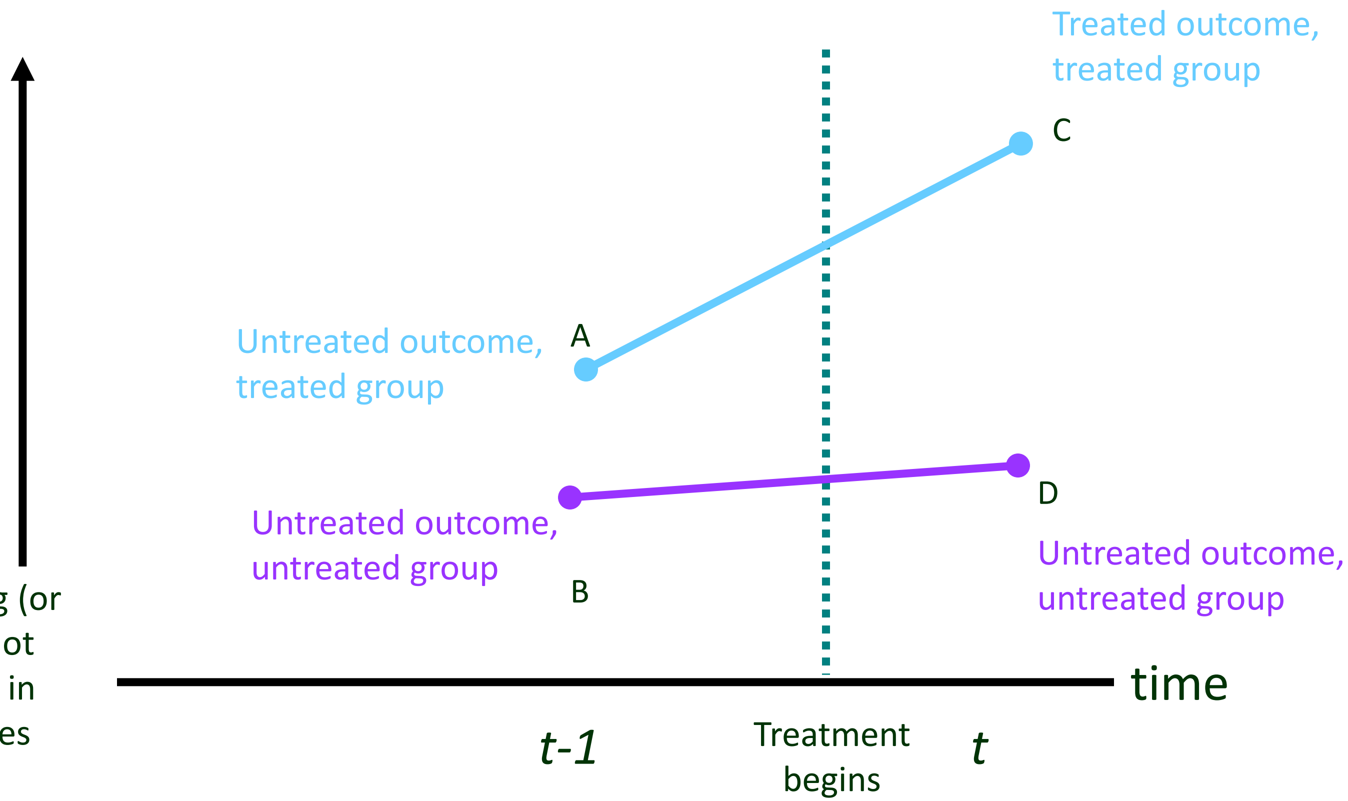
- Difference-in-differences (DiD) is a common approach to take to estimate the causal impact of a policy intervention, used frequently to exploit “natural experiments”
- Recent literature suggests DiD designs can pose big problems for inference (researchers falsely concluding policies are having an effects)
- Using Monte Carlo evidence, we show
  - controlling test size in DiD need not be big problem; key problem is low power
  - BC-FGLS combined with robust inference can help significantly

What is the difference-in-differences approach?

- A difference-in-differences (DiD) approach seeks to estimate causal impact of a policy intervention
- Usually have:
  - a treatment group (individuals exposed to treatment)
  - a comparison group (individuals not exposed to treatment)
- DiD usually used when:
  - we suspect untreated outcomes for treatment and comparison groups are different, even after matching (i.e. unconfoundedness does not hold; selection is on unobservables)
  - we have data from time when both groups are untreated
    - NB doesn't have to be the same individuals; DiD is more general than using longitudinal data

