

Methods for Difference-in-Differences Studies

Laura Hatfield, PhD | Associate Professor | Harvard Medical School
Fields Institute | Toronto, Ontario | September 25, 2018

Thanks to our team



Alyssa Bilinski



Alex McDowell



Nancy Beaulieu



Sherri Rose



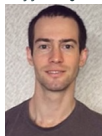
Laura Hatfield



Michael McWilliams



Toyya Pujol-Mitchell



Bret Zeldow



Megan Schuler



Jamie Daw



Andrew Hicks



Mike Chernew

Thanks to the National Institute on Aging and the Laura and John Arnold Foundation for funding.

I study policy interventions

Vox

One of Obamacare's big experiments to lower costs is working surprisingly well

The least-sexy news in health care these days — bundled payments — might also be the most important.

By Dylan Scott | @dylaniscott | dylan.scott@vox.com | Sep 4, 2018, 11:00am EDT

The New York Times

Would Americans Accept Putting Health Care on a Budget?

The intuitive appeal of such a system is growing, and it's getting a test in Maryland.

Forbes

1,926 views | Sep 18, 2018, 08:20am

Half Of U.S. Doctors Have Pay Tied To 'Value-Based Metrics'

AJMC

Managed Markets
Network

News

Newsroom – Published on: September 10, 2018

California Medicaid Expansion Had Mixed Effects for HIV Population

Jaime Rosenberg

With the expansion of Medicaid, eligible, low-income people living with HIV were transferred from the Ryan White Program to Medicaid, causing concern for interrupted access to care and treatment.

The Boston Globe

Beth Israel-Lahey merger raises a Medicaid issue

Low-income patients

If the merger is approved, Massachusetts' two largest health systems would be the new Beth Israel Lahey Health and Partners HealthCare. Both would treat a lower share of Medicaid patients than the state average.

Outcomes before an intervention



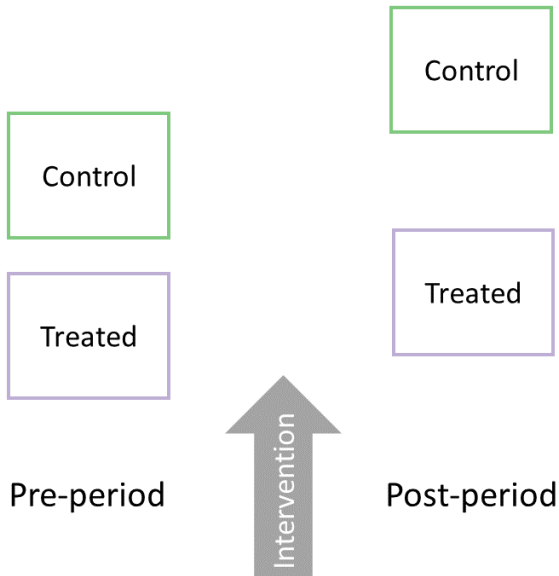
Control

The diagram consists of two vertically stacked rectangular boxes. The top box has a green border and contains the word 'Control'. The bottom box has a purple border and contains the word 'Treated'. Below these boxes is the text 'Pre-period'.

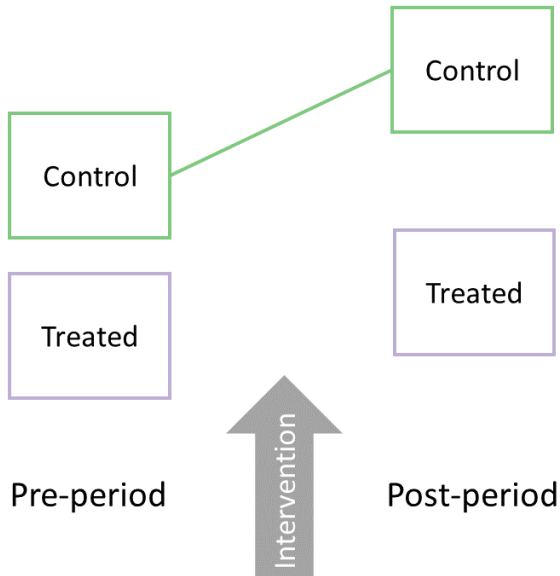
Treated

Pre-period

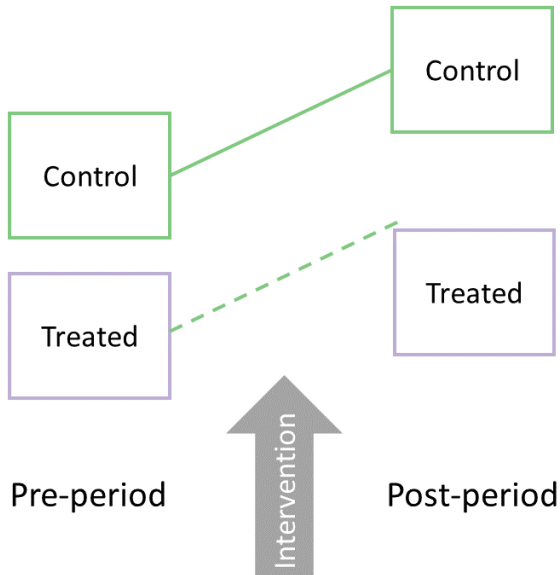
Outcomes after an intervention



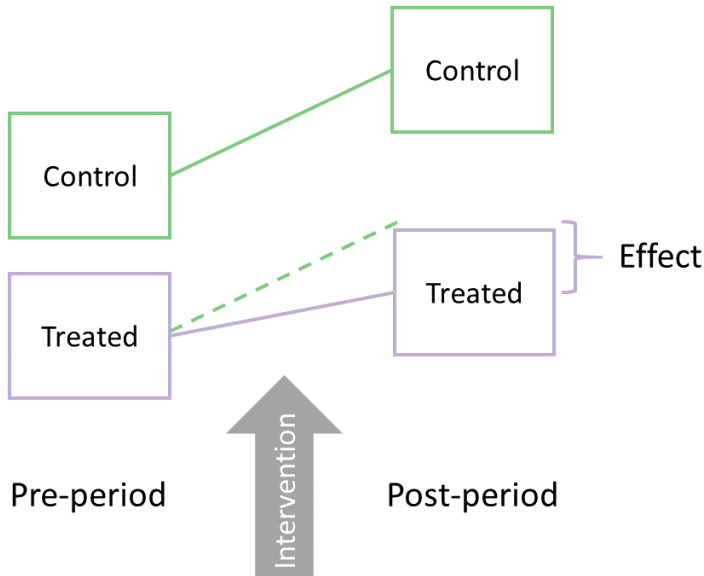
Change in the control group outcomes...



...is counterfactual for treated change



Compare actual and counterfactual changes



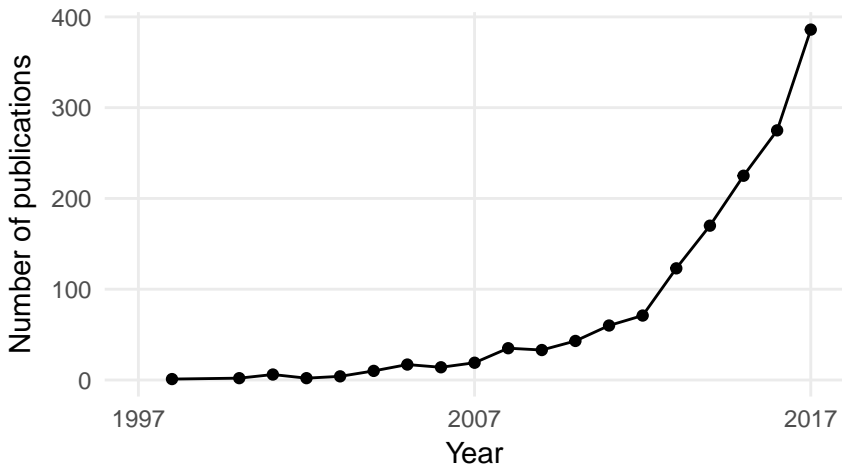
Key causal assumption

*The change in outcomes in the control group
from pre- to post-intervention
is a **valid counterfactual**
for the change that would have occurred
in the treated group
in the absence of intervention*

Put another way...

