NCRM National Centre for Research Methods

PEPA Programme Evaluation for Policy Analysis

Inference in difference-in-differences

Robert Joyce (IFS)

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

PEPA is based at the IFS and CEMMAP

Institute for Fiscal Studies

cemmap centre for microdata methods and practice

Introduction/motivation

- DiD evaluations extremely common; earnings and employment are most common outcomes of interest (Bertrand, Duflo, Mullainathan (henceforth BDM, 2004))

- As always, need to quantify uncertainty around central estimates
  - Want to test hypotheses, e.g. "how likely would patterns in our data be if this training program had no effect on earnings?"

- Emerging literature on :
  - how to do inference properly in common DiD situations
  - the fact that it can really matter if you do not do so!

Our aims

1. Provide accessible synthesis of the inference issues in DiD and guide to various practical solutions proposed in the literature

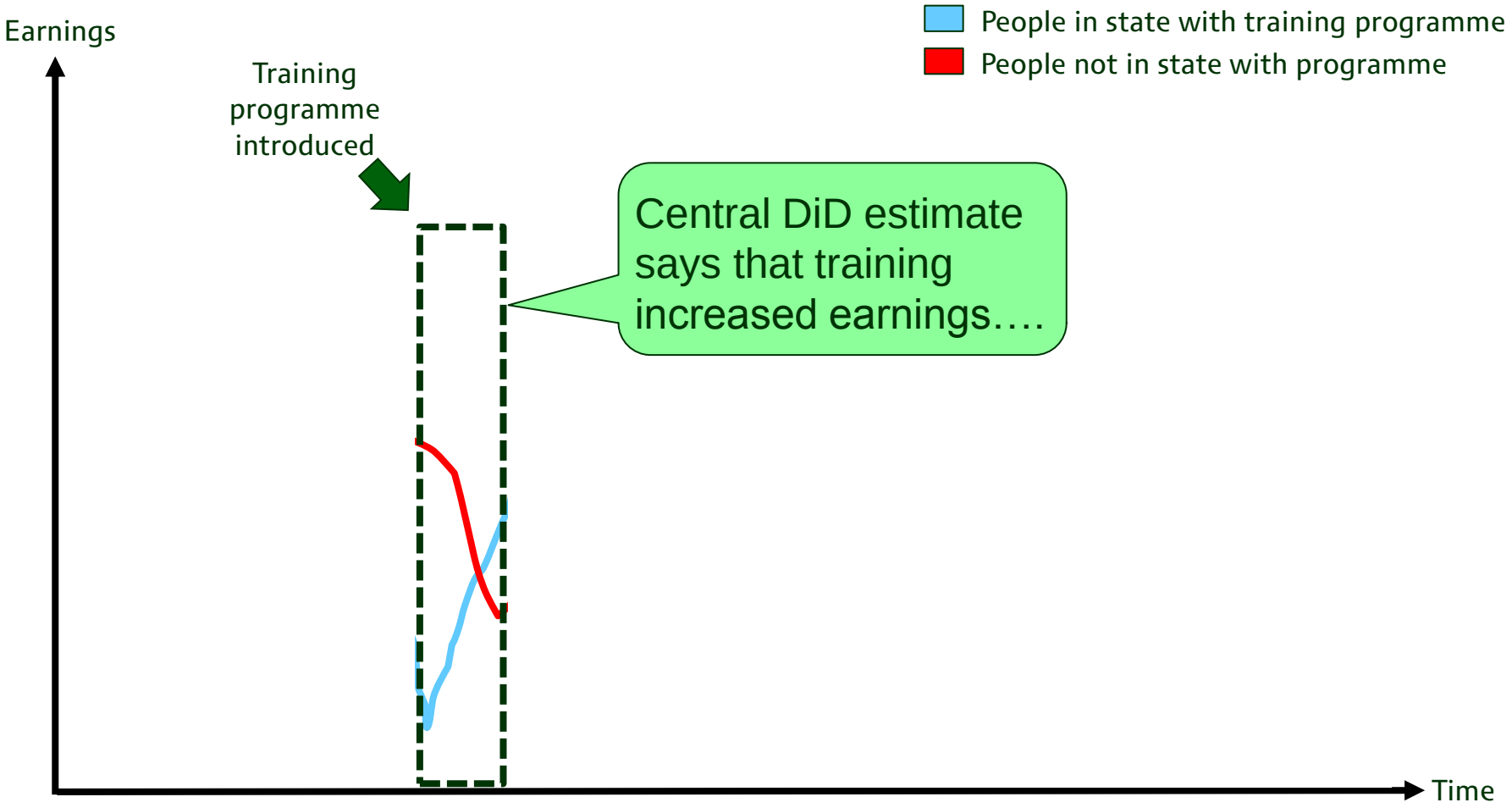
2. Illustrate performance of different methods in different contexts using simulation techniques