The difference-in-difference estimator

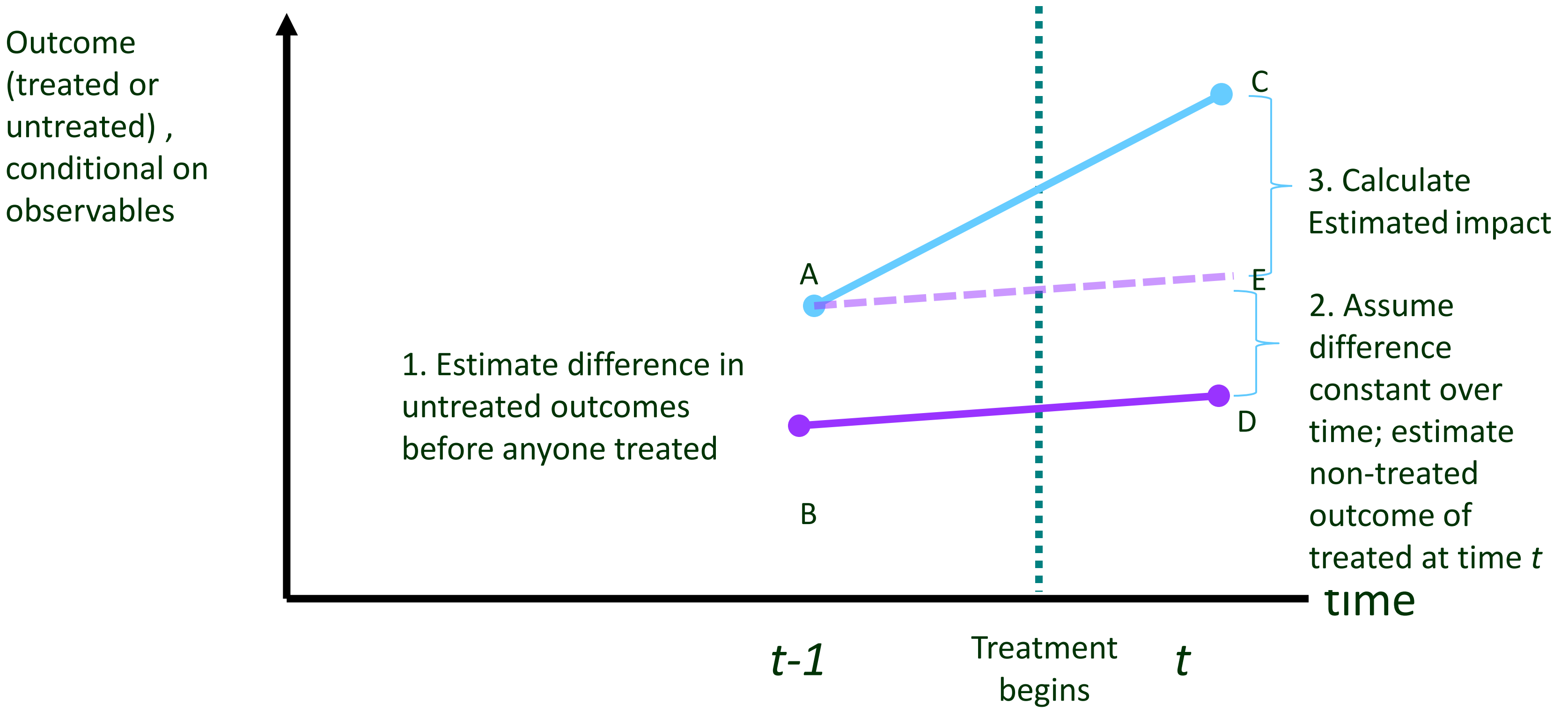
Outcome (treated or untreated) , conditional on observables



Problem: matching (or equivalent) does not remove difference in untreated outcomes

Solution: use data from period when both groups untreated

The difference-in-difference estimator



Generalise to many periods and many groups:

$$Y_{ict} = \alpha + \beta T_{ct} + \delta X_{ict} + \mu_c + \xi_t + u_{ict}$$

$$E(u_{ict} \mid T_{ct}, X_{ict}, \mu_c, \xi_t) = 0$$

(where $c \geq 2$ indexes groups, $t \geq 2$ indexes time, and T_{ct} is indicator for treatment)

Two non-standard error issues:

1. errors may be correlated within group, e.g. $u_{ict} = \eta_{ct} + \varepsilon_{ict}$
2. errors may be serially correlated.

These cause issues for inference as T_{ct} also (perfectly) correlated within groups, and (highly) serially-correlated

$$Y_{ict} = \alpha + \beta T_{ct} + \delta X_{ict} + \mu_c + \xi_t + u_{ict}$$

$$E(u_{ict} \mid T_{ct}, X_{ict}, \mu_c, \xi_t) = 0$$

$$u_{ict} = \eta_{ct} + \varepsilon_{ict}$$

Convenient approach is the two-step:

A: Partial out individual-level controls by regressing on individual-level controls and full set of group-time dummies

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

B. Regress estimated group-time dummies $\hat{\lambda}_{ct}$ on group dummies, time dummies and treatment dummy

$$\hat{\lambda}_{ct} = \alpha + \beta T_{ct} + \mu_c + \xi_t + \left(\eta_{ct} + \left(\hat{\lambda}_{ct} - \lambda_{ct}\right)\right)$$

Problem: how to do inference on β given serial correlation in error term

1. "Cluster-robust" standard errors (CRSEs)

Can take commonly-used formula for the covariance matrix that is robust to clustered errors of an arbitrary form (Liang and Zeger, 1986)

$$\hat{V}_{LZ} = (X'X)^{-1}\left(\sum_{c=1}^{C} X_c u_c u_c' X_c'\right)(X'X)^{-1}$$

- …so if you cluster at the group level (***not*** group-time level), you also allow for serial correlation within groups

But consistency of CRSEs applies as # clusters gets large, and number of clusters in typical DiD applications can be small

NB:

- Common to scale residuals by sqrt(G/(G-1)) before plugging into CRSE formula. Exact theoretical validity only under special circumstances. Stata does this (almost).
- We implement variant where we scale residuals AND compare resulting *t*-statistic to critical values from *t*(*G*-1) distribution (rather than *N*(0,1)). Stata does this with "regress", but not other commands.