*The **difference** between
treated and control
would remain **constant**
in the absence of intervention*

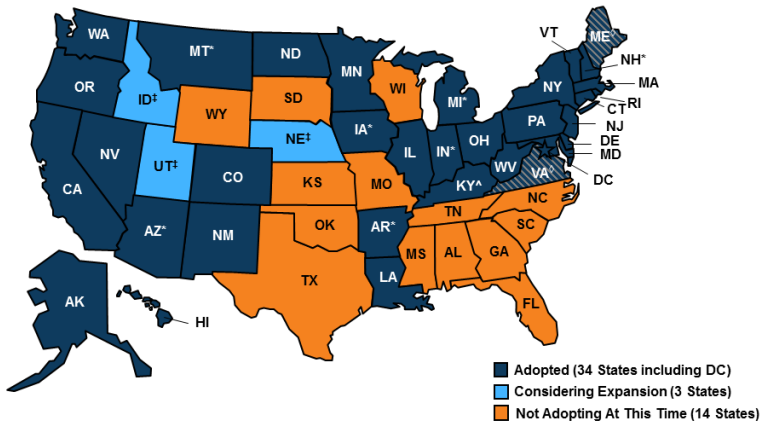
Diff-in-diff is increasingly popular



Source: PubMed search for "difference-in-difference" OR "difference-in-differences"

For example, the Medicaid expansion of the ACA...

Status of State Medicaid Expansion Decisions



NOTES: Current status for each state is based on KFF tracking and analysis of state activity. *AR, AZ, IA, IN, MI, MT, and NH have approved Section 1115 expansion waivers. †On June 29, 2018, the DC federal district court invalidated the Kentucky HEALTH expansion waiver approval and sent it back to HHS to reconsider the waiver program. ‡UT passed a law directing the state to seek CMS approval to partially expand Medicaid to 100% FPL using the ACA enhanced match. ID, NE, and UT have measures on their November ballots to fully expand Medicaid to 138% FPL. †Expansion is adopted but not yet implemented in VA and ME. (See the link below for more detailed state-specific notes.)

SOURCE: "Status of State Action on the Medicaid Expansion Decision," KFF State Health Facts, updated September 11, 2018.

<https://www.kff.org/health-reform/state-indicator/state-activity-around-expanding-medicaid-under-the-affordable-care-act/>

... which has been extensively studied

202 studies

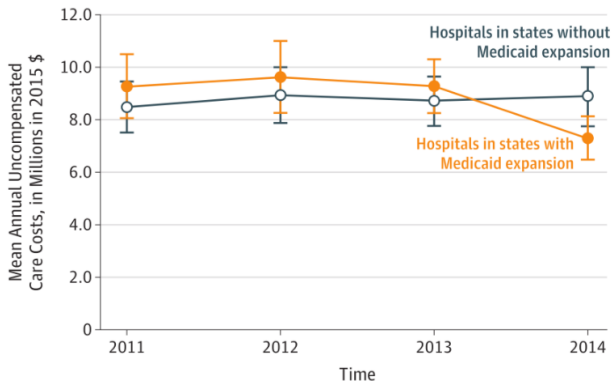
A review of the research shows that **Medicaid expansion** is associated with:

- *Significant coverage gains
- *Increased access to care, utilization & affordability
- *Reductions in uncompensated care costs for hospitals and clinics

KFF
HENRY J KAISER
FAMILY FOUNDATION

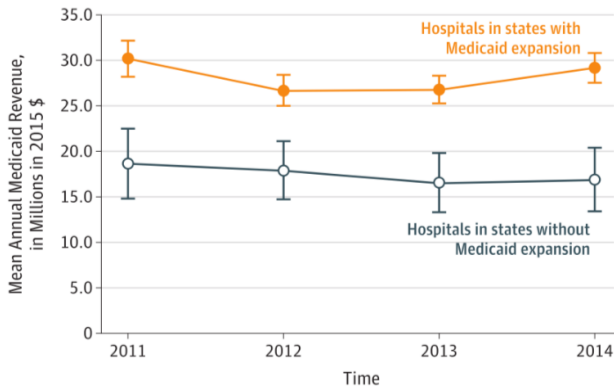
Some results are easy & clean

Figure 1. Trends in Mean Annual Uncompensated Care Costs for Hospitals in States With and Without Medicaid Expansion for Fiscal Years 2011-2014



Others are harder to see

Figure 2. Trends in Mean Annual Medicaid Revenue for Hospitals in States With and Without Medicaid Expansion for Fiscal Years 2011-2014



Source: Blavin et al JAMA 11 Oct 2016

Our research questions

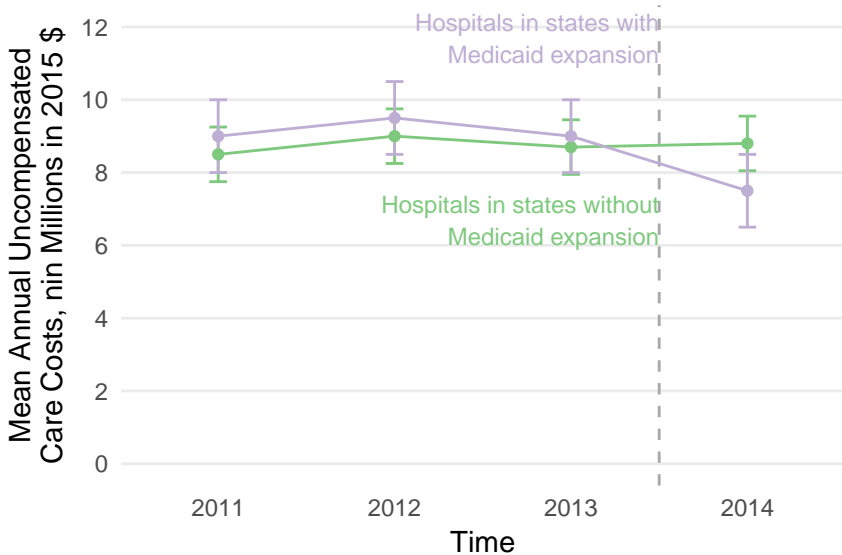
1. Should we match on pre-intervention variables?
2. How useful is a test of parallel pre trends?
3. What does confounding mean in diff-in-diff?
4. What impact do variance patterns have on inference?

[Q1: matching]

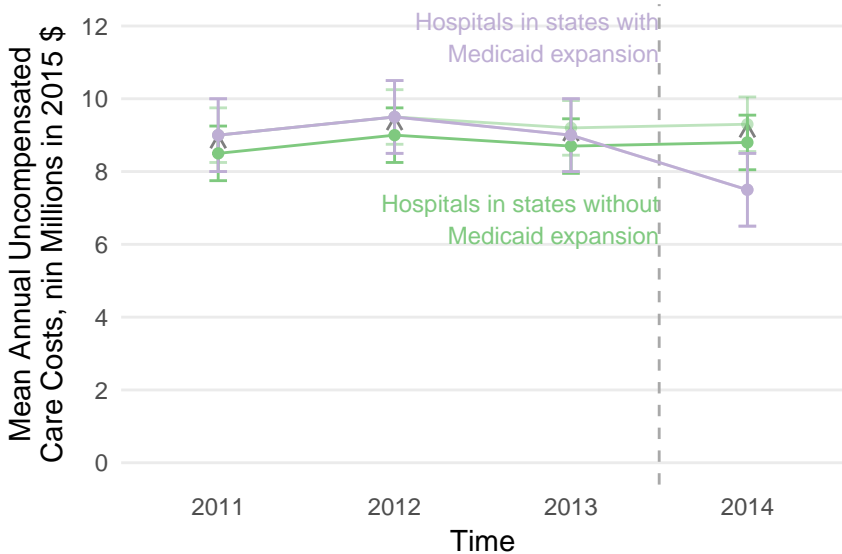
[Q1: matching]

1. **Should we match on pre-intervention variables?**
2. How useful is a test of parallel pre trends?
3. What does confounding mean in diff-in-diff?
4. What impact do variance patterns have on inference?

