

# Econometrics II Syllabus 2024

## Instructors

- Konrad Burchardi (5 lectures): [konrad.burchardi@iies.su.se](mailto:konrad.burchardi@iies.su.se)
- David Schönholzer (7 lectures): [david.schonholzer@iies.su.se](mailto:david.schonholzer@iies.su.se)
- David Strömborg (guest lecture): [david.stromberg@ne.su.se](mailto:david.stromberg@ne.su.se)
- Shuhei Kainuma (TA sessions): [shuhei.kainuma@su.se](mailto:shuhei.kainuma@su.se)

## Course Goals

This is a PhD level course on modern econometric techniques with a focus on causal inference. It covers identification, estimation and inference principles, the experimental ideal, matching and propensity scores, instrumental variables, difference-in-differences, and regression discontinuity designs; and it provides an introduction to machine learning. The goals are to equip students with:

1. Formal tools to understand and express issues in modern applied econometrics.
2. Translating formal econometric language into estimators.
3. Apply and test estimators in simulated data.

## Lectures Overview

- Lectures are 10:00-12:00 on Tuesdays and Thursdays, except Lecture 12.
- See the schedule [here](#).
- All lectures are in person. There is no hybrid option.
- 10-minute break every hour.
- Lecture content:
  1. Identification of Causal Effects (David)
  2. Estimation Principles (David)
  3. Inference Principles (Konrad)
  4. Experiments (Konrad)
  5. Matching (Konrad)
  6. IV/2SLS (Konrad)
  7. Heterogeneous IV (Konrad)
  8. Fixed effects (David)
  9. Static DID (David)
  10. Dynamic DID (David)
  11. Matched DID and Synthetic Controls (David)
  12. Regression Discontinuity Design (David)
  13. Machine Learning (David Str.)

## Communication and Materials

- *Slack*: Our primary platform for communication and links to materials.

- Click [here](#) to join our Workspace.
- We strongly encourage all of you to actively participate in discussions on Slack, both when you have doubts and questions about the material, and to help others out where you feel you can! The more actively you participate, the more others will also be inclined to participate.
- There are the following channels:
  - exam: information about the exam (close to exam date).
  - general: general course announcements.
  - lecture##: link to slides and general discussion about each lecture.
  - problemset##: link to problem set and discussion.
  - stata\_coding: questions and tips on Stata coding problems.
  - random: anything else.
- Also feel free to use Slack's direct messaging (DM) system to message each other. For questions to Konrad or David, it is generally better to ask publicly rather than DM so others can benefit from the discussion. We will reply to DMs when it is a personal question that applies only to you. Any other questions you send us in DMs we will re-post in the appropriate public channel and answer there.
- **Athena:** Athena will be used for problem set submission
  - Click [here](#) to access our Athena website. We will post the problem sets and lecture slides on Athena, but note that all relevant material on Athena (i.e. lectures and problem sets) will also be accessible through links on Slack, which will be our primary way of communicating with you.
  - Please make sure you are registered with the correct email on Athena so we have you on the class roster.
  - Submit your problem set solutions through Athena.
  - Other than an initial welcome message, we won't use Athena for communication.
- **Email:** Please do write to us on Slack instead.

## Problem Sets and TA Sessions

- Problem sets consist of both theory and coding questions (about 50/50). Coding exercises serve to help understand theory and to build coding skills.
- The primary coding language is Stata.
  - This is primarily due to its convenience for regressions and continued prevalence in economics (see [here](#)).
  - There is a 2-hour coding session (with Shuhei Kainuma, on April 4, 13:00-15:00) which will go through examples of code meant to equip you with the core knowledge needed for the problem set.
  - Make sure to download and install Stata before the session! You should be able to obtain Stata licenses through your institutions.
  - Please also see these [UCLA](#) and [Princeton](#) Stata guides.
  - You are also encouraged to use [ChatGPT](#) to help you write Stata code. It makes mistakes, but it is very helpful to get started. However, it is not reliable for theory questions.
  - You can use R if you prefer but TA sessions and solutions will all be using Stata.

- You are welcome to work on problem sets together and you are strongly encouraged to discuss them on the corresponding Slack channels. But you have to hand in your own solution.
- Solutions include both answers to theory questions as well as the code that generated your results. Make sure your code runs and produces the desired result. This contributes to your problem set grade.
- Problem set submission:
  - They are due at 4pm the day *before the TA session*. Submit your solution on the Athena course page.
  - Later submissions will not be graded (and hence receive zero points).
- Problem set grading:
  - You can receive between 0 and 4 points for a solution. Full credit will be awarded if a serious, independent effort for each question has been made, and you've had at least partial success with your answers. Getting every answer right is not a requirement for full credit. Answers that are copied from others or previous years receive no credit.
  - If you feel like we have seriously misjudged the effort and success of your PS solution, you can submit a regrading request by David (PS1, PS4, and PS5) or Konrad (PS1, PS2, and PS3). Requesting a regrade may increase or decrease the credit you get for your PS submission.
  - By default, no personalized feedback to solutions will be provided, but you are strongly encouraged to discuss questions and solutions both before and after submission on the relevant Slack channel.
- Recommended for writing solutions: LaTeX or LyX. Paper and pen/pencil are also fine.

Schedule for problem sets:

- PS1: Identification, Estimation and Inference Principles
  - available: Tuesday, 2/4.
  - deadline: Monday, 15/4, 4pm.
  - TA Session 1: Tuesday, 16/4.
- PS2: Experiments and Matching
  - available: Thursday, 9/4.
  - deadline: Wednesday, 22/4, 4pm.
  - TA Session 2: Thursday, 23/4.
- PS3: IV and 2SLS
  - available: Tuesday, 4/16.
  - deadline: Monday, 4/29, 4pm.
  - TA Session 3: Tuesday, 4/30.
- PS4: Fixed Effects and DID
  - available: Wednesday, 4/24.
  - deadline: Tuesday, 7/5, 4pm.
  - TA Session 4: Wednesday, 8/5.
- PS5: Synthetic Controls and RDD
  - available: Tuesday, 3/5.
  - deadline: Monday, 16/5, 4pm.
  - TA Session 5: Tuesday, 17/5.

## Quizzes

- Quizzes are short multiple choice questions, released shortly after each lecture.
- They are automatically graded, so you see which questions you answered correctly right away, and you can also see a suggested solution.
- Note that you get points for participating (one per quiz, for a total of 12% of the grade), and importantly, independently of how many answers you get right! Just try your best, not any worse, not any better.
- The quizzes serve (i) as an opportunity to work through the material of the lectures again and deepen your understanding, (ii) test your understanding and receive feedback, (iii) and for us to get some feedback which points have been understood well, and which ones less so.
- Time expectation: about one hour per quiz.
- You have one week to complete each quiz from when it is available (usually on the same day as the corresponding lecture).

## Readings

- Our philosophy regarding readings:
  - Prioritize problem sets, lecture slides, and quizzes; only go to the readings if those materials are not clear enough.
  - Among the readings, the starred material should help to clarify the basics of the content in the lectures. Other material is only relevant insofar you are interested in the topic.
  - You are not expected to read any of the material before class.
- Main textbooks: (note however that we do not follow any textbook closely)
  - Angrist and Pischke (2008): *Mostly Harmless Econometrics*, Princeton University Press, [unpublished version](#).
  - Imbens and Rubin (2015): *Causal Inference for Statistical, Social, and Biomedical Sciences: An Introduction*, Cambridge University Press, [e-book](#).
  - Cunningham (2022): *Causal Inference: The Mixtape*. Yale University Press, [e-book](#).
- Also highly recommended are Paul Goldsmith-Pinkham's [lecture notes](#).
- Access to NBER working papers and journals for further readings through [SU Library](#).

## Grading and Exam

- Distribution of points for this class:
  - You can earn up to 102 points on the course.
  - You need 50 points to pass the course.
  - You can earn up to 70 points on the exam, up to 20 points for the 5 problem sets, and up to 12 points for filling out the quizzes. *Notice that the points on the quizzes are awarded for filling them in, and are independent of which answers you provide.*
- Exam information:
  - Date and time: 27th of May 2024, 14:00-19:00.

- Only theory questions, i.e. there will be no coding questions on the exam.
- Retake exam at end of summer (date TBD).

## Lectures Details and Recommended Readings

### L1 (2nd of April, 10-12): Identification of Causal Effects

- *Topics:* descriptive versus causal; structural/reduced form; (non-)parametrics; identification; identification strategy / research design; potential outcomes, selection, Conditional Independence Assumption; assignment mechanisms
- *Readings:*
  - \*Imbens and Rubin (2015): Chapters 1 and 2 [introduction]; Chapter 3 [assignment mechanism]
  - \*Angrist and Pischke (2008): Chapter 1; Chapter 2.1 [selection bias]
  - Lewbel, Arthur (2019): “The Identification Zoo: Meanings of Identification in Econometrics”, *Journal of Economic Literature*, 57(4), 835-903 [identification]
  - Currie, Janet and Kleven, Henrik and Zwiers, Esmée (2020): “Technology and Big Data Are Changing Economics: Mining Text to Track Methods”, AEA Papers and Proceedings, 110, 42-48 [identification strategies]
  - Angrist, Joshua D. and Pischke, Jörn-Steffen (2010): “The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con out of Econometrics”, *Journal of Economic Perspectives*, 24(2), 3-30. [research design]
  - Card, David (2011): “Model-Based or Design-Based? Competing Approaches in “Empirical Micro””, [slides](#) [model-based versus design-based identification]
  - Haile, Phil (2020): “Structural vs. Reduced Form”, [slides](#) [structural identification]

### L2 (4th of April): Estimation Principles

- *Topics:* canonical estimation; nonparametric estimation; regression algebra; FWL
- *Readings:*
  - \*Angrist and Pischke (2008): Chapter 3.1 [regression fundamentals]
  - Newey, KW and McFadden, D (1994): “Large Sample Estimation and Hypothesis Testing”, Handbook of Econometrics, IV, 2112-2245. [canonical estimation]

### Coding Camp: STATA (4th of April, 13-15)

- Participation is optional for those who do not feel comfortable using STATA.
- If you are joining, make sure you are familiar with the [UCLA](#) and [Princeton](#) guides!
- Fabian will go through examples of code relevant to the problem sets.

### L3 (9th of April, 10-12): Inference Principles

- *Topics:* large sample standard errors; clustering; design vs. sampling based uncertainty.
- *Readings:*
  - Imbens and Rubin (2015): Chapter 5 [randomization inference]
  - Hansen (2022): Chapter 4.11 and 4.13 [variance estimators]
  - Angrist and Pischke (2009): Chapter 8 [variance estimators]

- Newey, KW and McFadden, D (1994): “Large Sample Estimation and Hypothesis Testing”, *Handbook of Econometrics*, IV, 2112-2245. [canonical estimation]
- Alberto Abadie and Susan Athey and Guido W. Imbens and Jeffrey M. Wooldridge (2020): “Sampling-based vs. Design-based Uncertainty in Regression Analysis”, *Econometrica* 88:1, 265-296. [inference]
- Cameron, A Colin and Miller, Douglas L (2015): “A Practitioner’s Guide to Cluster-Robust Inference”, *Journal of Human Resources*, 50(2), 317-372. [classic reference for clustering]
- Abadie, Alberto, Susan Athey, Guido Imbens and Jeffrey Wooldridge (2017): “When Should You Adjust Standard Errors for Clustering”. Working paper.

## L4 (11th of April, 10-12): Experiments

- *Topics:* balance; power; stratification; paired experiments; controls; **bad controls**; attrition; Fisher Inference; canonical experimental designs; adaptive trials
- *Readings:*
  - \*Angrist and Pischke (2008): Chapter 2 [introduction]; Chapter 3.2.3 [bad controls]
  - \*Imbens and Rubin (2015): Chapter 4 [introduction]; Chapter 5 [Fisher inference], Chapter 7.5 [controls]
  - Kasy, Maximilian (2016) “Why Experimenters Might Not Always Want to Randomize, and What They Could Do Instead” *Political Analysis*: 1-15. [stratification]
  - Lee, D. S. (2009) “Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects”. *Review of Economic Studies* 76: 1071–1102. [attrition]
  - Bai, Yuehao (2021) “Optimality of Matched-Pair Designs in Randomized Controlled Trials”. Working Paper. [paired experiments]
  - Kasy, Maximilian and Anja Sautman (2021): “Adaptive Treatment Assignment In Experiments For Policy” *Econometrica* 89:1, 113–132.

## L5 (16th of April, 10-12): Matching

- *Topics:* propensity score; achieving balance; matching; marginal treatment effects
- *Readings:*
  - \*Imbens and Rubin (2015): Chapters 12.1-12.3 [propensity scores; all of Part III of this book is good if you want to learn more]
  - Angrist and Pischke (2008): Sections 3.3.1-3.3.3 [relation of matching and regression analysis]
  - McKenzie, David (2021) “What do you need to do to make a matching estimator convincing? Rhetorical vs statistical checks” *WorldBank Development Impact* blog post [good practical advice on convincing matching approaches]
  - Dagan, Noa, et. al. (2021) “BNT162b2 mRNA Covid-19 Vaccine in a Nationwide Mass Vaccination Setting” *New England Journal of Medicine*. [please also read “Supplementary Methods 3”; shows a great use of pseudo outcomes]
  - LaLonde, Robert J. (1986): “Evaluating the Econometric Evaluations of Training Programs with Experimental Data”, *American Economic Review*, 76(4)

- Dehejia, Rajeev and Sadek Wahba (2002) "Propensity Score-Matching Methods for Nonexperimental Causal Studies" *The Review of Economics and Statistics*, February 2002, 84(1): 151–161
- Zhou, Xiang and Yu Xie (2019): "Marginal Treatment Effects from a Propensity Score Perspective" *Journal of Political Economy* 127:6.

## TA Session 1 (16th of April, 13-15)

### L6 (18th of April, 10-12): Instrumental Variables

- *Topics:* valid instruments: conditions and intuitions; common mistakes; understanding the IV bias; **weak instruments**; two-sample IV and split-sample IV; jackknife estimators; **shift-share instruments**.
- *Readings:*
  - \*Angrist and Pischke (2008): Chapters 4.1-4.3; 4.6.1; 4.6.4
  - Angrist, Joshua, and Alan Krueger (1995) "Split-Sample Instrumental Variable Estimates of the Returns to Education" *Journal of Business & Economic Statistics* 13:2.
  - Borusyak, Kirill, Peter Hull, and Xavier Jaravel (forthcoming) "Quasi-Experimental Shift-Share Research Designs" *Review of Economic Studies*
  - Goldsmith-Pinkham, Paul, Isaac Sorkin, and Henry Swift (forthcoming) "Bartik Instruments: What, When, Why and How" *American Economic Review*

### L7 (23rd of April, 10-12): IV: Heterogeneous Effects, Compliers, Marginal Effects

- *Topics:* Local Average Treatment Effects (LATE); compliers, always-takers and never-takers; characterizing compliers; generalizations; **judges design**.
- *Readings:*
  - \*Angrist and Pischke (2008): Chapters 4.4 and 4.5
  - Imbens, Guido and Joshua Angrist (1994) "Identification and Estimation of Local Average Treatment Effects", *Econometrica* 62:2, 467-475.
  - Imbens, Guido, Joshua Angrist and Donald B. Rubin (1996): "Identification of Causal Effects Using Instrumental Variables", *Journal of the American Statistical Association*, 91(434), 444-455.