2. FGLS

$$\hat{\lambda}_{ct} = \alpha + \beta T_{ct} + \mu_c + \xi_t + \left(\eta_{ct} + \left(\hat{\lambda}_{ct} - \lambda_{ct}\right)\right)$$

- Hansen (2007) proposes FGLS estimation having assumed errors follow an auto-regressive (AR) process

Aside: feasible GLS

In our case:

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

$$\eta_{ct} = \rho_1 \eta_{ct-1} + \rho_2 \eta_{ct-2} + \varepsilon_{ct} \text{ with } \varepsilon_{ct} \text{ serially uncorrelated}$$

Consider transformed model:

$$\begin{aligned}\hat{\lambda}_{ct} - \rho_1 \hat{\lambda}_{ct-1} - \rho_2 \hat{\lambda}_{ct-2} = & (\alpha + \mu_c)(1 - \rho_1 - \rho_2) \\ & + \beta (T_{ct} - \rho_1 T_{ct-1} - \rho_2 T_{ct-2}) + (\xi_t - \rho_1 \xi_{t-1} - \rho_2 \xi_{t-2}) \\ & + (\eta_{ct} - \rho_1 \eta_{ct-1} - \rho_2 \eta_{ct-2})\end{aligned}$$

This allows OLS since:

$$\eta_{ct} - \rho_1 \eta_{ct-1} - \rho_2 \eta_{ct-2} = \varepsilon_{ct} \text{ is serially uncorrelated}$$

In practice, estimate OLS of:

$$\begin{aligned}\hat{\lambda}_{ct} - \hat{\rho}_1 \hat{\lambda}_{ct-1} - \hat{\rho}_2 \hat{\lambda}_{ct-2} = & (\alpha + \mu_c)(1 - \hat{\rho}_1 - \hat{\rho}_2) \\ & + \beta (T_{ct} - \hat{\rho}_1 T_{ct-1} - \hat{\rho}_2 T_{ct-2}) + (\xi_t - \hat{\rho}_1 \xi_{t-1} - \hat{\rho}_2 \xi_{t-2}) \\ & + (\eta_{ct} - \hat{\rho}_1 \eta_{ct-1} - \hat{\rho}_2 \eta_{ct-2})\end{aligned}$$

2. FGLS

$$\hat{\lambda}_{ct} = \alpha + \beta T_{ct} + \mu_c + \xi_t + \left(\eta_{ct} + \left(\hat{\lambda}_{ct} - \lambda_{ct}\right)\right)$$

- Hansen (2007) proposes FGLS estimation having assumed errors follow an auto-regressive (AR) process
- Limitations:
  - Need an assumption on nature of serial correlation (as with all FGLS)
  - Estimate of AR parameter(s) biased because of fixed group effects and fixed *T;* Hansen derives a bias correction, but this is consistent as *G* goes to infinity (or becomes vanishingly small relative to *T*)
- We implement Hansen's method, but also implement variant where we allow for CRSEs even after FGLS has "removed" serial correlation

$$\hat{V}_{BC-FGLS-ROBUST} = (\mathbf{X}'\hat{\mathbf{\Omega}}^{-1}\mathbf{X})^{-1}(\sum_{c=1}^{C}\mathbf{X}_\mathbf{c}\hat{\mathbf{\Omega}}_c^{-1}\mathbf{u}_\mathbf{c}\mathbf{u}_\mathbf{c}'\hat{\mathbf{\Omega}}_c^{-1}\mathbf{X}_\mathbf{c}')(\mathbf{X}'\hat{\mathbf{\Omega}}^{-1}\mathbf{X})^{-1}$$

3. Wild cluster bootstrap-t

- Cameron et al (2008) suggests calculating the t-statistic using (inconsistent-with-fixed-*G*) CRSEs, and then using a cluster version of the wild bootstrap (aka "block bootstrap) to get p-values

- Implementation:
  i. repeatedly re-sample with replacement clusters (groups) of data, and re-compute (inconsistent-with-fixed-*G*) *t*-statistic each time
  ii. Compare original (inconsistent-with-fixed-*G*) t-statistic to empirical distribution of (inconsistent-with-fixed-*G*) t-statistics to get *p*-values

- Note:
  – Resampling scheme at (i) imposes the null hypothesis
  – Method robust to arbitrary heteroscedasticity and serial correlation within groups/clusters

With Monte Carlo simulations we make these points:

1. Test size is not the primary concern
   - Wild cluster bootstrap works in most cases, and CRSEs with t distribution works just as well, except where small fraction of G are (not) treated
2. A more pressing problem is the low power of DiD to detect genuine effects
3. BC-FGLS combined with robust inference can help a lot, especially with high T

Monte Carlo experiments

- Use data on women's log-earnings based on repeated cross-sections 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 15,000 times, varying G from 6 to 50:
  - Randomly choose G states with replacement
  - Randomly choose some (initially $G/2$) states to be 'treated'
  - Randomly choose a year from which 'treated' states will be treated
  - Estimate (non-existent) 'treatment effect'
  - Test (true) null of 'no effect' using nominal 5%-level test
- Report how often null is rejected (over 15,000 replications)

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 15,000 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 15,000 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 15,000 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 15,000 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 15,000 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.

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}				
Wild cluster bootstrap-t				
Effect on log-earn = 0.05				
CRSE*sqrt(G/(G-1)), t_{G-1}				
Wild cluster bootstrap-t				
Effect on log-earn = 0.10				
CRSE*sqrt(G/(G-1)), t_{G-1}				
Wild cluster bootstrap-t				
Effect on log-earn = 0.15				
CRSE*sqrt(G/(G-1)), t_{G-1}				
Wild cluster bootstrap-t				

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.

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			
Wild cluster bootstrap-t	0.225			
Effect on log-earn = 0.05				
CRSE*sqrt(G/(G-1)), t_{G-1}	0.822			
Wild cluster bootstrap-t	0.799			
Effect on log-earn = 0.10				
CRSE*sqrt(G/(G-1)), t_{G-1}	1.000			
Wild cluster bootstrap-t	0.999			
Effect on log-earn = 0.15				
CRSE*sqrt(G/(G-1)), t_{G-1}	1.000			
Wild cluster bootstrap-t	1.000			

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.