3. Refine/add to existing methods (hopefully!)

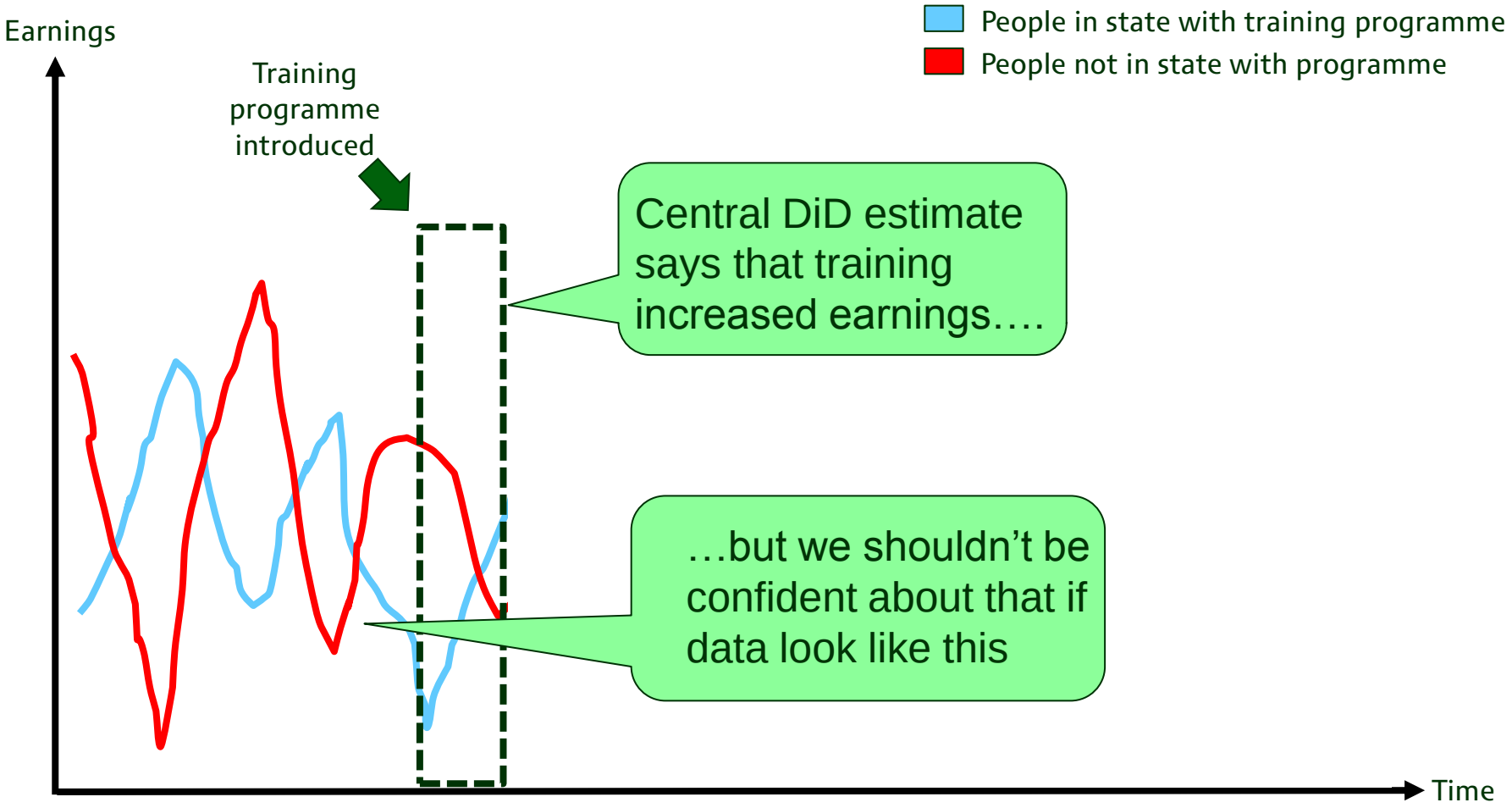
- Have worked on 1 and 2, and barely started 3.
- Intended audience: applied economists; other analysts who frequently come across policy evaluations