## TA Session 2 (23rd of April, 13-15)

### L8 (25th of April, 10-12): Fixed Effects

- *Topics:* generic FE; panel FE and incidental parameters problem; two-way FE; AKM; Empirical Bayes
- *Readings:*
  - \*Angrist and Pischke (2008): Chapter 5.1

- Ashenfelter, Orley and Alan Krueger (1994): “Estimates of the Economic Return to Schooling from a New Sample of Twins”, *American Economic Review*, 84(5), 1157--1173 [fixed effects]
- Bonhomme, Stéphane and Manresa, Elena (2015): “Grouped Patterns of Heterogeneity in Panel Data”, *Econometrica*, 83(3), 1147--1184 [group FE]
- Card, David and Heining, Jörg and Kline, Patrick (2013): “Workplace Heterogeneity and the Rise of West German Wage Inequality”, *Quarterly Journal of Economics*, 128(3), 967-1015 [AKM]
- Chetty, Raj and Friedman, John N. and Hilger, Nathaniel and Saez, Emmanuel and Schanzenbach, Diane Whitmore and Yagan, Danny (2011): “How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project Star”, *Quarterly Journal of Economics*, 126(4), 1593-1660 [random effects]
- Nickell, Stephen (1981): “Biases in Dynamic Models with Fixed Effects”, *Econometrica*, 49(6), 1417-1426. [incidental parameters]
- Morris, Carl N. (1983): “Parametric Empirical Bayes Inference: Theory and Applications”, *Journal of the American Statistical Association*, 78(381), 47-55 [Empirical Bayes]
- Searle, Shayle R and Casella, George and McCulloch, Charles E (2009): *Variance Components*, John Wiley & Sons [random & fixed effects]
- Wooldridge, Jeffrey M (2010): *Econometric Analysis of Cross Section and Panel Data*, MIT Press [within/between transformation]

## L9 (30th of April, 10-12): Static Difference-in-Differences

- *Topics*: panel SUTVA; anticipation and memory; group-time ATEs; identification in DID with potential outcomes; block structures; dynamic effects; staggered rollouts; inference issues
- *Readings*:
  - \*Angrist and Pischke (2008): Chapter 5.2
  - \*Card, David and Krueger, Alan B (1994): “Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania”, *American Economic Review*, 84(4), 772--793 [leading example]
  - Freyaldenhoven, Simon and Hansen, Christian and Shapiro, Jesse M. (2019): “Pre-event Trends in the Panel Event-Study Design”, *American Economic Review*, 109(9), 3307-38 [auxiliary IV correction of pre-trends]
  - Bertrand, Marianne and Duflo, Esther and Mullainathan, Sendhil (2004): “How Much Should We Trust Differences-In-Differences Estimates?”, *Quarterly Journal of Economics*, 119(1): 249: 275. [inference]
  - de Chaisemartin, Clément and D'Haultfœuille, Xavier (2020): “Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects”, *American Economic Review*, 110(9), 2964-96 [heterogeneous effects in panels]
  - Goodman-Bacon, Andrew (2018): “Difference-in-Differences with Variation in Treatment Timing”, *NBER Working Paper* 25018 [weighted 2x2 DIDs]

## TA Session 3 (30th of April, 13-15)

### L10 (3rd of May, 10-12): Dynamic DID

- *Topics:* Ashenfelter's Dip; pre-trend violations; functional form issues; heterogeneous treatment effects in panel data; event studies; heterogeneity in staggered rollout designs;
- *Readings:*
  - Abraham, Sarah and Liyang Sun (2021): "Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects", *Journal of Econometrics*, 225(2), 175-199. [heterogeneous event studies]
  - Borusyak, Kirill and Jaravel, Xavier and Spiess, Jann (2021): "Revisiting Event Study Designs: Robust and Efficient Estimation", *Working Paper*.
  - Callaway, Brantly and Sant'Anna, Pedro H.C. (2020): "Difference-in-Differences with multiple time periods", *Journal of Econometrics* [group-time ATEs]
  - Davis, Lucas W. (2004): "The Effect of Health Risk on Housing Values: Evidence from a Cancer Cluster", *American Economic Review*, 94(5), 1693-1704.
  - Duflo, Esther (2001): "Schooling and Labor Market Consequences of School Construction in Indonesia: Evidence from an Unusual Policy Experiment", *American Economic Review*, 91(4), 795-813.
  - Freyaldenhoven, Simon and Hansen, Christian and Shapiro, Jesse M. (2019): "Pre-event Trends in the Panel Event-Study Design", *American Economic Review*, 109(9), 3307-38.
  - Jackson, C. Kirabo and Johnson, Rucker C. and Persico, Claudia (2016): "The Effects of School Spending on Educational and Economic Outcomes: Evidence from School Finance Reforms", *Quarterly Journal of Economics*, 131(1), 157-218.
  - Jensen, R. (2007): "The Digital Provide: Information (Technology), Market Performance, and Welfare in the South Indian Fisheries Sector", *Quarterly Journal of Economics*, 122(3), 879-924.
  - Lester, R. A. (1937): "The Gold-parity Depression in Norway and Denmark, 1925-28", *Journal of Political Economy*, 45(4), 433-465.
  - Rambachan, Ashesh, and Roth, Jonathan (2022): "A More Credible Approach to Parallel Trends", *Working Paper*.
  - Roth, Jonathan (2022): "Pre-test with Caution: Event-study Estimates After Testing for Parallel Trends". American Economic Review: Insights (Forthcoming)
  - Roth, Jonathan and Sant'Anna, Pedro HC (2021): "When Is Parallel Trends Sensitive to Functional Form?", *Working Paper* [functional form]
  - Roth, Jonathan, Sant'Anna, Pedro HC, Bilinski, Alyssa, and Poe, John (2022): "What's Trending in Difference-in-Differences? A Synthesis of the Recent Econometrics Literature", *Working Paper*.
  - de Chaisemartin, Clément and D'Haultfoeuille, Xavier (2022): "Two-Way Fixed Effects and Differences-in-Differences with Heterogeneous Treatment Effects: A Survey", *Working Paper*.

### L11 (7th of May, 10-12): Matched DID and Synthetic Control Methods

- *Topics:* matched DID, synthetic control methods; horizontal and vertical regressions, matrix completion methods

- *Readings:*
  - Abadie, Alberto (2005): “Semiparametric Difference-in-Differences Estimators”, *The Review of Economic Studies*, 72(1), 1-19.
  - Abadie, Alberto and Gardeazabal, Javier (2003): “The Economic Costs of Conflict: A Case Study of the Basque Country”, *American Economic Review*, 93(1), 113-132 [basic SCM example]
  - Abadie, A. and Imbens, G.W. (2006): “Large Sample Properties of Matching Estimators for Average Treatment Effects”, *Econometrica*, 74: 235-267.
  - Andersson, Julius J. (2019): “Carbon Taxes and CO<sub>2</sub> Emissions: Sweden as a Case Study”, *American Economic Journal: Economic Policy*, 11(4), 1-30.
  - Ben-Michael, Eli and Feller, Avi and Rothstein, Jesse (2022): “Synthetic controls with staggered adoption”, *Journal of the Royal Statistical Society: Series B*, 84(2).
  - Doudchenko, Nikolay and Imbens, Guido W (2016): “Balancing, Regression, Difference-In-Differences and Synthetic Control Methods: A Synthesis”, *NBER Working Paper* 22791 [hybrid methods]
  - Goldschmidt, Deborah and Schmieder, Johannes F. (2017): “The Rise of Domestic Outsourcing and the Evolution of the German Wage Structure”, *The Quarterly Journal of Economics*, 132(3), 1165-1217.
  - Levy, Roee and Mattsson, Martin (2020): “The Effects of Social Movements: Evidence from #MeToo”, *Working Paper*.
  - Roth, Jonathan, Sant'Anna, Pedro HC, Bilinski, Alyssa, and Poe, John (2022): “What's Trending in Difference-in-Differences? A Synthesis of the Recent Econometrics Literature”, *Working Paper*.
  - Chernozhukov, Victor, Wuthrich, Kaspar, and Zhu Yinchu (2023): “A t-test for synthetic controls”, *Working Paper*.

## TA Session 4 (8th of May, 13-15)

### L12 (10th of May, 10-12): Regression Discontinuity Design

- *Topics:* assumptions for LATE identification, visualization, global and local estimation, sharp and fuzzy designs, spatial RD
- *Readings:*
  - \*Lee, David S. and Lemieux, Thomas (2010): “Regression Discontinuity Designs in Economics”, *Journal of Economic Literature*, 48(2), 281-355
  - Barreca, Alan I. and Guldi, Melanie and Lindo, Jason M. and Waddell, Glen R. (2011): “Saving Babies? Revisiting the effect of very low birth weight classification”, *The Quarterly Journal of Economics*, 126(4), 2117-2123.
  - Calonico, Sebastian and Cattaneo, Matias D. and Titiunik, Rocio (2014): “Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs”, *Econometrica*, 82(6), 2295--2326.
  - Card, David and Lee, David S. and Pei, Zhuan and Weber, Andrea (2015): “Inference on Causal Effects in a Generalized Regression Kink Design”, *Econometrica*, 83(6), 2453-2483.
  - Cattaneo, Matias D., Nicolas Idrobo and Rocio Titiunik (2020): “A Practical Introduction to Regression Discontinuity Designs: Foundations”, *Cambridge*

- Elements: Quantitative and Computational Methods for Social Science*, Cambridge University Press, February 2020, [final draft](#)
- Cattaneo, Matias D., Nicolas Idrobo and Rocio Titiunik (forthcoming): "A Practical Introduction to Regression Discontinuity Designs: Extensions", *Cambridge Elements: Quantitative and Computational Methods for Social Science*, Cambridge University Press, [preliminary draft](#)
  - Cheng, Ming-Yen (1997): "Boundary Aware Estimators of Integrated Density Derivative Products", *Journal of the Royal Statistical Society: Series B*, 59(1), 191-203.
  - Dong, Yingying (2015): "Regression Discontinuity Applications with Rounding Errors in the Running Variable", *Journal of Applied Econometrics*, 30(3), 422-446.
  - Fan, Jianqing and Gijbels, Irene (1992): "Variable Bandwidth and Local Linear Regression Smoothers", *The Annals of Statistics*, 20(4), 2008-2036.
  - Fredriksson, Peter and Öckert, Björn and Oosterbeek, Hessel (2012): "Long-Term Effects of Class Size", *The Quarterly Journal of Economics*, 128(1), 249-285.
  - Gelman, Andrew and Imbens, Guido (2019): "Why High-Order Polynomials Should Not Be Used in Regression Discontinuity Designs", *Journal of Business & Economic Statistics*, 37(3), 447-456.
  - He, Guojun and Wang, Shaoda and Zhang, Bing (2020): "Watering Down Environmental Regulation in China", *The Quarterly Journal of Economics*, 135(4), 2135-2185.
  - Kolesár, Michal and Rothe, Christoph (2018): "Inference in Regression Discontinuity Designs with a Discrete Running Variable", *American Economic Review*, 108(8), 2277-2304.
  - Londoño-Vélez, Juliana and Rodríguez, Catherine and Sánchez, Fabio (2020): "Upstream and Downstream Impacts of College Merit-Based Financial Aid for Low-Income Students: Ser Pilo Paga in Colombia", *American Economic Journal: Economic Policy*, 12(2), 193-227.
  - McCrary, Justin (2008): "Manipulation of the running variable in the regression discontinuity design: A density test", *Journal of Econometrics*, 142(2), 698 - 714.
  - Pei, Zhuan, Lee, David S., Card, David and Weber, Andrea (2021): "Local Polynomial Order in Regression Discontinuity Designs", *Journal of Business & Economic Statistics*, forthcoming.

## L13 (14th of May, 14-16): Machine Learning

- *Topics:* supervised and unsupervised learning methods used by economists
- *Readings:*
  - Mullainathan, Sendhil, and Jann Spiess (2017): "Machine learning: an applied econometric approach." *Journal of Economic Perspectives* 31.2, 87-106.

## TA Session 5 (17th of May, 10-12)

# Econometrics II

## Lecture 1: Identification of Causal Effects

David Schönholzer

Stockholm University

April 2, 2024

# Plan for Today

## 1 Econometric Models

- Parametric versus Nonparametric Models
- Descriptive versus Causal Models

## 2 Identification in Econometric Models

- Formal Definition
- Examples

## 3 Potential Outcomes Framework

- Assumptions
- Average Treatment Effects
- Assignment Mechanisms

# Table of Contents

## 1 Econometric Models

- Parametric versus Nonparametric Models
- Descriptive versus Causal Models

## 2 Identification in Econometric Models

- Formal Definition
- Examples

## 3 Potential Outcomes Framework

- Assumptions
- Average Treatment Effects
- Assignment Mechanisms

# Econometrics: Models for Data

Every empirical economics project comes down to this:

$Y_i$	$D_i$	$X_i$	$Z_i$	...
4	0	6	1	...
1	1	3	1	...
...	...	...	...	...

Econometric models provide a framework to analyze (i.e. summarize) economic data

To this end, they posit a **data generating process (DGP)**:  $(W_i, \varepsilon_i) \sim F_\theta$

- A DGP is a joint distribution of a data row:  $W_i = (Y_i, D_i, X_i, Z_i, \dots)$
- The DGP also includes unobservables that are not in the data,  $\varepsilon_i$
- We pretend all rows were drawn from this DGP with **parameter/structure  $\theta$**

A **model** is a restricted family of DGPs, i.e. we make assumptions about  $F_\theta$ :

- Parametric versus non-parametric: is  $\theta \in \mathbb{R}^K$  or is it a function?
- Descriptive versus causal: is  $\theta$  “factual” or “counterfactual”?

## Example: Parametric Model

Consider

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + \varepsilon_i \\ (X_i, \varepsilon_i) &\stackrel{\text{iid}}{\sim} N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}\right) \\ \theta &= (\beta_0, \beta_1, \mu_1, \mu_2, \log \sigma_1, \log \sigma_2, \sigma_{12}) \in \mathbb{R}^7 \end{aligned}$$

Everything there is to know about the data in 7 numbers!

## Example: A Semi-Parametric Model

$$Y_i = X'_i \beta + \varepsilon_i$$

$$\mathbb{E} [\varepsilon_i | X_i] = 0$$

$$\beta \in \mathbb{R}^K$$

- Note we did not restrict  $F_X(\cdot)$  of  $X_i$
- Only restricted mean of  $F_{\varepsilon|X}(\cdot)$  of  $\varepsilon_i$
- Both  $F_X(\cdot)$  and  $F_{\varepsilon|X}(\cdot)$  potentially infinite dimensional

## Another Semi-Parametric Example: Index Model

$$Y_i = g(X_i' \beta) + \varepsilon_i$$

$$\mathbb{E} [\varepsilon_i | X_i] = 0$$

$$\beta \in \mathbb{R}^K$$

$g(\cdot) : \mathbb{R} \rightarrow \mathbb{R}$  is monotone increasing

- We are interested in  $\beta$
- Functions  $\{g(\cdot), F_{\varepsilon|X}(\cdot), F_X(\cdot)\}$  are “nuisance” (i.e. not of direct interest)

# A Non-Parametric Model

$$Y_i = g(X_i, \varepsilon_i)$$

$$X_i \perp \varepsilon_i$$

$g(\cdot, \cdot) : \mathbb{R}^2 \rightarrow [0, 1]$  is monotonically increasing in both arguments

- Unrestricted marginals of  $\varepsilon_i$  and  $X_i$
- Interested in function  $g(\cdot, \cdot)$  or features like

$$h(X_i) = \mathbb{E}_\varepsilon [g(X_i, \varepsilon_i)]$$

# Descriptive Models

- $\theta$  is “factual”: capturing **moments of the data**, e.g. means, correlations, etc
- Goal: imagine I got a new  $(D_i, X_i, Z_i)$ . What would  $Y_i$  be?  $\rightarrow$  Want  $\hat{Y}_i$
- Examples:
  - ①  $\theta = \mathbb{E}[Y_i|X_i]$ : earnings by educational background
  - ②  $\theta = \text{Corr}(Y_i, X_i|G_i, R_i)$ : correlation of mortality and income by gender/race
  - ③  $\theta = F(Y_i, X_i|Z_i)$ : joint distribution of wealth of children and parents by city
- These are always valid and often interesting objects  
 $\rightarrow$  See much of Raj Chetty's recent work

# Causal Models

- $\theta$  is “counterfactual”: capturing (causal) effects of treatments on outcomes
- Goal: imagine I change  $D_i = 0$  to  $D_i = 1$  for some  $i$ . What would  $Y_i$  be?
  - Want to pin down  $\theta$  itself rather than just construct  $\hat{Y}$
  - Interested in causal mechanisms, not descriptive facts
- When do correlations speak to causality?
  - Need an explicit counterfactual: how the world would change if I manipulate data
  - Imagine our model is  $Y_i = f(X_i, U_i)$ . Causal effect of changing  $X_i$  from  $x'$  to  $x''$ :

$$\Delta(x'', x') = f(x'', U_i) - f(x', U_i)$$

- Can we change  $X_i$  without changing  $U_i$ ? Beware of FUQs

Examples (we will expand on all of these):

- ①  $\theta = \mathbb{E}[Y_i(1) - Y_i(0)]$  the average treatment effect of a college degree
- ②  $\theta = \mathbb{E}[Y_i(1) - Y_i(0)|D_i = 1, X_i = x]$ : conditional ATE of new law on large firms
- ③  $\theta = F_{Y(1)}(0.5) - F_{Y(0)}(0.5)$ : median treatment effect of a drug

# Table of Contents

## 1 Econometric Models

Parametric versus Nonparametric Models

Descriptive versus Causal Models

## 2 Identification in Econometric Models

Formal Definition

Examples

## 3 Potential Outcomes Framework

Assumptions

Average Treatment Effects

Assignment Mechanisms

# Observationally Equivalent Structures

When is a model (or its parameter/structure) identified?

A couple of preliminary definitions:

- $F_\theta(\cdot)$ : distribution function implied by  $\theta$  under the model

Definition (Observationally equivalent structures)

Two structures  $\theta'$  and  $\theta''$  are *observationally equivalent* if

$$F_{\theta'}(\cdot) = F_{\theta''}(\cdot)$$

for any value in the domain of  $F_\theta(\cdot)$

Two values of  $\theta$  could produce the same data  $\Rightarrow$  they are observationally equivalent

# The Identified Set

What values of  $\theta$  are consistent with the joint distribution of the data?