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		
Wild cluster bootstrap-t	0.225	0.125		
Effect on log-earn = 0.05				
CRSE*sqrt(G/(G-1)), t_{G-1}	0.822	0.513		
Wild cluster bootstrap-t	0.799	0.490		
Effect on log-earn = 0.10				
CRSE*sqrt(G/(G-1)), t_{G-1}	1.000	0.919		
Wild cluster bootstrap-t	0.999	0.898		
Effect on log-earn = 0.15				
CRSE*sqrt(G/(G-1)), t_{G-1}	1.000	0.995		
Wild cluster bootstrap-t	1.000	0.992		

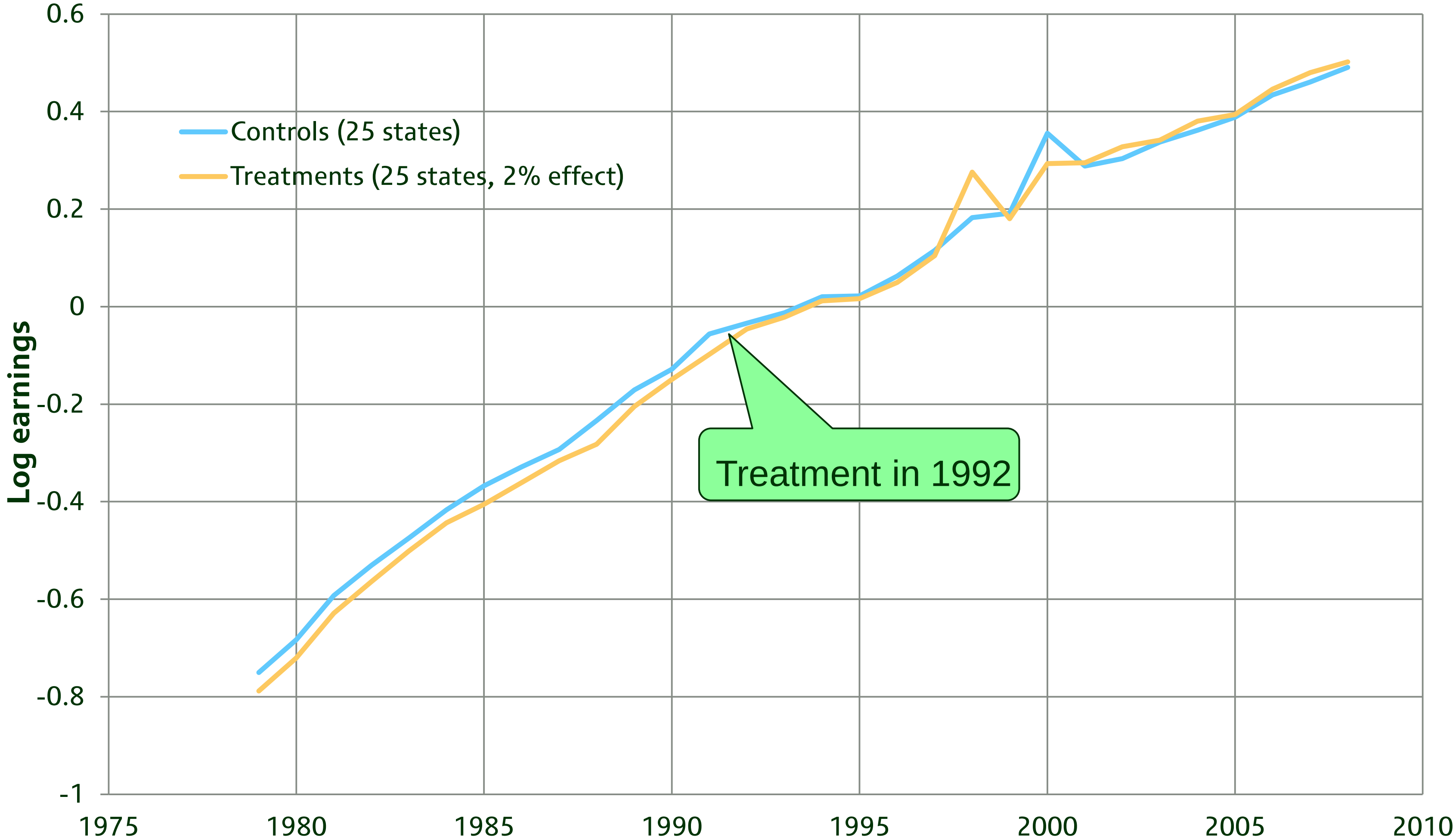
Note:
Following Davidson and Mackinnon (1998), the nominal significance level used to determine whether to reject the null hypothesis is that which gives a test of true size 0.05. This nominal significance level is obtained from the 5th percentile of the empirical distribution of p-values from Monte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

But what about power?