THE PROBLEM

Inference in DiD: first problem



Inference in DiD: first problem



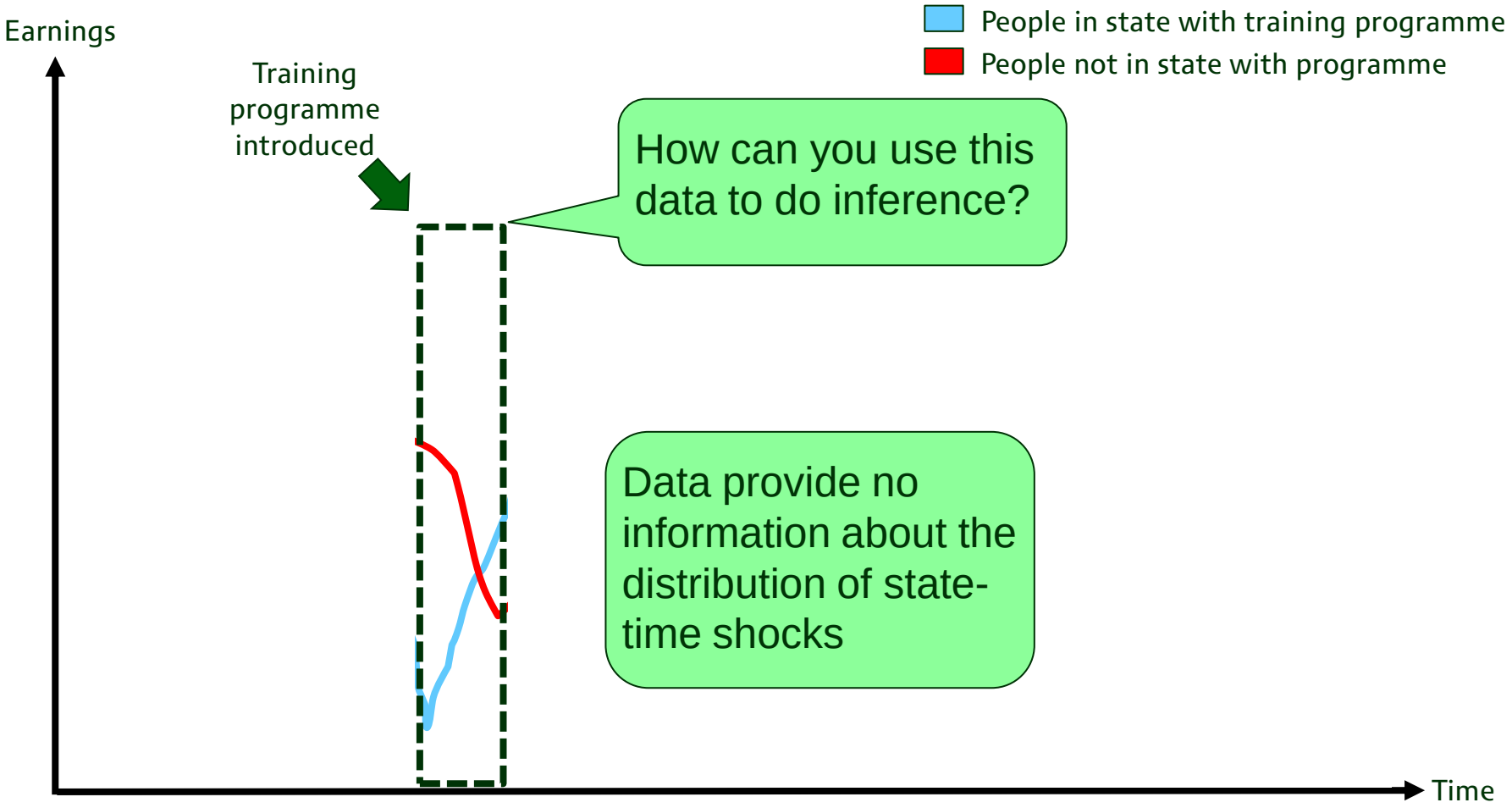
Why the volatile data pre-treatment?