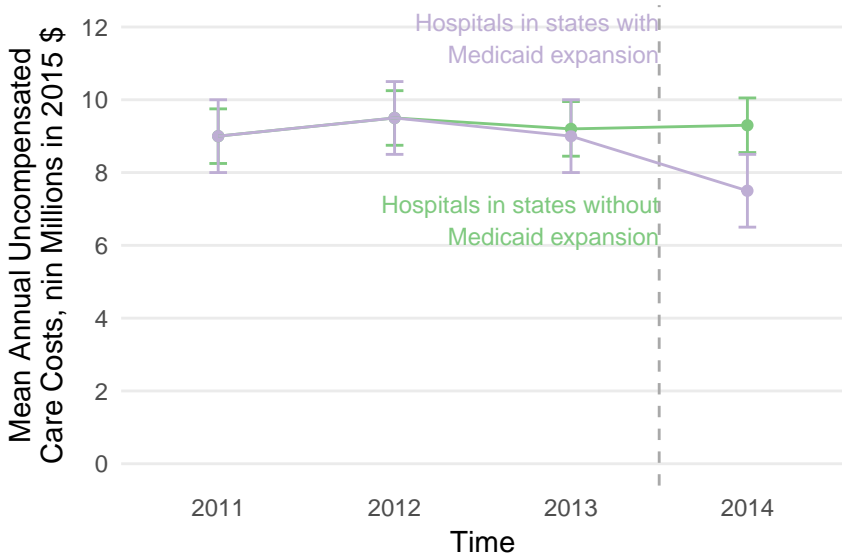
Recall the Medicaid expansion example



Many researchers want the groups to be more similar



Matching is one way to accomplish that



People often match in diff-in-diff

Our primary analytic approach is to examine dependent variables before and after the intervention alongside a contemporaneous comparison group (a “difference-in-difference” approach). To ensure a valid comparison, we further refine our cohorts so that the intervention and comparison groups are similar along important dimensions using a propensity score approach (Rosenbaum and Rubin 1984; D’Agostino 1998; Rubin 1998).

fifth grade (we refer to these as *baseline schools*). Specifically, baseline schools are considered matched when they have School Performance Scores (SPS) in the same five-point bin.¹³ In addition to baseline schools, the RSD comparison sample matches grandfathering-eligible and ineligible students on race, sex, baseline year, baseline special education status, and baseline subsidized lunch eligibility.

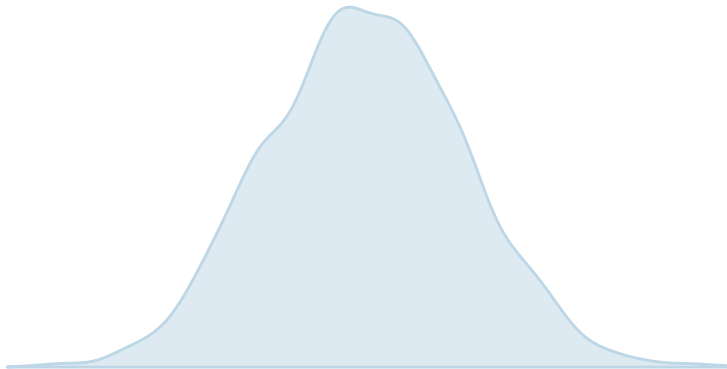
We used propensity score matching to select 7 groups of comparison practices — 1 for each region. We selected up to 5 comparison practices per initiative practice to ensure that there were similar characteristics across patients (e.g., age, sex, chronic conditions, and prior expenditures and use of services), practices (e.g., meaningful use of EHRs and number of clinicians), and markets (e.g., mean county income) (Section 3 in the Supplementary Appendix).¹⁵

The control population included employees in MarketScan who were also continuously enrolled over the study period, matched to employees in the intervention population using both exact matching and propensity score matching.

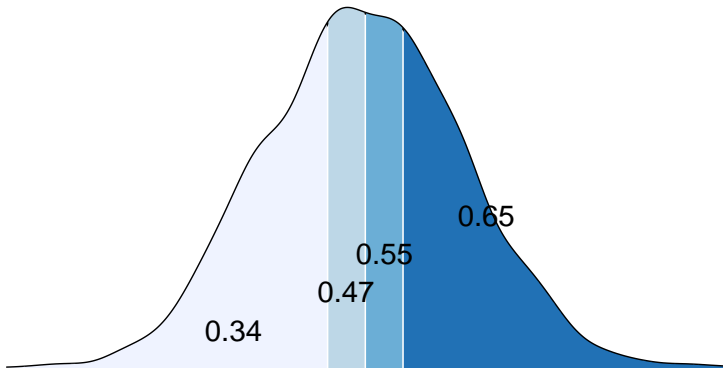
The REASON for the difference matters



Suppose a single population



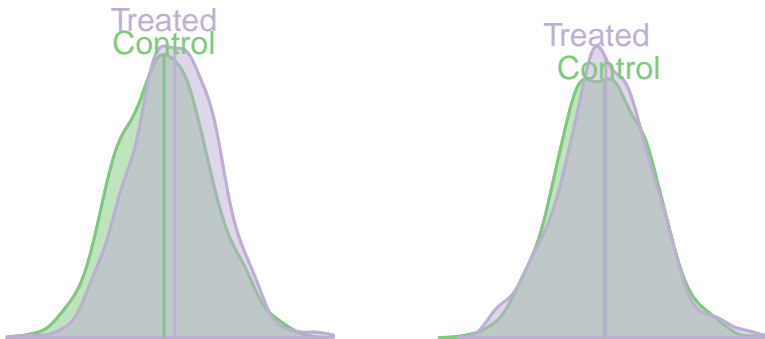
Probability of treatment increases with baseline level



In the unmatched analysis

Pre intervention, treated group has a higher mean

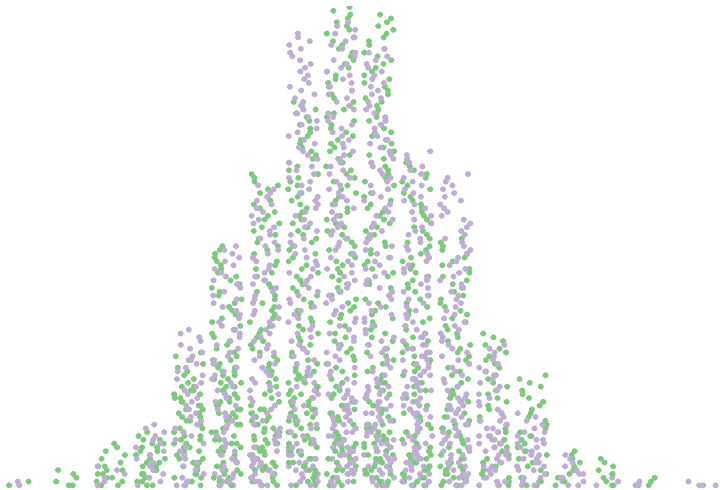
Post intervention, both groups regress back to their common mean