- $F_W(\cdot)$ : distribution function governing observed variables

## Definition (The identified set)

The *identified set* of  $\theta$  is the set of observationally equivalent structures:

$$\Omega(F_W, \Theta) = \{\theta \in \Theta : F_\theta(\cdot) = F_W(\cdot)\}$$

In words, the identified set is the subset  $\theta \in \Theta$  we can “isolate” with data

$\Omega(F_W, \Theta)$  could be e.g. a single point, a collection of points, an interval, or all of  $\mathbb{R}^K$

# Identification, Identification Strategy, and Research Design

## Definition (Point identification)

A model (or equivalently its parameter) is (*point*) *identified* if the identified set  $\Omega(F_W, \Theta)$  is a singleton, i.e. there are no observationally equivalent structures.

If we knew population (i.e. *infinite* sample size), would we be able to learn  $\theta$ ?

- Studies empirical implications of theoretical model
- Logically precedes question of how to estimate  $\theta$
- If  $\theta$  is not identified, not worth constructing estimator!

## Definition (Identification strategy or research design)

An *identification strategy* or *research design* consists of assumptions about the data and the model such that a (typically causal) parameter of interest  $\theta$  is identified.

## Example: Mixture Model

Consider this model with observed  $D_i$  and unobserved  $\varepsilon_i$ :

$$Y_i = (1 + D_i)\varepsilon_i, \quad (D_i, \varepsilon_i) \stackrel{iid}{\sim} \text{Bernoulli}(p) \times N(0, 1), p \in (0, 1)$$

### Claim

*The parameter  $p$  is point-identified.*

### Proof.

By Law of Total Probability:

$$f_Y(y) = \phi(y)(1 - p) + \frac{1}{2}\phi\left(\frac{y}{2}\right)p$$

where  $\phi(\cdot)$  is pdf of Standard Normal. Solving for  $p$  we have:

$$p = \frac{f_Y(y) - \phi(y)}{\frac{1}{2}\phi\left(\frac{y}{2}\right) - \phi(y)}$$

## Example: Additive Model

Consider the model:

$$Y_i = U_i + V_i, \quad (U_i, V_i) \stackrel{iid}{\sim} N\left(\begin{pmatrix} \alpha \\ \beta \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right)$$

### Claim

$(\alpha, \beta)$  is not point-identified. However,  $\alpha + \beta$  is.

### Proof.

Taking expectations,  $\mathbb{E}[Y_i] = \alpha + \beta$ , so the sum is identified.

To show that  $(\alpha, \beta)$  are not separately identified:

- Define  $\tilde{\alpha} = \alpha + x$  and  $\tilde{\beta} = \beta - x$  for some  $x \in \mathbb{R}$
- Then  $F_{(\alpha, \beta)}(y) = F_{(\tilde{\alpha}, \tilde{\beta})}(y)$  for all  $y \in \mathbb{R}$

□

# Example: Nonlinear Model

Consider the model

$$Y_i = (X_i - \theta)^2 + \varepsilon_i, \quad \mathbb{E}[\varepsilon_i] = 0$$

Claim

The identified set is  $\Omega(F_Y, \Theta) = \left\{ \mathbb{E}[X_i] \pm \sqrt{\mathbb{E}[Y_i] - \text{Var}(X_i)} \right\}$ .

Proof.

Taking expectations and solving for zero, we get:

$$\mathbb{E}[Y_i] - \mathbb{E}[X_i^2] + 2\theta\mathbb{E}[X_i] - \theta^2 = 0$$

Which yields  $\Omega(F_Y, \Theta)$  as the solutions for  $\theta$  (using quadratic formula). □

# Table of Contents

## 1 Econometric Models

Parametric versus Nonparametric Models

Descriptive versus Causal Models

## 2 Identification in Econometric Models

Formal Definition

Examples

## 3 Potential Outcomes Framework

Assumptions

Average Treatment Effects

Assignment Mechanisms

# Potential Outcomes Framework

- Statistics approach to causality (Neyman-Rubin) (“atheoretical” counterfactuals)
- **Potential outcome  $Y_i(D_i)$ :** outcome if  $D_i$  exposed to  $D_i \in \{0, 1\}$
- Typically binary  $D_i$  but could be richer
- Example: Drug trial
  - $D_i$ : 1 if treated, 0 if placebo
  - $Y_i(1)$ : health if treated
  - $Y_i(0)$ : health if placebo
- “Fundamental problem of causal inference” (Holland 1986):

$$Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0)$$

→ can only observe outcome  $Y_i$  under *one* treatment status

→ can never observe individual causal effects  $\tau_i = Y_i(1) - Y_i(0)$

# Design-Based versus Model-Based Identification

Identification can be either design-based or model-based

- 1 **Design-based:**  $D_i$  is random(ized), conditional on  $(Y_i(0), Y_i(1))$ 
  - Randomized control trials
  - Instrumental variables
  - Some event study designs
- 2 **Model-based:**  $(Y_i(0), Y_i(1))$  is random, conditional on fixed  $D_i$ 
  - Basic difference-in-differences design
  - Sharp regression discontinuity design

Useful distinction to understand the identification logic of an approach

# SUTVA

- When are potential outcomes well defined?
- Typically need **Stable Unit Treatment Value Assumption** (SUTVA, Rubin 1986)

## Definition (SUTVA)

A treatment satisfies SUTVA if

- ① *No hidden treatment variation:* treatment has consistent effect on unit  $i$   
→ e.g. if  $D_i \in \{0, 1\}$ , it cannot be that  $D_i = 2$
- ② *No interference:* my outcome only depends on my own treatment status:

$$Y_i(d_1, \dots, d_N) = Y_i(d_i)$$

# A Finite Population Example

$i$	$D_i$	$Y_i(0)$	$Y_i(1)$	$Y_i$
1	1	3	2	2
2	0	4	5	4
3	0	1	4	1
4	1	3	6	6
5	0	5	5	5
6	1	4	3	3

- Causal inference involves “imputing” missing potential outcomes
- Classic assumption: potential outcomes are missing at random (MAR):

$$D_i \perp \{Y_i(0), Y_i(1)\} \quad \forall i$$

- Conditional independence assumption (CIA):  $D_i | X_i \perp \{Y_i(0), Y_i(1)\}$

# PO Identification Example: Average Treatment Effect

Consider the model

$$Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0), \quad D_i \in \{0, 1\}$$

$$\mathbb{E}[Y_i(d)|D_i] = \mathbb{E}[Y_i(d)] \text{ for } d \in \{0, 1\}$$

## Claim

The average treatment effect  $\tau_{ATE} = \mathbb{E}[Y_i(1) - Y_i(0)]$  is identified.

## Proof.

Note that

$$\begin{aligned}\tau_{ATE} &= \mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)] \\ &= \mathbb{E}[Y_i(1)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 0] \quad (\text{due to mean indep. assumption}) \\ &= \mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[Y_i|D_i = 0]\end{aligned}$$

## Selection Bias

- A similar argument shows *average treatment effect on the treated* (ATT)

$$\tau = \mathbb{E}[Y_i(1) - Y_i(0)|D_i = 1]$$

is identified if the model includes  $\mathbb{E}[Y_i(d)|D_i] = \mathbb{E}[Y_i(d)]$

- Now suppose we lack this mean independence assumption
- Then identification fails because of *selection bias*:

### Definition (Selection Bias)

The gap between ATT and difference between treatment and control means is:

$$\mathbb{E}[Y_i|D_i = 1] - E[Y_i|D_i = 0] = \tau + \underbrace{\mathbb{E}[Y_i(0)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 0]}_{\text{Selection bias}}$$

# Assignment Mechanism

- Whether MAR or CIA holds depends on the *design*  $\{D_i\}_{i=1}^N$
- The design of a setting can be described through the *assignment mechanism*
- What determines why some units are treated and others not?
- Let  $\mathbf{D} = (D_1, \dots, D_N)'$  and  $\mathbf{Y}(\mathbf{d}) = (Y_1(d), \dots, Y_N(d))'$  for  $d \in \{0, 1\}$

## Definition (Assignment Mechanism)

For a population of  $N$  units, it is a function  $P(\mathbf{D}|\mathbf{X}, \mathbf{Y}(\mathbf{0}), \mathbf{Y}(\mathbf{1}))$  that satisfies

$$\sum_{\mathbf{D} \in \{0,1\}^N} P(\mathbf{D}|\mathbf{X}, \mathbf{Y}(\mathbf{0}), \mathbf{Y}(\mathbf{1})) = 1$$

for all  $\mathbf{X}$ ,  $\mathbf{Y}(\mathbf{0})$ , and  $\mathbf{Y}(\mathbf{1})$  where  $\mathbf{D} \in \{0, 1\}^N$  is all possible treatment assignments.

# Unit Assignment Probability and Propensity Score

## Definition (Unit assignment probability)

Unit  $i$  has the following probability of being treated for all  $\mathbf{X}$ ,  $\mathbf{Y(0)}$ , and  $\mathbf{Y(1)}$ :

$$p_i(\mathbf{X}, \mathbf{Y(0)}, \mathbf{Y(1)}) = \sum_{\mathbf{D}: D_i=1} P(\mathbf{D} | \mathbf{X}, \mathbf{Y(0)}, \mathbf{Y(1)})$$

## Definition (Propensity score)

The propensity score at  $X_i = x$  is for all  $\mathbf{X}$ ,  $\mathbf{Y(0)}$ , and  $\mathbf{Y(1)}$ :

$$e(x) = \frac{1}{N(x)} \sum_{i: X_i=x} p_i(\mathbf{X}, \mathbf{Y(0)}, \mathbf{Y(1)})$$

where  $N(x) = \#\{i = 1, \dots, N | X_i = x\}$

## Example: Completely Randomized Experiment

- Let us return to the finite-sample example
- Completely randomized experiment with  $n_1 < N$  treated units:

$$P(\mathbf{D}|\mathbf{Y(0)}, \mathbf{Y(1)}) = \binom{N}{n_1}^{-1}$$

if  $\sum_{i=1}^N D_i = n_1$  and  $P(\mathbf{D}|\mathbf{Y(0)}, \mathbf{Y(1)}) = 0$

- Treatment assignment with  $N = 3$  and  $n_1 = 2$ :

Assignment number:	1	2	3	4	5	6	7	8	$p_i$	$e(x)$
$i = 1$	0	1	0	0	1	1	0	1	$\frac{2}{3}$	$\frac{2}{3}$
$i = 2$	0	0	1	0	1	0	1	1	$\frac{2}{3}$	$\frac{2}{3}$
$i = 3$	0	0	0	1	0	1	1	1	$\frac{2}{3}$	$\frac{2}{3}$
$P(\mathbf{D} \mathbf{X}, \mathbf{Y(1)}, \mathbf{Y(0)})$	0	0	0	0	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	0		

# Econometrics II

## Lecture 2: Estimation Principles

David Schönholzer

Stockholm University

April 4, 2024

# Plan for Today

## 1 General Estimation Principles

Extremum Estimation

Examples of Extremum Estimators

## 2 Linear Regression Mechanics

The Relationship Between CEF and OLS

Using OLS to Estimate Means

## 3 Nonparametric Estimation and Visualization

Kernel Estimation

Applied Nonparametric CEF Estimation

## 4 Appendix: Semi-Parametric Efficiency of OLS

# Estimation Principles

- Last lecture: what can be learned?
- Today: how best to learn it?
- Goal is to find  $\theta$ : *parameter, estimand, population estimator*
- We do so using  $\hat{\theta}$ : *(sample) estimator*
- “Best” meaning:
  - Unbiased:  $\mathbb{E}[\hat{\theta}] = \theta$
  - Consistent:  $\hat{\theta} \xrightarrow{P} \theta$
  - Efficient:  $Var(\hat{\theta})$  as small as possible (but no smaller)
- Begin with *extremum estimators*
  - Covers large class of nonlinear estimators
  - Useful to illustrate general estimation principles

# Table of Contents

## 1 General Estimation Principles

Extremum Estimation

Examples of Extremum Estimators

## 2 Linear Regression Mechanics

The Relationship Between CEF and OLS

Using OLS to Estimate Means

## 3 Nonparametric Estimation and Visualization

Kernel Estimation

Applied Nonparametric CEF Estimation

## 4 Appendix: Semi-Parametric Efficiency of OLS

# Extremum Estimation

- Let  $\mathbf{Z}_i$  be a matrix of data on  $i$ , e.g.  $\mathbf{Z}_i = (Y_i, D_i, \mathbf{X}_i)$
- Want to maximize *population objective*  $Q_0(\theta)$
- $\theta \in \Theta$  is parameter vector
- *Sample objective*:  $\hat{Q}_N(\theta, \mathbf{Z}_1, \dots, \mathbf{Z}_N)$  with sample size  $N$
- Define parameter of interest as:

$$\theta_0 = \arg \max_{\theta \in \Theta} Q_0(\theta)$$

where we assume the max is unique

- Extremum estimator maximize sample *criterion function*:

$$\hat{\theta} = \arg \max_{\theta \in \Theta} \hat{Q}_N(\theta)$$

# Examples of Extremum Estimation

- Example 1: OLS
  - $\mathbf{Z}_i = (Y_i, \mathbf{X}_i)$
  - $\theta$  is projection coefficient of  $Y_i$  on  $\mathbf{X}_i$
  - $Q_0(\theta) = -\mathbb{E}[(Y_i - \mathbf{X}'_i \theta)^2]$  and  $\theta_0 = \mathbb{E}[\mathbf{X}_i \mathbf{X}'_i]^{-1} \mathbb{E}[\mathbf{X}_i Y'_i]$
  - $\hat{Q}_N(\theta) = -\frac{1}{N} \sum_i^N (Y_i - \mathbf{X}'_i \theta)^2$
- Example 2: Nonlinear LS
  - Nonlinear parametric model  $\mu(\mathbf{X}_i, \theta)$  for CEF
  - $Q_0(\theta) = -\mathbb{E}[(Y_i - \mu(\mathbf{X}_i, \theta))^2]$
  - $\hat{Q}_N(\theta) = -\frac{1}{N} \sum_{i=1}^N (Y_i - \mu(\mathbf{X}_i, \theta))^2$
- But could be any estimator expressed with  $Q_0(\theta)$

# Consistency of Extremum Estimators

Definition (Uniform convergence in probability)

$\hat{Q}_N(\theta)$  converges uniformly to  $Q_0(\theta)$  if

$$\sup_{\theta \in \Theta} \left| \hat{Q}_N(\theta) - Q_0(\theta) \right| \xrightarrow{P} 0.$$

Theorem (Consistency of Extremum Estimators)

If (i)  $Q_0(\theta)$  is uniquely maximized at  $\theta_0$ , (ii)  $\Theta$  is compact, (iii)  $Q_0(\theta)$  is continuous, and (iv)  $\hat{Q}_N(\theta)$  converges uniformly to  $Q_0(\theta)$ , then  $\hat{\theta} \xrightarrow{P} \theta_0$ .

# Table of Contents

## 1 General Estimation Principles

Extremum Estimation

Examples of Extremum Estimators

## 2 Linear Regression Mechanics

The Relationship Between CEF and OLS

Using OLS to Estimate Means

## 3 Nonparametric Estimation and Visualization

Kernel Estimation

Applied Nonparametric CEF Estimation

## 4 Appendix: Semi-Parametric Efficiency of OLS

# Extremum Estimator 1: Classical Minimum Distance

- Sample objective:

$$\hat{Q}_N(\theta) = -[\hat{\pi} - \mathbf{h}(\theta)]' \hat{\mathbf{W}} [\hat{\pi} - \mathbf{h}(\theta)],$$

- where  $\hat{\pi} \xrightarrow{P} \pi$  is a vector of “reduced form” moments, e.g.
  - means of some variables of interest
  - covariances (recall variance component estimation)
  - other functions of the data
- $\mathbf{h}(\theta)$  is a *structural function* from model predictions
- $\hat{\mathbf{W}} \xrightarrow{P} \mathbf{W}$  is a symmetric weighting matrix
  - e.g.  $\mathbf{W} = \mathbf{I}$  or inverse variance weighting
- Hence,  $\hat{Q}_N(\theta)$ : “squared distance” between data and model

## Example CMD Estimation

- Example from behavioral economics: Laibson et al. (2007)
- They document two facts:
  - ① Individuals borrow through credit cards with high interest
  - ② Accumulate wealth by the time they retire
- What preferences can explain this behavior?
- Present bias  $\beta$ , discounting  $\delta$ , risk aversion  $\rho$
- Model yields, given  $\theta = (\beta, \delta, \rho)$ , predictions for moments:
  - ① Share of 21-30 year olds with credit card:  $h_1(\theta)$
  - ② Share annual income borrowed with credit card:  $h_2(\theta)$
  - ③ Wealth of 51-60 year olds:  $h_3(\theta)$
- In the data, observe shares and wealth:  $\hat{\pi}_1, \hat{\pi}_2, \hat{\pi}_3$
- Optimal choice  $\hat{\theta}$  quantifies preference parameters

## Extremum Estimator 2: Generalized MM

- Generalized MM sample criterion function:

$$\hat{Q}_N(\theta) = -\hat{\mathbf{g}}(\theta)' \hat{\mathbf{W}} \hat{\mathbf{g}}(\theta)$$

where  $\hat{\mathbf{g}}(\theta) = \frac{1}{N} \sum_i f(\mathbf{Z}_i, \theta)$  and weights  $\hat{\mathbf{W}}$

- E.g. if  $f(\mathbf{Z}_i, \theta) = (Y_i - \mathbf{X}'_i \beta) \mathbf{X}_i$  would be OLS
- Population moment conditions:

$$\mathbf{g}(\theta) = \mathbb{E}[f(\mathbf{Z}_i, \theta)] = 0$$

- Often originates from economic FOC
  - Euler condition in macro
  - Nash equilibrium in game

## Extremum Estimator 3: Maximum Likelihood

- Call  $\ell(\mathbf{Z}_i, \theta)$  the *log likelihood* of observing  $\mathbf{Z}_i$  given  $\theta$
- Sample criterion:

$$\widehat{Q}_N(\theta) = \frac{1}{N} \sum_i \ell(\mathbf{Z}_i, \theta)$$

- Population criterion:  $Q(\theta) = \mathbb{E} [\ell(\mathbf{Z}_i, \theta)]$
- Maximizing  $\widehat{Q}_N(\theta)$  solves:

$$\frac{1}{N} \sum_i \mathbf{s}(\mathbf{Z}_i, \widehat{\theta}_{\text{ML}}) = 0$$

where  $\mathbf{s}(\mathbf{Z}_i, \theta) \equiv \nabla_{\theta} \ell(\mathbf{Z}_i, \theta_0)$  is the score

- Key element in MLE: fully characterize  $f(\mathbf{Z}_i, \theta)$
- More than just (mean) independence assumptions!