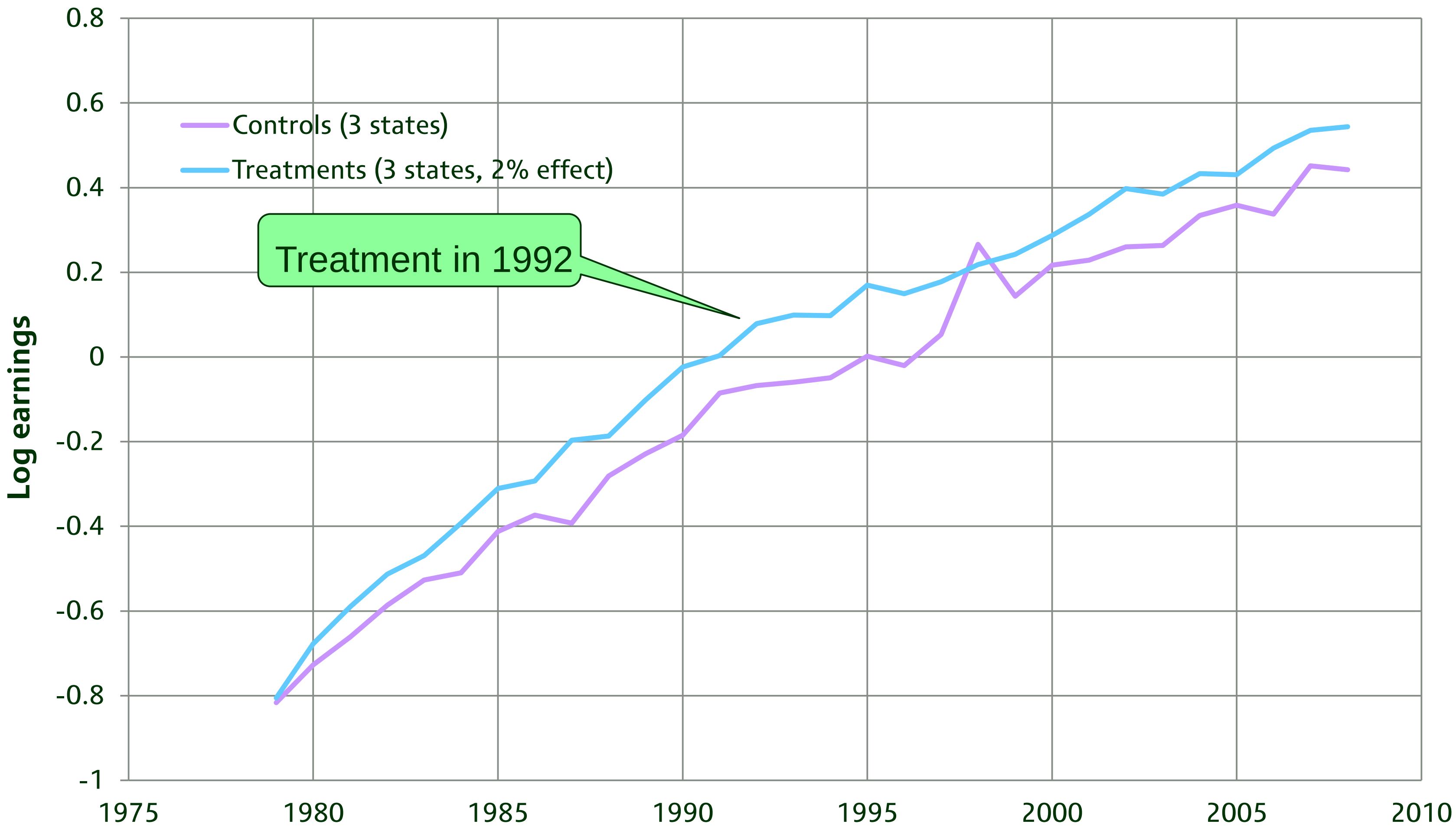
	Number of groups (US states), half of which are treated			
	50	20	10	6
Effect on log-earn = 0.02				
CRSE*sqrt(G/(G-1)), t_{G-1}	0.238	0.134	0.088	0.074
Wild cluster bootstrap-t	0.225	0.125	0.093	0.074
Effect on log-earn = 0.05				
CRSE*sqrt(G/(G-1)), t_{G-1}	0.822	0.513	0.273	0.168
Wild cluster bootstrap-t	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.

Simulated time series of log(earnings) for treatments and controls, with 2% treatment effect on earnings



Simulated time series of log(earnings) for treatments and controls, with 2% treatment effect on earnings



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		0.041		0.052	
FGLS						
FGLS, robust						
BC-FGLS						
BC-FGLS, robust						

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		0.041		0.052	
FGLS	0.106		0.101		0.124	
FGLS, robust	0.049		0.045		0.061	
BC-FGLS						
BC-FGLS, robust						

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		0.041		0.052	
FGLS	0.106		0.101		0.124	
FGLS, robust	0.049		0.045		0.061	
BC-FGLS	0.073		0.070		0.096	
BC-FGLS, robust	0.049		0.045		0.065	

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, robust	0.049	0.957	0.045	0.670	0.061	0.255
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).

FGLS under misspecification of error process (10 groups)

	Heterogeneous AR(2)		MA(1)	
	No effect	Effect of +0.05 log-points	No effect	Effect of +0.05 log-points
OLS, robust	0.041	0.536	0.052	0.597
FGLS, robust	0.055	0.703	0.053	0.580
BC-FGLS, robust	0.058	0.717	0.053	0.578

Note: FGLS is implemented assuming an AR(2) process for the state-time shocks. For the BC-FGLS procedure, see Hansen (2007). For the heterogeneous AR(2) process, the coefficient on the first lag (alpha) is drawn from a uniform distribution between zero and one for each state. The coefficient on the second lag is set equal to 0.5*min(alpha,1-alpha), which ensures stationarity. The MA(1) process has a lag parameter of 0.5. For both processes, the white noise is normally distributed. Its variance ensures that the error term has the same stationary variance as the log-earnings residuals in the CPS (0.04).