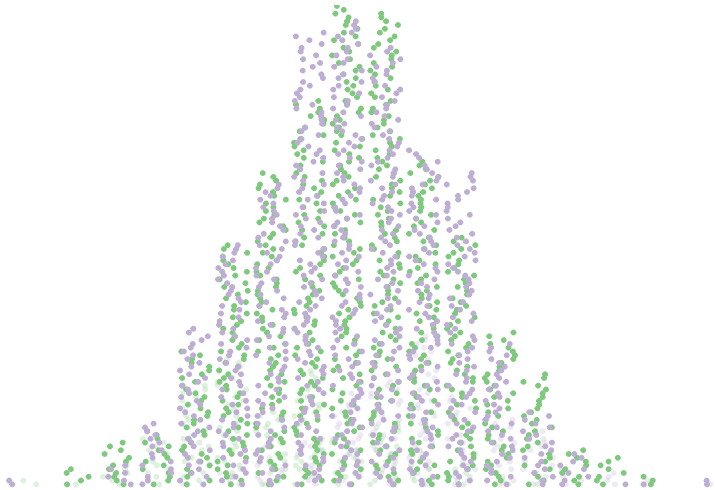
Diff-in-diff estimate is biased

term	estimate	std.error
(Intercept)	-0.178	0.026
trt	0.372	0.036
post	0.233	0.036
post.trt	-0.399	0.051

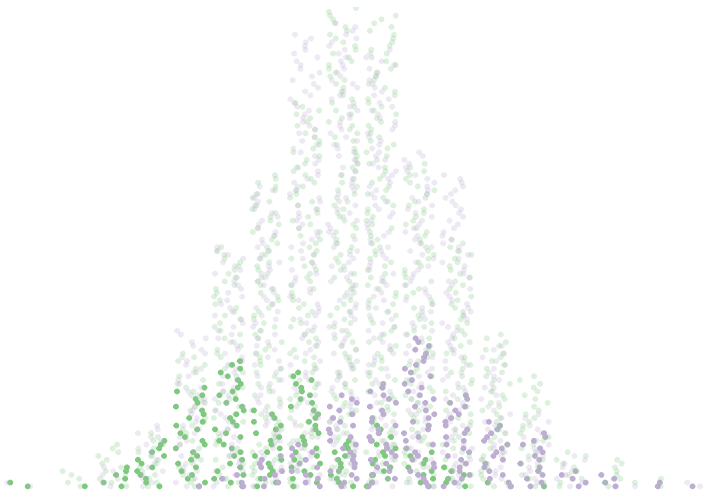
Can match on covariates or outcomes



Choose units from each that are relatively closer together



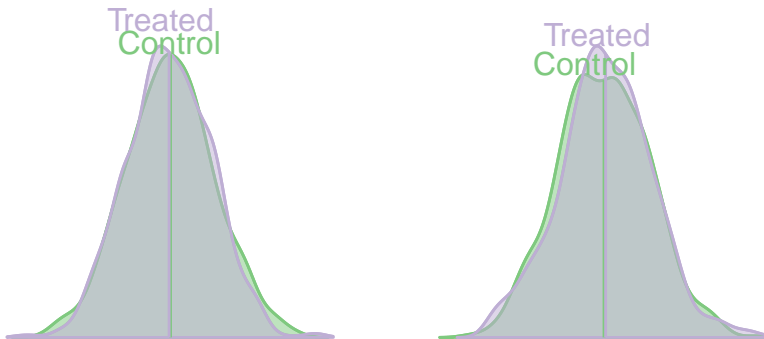
Unmatched units are more separated



In the matched analysis

Pre intervention, means are the same (by construction)

Post intervention, there's no regression to the mean



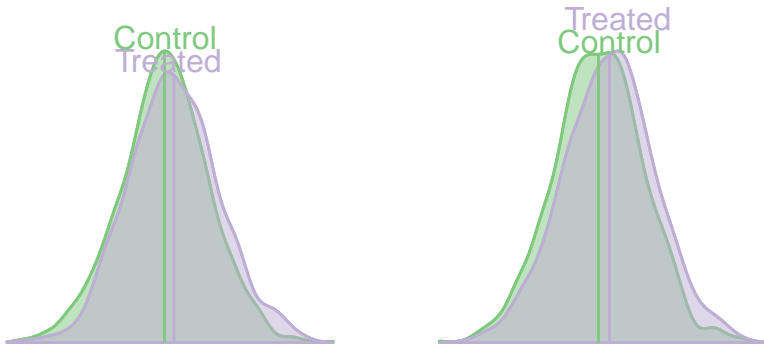
Diff-in-diff estimate is unbiased

term	estimate	std.error
(Intercept)	0.007	0.028
trt	0.012	0.040
post	0.040	0.040
post.trt	-0.043	0.056

What if underlying populations are different?

Pre intervention, treated group has different distribution

Post intervention, there's no regression to the mean



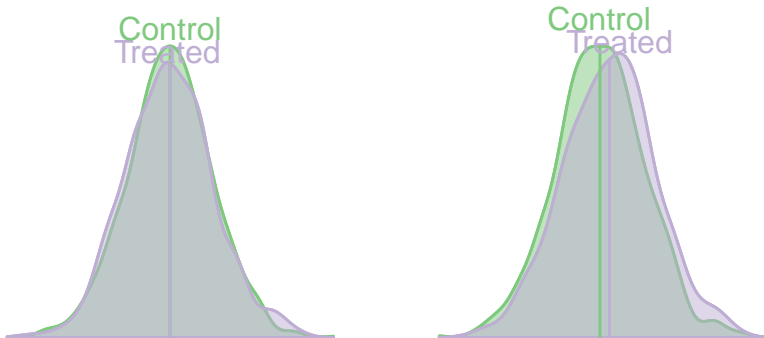
And the diff-in-diff estimate is unbiased

term	estimate	std.error
(Intercept)	-0.155	0.026
trt	0.336	0.036
post	0.016	0.036
post.trt	0.018	0.051

In the matched analysis

Pre intervention, the two groups have similar distributions

Post intervention, they regress back to their respective means

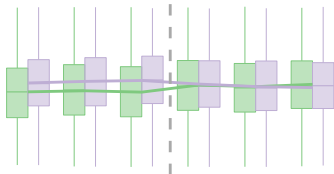


And the diff-in-diff estimate is biased

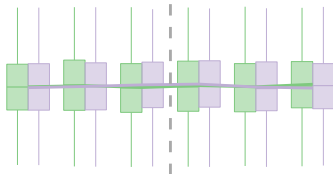
term	estimate	std.error
(Intercept)	0.023	0.028
trt	0.012	0.039
post	-0.155	0.039
post.trt	0.338	0.055

We cannot tell the difference from the data

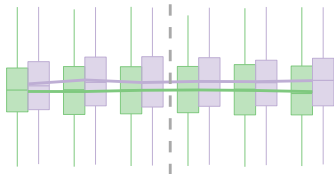
Same pop, unmatched



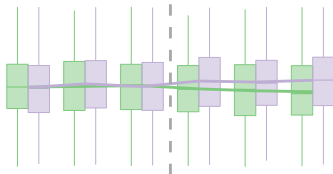
Same pop, matched



Diff pops, unmatched



Diff pops, matched



Lest you think small effect sizes aren't important

Adjusted Medicare spending and spending trends were similar in the ACO cohorts and the control group during the precontract period. In 2013, the differential change (i.e., the between-group difference in the change from the precontract period) in total adjusted annual spending was $-\$144$ per beneficiary in the 2012 ACO cohort as compared with the control group ($P=0.02$), consistent with a 1.4% savings, but only $-\$3$ per beneficiary in the 2013 ACO cohort as compared with the control group ($P=0.96$). Estimated savings were consistently greater in independent primary care groups than in hospital-integrated groups among 2012 and 2013 MSSP entrants ($P=0.005$ for interaction). MSSP contracts were associated with improved performance on some quality measures and unchanged performance on others.

