NCRM National Centre for Research Methods

PEPA Programme Evaluation for Policy Analysis

Inference in difference-in-differences approaches

Mike Brewer (University of Essex & IFS) and Robert Joyce (IFS)

PEPA is based at the IFS and CEMMAP

Institute for Fiscal Studies

cemmap centre for microdata methods and practice

Introduction

- Often want to estimate effects of policy/programme (“treatment”)
  - effect of unemployment benefits on unemployment duration
  - effect of gun laws on crime

- Typically do this using data on a **sample** from population

- This talk is *not* about estimating treatment effects in an unbiased and consistent way

- It is about quantifying the uncertainty around those estimates

Motivation: the importance of inference

- Unbiased estimation implies that if we applied method repeatedly on different samples, we would get the right answer **on average**

- But no estimate based on one population sample is exactly right

- So it is crucial to quantify uncertainty around central estimate, otherwise, cannot test any hypotheses (conduct "inference")

  - *"How surprising would the patterns in this sample be if in fact the treatment had no effect?"*

We focus on settings where inference is tricky

1. Data contains (small number of) “clusters”, where unobserved determinants of the outcome are correlated within clusters
   - e.g. People in same region affected by same economic shocks

2. Treatment status the same for everyone within clusters
   - e.g. Gun laws the same for everyone in the same US state

3. Longitudinal settings: unobserved determinants of the outcome are persistent over time for a given cluster (“serial correlation”)
   - e.g. Regional economic shocks are persistent, as they reflect the state of the business cycle

- These issues arise frequently in data used for policy evaluation but are “non-standard” in statistical/econometric theory

Outline of rest of talk

- The standard approach to inference in OLS
- Intuition behind the clustering problem
- Standard models of, and solutions to, the clustering problem
- Adding a serial correlation problem: difference-in-differences
- Methods for inference with both clustering and serial correlation

STANDARD INFERENCE IN OLS

Reminder: standard settings

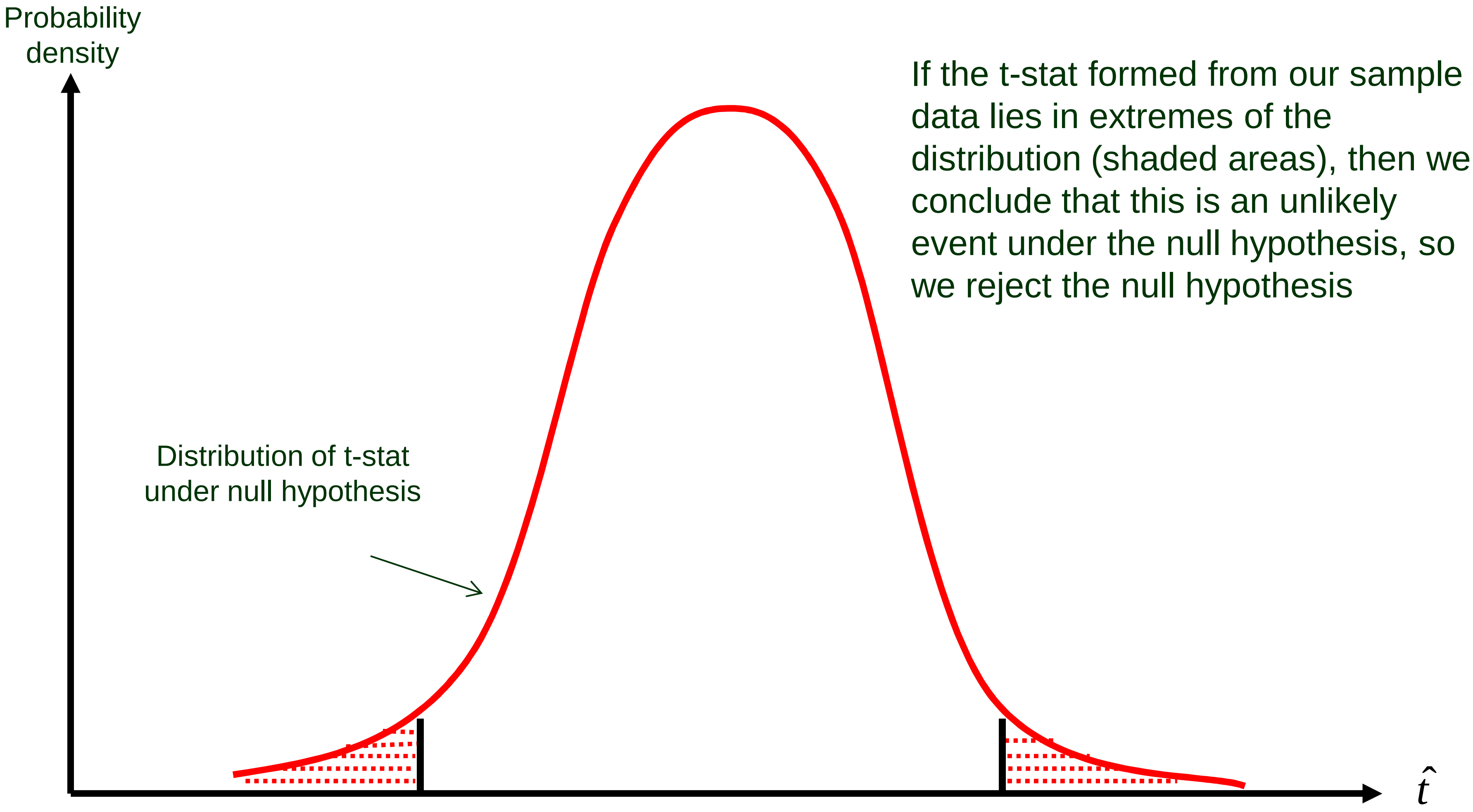
- Simple model : $Y_i = \alpha + \beta T_i + \delta X_i + u_i$ $E(u_i | T_i, X_i) = 0$
  - Y_i is outcome for individual i (e.g. their earnings)
  - T_i equals 1 if individual i is treated (e.g. in training programme); 0 if not
  - X_i is other control variables for individual i (e.g. education)
  - β is the key unknown parameter: the treatment effect
  - u_i is the "error term" or "shock": unobserved determinants of Y for individual i

- Can easily estimate β in unbiased way (because $E(u_i | T_i, X_i) = 0$)
  - Reminder: what does "unbiased" mean? If repeatedly estimated β using different samples from population, would, on average, get right answer