FGLS with varying panel length (10 groups)

	T=30		T=20		T=10	
	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	0.044	0.280	0.049	0.282	0.041	0.346
FGLS, robust	0.051	0.401	0.052	0.352	0.046	0.328
BC-FGLS, robust	0.054	0.419	0.055	0.367	0.046	0.327

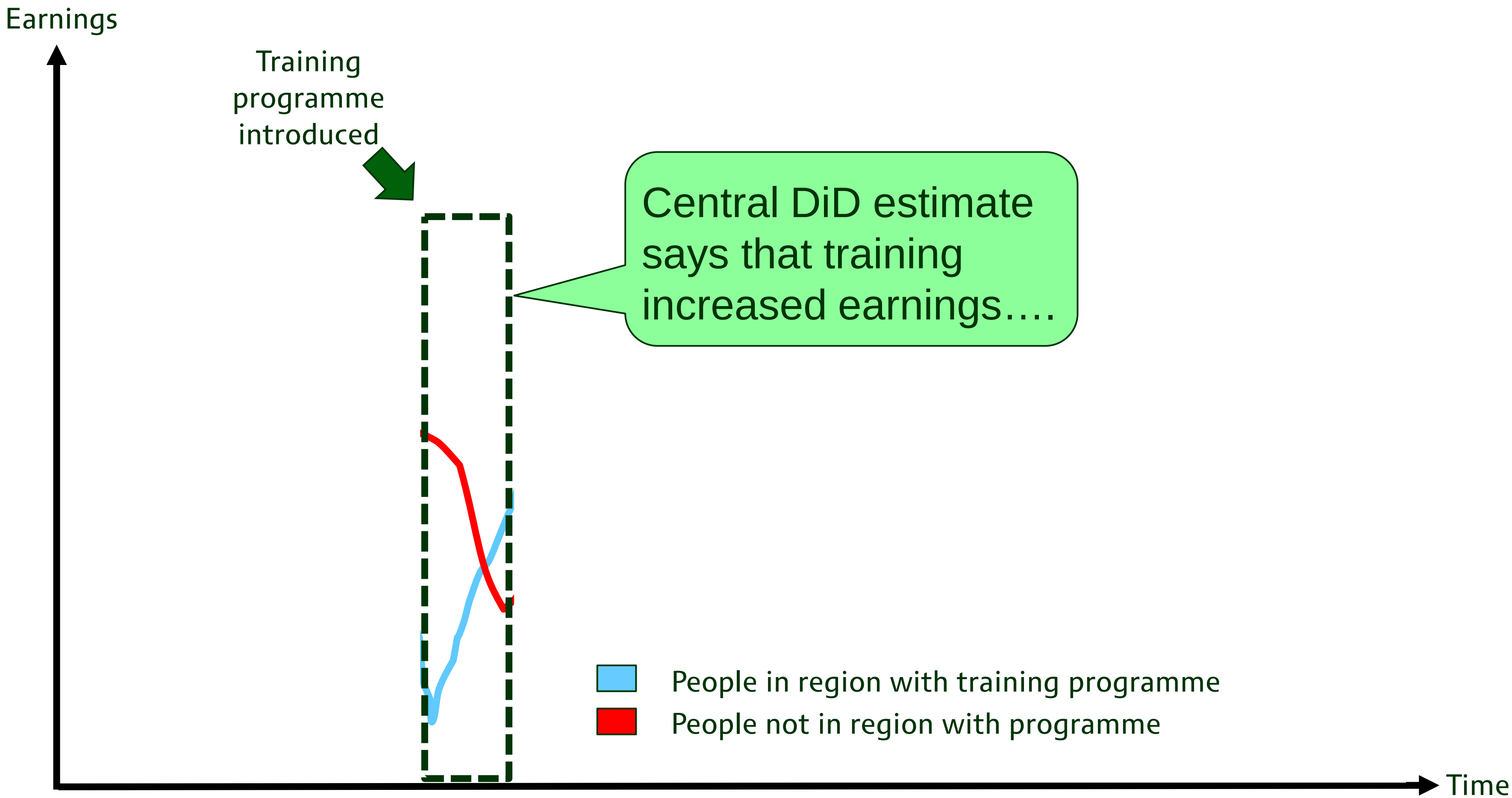
Note: FGLS is implemented assuming an AR(2) process for the state-time shocks. For the BC-FGLS procedure, see Hansen (2007).