Conclusions


- Matching on levels
 - corrects bias
 - creates bias
 - has nothing to do with bias

Further reading on matching in diff-in-diff

HSR


Health Services Research

Matching and Regression to the Mean in Difference-in-Differences Analysis

Jamie R. Daw  and Laura A. Hatfield

Editorial

Well-Balanced or too Matchy-Matchy? The Controversy over Matching in Difference-in-Differences

Andrew M. Ryan 

Editorial

Matching in Difference-in-Differences: between a Rock and a Hard Place

Jamie R. Daw 
Laura A. Hatfield

Further reading on matching in diff-in-diff

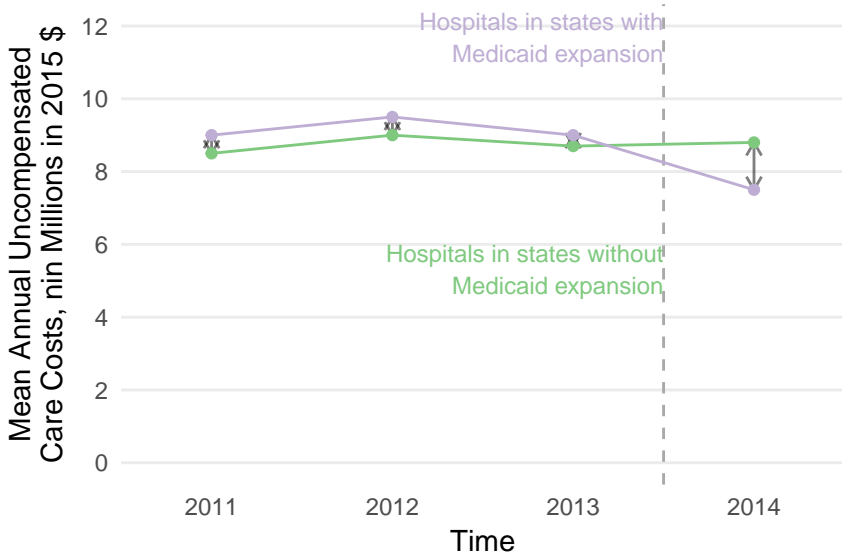
- Chabé-Ferret, S. Should we combine difference in differences with conditioning on pre-treatment outcomes? (Toulouse School of Economics, 2017).
- Daw, J. R. & Hatfield, L. A. Matching and regression-to-the-mean in difference-in-differences analysis. Health Services Research (2018). doi:10.1111/1475-6773.12993
- Daw, J. R. & Hatfield, L. A. Matching in difference-in-differences: Between a rock and a hard place. Health Services Research (2018). doi:10.1111/1475-6773.13017
- Ryan, A. M. Well-balanced or too matchy-matchy? The controversy over matching in difference-in-differences. Health Services Research (2018). doi:10.1111/1475-6773.13015
- Ryan, A. M., Burgess, J. F. & Dimick, J. B. Why we should not be indifferent to specification choices for difference-in-differences. Health Services Research (2015). doi:10.1111/1475-6773.12270
- Stuart, E. A. et al. Using propensity scores in difference-in-differences models to estimate the effects of a policy change. Health Services and Outcomes Research Methodology 14, 166-182 (2014).

[Q2: testing trends]

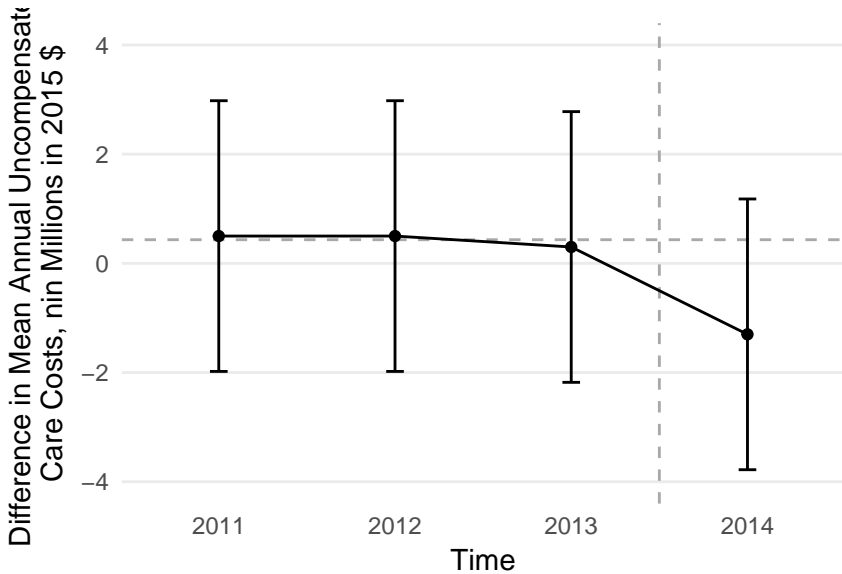
[Q2: testing trends]

1. Should we match on pre-intervention variables?
2. **How useful is a test of parallel pre trends?**
3. What does confounding mean in diff-in-diff?
4. What impact do variance patterns have on inference?

Pre-period differences are felt to be important



“Parallel trends” implies constant difference in the pre-period



Usual statement: two assumptions

JAMA Guide to Statistics and Methods

Methods for Evaluating Changes in Health Care Policy The Difference-in-Differences Approach

Justin B. Dimick, MD, MPH; Andrew M. Ryan, PhD

The 2 main assumptions of difference-in-differences analysis are parallel trends and common shocks.⁴ The *parallel trends assumption* states that the trends in outcomes between the treated and comparison groups are the same prior to the intervention (Figure). If true, it is reasonable to assume that these parallel trends would continue for both groups even if the program was not implemented. This is tested empirically by examining the trends in both groups before the policy was implemented.

In economics, a *shock* is an unexpected or unpredictable event (unrelated to the policy) that affects a system. *The common shocks assumption* states that any events occurring during or after the time the policy changed will equally affect the treatment and compari-

Problems with the parallel pre-trends “assumption”

1. Fundamental problem: not an assumption at all
2. Statistical problem: wrong null hypothesis
3. Conceptual problem: neither necessary nor sufficient

Fundamental problem: people often test pre trends

We tested our data for parallel trends in prescribing between “never-MCL” states and pre-MCL years for states that implement the policy during our study period; we cannot reject the null hypothesis of parallel trends, which supports the use of our models (see online eAppendix eTable 8 in the Supplement).

Thirdly, admission trends for incentivised and non-incentivised conditions were divergent before the introduction of the pay for performance scheme. This violates the assumption of parallel trends required for a simple difference in difference analysis, and we therefore adjusted admission rates for underlying trends in the pre-Quality and Outcomes Framework period.

Figure 2 shows parallel pre-takeover trends in years up to, but not including, the last grade of legacy school enrollment

¹ We tested the validity of the parallel trends assumption by comparing the changes in outcomes between adult learners and regular students before the blended programme was implemented. In the absence of treatment one would expect both groups to move in tandem. This is exactly what we found in the data for the bridge programme in 2009 and 2010 (in the bridge programme the blended condition was only implemented in 2011). In other words, a difference-in-difference model using the data for 2009 and 2010 in the bridge programme only, showed no significant “treatment” effects on either dropout rates, exam pass rates or course pass rates (results are available upon request).

The usual test of parallel trends

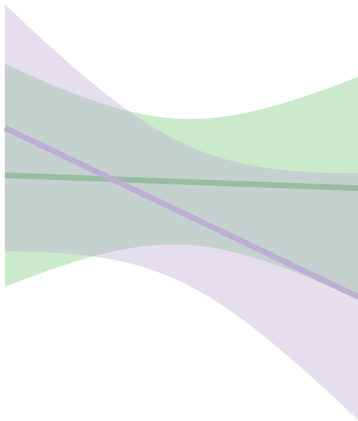
Fit a regression for pre-period data

$$E(y_{it}) = \beta_0 + \beta_1 \text{trt}_i + \beta_2 \text{time}_t + \beta_3 \text{trt}_i \text{time}_t$$

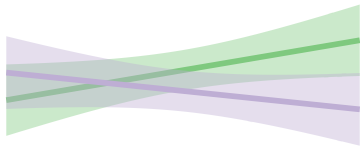
Test a null of no slope difference $H_0 : \beta_3 = 0$

Easiest way to “pass” this test is low power

High variance



Low variance



The statistical problem: wrong null hypothesis

Under-powered studies "pass"



Well-powered studies "fail"¹



¹Recall, as $N \rightarrow \infty$, $Pr(\text{reject } H_0) \rightarrow 1$

Non-inferiority formulation of the test

Fit a regression to the pre-period data only

$$E(y_{it}) = \beta_0 + \beta_1 trt_i + \beta_2 time_t + \beta_3 trt_i time_t$$

Test a null of a large slope difference $H_0 : |\beta_3| > \delta$

This fixes the paradox...