## Extremum Estimator 4: OLS

- Population criterion:  $Q_0(\theta) = -\mathbb{E} \left[ (Y_i - \mathbf{X}'_i \theta)^2 \right]$
- Sample criterion:  $\hat{Q}_N(\theta) = -\frac{1}{N} \sum_i^N (Y_i - \mathbf{X}'_i \theta)^2$
- Unlike general case, this criterion has explicit solution:

$$\begin{aligned}\hat{\theta} &= \left( \sum_{i=1}^N \mathbf{X}_i \mathbf{X}'_i \right)^{-1} \left( \sum_{i=1}^N \mathbf{X}_i Y_i \right) \\ &= (\mathbf{X}' \mathbf{X})^{-1} (\mathbf{X}' \mathbf{Y})\end{aligned}$$

for  $\mathbf{X} = [\mathbf{X}_1, \dots, \mathbf{X}_N]'$  and  $\mathbf{Y} = [Y_1, \dots, Y_N]'$

- Corresponds to model  $Y_i = \mathbf{X}'_i \theta_0 + \varepsilon_i$  with restrictions
  - Specifically,  $\mathbb{E}[X_i \varepsilon_i] = 0$  and  $\dim(\mathbf{X}) = K$
- Under some conditions (Econometrics I),  $\hat{\theta} \xrightarrow{P} \theta_0$  (consistent)
- Side note: in general,  $\mathbb{E}[\hat{\theta}] \neq \theta_0$  (biased)
  - Unless either (a) CEF is linear or (b)  $\mathbf{X}_i$  are fixed
  - This is not of great practical importance

# Table of Contents

## 1 General Estimation Principles

Extremum Estimation

Examples of Extremum Estimators

## 2 Linear Regression Mechanics

The Relationship Between CEF and OLS

Using OLS to Estimate Means

## 3 Nonparametric Estimation and Visualization

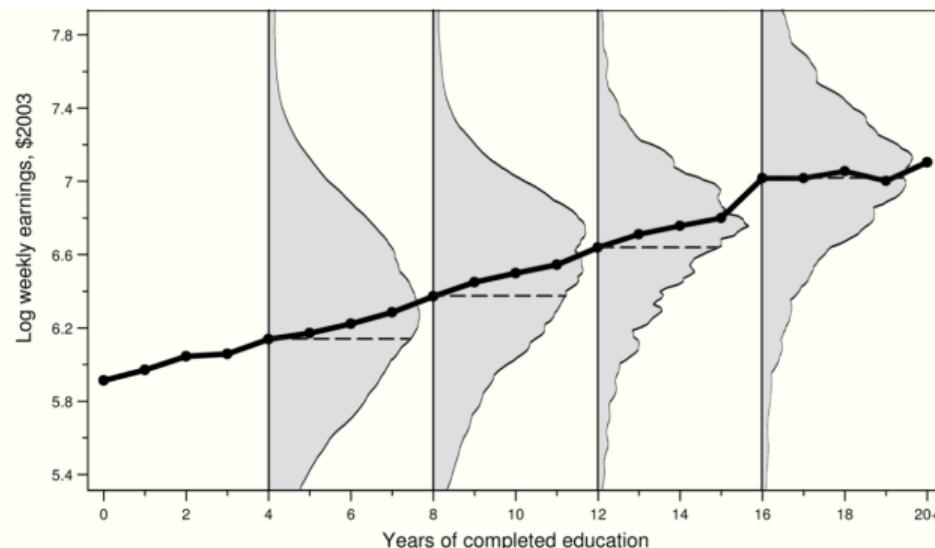
Kernel Estimation

Applied Nonparametric CEF Estimation

## 4 Appendix: Semi-Parametric Efficiency of OLS

## Reminder: CEF

- Central object to summarize data:  $\mathbb{E}[Y_i|X_i]$
- Population average association of outcome  $Y_i$  with  $X_i$
- Recall  $\mathbb{E}[Y_i|X_i]$  is random but  $\mathbb{E}[Y_i|X_i = x]$  is fixed



Why do economists love CEF and OLS? Many useful properties

# The CEF Decomposition Property

Define  $\varepsilon_i \equiv Y_i - \mathbb{E}[Y_i|X_i]$ . Then:

Theorem (The CEF Decomposition Property)

If we write

$$Y_i = \mathbb{E}[Y_i|X_i] + \varepsilon_i$$

it holds by definition that

- (a)  $\mathbb{E}[\varepsilon_i|X_i] = 0$ , and therefore
- (b)  $\text{Cov}(\varepsilon_i, X_i) = 0$

→ Any  $Y_i$  can be decomposed into:

- ① A piece “explained” by  $X_i$ : the CEF
- ② A piece uncorrelated with (any function of)  $X_i$

# The CEF Prediction Property

## Theorem (The CEF Prediction Property)

Let  $m(X_i)$  be any function of  $X_i$  with finite second moment. The CEF solves:

$$\mathbb{E}[Y_i|X_i] = \arg \min_{m(X_i)} \mathbb{E} \left[ (Y_i - m(X_i))^2 \right],$$

so it minimizes MSE of prediction of  $Y_i$  given  $X_i$

→ CEF is the best function of  $X_i$  to predict  $Y_i$

# OLS Justification 1: Linear CEF Theorem

It turns out population OLS is a great estimator of the CEF

Recall population regression:  $\beta_{OLS} \equiv \mathbb{E}[X_i X_i']^{-1} \mathbb{E}[X_i Y_i]$

- Defines linear projection  $\mathbb{E}^*[Y_i|X_i] \equiv X_i' \beta_{OLS}$

## Theorem (The Linear CEF Theorem)

*Suppose the CEF is linear. Then*

$$\mathbb{E}[Y_i|X_i] = \mathbb{E}^*[Y_i|X_i]$$

→ OLS is great for linear CEF. But when is it linear?

- Multivariate Normal distributions
- Saturated models (see later today): one dummy for each possible value of CEF

## OLS Justification 2: Best Linear Predictor

OLS is also good at predicting  $Y_i|X_i$  directly:

**Theorem (The Best-Linear-Predictor Theorem)**

$\mathbb{E}^*[Y_i|X_i]$  minimizes MSE of linear prediction of  $Y_i$  given  $X_i$

- CEF is best function predicting  $Y_i|X_i$
- OLS is best *linear* function predicting  $Y_i|X_i$

## OLS Justification 3: Regression-CEF Relationship

Even when CEF is nonlinear, OLS is still good at predicting it:

Theorem (Regression-CEF Theorem)

$\mathbb{E}^*[Y_i|X_i]$  minimizes MSE of any linear approximation of CEF, i.e.

$$\beta_{OLS} = \arg \min_b \mathbb{E} \left[ (\mathbb{E}[Y_i|X_i] - X_i' b)^2 \right]$$

## OLS Justification 4: Law of Iterated Projections

Linear projections have equivalent property to LIE:

- ① Long regression:  $\mathbb{E}^*[Y_i|W_i, Z_i] = W_i\beta + Z_i\gamma$
- ② Short regression:  $\mathbb{E}^*[Y_i|W_i] = W_i\delta$
- ③ Auxiliary regression:  $\mathbb{E}^*[Z_i|W_i] = W_i\pi$

### Theorem (Law of Iterated Projections)

$$\mathbb{E}^*[Y_i|W_i] = \mathbb{E}^*[\mathbb{E}^*[Y_i|W_i, Z_i]|W_i] \text{ which implies } \delta = \beta + \pi\gamma$$

Proof of implication:

$$\begin{aligned}\mathbb{E}^*[Y_i|W_i] &= \mathbb{E}^*[W_i\beta + Z_i\gamma|W_i] \\ &= \mathbb{E}^*[W_i|W_i]\beta + \mathbb{E}^*[Z_i|W_i]\gamma \\ &= W_i\beta + (W_i\pi)\gamma = W_i(\beta + \pi\gamma)\end{aligned}$$

## Illustration of LIP

```
clear  
set seed 1234  
set obs 1000  
gen z = rnormal()  
gen w = z + rnormal()  
gen y = .5*w + .5*z + rnormal()  
eststo lr: reg y w z // long regression  
local beta = _b[w]  
local gamma = _b[z]  
eststo sr: reg y w // short regression  
local delta = _b[w]  
eststo ar: reg z w // auxiliary regression  
local pi = _b[w]  
esttab lr sr ar, cells(b(fmt(a2)) se(par))
```

## Results from LIP Simulation

	(1)	(2)	(3)
	y	y	z
w	0.44 (0.033)	0.76 (0.024)	0.53 (0.016)
z	0.59 (0.044)		
_cons	0.022 (0.031)	0.037 (0.034)	0.025 (0.022)
N	1000	1000	1000

- As predicted by the LIP:

$$0.76 = 0.44 + 0.53 \times 0.59$$

$$\widehat{\delta} = \widehat{\beta} + \widehat{\pi} \times \widehat{\gamma}$$

- Useful to think about omitted variable bias

## OLS Justification 5: Frisch-Waugh-Lovell

- Recall the long regression:  $\mathbb{E}^*[Y_i|W_i, Z_i] = W_i\beta + Z_i\gamma$
- Residuals:  $\tilde{Y}_i \equiv Y_i - \mathbb{E}^*[Y_i|Z_i]$
- $\tilde{W}_i \equiv W_i - \mathbb{E}^*[W_i|Z_i]$

### Theorem (Frisch-Waugh-Lovell)

$$\beta = \frac{\mathbb{E}[\tilde{W}_i \tilde{Y}_i]}{\mathbb{E}[\tilde{W}_i]^2}$$

- Recover  $\beta$  from long reg by running a residualized short reg
  - Extremely useful to visualize conditional relationships
  - Multivariate versions of LIP and FWL also exist
  - Both are mechanical results of OLS – work in every dataset!

# Illustration of FWL

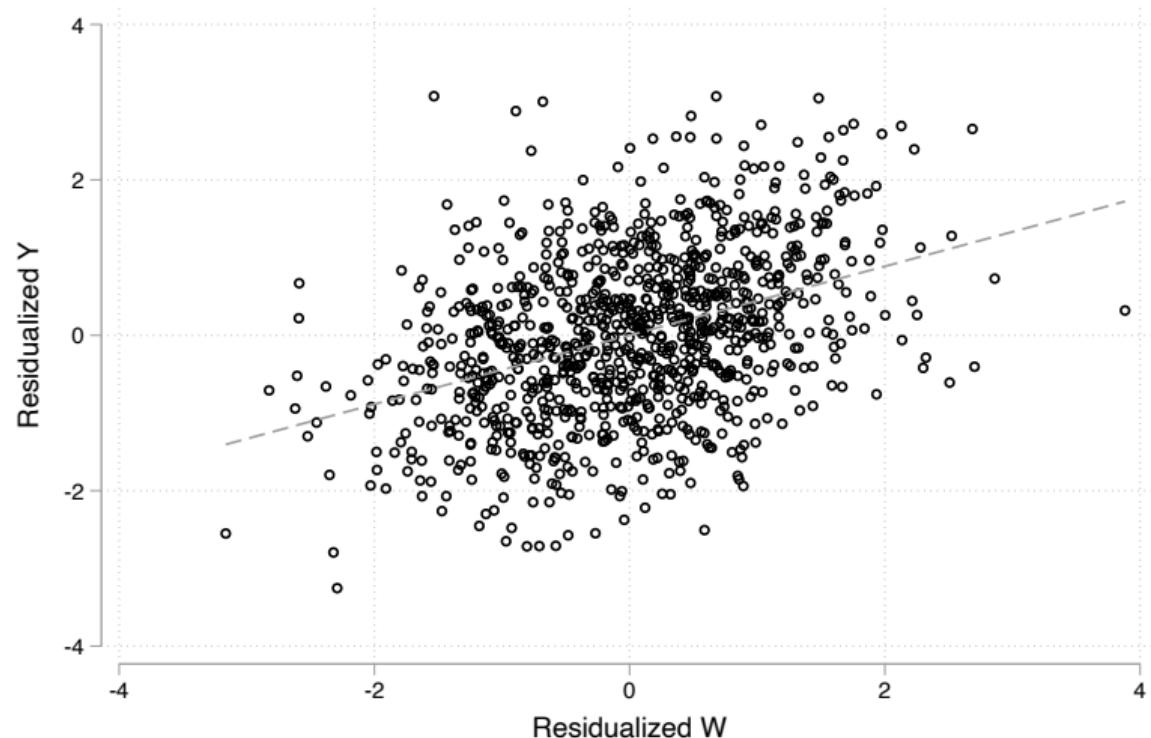
```
clear
set seed 1234
set obs 1000
gen z = rnormal()
gen w = z + rnormal()
gen y = .5*w + .5*z + rnormal()
eststo lr: reg y w z // long regression
eststo far: reg w z // flipped auxiliary regression
predict wres, res
eststo arr: reg y z // other short regression
predict yres, res
eststo rr: reg yres wres // residual regression
esttab lr far arr rr, cells(b(fmt(a2)) se(par))
```

# Results from FWL Simulation

	(1)	(2)	(3)	(4)
	y	w	y	yres
w	0.44 (0.033)			
z	0.59 (0.044)	0.99 (0.030)	1.03 (0.033)	
wres				0.44 (0.032)
_cons	0.022 (0.031)	-0.047 (0.030)	0.0015 (0.034)	5.4e-11 (0.031)
N	1000	1000	1000	1000

→ Useful when there are many controls

# Application of FWL: Residualized Scatterplots



# Table of Contents

## 1 General Estimation Principles

Extremum Estimation

Examples of Extremum Estimators

## 2 Linear Regression Mechanics

The Relationship Between CEF and OLS

Using OLS to Estimate Means

## 3 Nonparametric Estimation and Visualization

Kernel Estimation

Applied Nonparametric CEF Estimation

## 4 Appendix: Semi-Parametric Efficiency of OLS

# OLS on Constant

- Important use of OLS: estimating means
- Simplest case:  $Y_i = \mu + \varepsilon_i$
- Population OLS of this is  $\beta_{\text{OLS}} = \mathbb{E}[Y_i]$ 
  - Convince yourself:  $\mathbb{E}[X_i^2]^{-1} \mathbb{E}[X_i Y_i]$  with  $X_i = 1$
- Sample OLS:  $\hat{\beta}_{\text{OLS}} = \frac{1}{N} \sum_{i=1}^N Y_i$ 
  - Again good exercise to evaluate  $(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{Y}$
  - $N \times 1$  vectors  $\mathbf{X} = (1, \dots, 1)$  and  $\mathbf{Y} = (Y_1, \dots, Y_N)$

# Analysis of Variance

- R.A. Fisher: do means across groups differ?
- Suppose we have a sample of wages  $Y_i$
- We also have  $X_i = 1$  [foreign] and  $W_i = 1$  [female]
- Suppose we run

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_i + \beta_3 X_i \times W_i + \varepsilon_i$$

- This is called a *saturated model*
- Number of coefficients = number of possible RHS values

	$X = 0$	$X = 1$
$W = 0$	(domestic, male)	(foreign, male)
$W = 1$	(domestic, female)	(foreign, female)