Inference in standard settings

1. Specify ‘null hypothesis’ we want to test
   - e.g. $\beta_0 = 0$ (“treatment has no effect, on average”)
2. Form some ‘test statistic’ based on the sample data
   - e.g. t-stat $t = (\hat{\beta} - \beta_0) \big/ \hat{se}(\hat{\beta})$
3. Use statistical theory to tell us what distribution of test statistics we would expect (if replicated estimation many times on fresh samples) if null hypothesis were true
   - e.g. t-stat has t-distribution if errors are normally-distributed
   - e.g. t-stat has normal distribution if sample size is very large (technically: distribution of t-stat gets closer to normal distribution as sample size grows; called an “asymptotic result”)
4. Can then judge how ‘surprising’ our test statistic would be if null hypothesis were true
   - E.g. $|t| > 1.96$ should occur with 5% probability under null

Inference in standard settings: graphically



If the t-stat formed from our sample data lies in extremes of the distribution (shaded areas), then we conclude that this is an unlikely event under the null hypothesis, so we reject the null hypothesis

Recap, and the relevance of what's to come

Standard inference proceeds based on two things:

1. Forming a test statistic
2. Knowing distribution of this statistic under null hypothesis

Problems considered in this talk can make both more difficult

1. Clustering and serial correlation make the usual estimators of $se(\hat{\beta})$ inappropriate: an obstacle to forming the proper test statistic
2. When data are clustered, it can be harder to know what distribution we would expect a t-stat to have under the null
   - Why? If data are clustered, then t-stat gets closer to being normally distributed as number of clusters (**not** observations) grows. With few clusters, hard to know what distribution we would expect a t-stat to have under the null
   - More observations does not necessarily mean more clusters

AN EXAMPLE OF THE CLUSTERING PROBLEM

Example of clustering: training and earnings

- Want to know effect of training scheme ("treatment") on earnings ("outcome")
- Some US states implement the training scheme, so individuals in these states are "treated"; other US states do not
- Use data on large sample of US citizens from a few US states

- An OLS regression (or similar) effectively compares earnings of individuals in different states.

$$Y_i = \alpha + \beta T_i + \delta X_i + u_i$$

  - If $E\left(u_i \mid T_i\right) = 0$ this is an unbiased estimate of the impact of training on earnings
  - But what about inference?

Why might clustering make inference tricky? (1)

$$Y_i = \alpha + \beta T_i + \delta X_i + u_i \qquad E\left(u_i \mid T_i, X_i\right) = 0$$

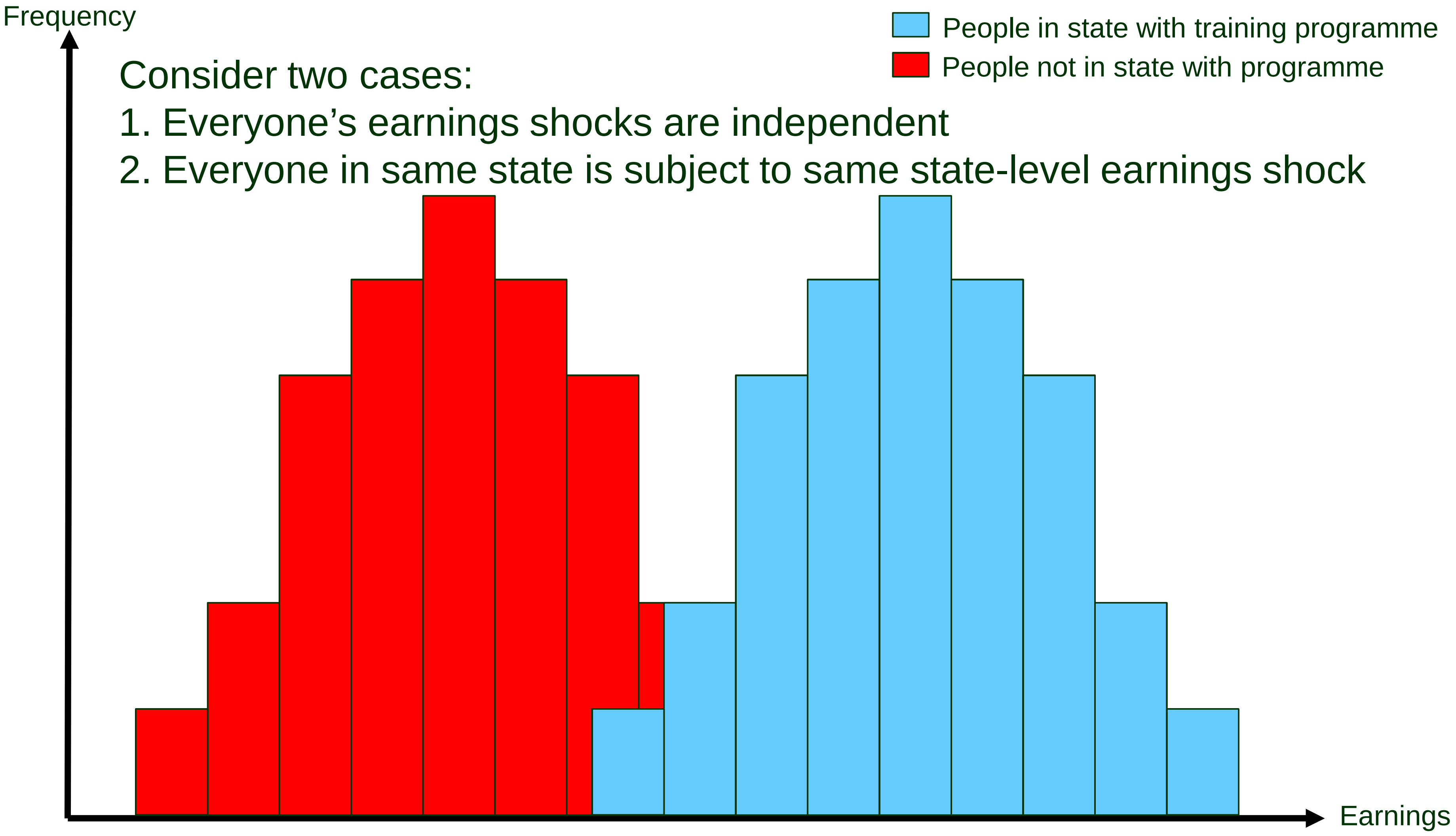
- For inference about effect of training scheme, need to ask:
  - "How likely is it that differences in earnings across states in our sample is due to different earnings shocks (rather than the training scheme)?"