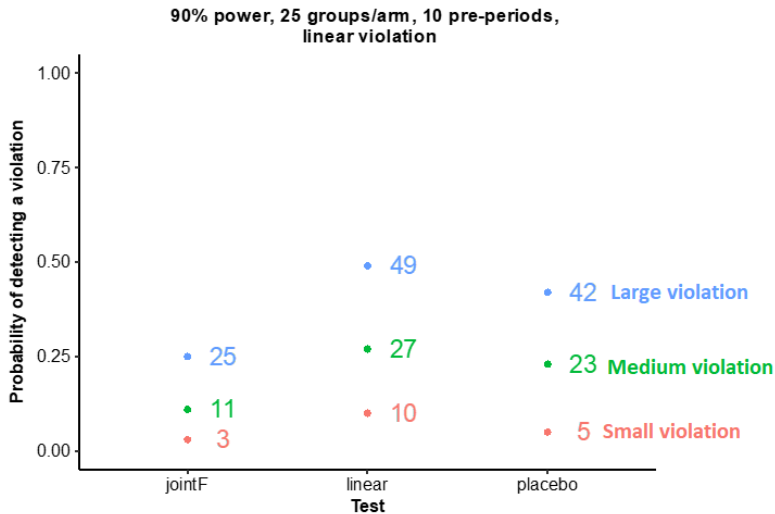
Under-powered tests fail to reject



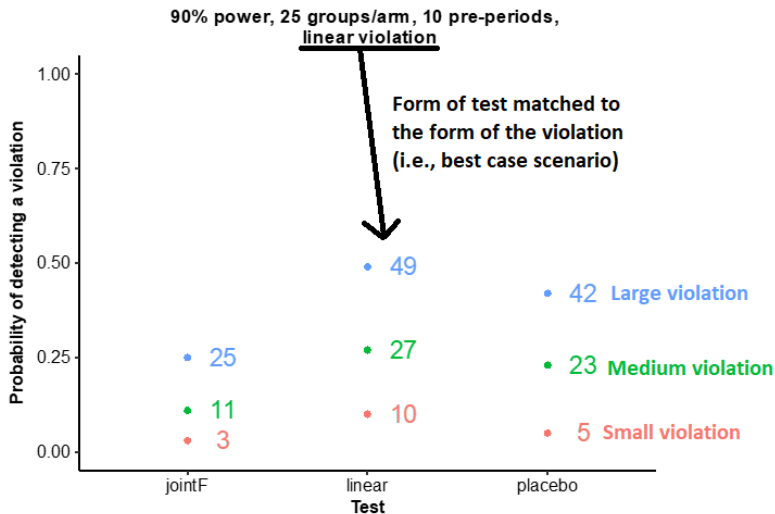
Rejection implies the violation not "too big"



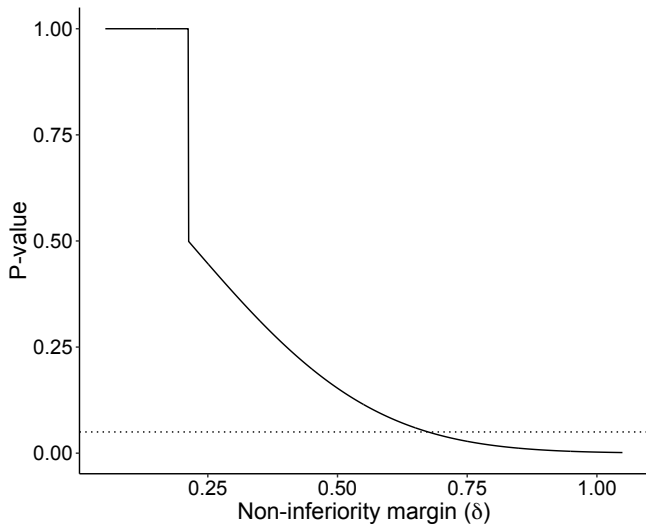
...but has terrible power...



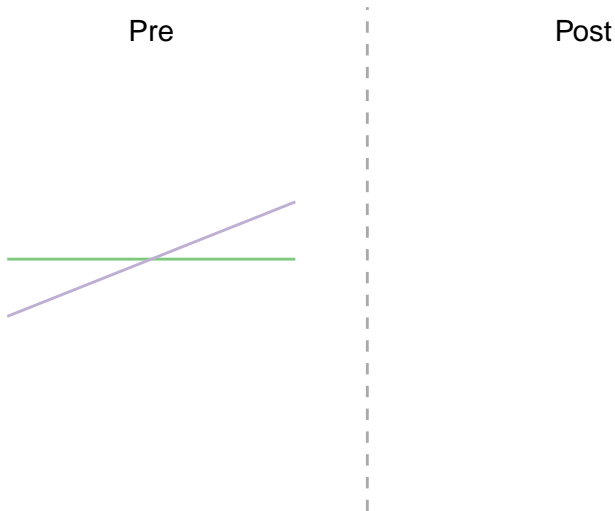
...and requires us to choose a threshold



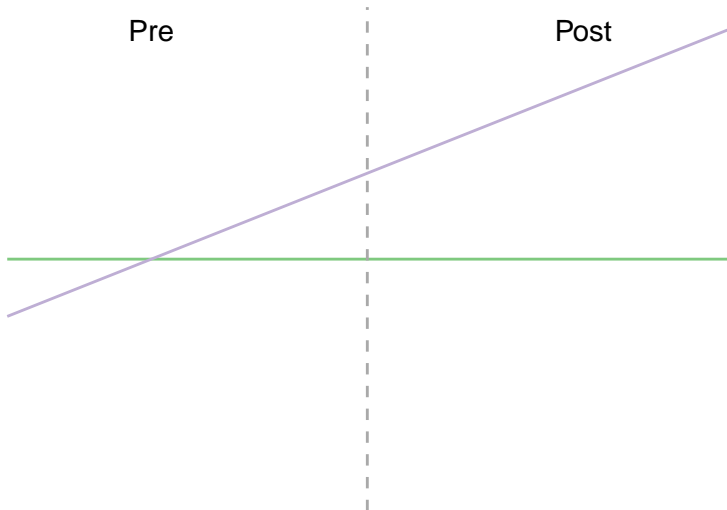
Varying the threshold



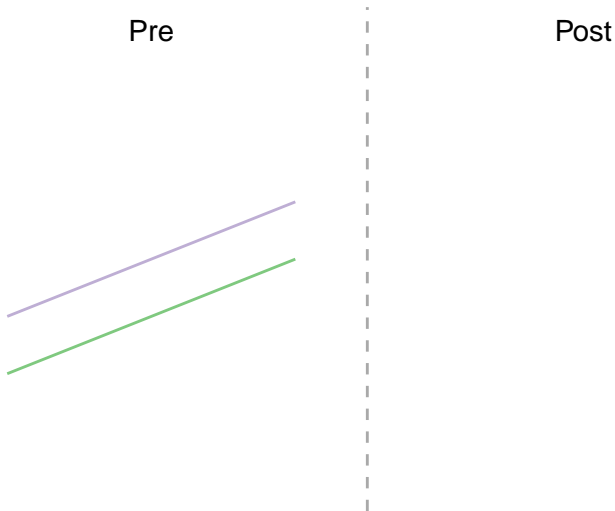
Conceptual problem: neither necessary nor sufficient