- With large sample, earnings shocks should average 0 in all time periods in all states…
  - …***unless*** people in same state at same time affected by common shock
- So the ***clustering*** of the errors/shocks increases uncertainty around estimates of policy impact
  - Need to account for this when estimating standard error

Accounting for volatility in state-time shocks

- Basically this means either:
  - Making an assumption about their distribution. (Then your conclusions may be only as reliable as your assumption.)
  - Using info in the data (most likely pre-treatment data) about their distribution. Hence, the more such information you have the better.

The Donald-Lang (2007) critique of a 2x2 DiD



Second problem: serial correlation

- So having lots of states and/or time periods should be useful…
- But if state-time shocks are serially correlated, then adding more time periods is less useful
- Can seriously under-estimate the uncertainty if just allow for clustered shocks at the state-time level, ignoring serial correlation
  - BDM create ‘placebo’ treatments and find 44% rejection rates for a nominal 5% level test.

And standard errors are not the only problem

Standard hypothesis testing relies on two things

1. Forming a test statistic
   - Typically a t-statistic, for which you need to estimate standard error

2. *Knowing distribution of this statistic under null hypothesis*
   - In large samples, use asymptotic results from statistical theory (t-stat converges to standard normal)
   - But with clustered errors, asymptotics generally apply only as the number of clusters gets large

SOLUTIONS

Just use “cluster-robust” standard errors?

- Ideally, could take commonly-used formula for the covariance matrix that is robust to clustered errors of an arbitrary form
  - …so if you cluster at the group level (***not*** group-time level) you allow for serial correlation within groups
- Trivial to implement, e.g. in Stata just use “cluster (vce *clustvar*)”
- But consistency of CRSEs applies as number of clusters gets large. With few clusters they can be biased.

Just use "cluster-robust" standard errors?

- Few-clusters bias corrections proposed (e.g. Bell and McCaffrey, 2002)

- But asymptotic normality of the t-stat (even if it uses a bias-corrected CRSE) also depends on having lots of clusters
  - If few clusters, might not know what critical values to compare t-stat to

- So: if you have enough groups (roughly 50+) just use cluster-robust SEs, clustering at the group level
  - But otherwise the best solution *may* be less straightforward

Wild cluster bootstrap-t (Cameron et al, 2008)

- Compute t-statistic using cluster-robust standard error…
- …then repeatedly resample clusters of data with replacement, compute t-statistic again, and compare original t-statistic to distribution of t-stats from bootstrap samples
- Resampling scheme imposes null and allows for arbitrary heteroscedasticity and serial correlation within clusters (but relies on additive errors)
  - For full details of implementation, see Cameron et al Appendix B, and Bansi Malde's ado file at **http://tinyurl.com/c8vz3br**
- Using similar 'placebo treatment'-type simulations to BDM, they find that this method rejects the null hypothesis with about right probability
  - Even with as few as six groups in the data