- If individual earnings shocks are independent of each other, they will average to (close to) zero within each state
  - Large differences in earnings between states with/without training scheme could then be confidently attributed to training scheme
- But clustering breaks independence: imagine that individual shocks within a given state share a common component (a state-level shock)
  - Now, individual earnings shocks within a state will not average to zero, unless the state-level shock happens to be zero
  - So harder to distinguish an impact of training from the impacts of earnings shocks

How confident are we that training had an effect?

Consider two cases:

1. Everyone’s earnings shocks are independent
2. Everyone in same state is subject to same state-level earnings shock



Why might clustering make inference tricky? (2)

- Groups with training will have different earnings shocks from groups without training, because:
  - There are state-level earnings shocks
  - Groups with training live in different states from groups without training

- If lots of states, *on average* there should be little difference between state-level earnings shocks in treated/untreated states…
  - …but not if few states (regardless of how many people in each state)

- If we know how state-level shocks are distributed, can judge whether differences in earnings between those in states with/without training are likely to be due to differences in state-level earnings shocks…
  - …but have little information about distribution of state-level shocks if only a few states (no matter how many people in each state)

Clustering is like a (big) reduction in sample size

- Say Ohio is one of the sampled states. How useful would it be to sample another 1 million people from Ohio?
  - Not very! They all have the same treatment status (trained/not trained) and the same Ohio-level earnings shock as the existing sample in Ohio
  - So doesn't help us distinguish the impact of training from the impact of Ohio's earnings shock

- Note the problem is particularly severe because we have a cluster-level treatment (everyone in Ohio is either trained or not trained)
  - If some in Ohio had access to training and some did not, more data on people from Ohio would be more useful: researcher can see what happens when treatment status changes but the state-level earnings shock stays same

- No matter how many individuals we have data on, we are really trying to separate state-level earnings shocks from effects of state-level treatments
  - So need lots of states, not lots of individuals

What does clustering mean in practice?

- Having clustered data always affects inference, unless regressors are completely uncorrelated with cluster membership
  - In evaluation settings, treatment often perfectly correlated within cluster

- If we ignore clustering, estimated standard errors will tend to be too small, so we will tend to over-reject null hypotheses (i.e. more likely to incorrectly-reject true nulls; confidence intervals too small)

- Even if we have correct standard errors, usual t-stat distributed normally only as number of clusters (NOT observations) gets large
  - So, even if have very large sample of individuals, we may not know distribution of t-stat under the null hypothesis, so will not be able to decide whether or not actual t-stat is especially unlikely (under the null)

CLUSTERING: SIMPLE MODEL, AND SOLUTIONS

What might a model with clustering look like?

- Model: $Y_{ic} = \alpha + \beta T_c + \delta X_{ic} + \mu_c + \varepsilon_{ic}$
  - Y_{ic} is outcome for individual i in cluster c
  - T_c is treatment status: same for everyone within a cluster
  - Unobserved "error" or "shock" now has TWO components: cluster-level shock μ_c and individual-level shock ε_{ic}
  - We'll assume that the cluster-level shock is independent across clusters, and individual-level shock independent across individuals
  - **But cluster-level component implies shocks are correlated within clusters**

- If $E(\mu_c + \varepsilon_{ic} \mid T_c) = 0$, straightforward to use OLS to get unbiased estimate of β

The "Moulton correction" (1)

- Say that the correlation between all shocks within all clusters is ρ

- Moulton (1986) showed that the clustering increases the variance of estimates of the treatment effect ($\hat{\beta}$) by factor of $1+\rho\frac{N}{C}$

  (N is sample size; C is number of clusters)

- So can scale up normal OLS standard error using (square root of) the Moulton factor…
  - …using an estimate of ρ : just compute intra-cluster correlation of residuals ("estimated shocks"), $\hat{u}_{ic} \equiv Y_{ic} - \hat{\alpha} - \hat{\beta} T_c - \hat{\delta} X_{ic}$

- Then form t-stat in usual way, but using corrected standard error

The “Moulton correction” (2)

- Correction relies crucially on the assumption that the correlation between all shocks within all clusters is the same (ρ)
- This assumption is true under a combination of:
  - Variance of shock components the same across clusters (which in particular requires homoscedasticity)
  - Independence between two shock components

- Under those conditions:

$$Var(\mu_c + \varepsilon_{ic} \mid T_c, X_{ic}) = \sigma_\mu^2 + \sigma_\varepsilon^2$$

$$\rho = \frac{\sigma_\mu^2}{\sigma_\mu^2 + \sigma_\varepsilon^2}$$

- If those conditions are not satisfied, then the Moulton correction can not be expected to yield correct standard errors

Moulton correction: discussion

- Note that square root of $1 + \rho \frac{N}{C}$ can be big!
  - Angrist and Lavy (2007) study school-level intervention with 4000 students in 40 schools
  - N/C = 100, and ρ estimated to be 0.1: implies true standard errors three times larger than what you would estimate if you (incorrectly) ignore clustering
- Correction requires strong assumption that ρ same for all clusters
- And correction (might) get you correct test statistic, but also need to know ***distribution of test statistic under null hypothesis***
  - With a large number of clusters, the t-stat will be close to normally distributed, but what if few clusters?
  - Can make one more assumption: the cluster-level shocks have a **normal** distribution (Donald and Lang (2007))

Two-step procedure (Donald and Lang, 2007)

Model: $Y_{ic} = \alpha + \beta T_c + \delta X_{ic} + \mu_c + \varepsilon_{ic}$

1. Estimate ‘cluster effects’, $\lambda_c \equiv \beta T_c + \mu_c$, using OLS on full sample: $$Y_{ic} = \alpha + \lambda_c + \delta X_{ic} + \varepsilon_{ic}$$
2. Estimate the treatment effect, β, using the estimated cluster effects: $$\hat{\lambda}_c = \beta T_c + \omega_c$$

$$\omega_c = \mu_c + \left(\hat{\lambda}_c - \lambda_c\right)$$

- Note that
  - 2nd component of error term is the ‘estimation error’ for the cluster effect from 1st stage. If sample sizes *within* clusters are large, this will be roughly normally distributed.
  - So if the cluster-level shocks (μ_c) also have normal distribution, then ω_c has a normal distribution
- This helps because, with normal errors, t-stats are known to have a t distribution even in small samples (i.e. with few clusters, in 2nd stage)