# Interpreting Group Indicator Coefficients

- How do we interpret  $(\beta_0, \beta_1, \beta_2, \beta_3)$ ?
- CEF is necessarily linear, and thus OLS = CEF
- CEF:  $\mathbb{E}[Y_i | X_i = x, W_i = w]$
- Specifically:

$$\beta_0 = \mathbb{E}[Y_i | X_i = 0, W_i = 0]$$

$$\beta_0 + \beta_1 = \mathbb{E}[Y_i | X_i = 1, W_i = 0]$$

$$\beta_0 + \beta_2 = \mathbb{E}[Y_i | X_i = 0, W_i = 1]$$

$$\beta_0 + \beta_1 + \beta_2 + \beta_3 = \mathbb{E}[Y_i | X_i = 1, W_i = 1]$$

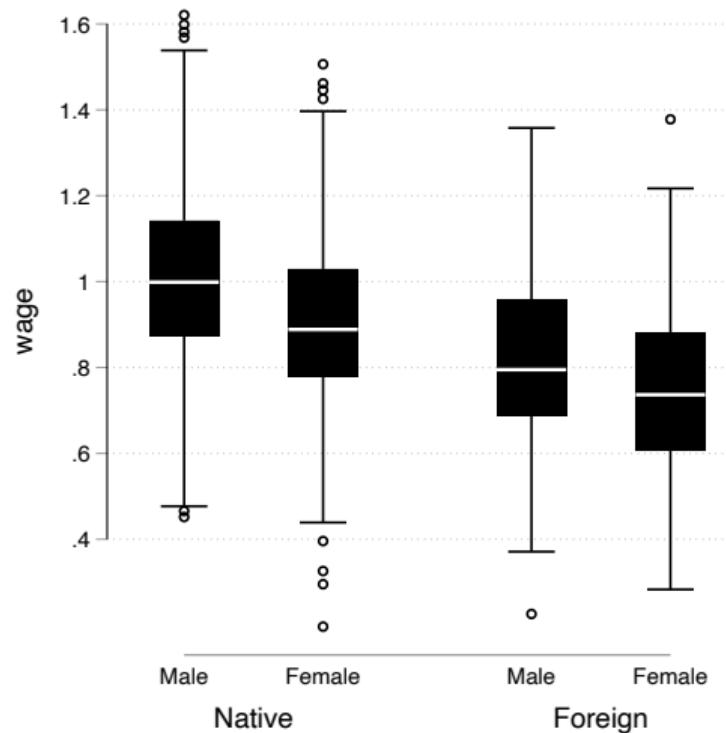
- There are other ways to parameterize same model, e.g.

$$Y_i = \gamma_0 X_i + \gamma_1 W_i + \gamma_2 (1 - X_i) \times W_i + \gamma_3 X_i \times W_i + \varepsilon;$$

## Illustration of Saturated Model

```
clear
set seed 123
set obs 1000
gen foreign = runiform() < .2
gen female = runiform() < .5
tab foreign female
gen wage = 1 - .1*female - .2*foreign + .05*foreign*female + .2*rnorma()
graph box wage, over(female, relabel(1 "Male" 2 "Female")) ///
    over(foreign, relabel(1 "Native" 2 "Foreign")) ///
    ylabel(.4(.2)1.6) xsize(4)
graph export figures/boxplot.pdf, replace
eststo sat: reg wage foreign##female
esttab sat, cells(b(fmt(2)) se(par)) ///
    keep(1.foreign 1.female 1.foreign#1.female _cons) ///
    label
```

# Results from Saturated Model Simulation



	(1)
wage	
b/se	
foreign=1	-0.20 (0.02)
female=1	-0.11 (0.01)
foreign=1 × female=1	0.05 (0.03)
Constant	1.01 (0.01)
Observations	1000

# Many Means

- Consider now  $X_i \in \{\xi_1, \dots, \xi_J\}$  for large  $J$  (but  $J < N$ )
- $\xi_j$  could be firm, or demographic group e.g. (foreign, female)
- All realizations of  $X_i$ :  $\Pr(X_i = \xi_j) = \pi_j > 0$  and  $\sum_j \pi_j = 1$
- We know that OLS = CEF if linear
- Thus OLS is  $\mathbb{E}[Y_i | X_i = x]$  for all  $x \in \{\xi_1, \dots, \xi_J\}$

# Method of Moments for Cell Means

- Can estimate using “cell means” (MM):

$$\hat{\mathbb{E}} [Y_i | X_i = x] = \frac{\sum_i 1[X_i = x] Y_i}{\sum_i 1[X_i = x]} = \frac{\frac{1}{N} \sum_i 1[X_i = x] Y_i}{\frac{1}{N} \sum_i 1[X_i = x]}$$

- With a LLN:

$$\frac{1}{N} \sum_i 1[X_i = x] \xrightarrow{P} \mathbb{E}[1(X_i = x)] = \Pr(X_i = x) = \pi_j$$

$$\frac{1}{N} \sum_i 1[X_i = x] Y_i \xrightarrow{P} \mathbb{E}[Y_i \cdot 1(X_i = x)] = \mathbb{E}[Y_i | X_i = x] \pi_j$$

where the last step uses the LIE

# OLS Estimates Cell Means

- So with continuity theorem

$$\frac{\frac{1}{N} \sum_i 1[X_i = x] Y_i}{\frac{1}{N} \sum_i 1[X_i = x]} \xrightarrow{p} \mathbb{E}[Y_i | X_i = x]$$

- Compare cell means to OLS of  $Y_i$  on  $1[X_i = x]$  for all  $x$ :

$$\hat{\beta}_{OLS} = \begin{bmatrix} \frac{\sum_i 1[X_i=\xi_1] Y_i}{\sum_i 1[X_i=\xi_1]} \\ \vdots \\ \frac{\sum_i 1[X_i=\xi_J] Y_i}{\sum_i 1[X_i=\xi_J]} \end{bmatrix}$$

- They are the same!
- So OLS estimates cell means for many groups

# Table of Contents

## 1 General Estimation Principles

Extremum Estimation

Examples of Extremum Estimators

## 2 Linear Regression Mechanics

The Relationship Between CEF and OLS

Using OLS to Estimate Means

## 3 Nonparametric Estimation and Visualization

Kernel Estimation

Applied Nonparametric CEF Estimation

## 4 Appendix: Semi-Parametric Efficiency of OLS

# Constructing Cells with a Window

- Consider scalar  $X_i$  but continuous with density  $f(x)$
- Logic from before hard because  $\Pr(X_i = x) = 0$
- So how can we approximate  $\mathbb{E}[Y_i | X_i = x]$  best?
- We imitate the cell means logic
- Let's construct a small window  $[x - h, x + h]$  for small  $h > 0$
- $h$  is called *bandwidth* or *window* – chosen/known by us

# Bandwidth Estimation

- Let's estimate these "window cell means"

$$\hat{\mathbb{E}}[Y_i|X_i = x] = \frac{\sum_i \mathbf{1}[x - h \leq X_i \leq x + h] \cdot Y_i}{\sum_i \mathbf{1}[x - h \leq X_i \leq x + h]}$$

- $\hat{\mathbb{E}}[Y_i|X_i] \xrightarrow{P} \mathbb{E}[Y_i|X_i]$  as  $N$  gets large and  $h$  small
- But unless  $\mathbb{E}[Y_i|X_i]$  constant in window,  $\hat{\mathbb{E}}[Y_i|X_i]$  biased
- On the other hand, variance increases as  $h$  shrinks
  - Intuitive: less observations in window
- Optimal  $h$  minimizing MSE infeasible: requires knowing  $f(x)$
- Solution: use auxiliary density  $K(x)$  (the "kernel")

# Univariate Density Estimation

- Alternative approach for  $\widehat{\mathbb{E}}[Y_i|X_i]$ : for continuous  $Y_i$  and  $X_i$

$$\mathbb{E}[Y_i|X_i = x] = \frac{\int y f_{X,Y}(x,y) dy}{\int f_{Y,X}(y,x) dy} = \frac{\int y f_{X,Y}(x,y) dy}{f(x)}$$

so can estimate  $f_{X,Y}(x,y)$  and  $f(x)$  to get CEF too

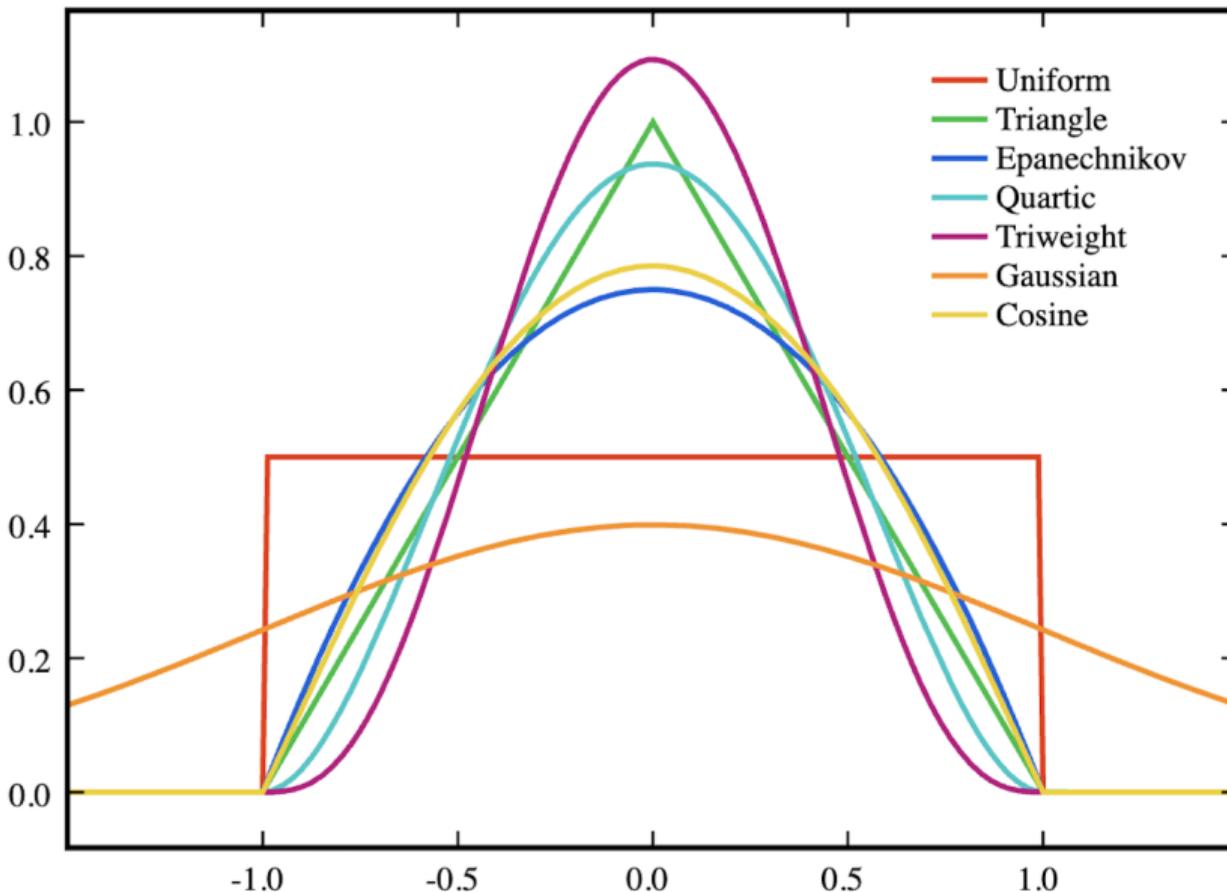
- May also be interested in  $f(x)$  in its own right
- CDF  $F(x) = \Pr(X_i \leq x)$  and  $\widehat{F}(x) = \frac{1}{N} \sum_{i=1}^N \mathbf{1}[X_i \leq x]$
- Definition of derivative:  $f(x) = \lim_{h \rightarrow 0} \frac{F(x+h) - F(x)}{h}$
- Empirical equivalent: *histogram*

$$\widehat{f}(x) = \frac{1}{nh} \sum_{i=1}^N \mathbf{1}[x < X_i \leq x + h]$$

- Can use  $K(\cdot)$  to construct continuous versions:

$$\widehat{f}(x) = \frac{1}{nh} \sum_{i=1}^N K\left(\frac{x - X_i}{h}\right)$$

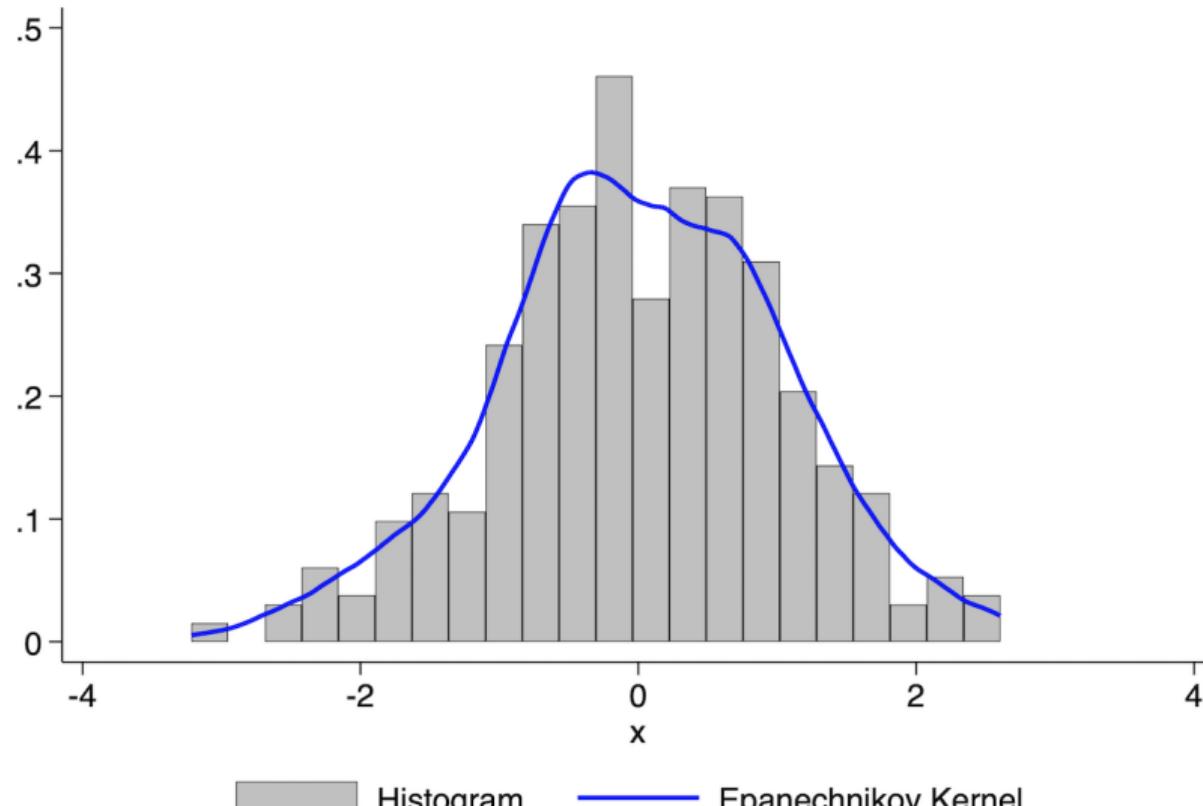
# Many Choices for Smoothers $K(\cdot)$ (i.e. Kernels)