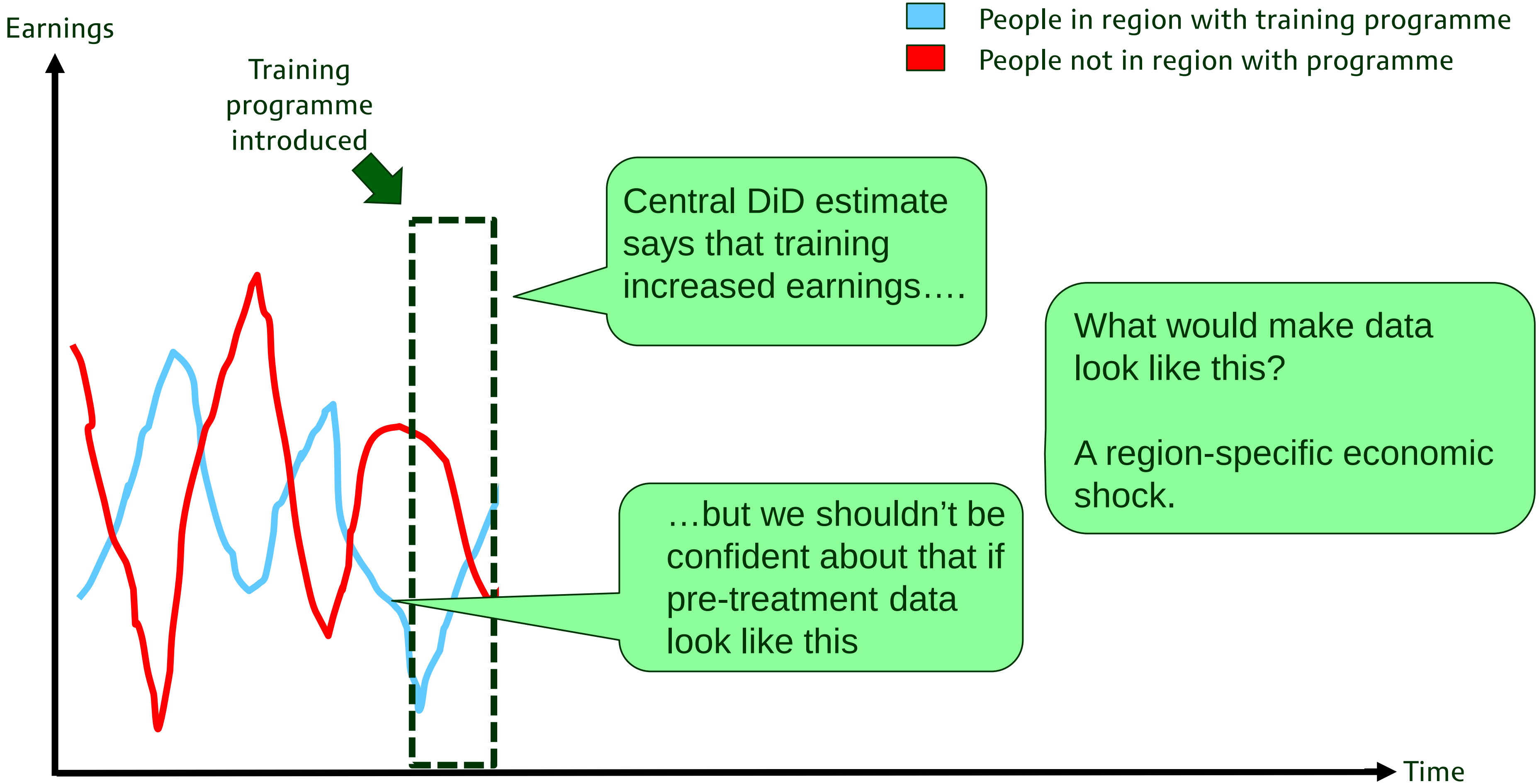
Summary and conclusions

- 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,* with power gain over CRSEs increasing in T

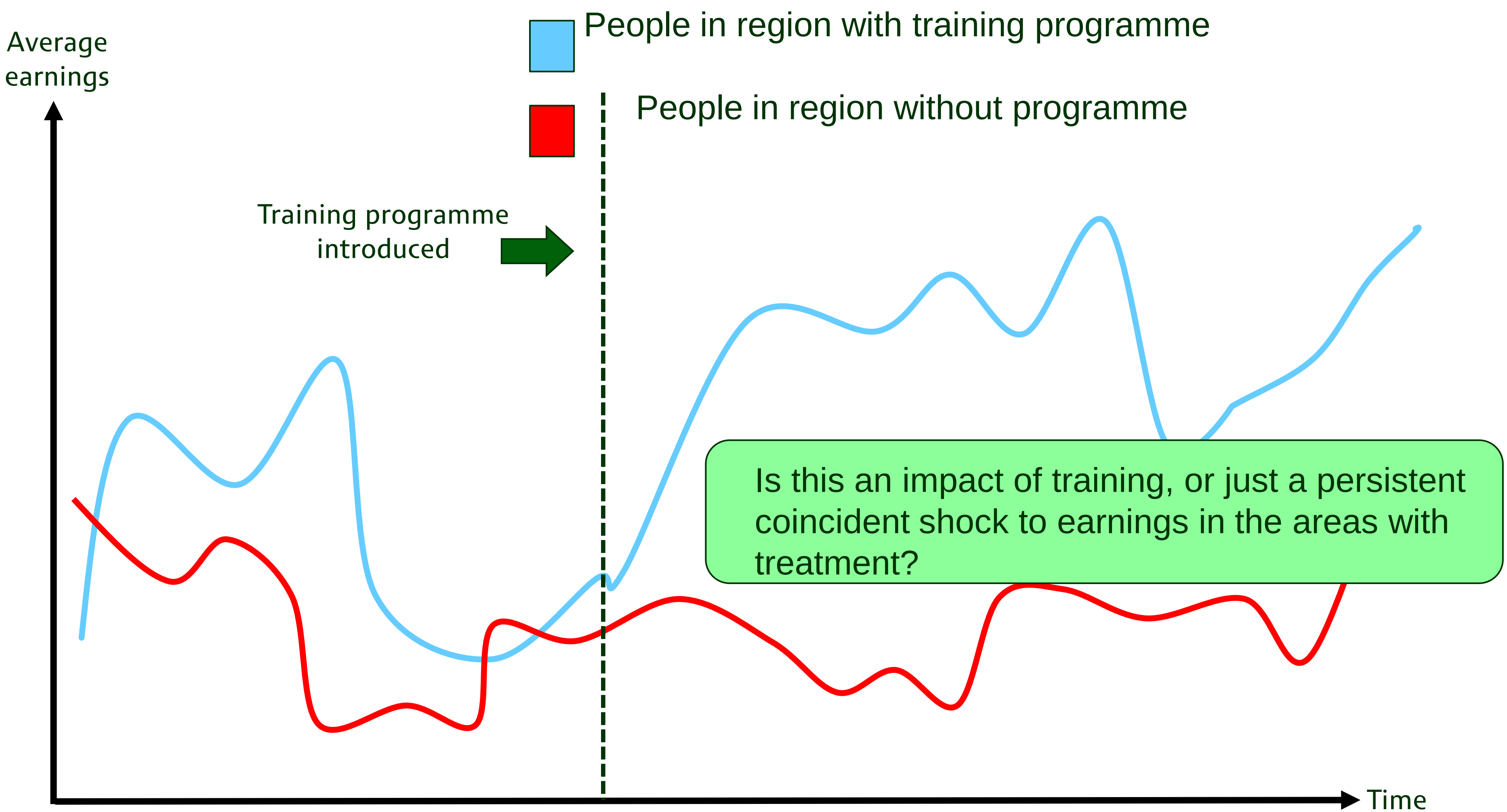
Spare


Earnings
Training programme introduced
Central DiD estimate says that training increased earnings….
People in region with training programme
People not in region with programme
Time



What would make data look like this?