Donald-Lang 2-step procedure: discussion

- Can deal with relaxation of some of the strong Moulton assumptions
  - With large cluster sizes, don't need homoscedasticity
  - The cluster-level shock could instead just be a shock correlated within clusters (i.e. it doesn't have to have intra-cluster correlation of one):
    - Technically, we require that the variance of its average within a cluster depends only on cluster size

- But, to do inference properly in situations with few clusters, also need to add assumption that cluster-level shock is normally distributed

- Under the required assumptions, it is efficient even with few clusters (it is equivalent to feasible GLS)

MAXIMUM FLEXIBILITY: CLUSTER-ROBUST STANDARD ERRORS

Relaxing the assumptions

- Methods discussed so far have required making assumptions about the unobserved shock

  - Moulton: constant correlation of shocks within clusters, for all clusters
  - Donald and Lang: cluster-level shock is normally distributed

- More flexible, and more commonly used, are “cluster-robust” standard errors (Liang and Zeger, 1986)

  - Generalisation of White’s heteroscedasticity-robust standard errors
  - Implemented in Stata regressions using “vce(cluster *clustvar*)” option

Cluster-robust standard errors

- These can solve inference problems without making any assumptions about the way that shocks are correlated within clusters
  - Still require that the cluster-level shocks are uncorrelated across clusters (i.e. that you have defined the clusters correctly)

- Cluster-robust standard errors obtained by plugging residuals into formula which is a generalisation of White's heteroscedasticity-robust formula to allow also for clustering

Again, we have problems with few clusters

- Cluster-robust estimates of standard errors only consistent as number of *clusters* (not observations) gets large
  - Some bias corrections for cluster-robust SEs have been proposed (see Angrist and Lavy, 2002) for setting with few clusters, but…
- …even if cluster-robust standard errors can be estimated with no bias, usual t-stat may have unknown distribution under the null hypothesis with small number of clusters

Recap

- Having clustered data always affects inference, unless explanatory variables completely uncorrelated with cluster membership
  - “Treatments” are often perfectly correlated within clusters
- Have considered
  - solutions which impose and exploit certain assumptions about the nature of the clustering
  - cluster-robust SEs , the most common solution, which assumes nothing about structure of shocks within cluster
- But these solutions:
  1. Do not work well with few clusters

  And/or:

  2. Make strong assumptions about the form of the clustering (i.e. about the nature of the errors within clusters)

What’s to come

- We now extend the problem to cover most common evaluation setting: where we have observations at multiple points in time
  - This should help (more data-points!), but often introduces another complication

- We then discuss solutions that can handle this extension

- Some of the solutions will be relevant for the simpler case without the time dimension as well, particularly for small number of clusters

INFERENCE WITH A TIME DIMENSION: DIFFERENCE-IN-DIFFERENCES

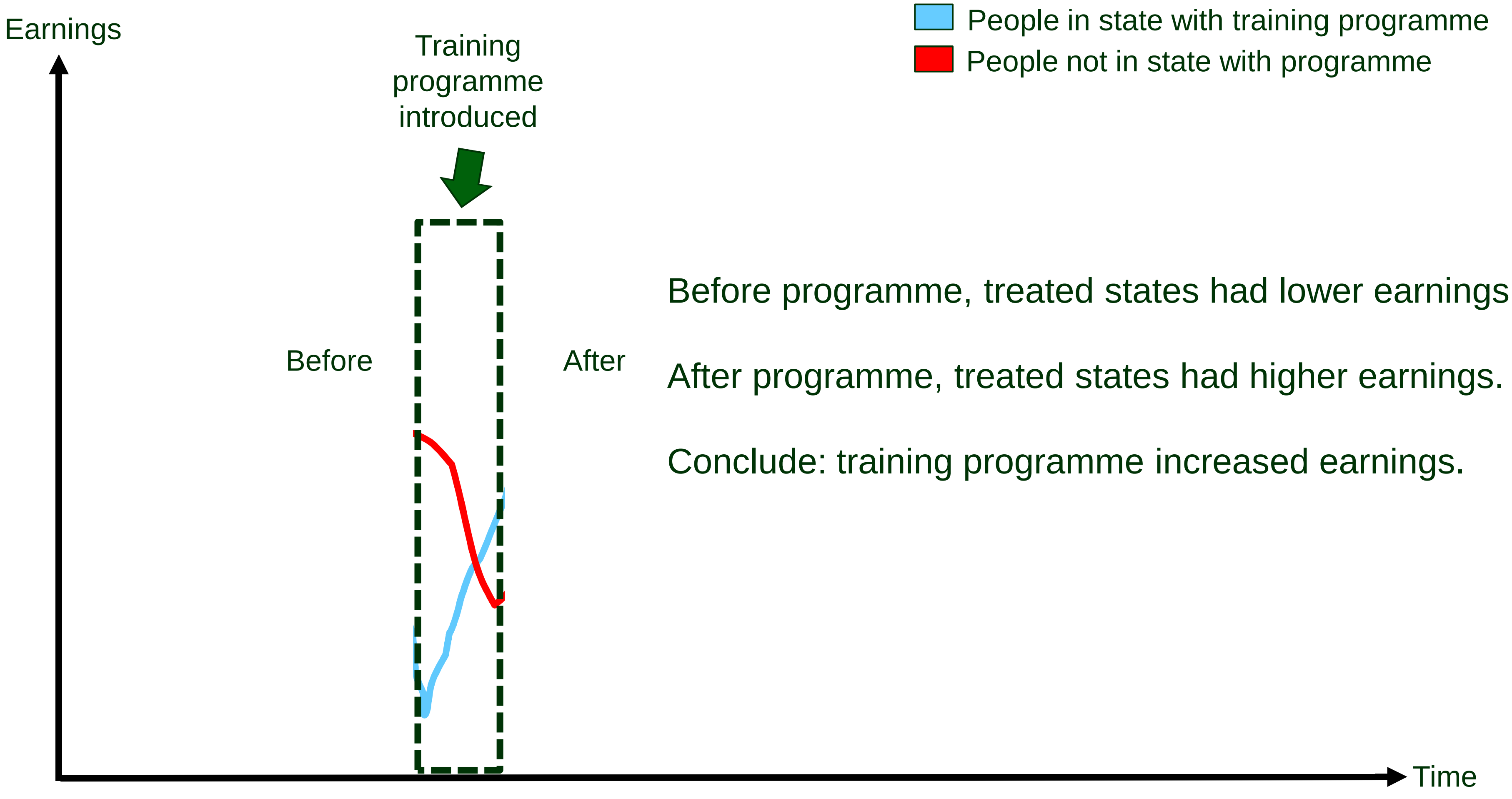
Effects of training with (slightly) different data