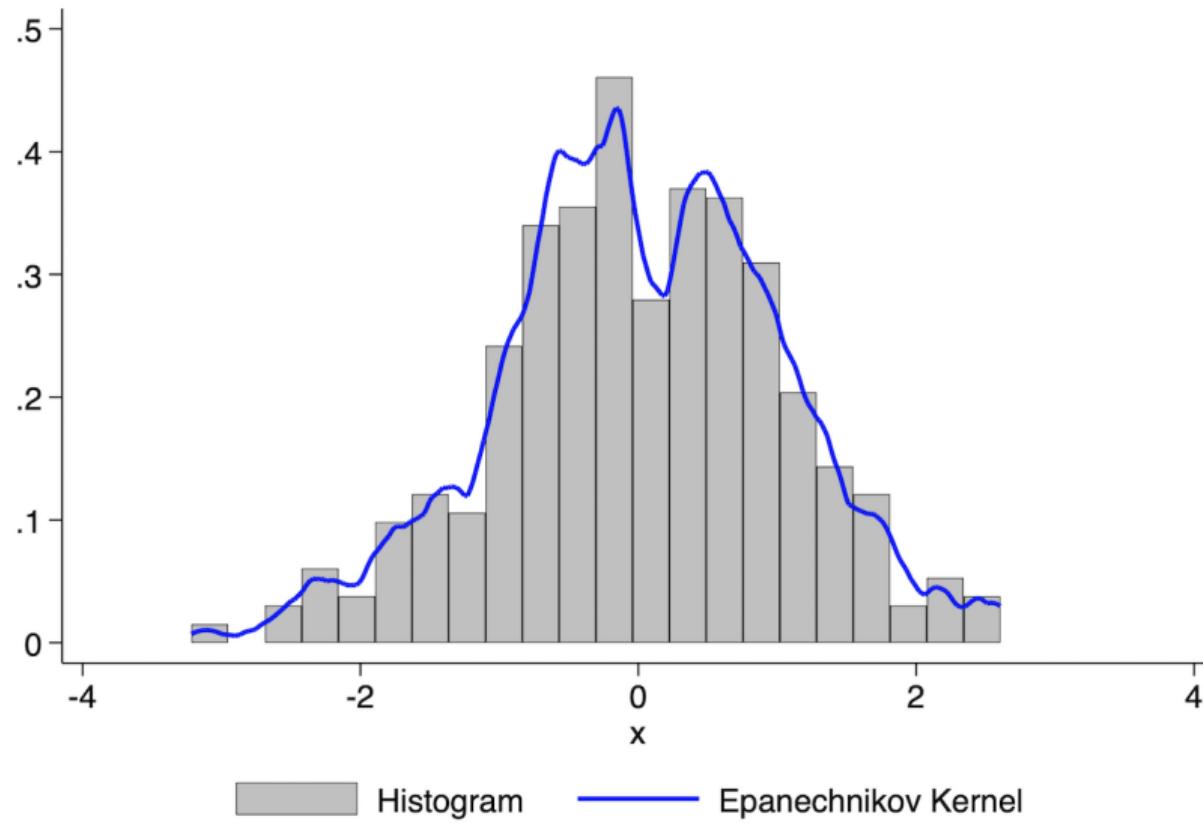
# Examples of Density Estimation

```
clear  
set seed 1234  
set obs 500  
gen x = rnormal()  
tw (histogram x, fc(gs12) lw(.1)) ///  
(kdensity x, lc(blue) lw(.5)), ///  
legend(label(1 "Histogram") label(2 "Epanechnikov Kernel")) ///  
note(Bandwidth: optimal ( 0.25))
```

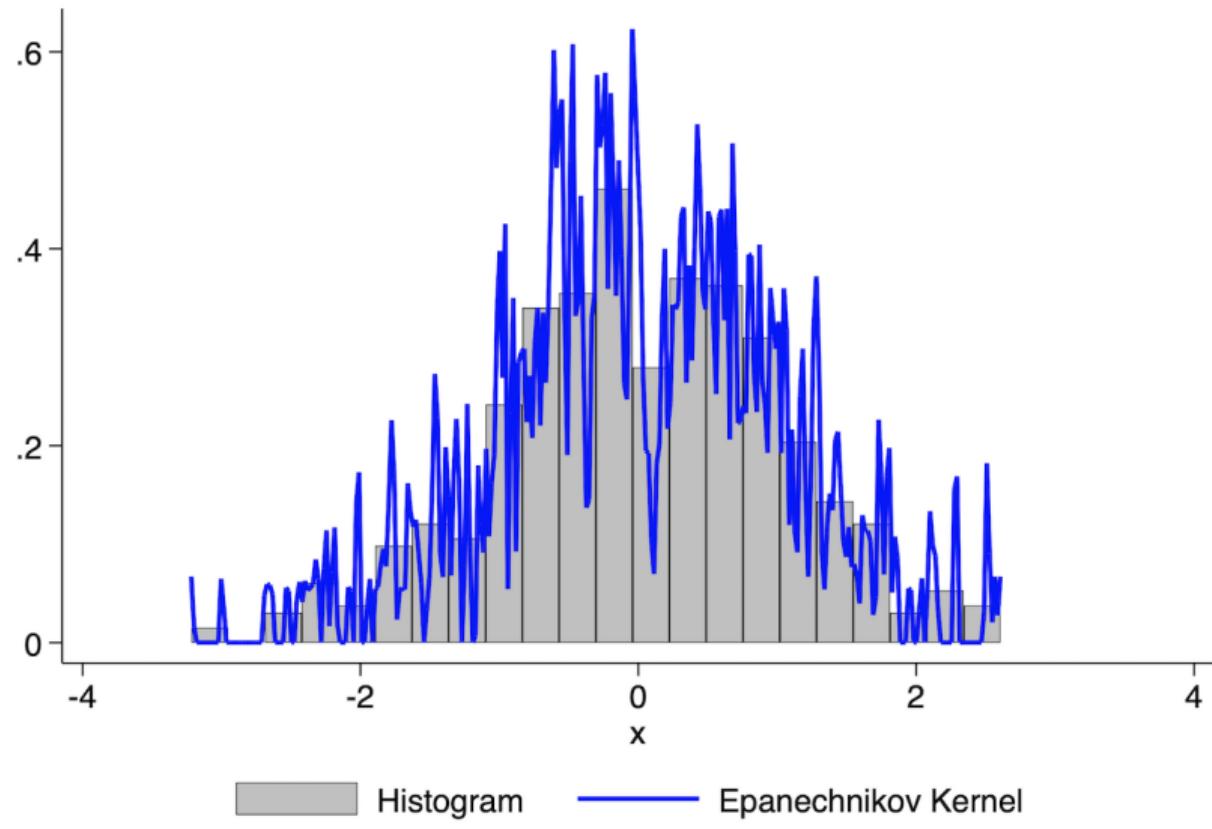
# Optimal Bandwidth Kernel



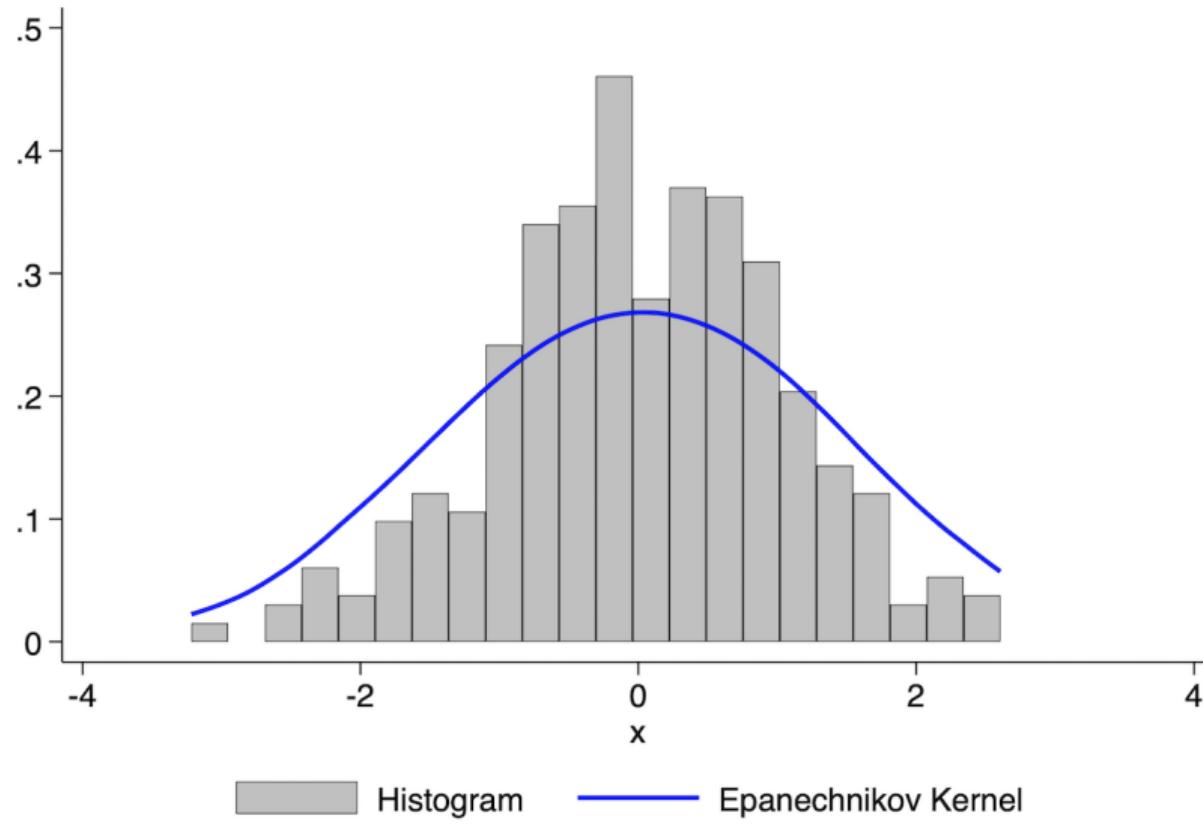
# Smaller Bandwidth



# Even Smaller



# Large Bandwidth



# Table of Contents

## 1 General Estimation Principles

Extremum Estimation

Examples of Extremum Estimators

## 2 Linear Regression Mechanics

The Relationship Between CEF and OLS

Using OLS to Estimate Means

## 3 Nonparametric Estimation and Visualization

Kernel Estimation

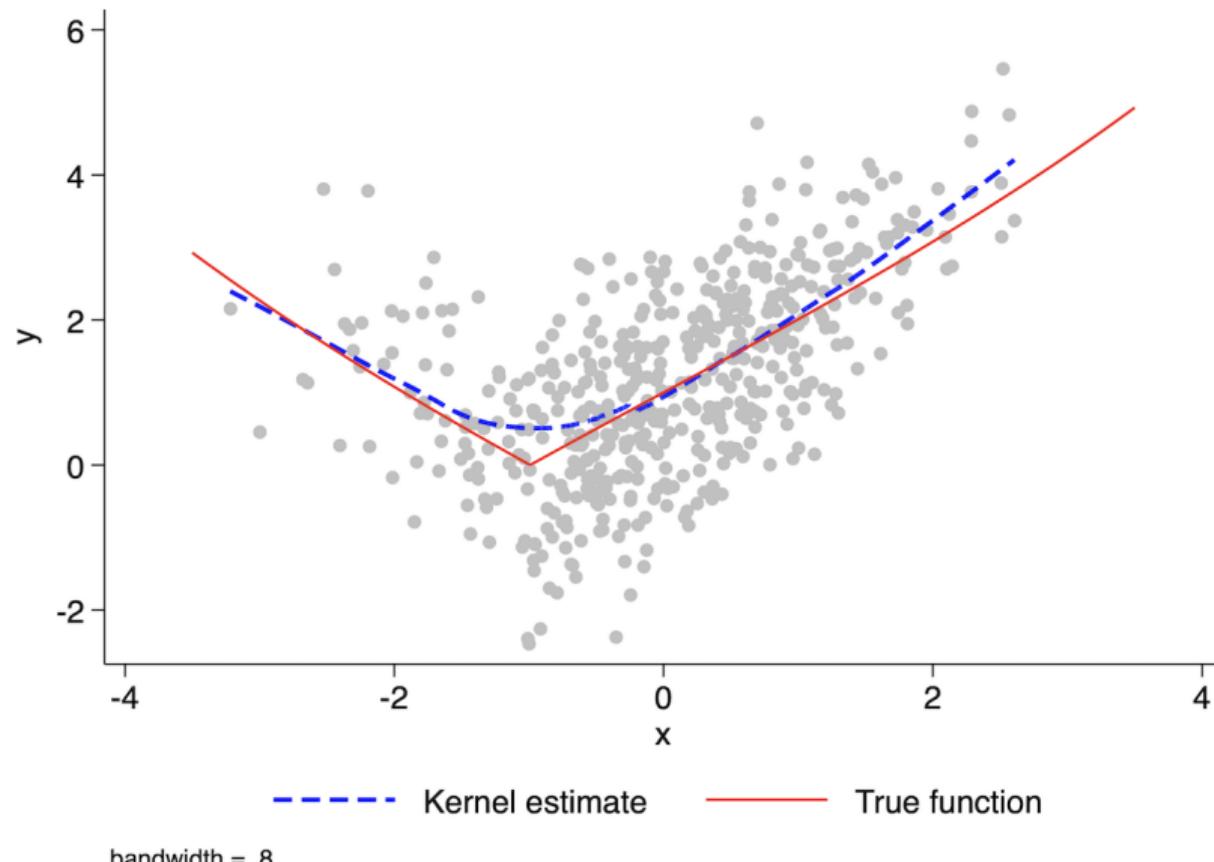
Applied Nonparametric CEF Estimation

## 4 Appendix: Semi-Parametric Efficiency of OLS

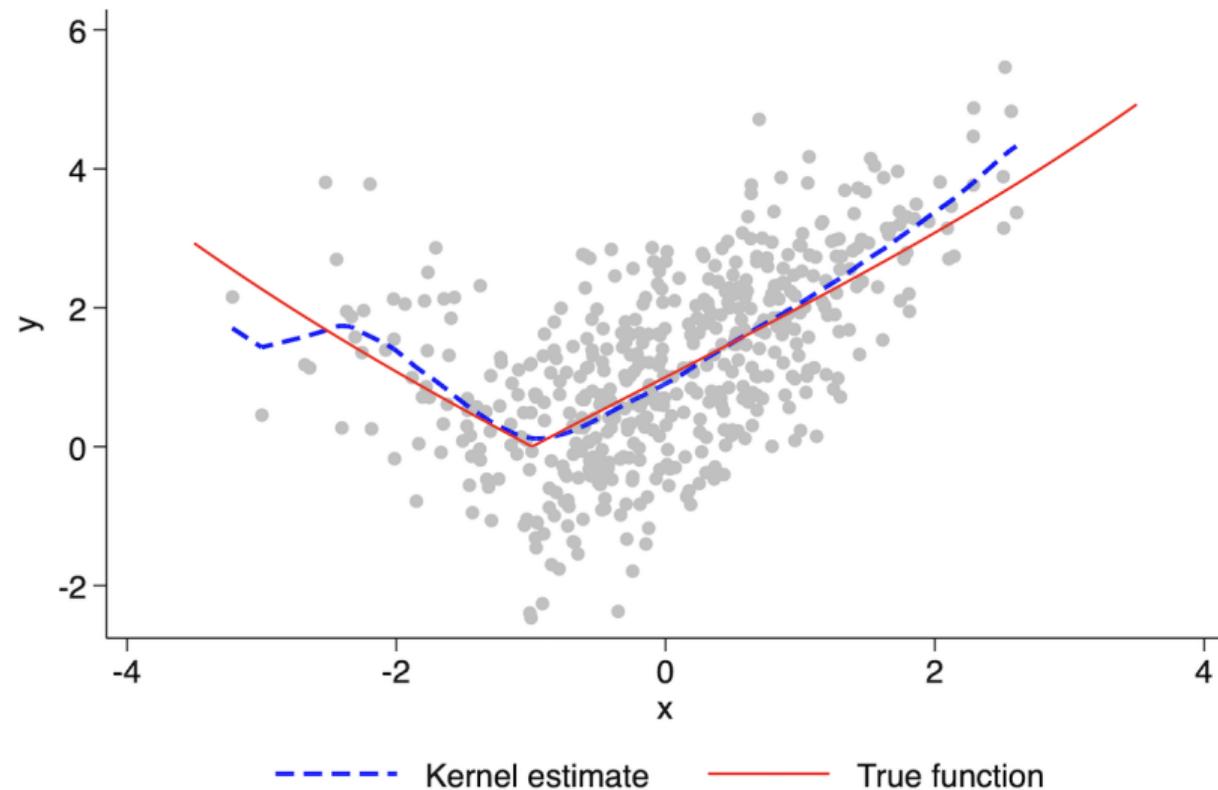
# Simulating a Nonlinear CEF

```
clear  
set seed 1234  
set obs 500  
gen x = rnormal()  
gen y = abs(1 + x + .01*x^3) + rnormal()  
* traditional LOWESS  
lowess y x, ///  
m(o) mc(gs12) lineopts(lc(blue) lw(.5)) ///  
addplot(function y = abs(1 + x + .01*x^3), range(-3.5 3.5) lc(red))  
///  
legend(order(2 3) label(2 Kernel estimate) label(3 True function))  
///  
title("")
```