Randomization/permutation-type techniques (1)

- Mainly used outside economics e.g. political science (see Helland and Tabarrok, 2004; Erikson et al, 2010; Abadie et al, 2010)
- Similar to bootstraps in that they attempt to learn about the distribution of (e.g.) t-stats without relying on asymptotic results
- (Repeatedly) randomly 're-assigns' time series of treatment indicators to different groups, and re-computes t-stat each time
  - Breaks any relationship between treatment and outcomes, recovering distribution of t-stat under the null hypothesis of no treatment effect

Randomization/permutation-type techniques (2)

- Assumption is 'exchangeability': no systematic differences in distribution of shocks between treated and untreated groups
  - Could be violated if (e.g.) policy rule used to allocate treatments meant that groups with more/less volatile outcomes were treated

- They test stronger null hypotheses than other methods discussed (which could be a good/bad thing)
  - The null you're testing is that treatment effect was zero for **everyone**
  - Other methods test nulls relating to **parameters,** e.g. that the treatment effect **averages** zero among some group

Rejection rates with 5% level tests from 5000 'placebo law' simulations using 30 years of CPS log-earnings data

	Number of groups (US states), half of which are treated			
Inference method	**6**	**10**	**20**	**50**
OLS SEs	.415*	.424*	.425*	.424*
Cluster-robust SEs, critical values from N(0,1) dist.	.104*	.073*	.054	.048
Cluster-robust SEs, critical values from t(G-1) dist.	.051	.044*	.038*	.043*
Wild cluster bootstrap-t	.067*	.054	.055	.055
Permutation	.041*	.055	.054	.057*

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.

But what about power to detect real effects?

Rejection rates when using cluster-robust SEs and critical values from t(G-1) distribution				
	Number of groups (US states), half of which are treated			
True effect size	**6**	**10**	**20**	**50**
No effect	.051	.044	.038	.043
2% increase in earnings	.082	.081	.116	.222
10% increase in earnings	.458	.700	.904	.999

Issues worthy of more exploration…

- Power
  - Literature has focused mainly on making tests the right size, but methods which achieve this may also be unlikely to detect real effects
- The case with very small number of treated groups but relatively large number of controls (Conley and Taber, REStat, 2010)
- Clustering at the right level, and multi-way clustering (see Cameron et al, 2011)
- Inference in non-linear DiD-style models

References

- Abadie, Diamond and Hainmueller, "Synthetic Control Methods for Comparative Case Studies: Estimating the effect of California's Tobacco Control Program", *Journal of the American Statistical Association* (2010)
- Bell, R. M., and D. F. McCaffrey, "Bias Reduction in Standard Errors for Linear Regression with Multi-Stage Samples," *Survey Methodology* 28:2 (2002), 169–179.
- Bertrand, Duflo, and Mullainathan, "How Much Should We Trust Differences-in-Differences Estimates?" *Quarterly Journal of Economics* 119 (2004), 249–275.
- Cameron, Gelbach and Miller, "Bootstrap- based improvements for inference with clustered errors", *Review of Economics and Statistics* 90:3 (2008), 414-427
- Card, D., and Krueger, A. B. (1994), "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania,"*American Economic Review 84, 772–793*
- Donald and Lang, "Inference with Difference-in-Differences and Other Panel Data," *Review of Economics and Statistics 89:2 (2007), 221–233*.
- Erikson, Pinto and Rader, "Randomization Tests and Multi-Level Data in U.S. State Politics", *State Politics & Policy Quarterly (2010)*
- Hansen, "Generalized Least Squares Inference in Panel and Multilevel Models with Serial Correlation and Fixed Effects", *Journal of Econometrics,* 140:2 (2007), 670-94
- Helland and Tabarrok, "Using Placebo Laws to Test 'More Guns, Less Crime'", *Advances in Economic Analysis and Policy* (2004)