- In most common data setting for policy evaluations, would observe earnings from *before and after* introduction of training programme

- Estimation (in effect) compares
  - *change in earnings* in treated states (before/after training programme)
  - with *change in earnings* in untreated states
- The "difference-in-differences" estimator (DiD)

- DiD uses information on pre-programme differences in earnings
  - Hope is that ***pre-programme differences in earnings between states*** are a good guide to what ***post-programme differences in earnings would have been in the absence of training***

The “difference-in-differences” estimator (1)



Before programme, treated states had lower earnings.

After programme, treated states had higher earnings.

Conclude: training programme increased earnings.

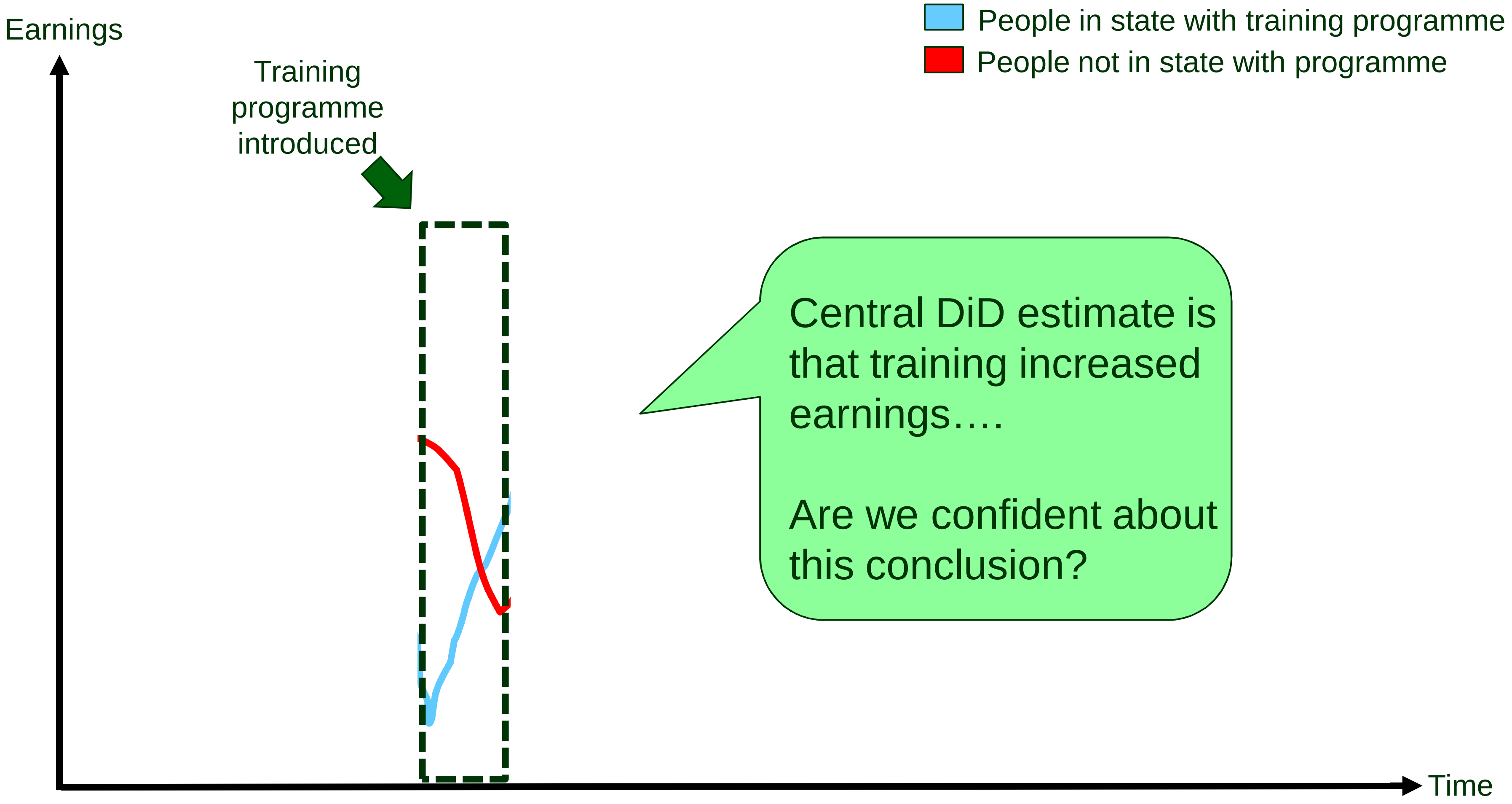
The “difference-in-differences” estimator (2)

- DiD assumes ***pre-programme differences in earnings between states*** are a good guide to ***what post-programme differences in earnings would have been in the absence of training***

- Now suppose that earnings of workers in a given state at a given time subject to some common shock (state-specific business cycle)
  - These shocks will NOT average to zero within a state-time period

- Differences in average earnings between states will then reflect:
  - Before: between-state difference in state-time shocks at time *t-1*
  - After: between-state difference in state-time shocks at time *t*
    + impact of training programme