# Locally Weighted Scatterplot Smoothing (LOWESS)

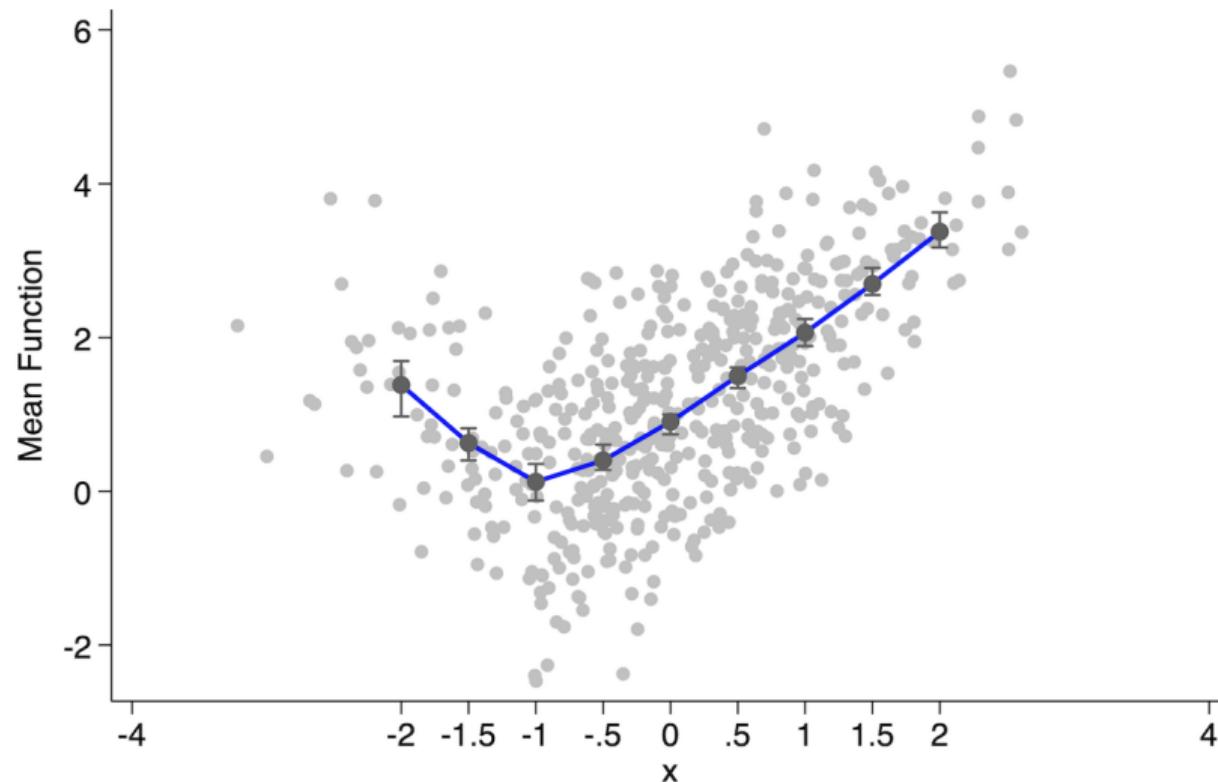


# Modern Cell Means Smoother: npregress



Local-linear estimates  
kernel = epanechnikov bandwidth = .2965328

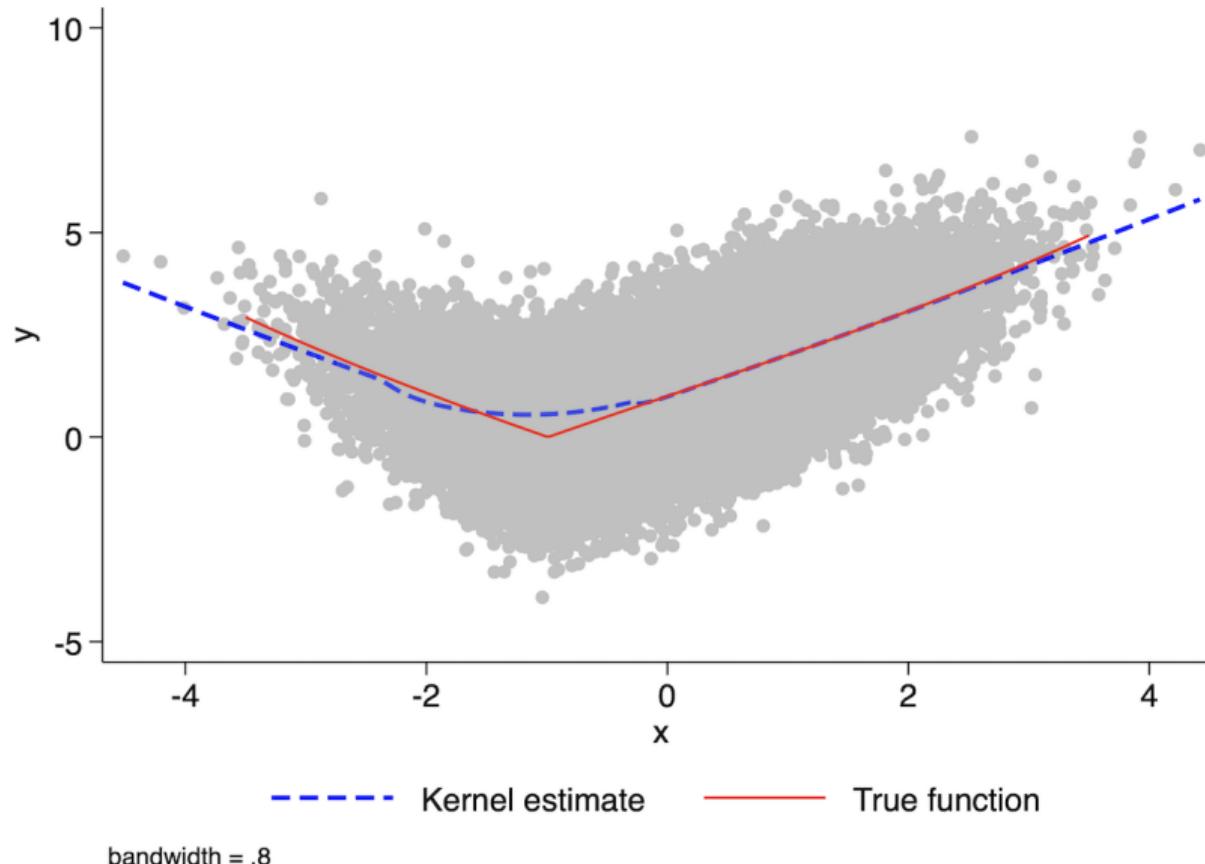
## npregress Also Estimates Confidence Intervals



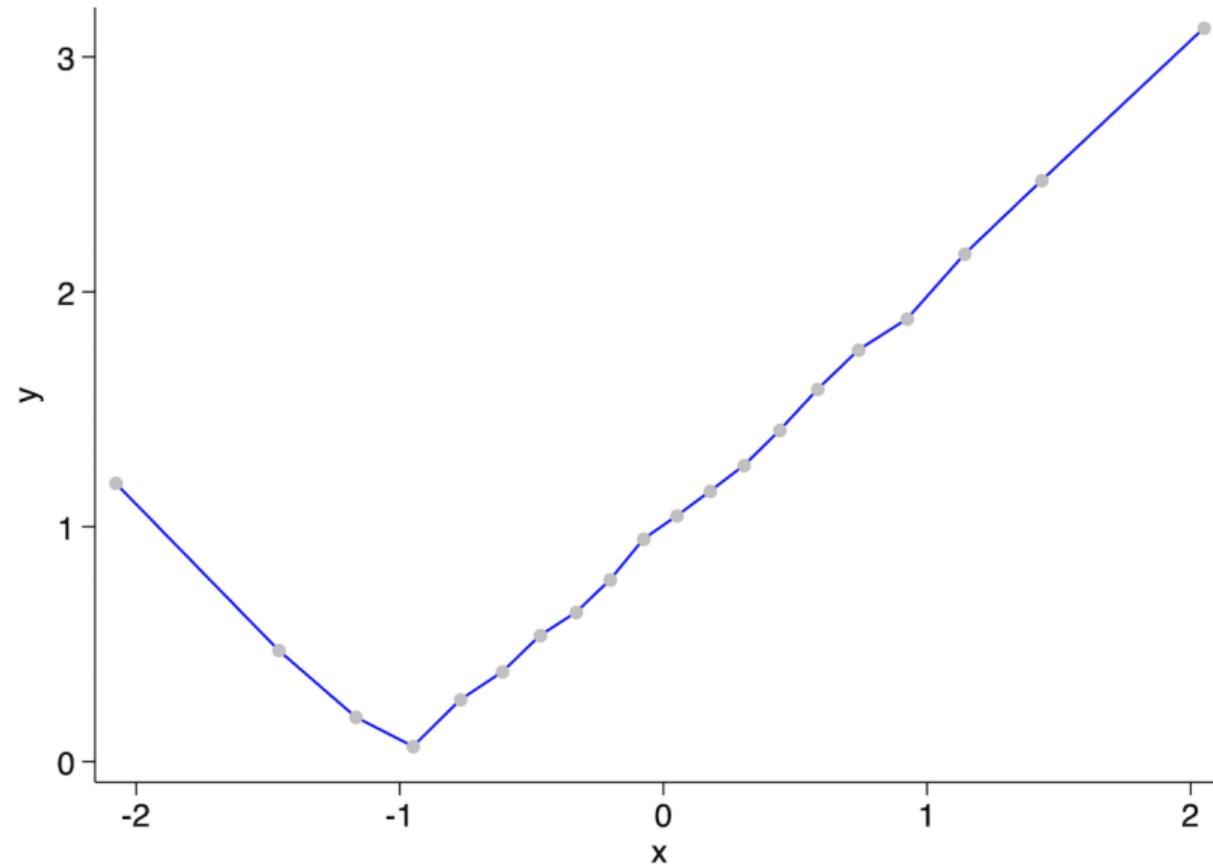
Mean Function

1

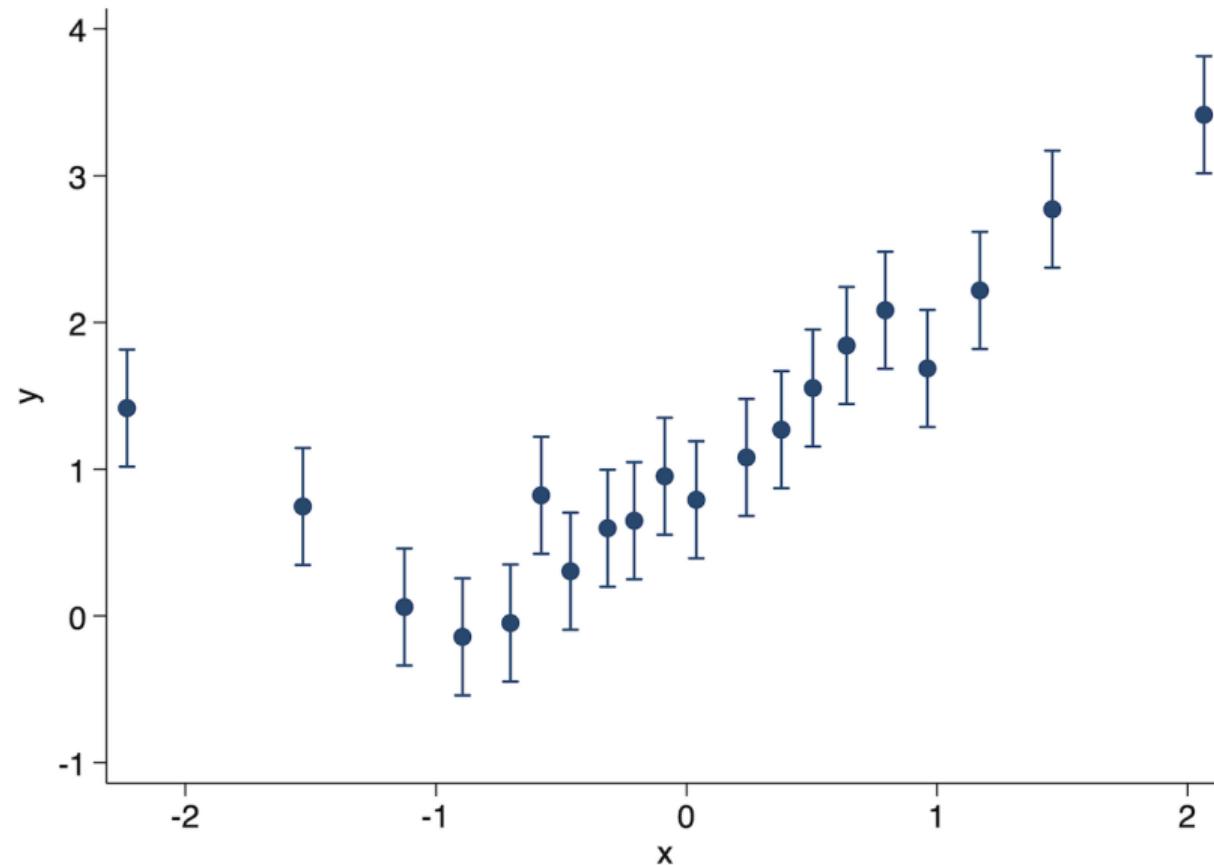
# Big Data: Most Smoothers Are Slow (here: LOWESS)



# Most Important Technique: binscatter



## Cutting Edge: binsreg



## Appendix: Semi-parametric Efficiency of OLS

# Efficiency of Cell Means Estimation

- How efficient is OLS?
- Recall BLUE (Gauss-Markov Theorem):
  - OLS is most efficient (i.e. lowest variance)...
  - ... among all linear unbiased estimators...
  - ... assuming  $\mathbb{E} [\varepsilon_i | \mathbf{X}_i] = 0$  and  $\mathbb{E} [\varepsilon \varepsilon' | \mathbf{X}_i] = \sigma^2 I$
- But what about general, nonlinear estimators?
  - MLE reaches Cramér-Rao Lower Bound (minimal variance)
  - Can OLS compete?
- Side note: recent work shows OLS is actually BUE... (Hansen 2022, ECMA)

# Semi-Parametric Efficiency of OLS

- It turns out the answer is yes (Chamberlain 1987)
- OLS is semi-parametrically efficient
  - We do not need errors to be homoskedastic
  - Using cell-means logic can show OLS = MLE
  - So OLS reaches Cramér-Rao Lower Bound as well
- Suppose i.i.d. random sample  $\mathbf{Z}_i = (Y_i, \mathbf{X}'_i)'$
- Because it is a sample,  $Y_i$  and  $\mathbf{X}_i$  are discrete
- Take on values  $z_j = (y_j, \mathbf{x}'_j)'$  for  $j = 1, \dots, J$  with

$$\mathbb{E} [1(\mathbf{Z}_i = z_j)] = \Pr (\mathbf{Z}_i = z_j) = \pi_j$$

# Population OLS of Cell Means

- Population OLS:

$$\begin{aligned}\beta_{OLS} &= \mathbb{E} [\mathbf{X}_i \mathbf{X}'_i]^{-1} \mathbb{E} [\mathbf{X}_i Y_i] \\ &= \mathbb{E} \left[ \sum_{j=1}^J 1[\mathbf{Z}_i = z_j] \mathbf{x}_j \mathbf{x}'_j \right]^{-1} \mathbb{E} \left[ \sum_{j=1}^J 1[\mathbf{Z}_i = z_j] \mathbf{x}_j y_j \right] \\ &= \left[ \sum_{j=1}^J \pi_j \mathbf{x}_j \mathbf{x}'_j \right]^{-1} \left[ \sum_{j=1}^J \pi_j \mathbf{x}_j y_j \right]\end{aligned}$$

- Unknown parameters:  $\pi = (\pi_1, \dots, \pi_J)'$

# Log Likelihood of Cell Means

- Fact:  $\mathbf{Z}_i \sim \text{Multinomial}(\pi_1, \dots, \pi_J)$
- Hence, log likelihood of data (dropping constant):

$$\log f(\mathbf{Z}_1, \dots, \mathbf{Z}_N, \pi) = \sum_{i=1}^N \sum_{j=1}^J 1[\mathbf{Z}_i = z_j] \log \pi_j$$

- Maximize this subject to  $\pi_j \geq 0$  and  $\sum_j \pi_j = 1$  yields

$$\hat{\pi}_{\text{MLE}} = \begin{bmatrix} \frac{1}{N} \sum_{i=1}^N 1[\mathbf{Z}_i = z_1] \\ \vdots \\ \frac{1}{N} \sum_{i=1}^N 1[\mathbf{Z}_i = z_J] \end{bmatrix}$$

## Cell Means OLS is MLE

- *Invariance Property of MLE:* For any  $\mu = f(\theta)$ , the MLE is

$$\hat{\mu}_{\text{MLE}} = f(\hat{\theta}_{\text{MLE}})$$

- Plugging MLE into population OLS:

$$\begin{aligned}\hat{\beta}_{\text{MLE}} &= \left[ \sum_j \hat{\pi}_j \mathbf{x}_j \mathbf{x}'_j \right]^{-1} \left[ \sum_j \hat{\pi}_j \mathbf{x}_j y_j \right] \\ &= \left[ \frac{1}{N} \sum_{j=1}^J \sum_{i=1}^N \mathbf{1}[\mathbf{z}_i = z_j] \mathbf{x}_j \mathbf{x}'_j \right]^{-1} \left[ \frac{1}{N} \sum_{j=1}^J \sum_{i=1}^N \mathbf{1}[\mathbf{z}_i = z_j] \mathbf{x}_j y_j \right] \\ &= \left[ \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i \mathbf{x}'_i \right]^{-1} \left[ \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i Y_i \right]\end{aligned}$$

- Hence, OLS is MLE! MLE reaches CRLB, and so does OLS

# Econometrics II

## Lecture 3: Inference Principles

Konrad Burchardi

Stockholm University

9th of April 2024

# Plan for Today

- 1 Inference Principles: Introduction
- 2 Classic Approach: Analytic Standard Errors
- 3 Sampling- and Design-Based Uncertainty
- 4 Bootstrap
- 5 Randomisation Inference

# Inference Principles: Introduction

**Goal:** “How certain is my estimate?”

**Focus:** What is standard deviation of estimator  $\hat{\beta}$ ,  $\sqrt{V(\hat{\beta})}$ ?

→ Estimator thereof is the “standard error of  $\hat{\beta}$ ”:  $\sqrt{\hat{V}(\hat{\beta})}$ .<sup>1</sup>

**Today:** Some answers, and many questions.

Very active research area!