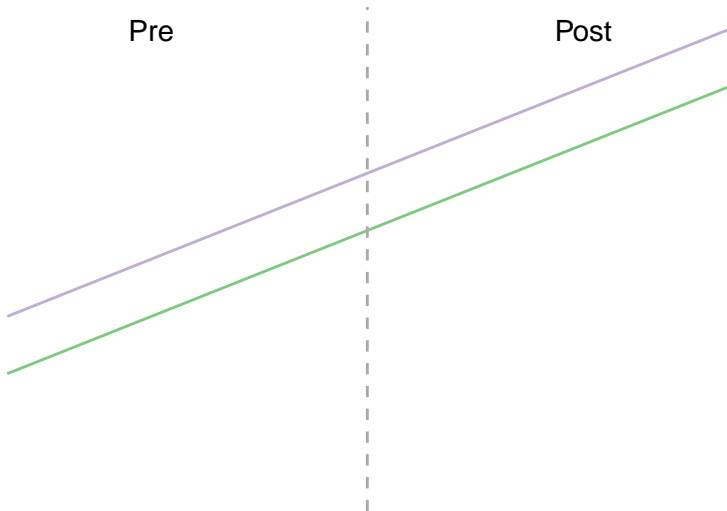
Conceptual problem: neither necessary nor sufficient



Conceptual problem: neither necessary nor sufficient



Conceptual problem: neither necessary nor sufficient



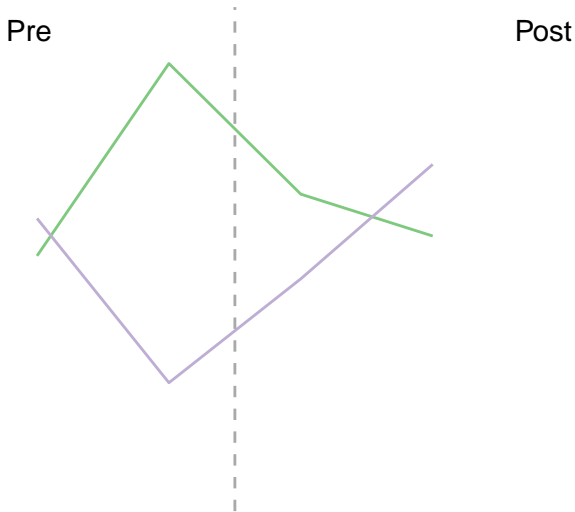
Conceptual problem: neither necessary nor sufficient

Pre

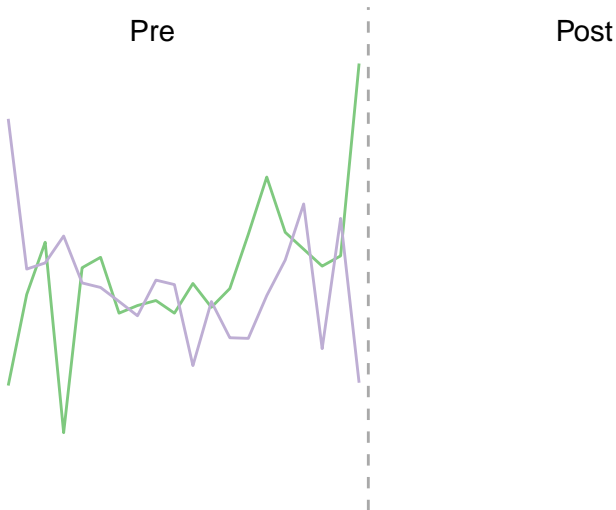
Post



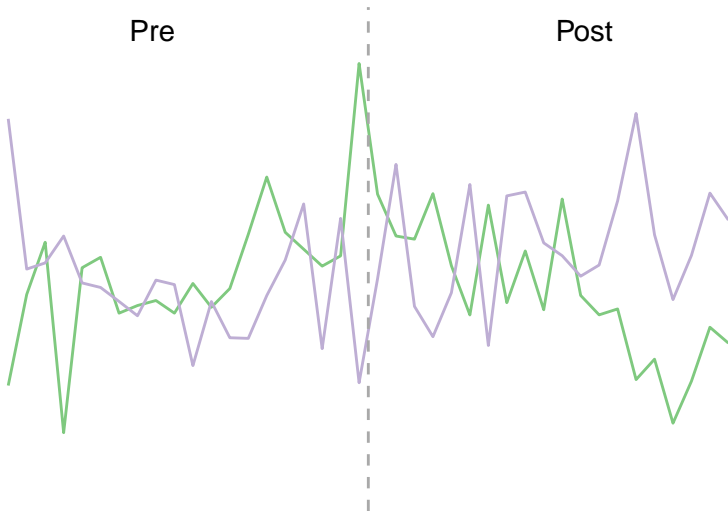
Conceptual problem: neither necessary nor sufficient



Conceptual problem: neither necessary nor sufficient



Conceptual problem: neither necessary nor sufficient



Further reading on trends and testing assumptions

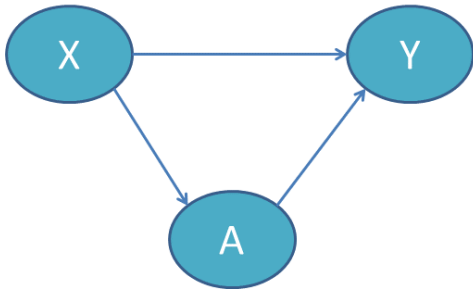
- Bilinski, A. & Hatfield, L. A. Seeking evidence of absence: Reconsidering tests of model assumptions. arXiv:1805.03273 [stat] (2018).
- Freyaldenhoven, S., Hansen, C. & Shapiro, J. M. Pre-event trends in the panel event-study design. (National Bureau of Economic Research, 2018).
- Hartman, E. & Hidalgo, F. D. An equivalence approach to balance and placebo tests. American Journal of Political Science (To appear). doi:10.7910/dvn/rynsdg
- Kahn-Lang, A. & Lang, K. The promise and pitfalls of differences-in-differences: reflections on '16 and Pregnant' and other applications. (National Bureau of Economic Research, 2018). doi:10.3386/w24857
- Roth, J. Should we adjust for the test for pre-trends in difference-in-difference designs? arXiv:1804.01208 [econ, math, stat] (2018).

[Q3: confounding]

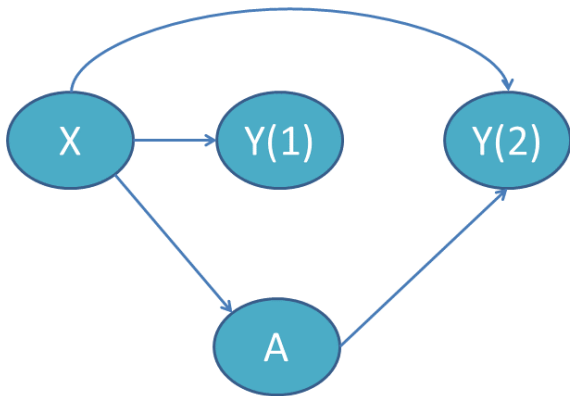
[Q3: confounding]

1. Should we match on pre-intervention variables?
2. How useful is a test of parallel pre trends?
3. **What does confounding mean in diff-in-diff?**
4. What impact do variance patterns have on inference?