The inference problem in DiD with state-time shocks: graphically



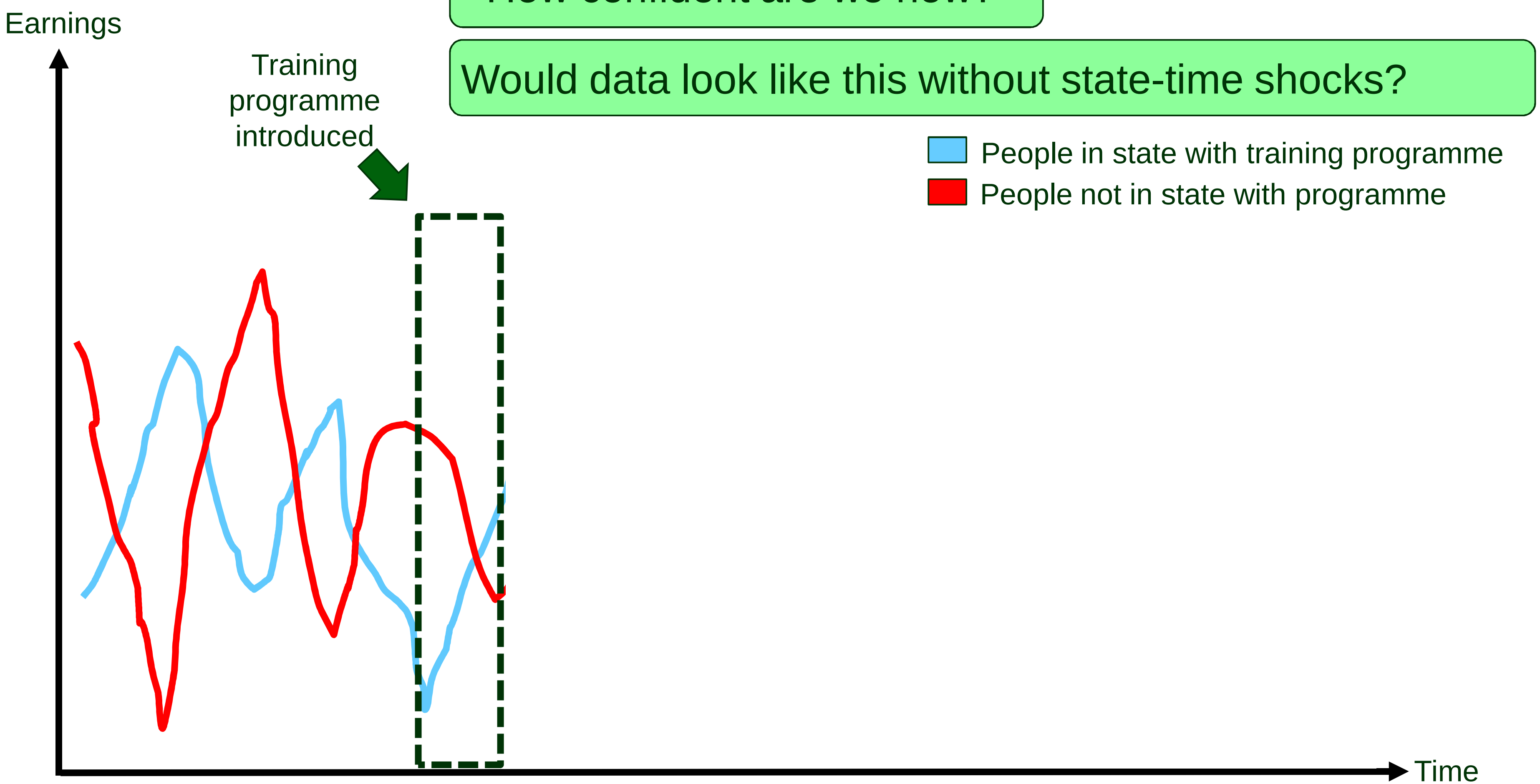
Central DiD estimate is that training increased earnings….

Are we confident about this conclusion?

The inference problem in DiD with state-time shocks: graphically

How confident are we now?

Would data look like this without state-time shocks?



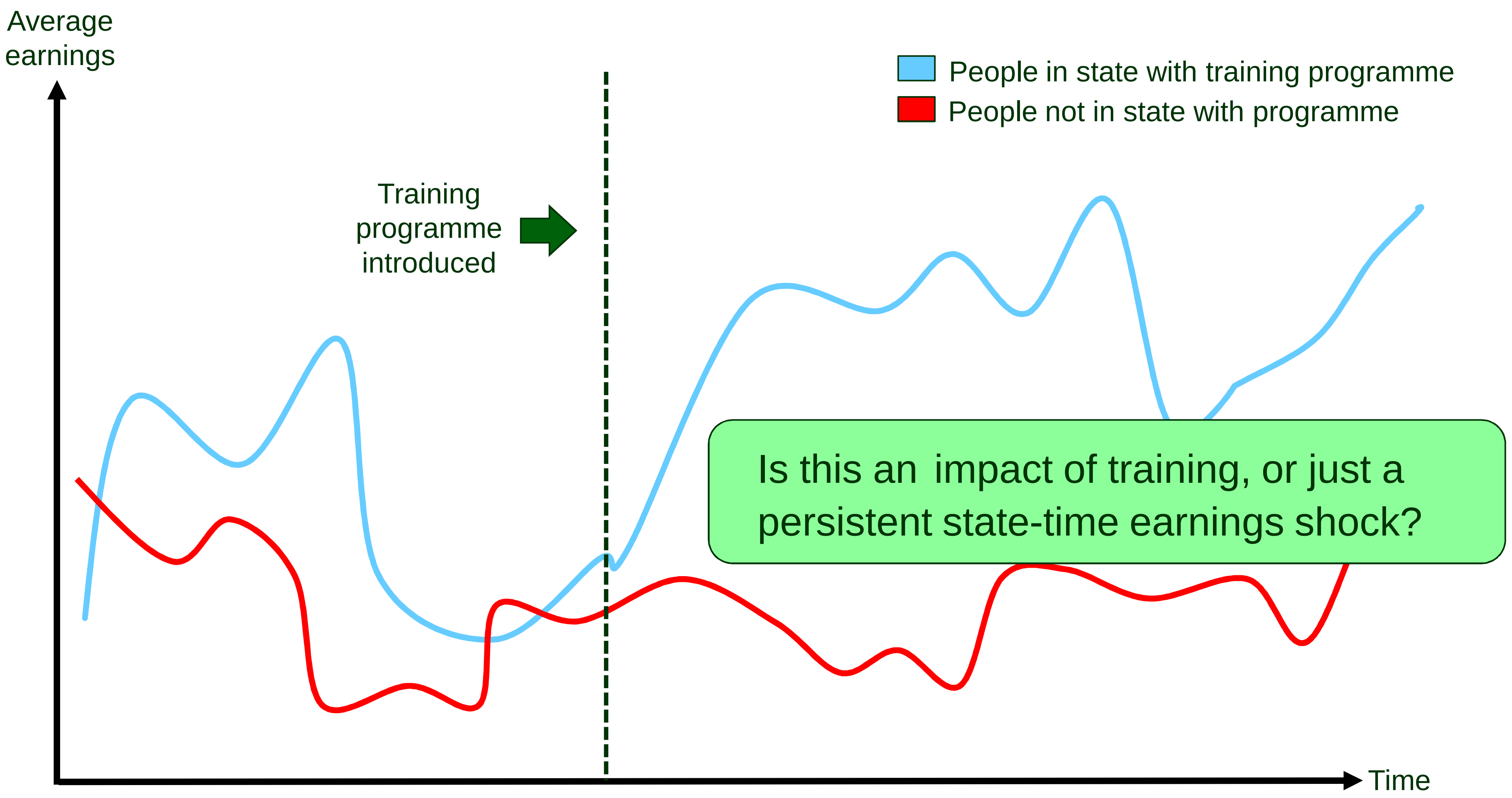
The inference problem in DiD with state-time shocks: in words