A region-specific economic shock.



Average earnings
People in region with training programme
People in region without programme
Training programme introduced
Is this an impact of training, or just a persistent coincident shock to earnings in the areas with treatment?
Time

Aside: GLS and feasible GLS

Justification: let $\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$ and error variance matrix $\mathbf{\Omega} \neq \sigma^2\mathbf{I}$

Consider: $\left(\mathbf{\Omega}^{-1/2}\mathbf{y}\right) = \left(\mathbf{\Omega}^{-1/2}\mathbf{X}\right)\beta + \left(\mathbf{\Omega}^{-1/2}\mathbf{u}\right)$

This model meets standard conditions for OLS since $Var\left(\mathbf{\Omega}^{-1/2}\mathbf{u}\right) = \mathbf{I}$

$\hat{\beta}_{GLS} = \left(\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{y}$, where error variance matrix $\mathbf{\Omega} \neq \sigma^2\mathbf{I}$

$\hat{\beta}_{FGLS} = \left(\mathbf{X}'\hat{\mathbf{\Omega}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\hat{\mathbf{\Omega}}^{-1}\mathbf{y}$, where $\hat{\mathbf{\Omega}}$ estimates unknown error variance matrix $\mathbf{\Omega} \neq \sigma^2\mathbf{I}$

In our case:

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

$$\eta_{ct} = \rho_1\eta_{ct-1} + \rho_1\eta_{ct-2} + \varepsilon_{ct} \text{ with } \varepsilon_{ct} \text{ serially uncorrelated}$$

Consider transformed model:

$$\begin{aligned}\hat{\lambda}_{ct} - \rho_1\hat{\lambda}_{ct-1} - \rho_2\hat{\lambda}_{ct-2} = &\left(\alpha + \mu_c\right)\left(1 - \rho_1 - \rho_2\right) \\ &+ \beta\left(T_{ct} - \rho_1 T_{ct-1} - \rho_2 T_{ct-2}\right) + \left(\xi_t - \rho_1\xi_{t-1} - \rho_2\xi_{t-2}\right) \\ &+ \left(\eta_{ct} - \rho_1\eta_{ct-1} - \rho_2\eta_{ct-2}\right)\end{aligned}$$

This allows OLS since:

$$\eta_{ct} - \rho_1\eta_{ct-1} - \rho_2\eta_{ct-2} = \varepsilon_{ct}$$

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).

FGLS under misspecification of error process (10 groups)

	Heterogeneous AR(2)		MA(1)	
	No effect	Effect of +0.05 log-points	No effect	Effect of +0.05 log-points
OLS, robust	0.041	0.536	0.052	0.597
FGLS	0.100	0.775	0.088	0.675
FGLS, robust	0.055	0.703	0.053	0.580
BC-FGLS	0.070	0.803	0.071	0.675
BC-FGLS, robust	0.058	0.717	0.053	0.578

Note: FGLS is implemented assuming an AR(2) process for the state-time shocks. For the BC-FGLS procedure, see Hansen (2007). For the heterogeneous AR(2) process, the coefficient on the first lag (alpha) is drawn from a uniform distribution between zero and one for each state. The coefficient on the second lag is set equal to 0.5*min(alpha,1-alpha), which ensures stationarity. The MA(1) process has a lag parameter of 0.5. For both processes, the white noise is normally distributed. Its variance ensures that the error term has the same stationary variance as the log-earnings residuals in the CPS (0.04).

FGLS with varying panel length (10 groups)

	T=30		T=20		T=10	
	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	0.044	0.280	0.049	0.282	0.041	0.346
FGLS	0.115	0.418	0.128	0.370	0.102	0.333
FGLS, robust	0.051	0.401	0.052	0.352	0.046	0.328
BC-FGLS	0.084	0.420	0.093	0.376	0.087	0.337
BC-FGLS, robust	0.054	0.419	0.055	0.367	0.046	0.327

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

PEPA Programme Evaluation for Policy Analysis

NCRM National Centre for Research Methods

FGLS with varying panel length (10 groups)

	T=30		T=20		T=10	
	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	0.044	0.280	0.049	0.282	0.041	0.346
FGLS	0.115	0.418	0.128	0.370	0.102	0.333
FGLS, robust	0.051	0.401	0.052	0.352	0.046	0.328
BC-FGLS	0.084	0.420	0.093	0.376	0.087	0.337
BC-FGLS, robust	0.054	0.419	0.055	0.367	0.046	0.327

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

PEPA Programme Evaluation for Policy Analysis

NCRM National Centre for Research Methods

Why does power decrease with *T* for OLS+CRSE?

- Diff-in-Diff estimates the relative (ie between-group) difference in pre- and post-treatment averages
- V[Diff] = V[Pre]+V[Post] – Cov[Pre,Post]
- With serially correlated shocks, Cov[Pre, Post] important
- As we add more years of data
  - V[Pre], V[Post] fall, decreasing V[Diff]
  - Cov[Pre, Post] falls, **increasing** V[Diff]
- In these simulations, the second effect dominates
  - Similar phenomena apparent in Hansen's (2007) simulations, but he does not discuss