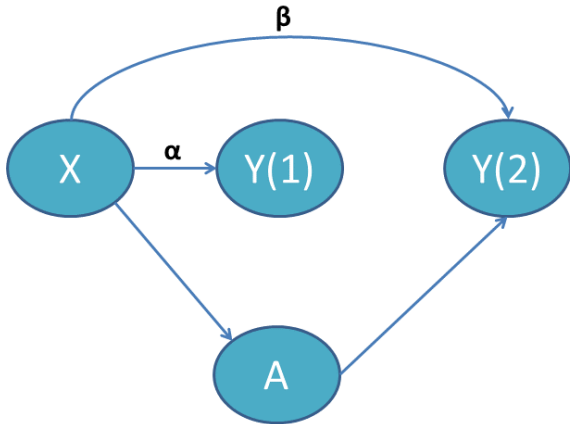
Confounding in cross-sectional setting



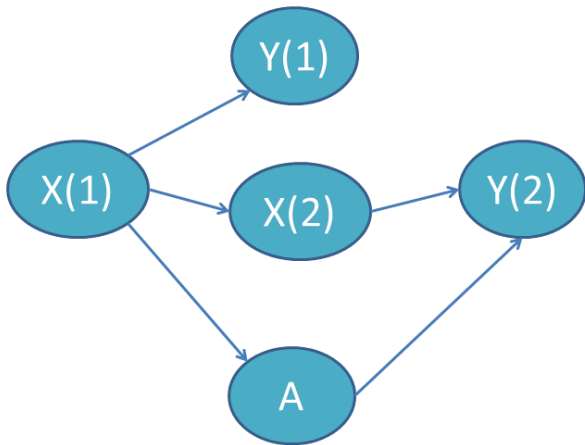
Confounding in two-period settings



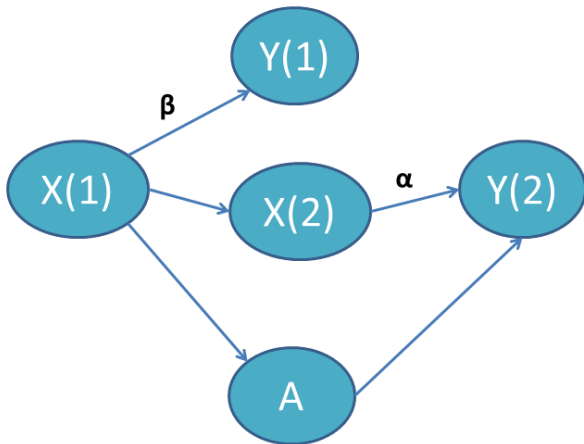
Confounding in two-period settings



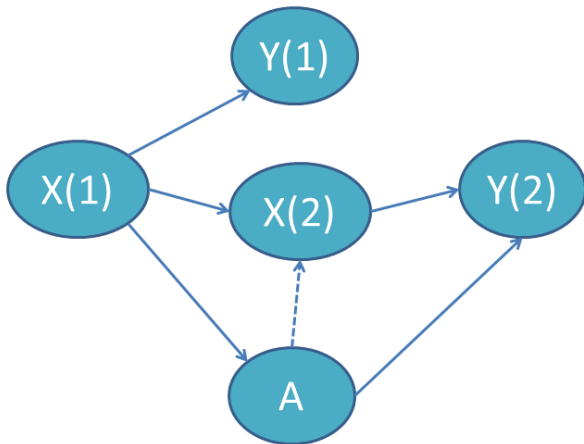
Confounding with time-varying covariates



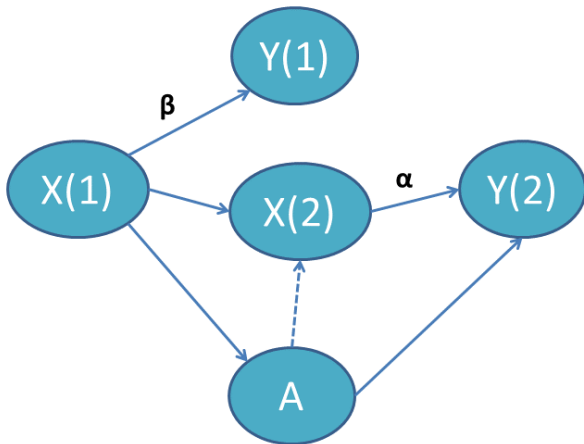
Confounding with time-varying covariates



Confounding with time-varying covariates



Confounding with time-varying covariates



Further reading on model specification and confounding

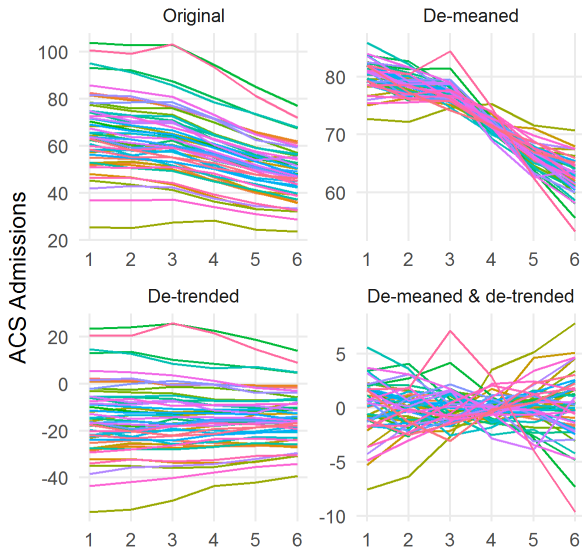
- Doudchenko, N. & Imbens, G. W. Balancing, regression, difference-in-differences and synthetic control methods: A synthesis. (National Bureau of Economic Research, 2016).
- Goodman-Bacon, A. Difference-in-differences with variation in treatment timing. (2018).
- Imai, K. & Kim, I. S. When should we use fixed effects regression models for causal inference with longitudinal data? (2017).
- Kropko, J. & Kubinec, R. Why the two-way fixed effects model is difficult to interpret, and what to do about it. (2018).
- Mummolo, J. & Peterson, E. Improving the interpretation of fixed effects regression results. *Political Science Research and Methods* 1-7 (2018).
doi:10.1017/psrm.2017.44
- Mora, R. & Reggio, I. Treatment effect identification using alternative parallel assumptions. (Universidad Carlos III de Madrid, 2012).

[Q4: error variance]

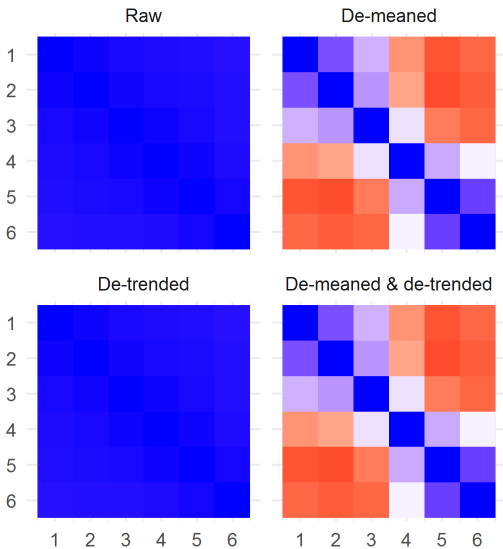
[Q4: error variance]

1. Should we match on pre-intervention variables?
2. How useful is a test of parallel pre trends?
3. What does confounding mean in diff-in-diff?
4. **What impact do variance patterns have on inference?**

Error variance depends on the model



Error variance in real data doesn't follow standard models



Further reading on error structures and inference

- Bertrand, M., Duflo, E. & Mullainathan, S. How much should we trust differences-in-differences estimates? *Quarterly Journal of Economics* 119, 249-275 (2004).
- Conley, T. G. & Taber, C. R. Inference with "difference in differences" with a small number of policy changes. *The Review of Economics and Statistics* 93, 113-125 (2011).
- Donald, S. G. & Lang, K. Inference with difference-in-differences and other panel data. *The Review of Economics and Statistics* 89, 221-233 (2007).
- Rokicki, S., Cohen, J., Fink, G., Salomon, J. A. & Landrum, M. B. Inference with difference-in-differences with a small number of groups: A review, simulation study, and empirical application using SHARE data. *Med Care* 56, 97-105 (2018).

Thank you!



`hatfield@hcp.med.harvard.edu`



`@hpdslab`

`@laura_tastic`



`HealthPolicyDataScience.org`