- After programme, between-state differences in earnings due to:
  1. Training programme (the treatment)
  2. Different state-time earnings shocks in different states
- To distinguish between 1 and 2, need to know distribution of state-time earnings shocks
- To learn about this, need as many state-time cells as possible
  - Adding observations with same state-time combination as existing data does not help, but more states or more time periods should help
  - Directly analogous to cross-section case with clustered data, except now want data on additional states or additional time periods
- NB Donald and Lang argue that a 2-group, 2-period DiD with group-level treatment is effectively useless in presence of state-time shocks
  - Why? No way of knowing importance of state-time shocks

But ....

- In practice, highly likely that (e.g.) Ohio’s earnings shock in one period is correlated with its earnings shock in next period
  - This would occur if shocks were persistent

The inference problem in DiD with serial correlation in state-time shocks: graphically



Is this an impact of training, or just a persistent state-time earnings shock?

The inference problem in DiD with serial correlation in state-time shocks: in words

- If earnings in Ohio rise after training programme starts, and stay high for many periods, is this because of earnings shock?
  - If no serial correlation, then unlikely: shock would have effect for only one period. Would conclude that training programme had an effect
  - But with serial correlation, earnings shock at same time as training programme could affect earnings in Ohio for some time afterwards. So less clear-cut
- Persistent shocks mean that shocks at different times in the same state are not independent
  - Adding extra time periods is ***not*** like adding new clusters
  - Adding extra time periods may provide less new information about distribution of state-time shocks than adding extra states
- Informally can see that serially-correlated shocks increase uncertainty over treatment impact

Serial correlation in state-time shocks: practical implications and solutions

- Harder to be certain about impact of policy if state-time shocks are serially correlated
- Flipside is that, if shocks are positively serially correlated, then estimates of standard errors which ignore this will be too small (estimates will appear overly-precise)
  - Bertrand, Duflo and Mullainathan (BDM, 2004) look at data that exhibits serial-correlation but that contains no treatment effects. They found that researchers testing the (true) null of “no treatment effect” would incorrectly reject the null 44% of the time when using a (nominal) 5% level test
    - Why? If we assume no serial correlation, then persistent differences between states get attributed to (non-existent) policy
- To solve problem of serially-correlated shocks, will need either to:
  - specify (and estimate) time-series process for the state-time shocks
  - use method that is flexible about within-state correlations

CLUSTERING AND SERIAL CORRELATION IN DiD: SOLUTIONS

Another 2-step procedure (Hansen, 2007)

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

- Hansen (2007) proposes 2-step procedure similar to Donald and Lang

1. Estimate ‘state-time’ effects using OLS on full sample

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

2. Estimate treatment effect using the estimated state-time effects:

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

- Error in 2nd stage is serially correlated; OLS would give unbiased estimate of coefficient , but naive standard errors would be wrong
- Hansen proposes:
  - assume 2nd-stage shock follows particular autoregressive (AR) process
  - estimate parameters of this process using 2nd-stage OLS residuals, plus a ‘bias correction’ (to account for inconsistency of these estimates with fixed state effects and fixed number of time periods)
  - Can then correct standard errors (equivalent to feasible GLS)

Specifying time-series process: assessment

- Method rejects null hypothesis correct proportion of the time when data contains no treatment effect

- But relies on
  - correctly specifying AR(p) process for 2nd stage shock
  - homoscedasticity in state-time shock
  - same AR process for shock in all states
- And method for estimating AR(p) process is consistent only as number of states goes to infinity, or as it becomes vanishingly small relative to number of time periods
  - Hansen tests with data from 51 states; but could be problems with few states (unless huge number of time periods)

Full flexibility again: cluster-robust SEs

- Earlier: cluster-robust SEs make no assumptions about within-cluster correlation structure
- Very important: in DiD, usually NOT be appropriate to estimate standard errors with clustering at the state-time level
  - Why not? Errors have to be independent between clusters, but serially-correlated errors break this
- But can define a cluster to be a state, each with its own time-series of observations
  - This estimates standard errors without making any assumption about nature of time-series process within each state
- Easy to implement, but not appropriate with few clusters
  - BDM find that, with serially-correlated data that contains no treatment effect, and when using cluster-robust SEs where clusters=states, they incorrectly reject a true null hypothesis of “no treatment effect” 8% (12%) of the time with 10 (6) states using a (nominal) 5% level test

USING THE BOOTSTRAP TO DEAL WITH CLUSTERED DATA, SERIALLY-CORRELATED SHOCKS AND FEW CLUSTERS

Using the bootstrap for inference: overview

- Classical hypothesis test: compare calculated test statistic (from data) to known distribution of test statistic under the null
- What if do not know distribution of test statistic under the null?
  - e.g. t-statistic with non-normal errors and few observations/clusters
- Bootstrap is a method of estimating the distribution of a test statistic by resampling from one's data. It treats the data as if they were the population (Horowitz, 2001).
- If used properly, bootstraps can be extremely useful
  - Can estimate the distribution of a test statistic under few assumptions (about, e.g., nature of unobserved shocks)
  - May be more accurate in small samples (formally: some test statistics based on the bootstrap converge more quickly than those based on standard statistics; "asymptotic refinement")

Using the bootstrap for hypothesis testing: example of “pairs bootstrap”

- Assume testing hypothesis $\beta = 0$ (at 5% level) using t-stat
  - Using original sample, estimate coefficient β_0 and calculate t-stat t_0
  - For b = 1 to B
    - Construct new dataset by sampling with replacement rows (i.e. “pairs” of Y and X) of existing dataset
    - Estimate coefficient β_b and calculate t-stat, t_b, **centred on original estimate β_0**
  - $\{t_1 \ldots t_B\}$ is your estimated distribution of t-statistics. Calculate the 2.5th and 97.5th quantiles
  - Reject hypothesis if original test statistic <= 2.5th quantile or >= 97.5th quantile

- NB this method is preferred to one which uses the bootstrap only to estimate the standard-error of estimated β_0

Using the bootstrap with clustered data

- Can account for clustered errors if we create our new samples in a way that preserves key features of original sampling scheme
- E.g. pairs bootstrap with clustered data
  - Key feature: shocks within a cluster might be correlated
  - Amend previous algorithm to re-sample (with replacement) entire clusters of data, not individual observations
  - Means that new samples also have errors correlated within clusters
  - Called the “pairs block bootstrap” (or “pairs clustered bootstrap”)

- Problem:
  - with few clusters, some generated samples will, by chance, contain limited or no variation in treatment
  - Cameron et al (2008) find pairs block bootstrap OK for >=30 groups

Using the bootstrap by drawing from residuals (1)

- Alternative to pairs bootstrap is to create new samples by holding constant *X*, drawing residuals (from some estimated or known distribution), and generating new *Y*.
- In simplest case
  - Residuals, *e*, estimated using full sample
  - Data for person *i* in sample *b* $y_i^b = \beta X_i + e_j$
    - *j* is chosen randomly (with replacement) from entire sample of individuals.
  - (Called “residual bootstrap”. Uses estimated distribution of residuals as guide to distribution of underlying unobserved error term)
- However, method unlikely to work in our case
  - Assumes errors are homoscedastic
  - Can allow for clustering, but only if all clusters the same size

Using the bootstrap by drawing from residuals (2)

- More complicated case: wild bootstrap
  - Residuals, *e*, estimated using full sample
  - To create data for person *i* in sample *b*, hold *X* constant and generate new *Y*:
    - With probability ½: $y_i^b = \beta X_i + e_i$
    - With probability ½: $y_i^b = \beta X_i - e_i$
  - (Called “wild bootstrap”. Uses each individual’s residual as guide to each individual’s distribution of underlying unobserved error term)

- Easy to modify for clustered data
  - When choosing whether to add or subtract residual to create new values of *y*, make same choice for all individuals in given cluster
  - Does not assume homoscedasticity, and group size irrelevant

Using the bootstrap with clustered data (3)

- Cameron et al (2008) find best performance given by “wild cluster bootstrap”
  - gets the rejection rate for true null with as few as six clusters
  - NB they also suggest that should estimate residuals having imposed the null. If null is that “coefficient is zero”, this means that residuals should be estimated from model which omits that variable

- For full details of implementation, see their Appendix B. See also
  - example Stata code from one of the authors at **http://www.econ.ucdavis.edu/faculty/dlmiller/statafiles/**
  - Bansi Malde’s ado file at **http://tinyurl.com/c8vz3br**

Recap

- Clustering always affects inference, and is especially problematic when the treatment is applied at the cluster level
- Having clustered data means
  - standard errors are (a lot) larger
  - need many clusters for t-statistics to have expected normal distribution
- Longitudinal data can help, but helps less if data are serially correlated

- Main solutions to clustering and serial correlation either
  - impose more structure, or make further assumptions
  - implement the cluster-robust variance estimator
  - use a bootstrap that accounts for clustering
- With few clusters, only the wild cluster bootstrap seems to perform reliably well

Further reading

- Angrist and Pischke, *Mostly Harmless Econometrics: An Empiricist's Companion* (Princeton University Press, 2008) ,Chapter 8
- Arceneaux and Nickerson, 2009, "Modeling Certainty with Clustered Data: A Comparison of Methods", Political Analysis, doi:10.1093/pan/mpp004
- 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
- Cameron and Trivedi, *Microeconometrics: Methods and Applications (*Cambridge: Cambridge University Press, 2005).
- Wooldridge, J. M., "Cluster-Sample Methods in Applied Econometrics," *American Economic Review* 93 (2003), 133–138.

Other work cited in this presentation, or advanced reading

- Angrist, J., and V. Lavy, (2007) “The Effects of High Stakes High School Achievement Awards: Evidence from a Group-Randomized Trial”
- Donald and Lang, “Inference with Difference-in-Differences and Other Panel Data,” *Review of Economics and Statistics 89:2 (2007), 221–233*.
- Hansen, “Generalized Least Squares Inference in Panel and Multilevel Models with Serial Correlation and Fixed Effects”, *Journal of Econometrics,* 140:2 (2007), 670-94
- Horowitz,“The Bootstrap**”**, in Heckman and Leamer (eds.), *Handbook of Econometrics*, edition 1, volume 5, chapter 52, pages 3159-3228 (Elsevier, 2001)
- Ibragimov and Mueller, 2010, “t -Statistic Based Correlation and Heterogeneity Robust Inference”, JASA, DOI: 10.1198/jbes.2009.08046
- Liang and Zeger, “Longitudinal Data Analysis Using Generalized Linear Models,” *Biometrika* 73 (1986), 13–22
- Moulton, “Random Group Effects and the Precision of Regression Estimates,” *Journal of Econometrics* 32 (1986), 385–397
- Moulton,*“An Illustration of a Pitfall in Estimating the Effects of Aggregate Variables on Micro Units,” REStat,* 72 (1990), 334–338
- White, H., “A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity,” Econometrica 48 (1980), 817–838.

Using the bootstrap for hypothesis testing: example

- Suppose want to test that β = 0 using a t-statistic. Three broad methods to create new samples

1. “Pairs bootstrap”
   - Re-sample data (i.e. rows of dataset) with replacement
2. “Residual bootstrap”
   - Keep *X*s the same but create sample with new *Y* variable by drawing from set of empirical residuals
   - Detail: residuals can be estimated using the full model, or (better) a restricted model which imposes the null
   - Requires errors to be homoscedastic
3. “Wild bootstrap”
   - Keep *X*s the same but create sample with new *Y* variable by either adding or subtracting (with equal probability) actual empirical residual *for that observation*
   - Detail: residuals can be estimated using the full model, or (better) a restricted model which imposes the null
   - Does not requires errors to be homoscedastic