Basic insights I thought were true turn out to be misleading.

---

<sup>1</sup>Confusing: In statistics the standard deviation of an estimator is often called “standard error”.

# Plan for Today

- 1 Inference Principles: Introduction
- 2 Classic Approach: Analytic Standard Errors
- 3 Sampling- and Design-Based Uncertainty
- 4 Bootstrap
- 5 Randomisation Inference

## Setup<sup>2</sup>

Suppose we have a sample of  $N$  individuals and estimate by OLS:

$$Y_i = \beta' X_i + \epsilon_i$$

where  $\beta$  and  $X_i$  are  $k \times 1$  vectors.

We have  $\hat{\beta}_{OLS} = (X'X)^{-1}X'Y$ . Assume  $\mathbb{E}[\epsilon|X] = 0$  and denote  $\Omega := \mathbb{E}[\epsilon\epsilon'|X]$ .

Then:

$$\mathbb{V}(\hat{\beta}|X) = (X'X)^{-1}(X'\Omega X)(X'X)^{-1}.$$

Denote  $\Omega_{ij} := \text{Cov}(\epsilon_i, \epsilon_j|X)$ .

---

<sup>2</sup>Reading suggestions: Angrist and Pischke, Chapter 8; Hansen, Chapter 4.

## Case 1: Homoskedastic Errors

Assume homoskedasticity:  $\Omega_{ij} = 0, \forall i \neq j$  and  $\Omega_{ii} = \sigma^2, \forall i$ .

Then

$$\begin{aligned}\mathbb{V}_{Homosc.}(\hat{\beta}|X) &= (X'X)^{-1}(X'\Omega X)(X'X)^{-1} \\ &= (X'X)^{-1}(X'\sigma^2 IX)(X'X)^{-1} \\ &= (X'X)^{-1}(\sigma^2 X'X)(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1}\end{aligned}$$

Two consistent estimators:

$$\hat{\mathbb{V}}_{HM0}(\hat{\beta}|X) = \frac{\sum_{i=1}^N \hat{\epsilon}_i^2}{N} (X'X)^{-1} \quad \text{or} \quad \hat{\mathbb{V}}_{HM1}(\hat{\beta}|X) = \frac{\sum_{i=1}^N \hat{\epsilon}_i^2}{N-k} (X'X)^{-1}.$$

## Case 1: Homoskedastic Errors (Bias Correction)

Two consistent estimators:

$$\hat{\mathbb{V}}_{HM0}(\hat{\beta}|X) = \frac{\sum_{i=1}^N \hat{\epsilon}_i^2}{N} (X'X)^{-1} \quad \text{or} \quad \hat{\mathbb{V}}_{HM1}(\hat{\beta}|X) = \frac{\sum_{i=1}^N \hat{\epsilon}_i^2}{N - k} (X'X)^{-1}.$$

It can be shown that  $\hat{\mathbb{V}}_{HM0}(\hat{\beta}|X)$  is biased.

Intuition: OLS overfits and hence  $\hat{\epsilon}_i$  underestimates  $\epsilon_i$ .

In contrast,  $\hat{\mathbb{V}}_{HM1}(\hat{\beta}|X)$  is unbiased.<sup>3</sup> (It is the default in STATA.)

Intuition: More  $k$ , more overfitting. Turns out  $N - k$  exactly right correction.

---

<sup>3</sup>For proofs check Hansen (2022), Chapters 4.11 and 4.13.

## Case 2: Heteroskedastic Errors

More reasonable assumption:  $\Omega_{ii} \neq \Omega_{jj}$  for at least some  $i, j$ .

$$\mathbb{V}_{\text{Heterosc.}}(\hat{\beta}|X) = (X'X)^{-1} \left( \sum_{i=1}^N \Omega_{ii} X_i X_i' \right) (X'X)^{-1}.$$

In case of heteroskedasticity,  $\hat{\mathbb{V}}_{HM1}(\hat{\beta}|X)$  is inconsistent for  $\mathbb{V}_{\text{Heterosc.}}(\hat{\beta}|X)$ .<sup>4</sup>  
Eicker-Huber-White (EHW) estimator consistent for  $\mathbb{V}_{\text{Heterosc.}}(\hat{\beta}|X)$ :

$$\hat{\mathbb{V}}_{EHW}(\hat{\beta}|X) = a \cdot (X'X)^{-1} \left( \sum_{i=1}^N \hat{\epsilon}_i^2 X_i X_i' \right) (X'X)^{-1},$$

where  $a$  is a bias correction factor.

---

<sup>4</sup>See Hansen (2022), Chapter 4.13.

## Case 2: Heteroskedastic Errors (Bias Correction)

$$\hat{\mathbb{V}}_{EHW}(\hat{\beta}|X) = a \cdot (X'X)^{-1} \left( \sum_{i=1}^N \hat{\epsilon}_i^2 X_i X_i' \right) (X'X)^{-1}.$$

Again, **bias correction**, **different versions**:

HC0:  $a = 1$ , poor performance in small samples.

HC1:  $a = N/(N - k)$ , ad hoc correction. STATA: , robust.

HC2:  $(X'X)^{-1} \left( \sum_{i=1}^N (1 - h_{ii})^{-1} \hat{\epsilon}_i^2 X_i X_i' \right) (X'X)^{-1}$ . STATA: , vce(hc2).

HC3:  $(X'X)^{-1} \left( \sum_{i=1}^N (1 - h_{ii})^{-2} \hat{\epsilon}_i^2 X_i X_i' \right) (X'X)^{-1}$ . STATA: , vce(hc3).

In case of interest, check also Young (2019, QJE).

# Non-diagonal $\Omega$

So far we assumed  $\Omega$  was diagonal. Why might it not be?

① Clusters in the data, within which  $\epsilon$ s are correlated:

- Students within schools,
- Households within villages,
- Firms within states.

Errors may be correlated b/c of common shocks / unobserved characteristics.

② Serial correlation in  $\epsilon$ s

- Dataset consists of individuals / firms / ... observed on multiple occasions.

Errors correlated with serially correlated shocks / persistent unobserved characteristics.

# Non-diagonal $\Omega$

So far we assumed  $\Omega$  was diagonal. Why might it not be?

① **Clusters** in the data, within which  $\epsilon_s$  are correlated:

- Students within schools,
- Households within villages,
- Firms within states.

Errors may be correlated b/c of common shocks / unobserved characteristics.

② **Serial correlation** in  $\epsilon_s$

- Dataset consists of individuals / firms / ... observed on multiple occasions.

Errors correlated with serially correlated shocks / persistent unobserved characteristics.

# Non-diagonal $\Omega$

So far we assumed  $\Omega$  was diagonal. Why might it not be?

① **Clusters** in the data, within which  $\epsilon_s$  are correlated:

- Students within schools,
- Households within villages,
- Firms within states.

Errors may be correlated b/c of common shocks / unobserved characteristics.

② **Serial correlation** in  $\epsilon_s$

- Dataset consists of individuals / firms / ... observed on multiple occasions.

Errors correlated with serially correlated shocks / persistent unobserved characteristics.

## Case 3: Non-diagonal $\Omega$ (Kloek-Moulton)

- Each unit is observed once and belongs to one of  $C$  clusters of equal size  $M$ , denoted by  $C_i \in \{1, \dots, C\}$ .<sup>5</sup>
- Error structure (note: homoskedastic-like):

$$\epsilon_{ic} = \alpha_c + \varepsilon_i$$

$$\Leftrightarrow \Omega_{ij} = \begin{cases} 0 & C_i \neq C_j \\ \rho_\epsilon \sigma^2 & C_i = C_j, i \neq j \\ \sigma^2 & i = j \end{cases}$$

We say that  $\Omega$  is “block diagonal”

$$\Omega = \begin{bmatrix} \Omega_1 & 0 & \cdots & 0 \\ 0 & \Omega_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Omega_C \end{bmatrix} \quad \Omega_c = \begin{bmatrix} \sigma^2 & \rho_\epsilon \sigma^2 & \cdots & \rho_\epsilon \sigma^2 \\ \rho_\epsilon \sigma^2 & \sigma^2 & \cdots & \rho_\epsilon \sigma^2 \\ \vdots & \vdots & \ddots & \vdots \\ \rho_\epsilon \sigma^2 & \rho_\epsilon \sigma^2 & \cdots & \sigma^2 \end{bmatrix}$$

---

<sup>5</sup>See Angrist and Pischke for version with heterogeneous cluster sizes.

## Case 3: Non-diagonal $\Omega$ (Kloek-Moulton)

Then

$$\mathbb{V}_{ClustersSpecial}(\hat{\beta}|X) = \mathbb{V}_{Homosc.}(\hat{\beta}|X) \times \underbrace{(1 + \rho_\epsilon \rho_X (M - 1))}_{\text{"Moulton Factor"}},$$

where  $\rho_X$  is the intra-cluster correlation of  $X$ .

### Insights:

bias: Expect  $\rho_\epsilon > 0$ , so if  $\rho_X > 0$ ,  $\mathbb{V}_{Homosc.} <$  true variance.

$\rho_\epsilon = 1$ : If other covariates constant in cluster, adding new observations adds no new information.

$\rho_X = 0$ : Treatment assignment fully independent of cluster, e.g. “completely randomized” experiments.

$\rho_X = 1$ : Treatment assigned to whole clusters, e.g. “cluster randomized” experiments, school-level programs,...

# clusters: More severe with fewer clusters (big  $M$  given  $N$ ).

## Case 3: Non-diagonal $\Omega$ (Kloek-Moulton)

Then

$$\mathbb{V}_{ClustersSpecial}(\hat{\beta}|X) = \mathbb{V}_{Homosc.}(\hat{\beta}|X) \times \underbrace{(1 + \rho_\epsilon \rho_X (M - 1))}_{\text{"Moulton Factor"}},$$

where  $\rho_X$  is the intra-cluster correlation of  $X$ .

**Insights:**

bias: Expect  $\rho_\epsilon > 0$ , so if  $\rho_X > 0$ ,  $\mathbb{V}_{Homosc.} <$  true variance.

$\rho_\epsilon = 1$ : If other covariates constant in cluster, adding new observations adds no new information.

$\rho_X = 0$ : Treatment assignment fully independent of cluster, e.g. “completely randomized” experiments.

$\rho_X = 1$ : Treatment assigned to whole clusters, e.g. “cluster randomized” experiments, school-level programs,...

# clusters: More severe with fewer clusters (big  $M$  given  $N$ ).

## Case 3: Non-diagonal $\Omega$ (Kloek-Moulton)

Then

$$\mathbb{V}_{ClustersSpecial}(\hat{\beta}|X) = \mathbb{V}_{Homosc.}(\hat{\beta}|X) \times \underbrace{(1 + \rho_\epsilon \rho_X (M - 1))}_{\text{"Moulton Factor"}},$$

where  $\rho_X$  is the intra-cluster correlation of  $X$ .

**Insights:**

bias: Expect  $\rho_\epsilon > 0$ , so if  $\rho_X > 0$ ,  $\mathbb{V}_{Homosc.} <$  true variance.

$\rho_\epsilon = 1$ : If other covariates constant in cluster, adding new observations adds no new information.

$\rho_X = 0$ : Treatment assignment fully independent of cluster, e.g. “completely randomized” experiments.

$\rho_X = 1$ : Treatment assigned to whole clusters, e.g. “cluster randomized” experiments, school-level programs,...

# clusters: More severe with fewer clusters (big  $M$  given  $N$ ).

## Case 3: Non-diagonal $\Omega$ (Kloek-Moulton)

Then

$$\mathbb{V}_{ClustersSpecial}(\hat{\beta}|X) = \mathbb{V}_{Homosc.}(\hat{\beta}|X) \times \underbrace{(1 + \rho_\epsilon \rho_X (M - 1))}_{\text{"Moulton Factor"}},$$

where  $\rho_X$  is the intra-cluster correlation of  $X$ .

**Insights:**

bias: Expect  $\rho_\epsilon > 0$ , so if  $\rho_X > 0$ ,  $\mathbb{V}_{Homosc.} <$  true variance.

$\rho_\epsilon = 1$ : If other covariates constant in cluster, adding new observations adds no new information.

$\rho_X = 0$ : Treatment assignment fully independent of cluster, e.g. “completely randomized” experiments.

$\rho_X = 1$ : Treatment assigned to whole clusters, e.g. “cluster randomized” experiments, school-level programs,...

# clusters: More severe with fewer clusters (big  $M$  given  $N$ ).

## Case 4: Non-diagonal $\Omega$ (Liang-Zeger)

Clustered errors typically estimated assuming more general error structure:

- Let  $X_c$  correspond to the submatrix of  $X$  with  $C_i = c$ .
- Allow for unrestricted  $\Omega_{ij}$  **within clusters**.
- Impose  $\Omega_{ij} = 0$  for  $C_i \neq C_j$ .

Then:

$$\mathbb{V}_{ClustersGeneral}(\hat{\beta}) = (X'X)^{-1} \left( \sum_{c=1}^C X_c' \Omega_c X_c \right) (X'X)^{-1}$$

$$\hat{\mathbb{V}}_{LZ}(\hat{\beta}) = a \cdot (X'X)^{-1} \left( \sum_{c=1}^C X_c' \hat{\epsilon}_c \hat{\epsilon}_c' X_c \right) (X'X)^{-1}$$

$\hat{\mathbb{V}}_{LZ}(\hat{\beta})$  is consistent for  $\mathbb{V}_{ClustersGeneral}(\hat{\beta})$  (as  $C \rightarrow \infty$ ), and  $a$  is bias correction.

STATA: `cluster(cluster_id)` or `vce(cluster cluster_id)`, with  $a = \frac{N-1}{N-k} \frac{C}{C-1}$ .

# “Classic” Advice: When to Cluster?

## “Classic” recommendations:

- Cluster if there could be intra-cluster correlation in the error term.
- Compare robust and clustered standard errors, and pick the bigger ones: If clustering increases the standard errors then it is conservative to do it, if not then no harm done.
- Cluster at the highest level, subject to having “sufficiently many” clusters.

I am afraid those recommendations **might not age well**, see later.

## Case 5: Non-diagonal $\Omega$ (Serial Correlation)

Often units are observed on multiple occasions over time.

- Typical case: panel data,
  - e.g. individuals in different states in annual tax data,
  - e.g. schools pre/post education reform,
  - e.g. an individual's sequence of decisions in a lab experiment.
- Serially correlated shocks or unobservables: correlation between the residuals.
- Conceptually very similar to correlation between disturbances within clusters.
- There exist variance estimators designed for serial correlation (Newey-West).
- Common to just cluster at the unit level or higher (e.g. person, state, school) which allows for more general variance-covariance structure.

## Bertrand, Duflo, Mullainathan (QJE, 2004)

*“How Much Should We Trust Difference-In-Difference Estimates?”*

Bertrand et al. (2004) focus on the case of D-in-D estimation, with a treatment that affects some units (e.g. states) at some point in time.

Influential: by far Esther Duflo's most cited paper!

- Outcomes within a state correlated over time, so over-time observations are not independent measures of state.
- Show that failing to correct for serial correlation leads to over-rejection of the null of no effect.
- Clustering performs well with “sufficiently many” clusters.

Popularised clustering.

# Plan for Today

- 1 Inference Principles: Introduction
- 2 Classic Approach: Analytic Standard Errors
- 3 Sampling- and Design-Based Uncertainty
- 4 Bootstrap
- 5 Randomisation Inference

# Sources of Uncertainty

Abadie, Athey, Imbens and Wooldridge:

*Where is uncertainty about estimate coming from?*

Think about some scenarios:

- ① Estimate is average age in this room...

...and you have data on age of all of us.

- ② Estimate is average age in this room...

...and you have data on age of randomly selected 5 of us.

- ③ Estimate is effect of treatment  $D$  for those in this room...

...and you have data on  $D$  and  $Y$  for all of us.

# Sources of Uncertainty

Abadie, Athey, Imbens and Wooldridge:

*Where is uncertainty about estimate coming from?*

Think about some scenarios:

- 1 Estimate is average age in this room...

...and you have data on age of all of us.

- 2 Estimate is average age in this room...

...and you have data on age of randomly selected 5 of us.

- 3 Estimate is effect of treatment  $D$  for those in this room...

...and you have data on  $D$  and  $Y$  for all of us.

# Sources of Uncertainty

Abadie, Athey, Imbens and Wooldridge:

*Where is uncertainty about estimate coming from?*

Think about some scenarios:

- 1 Estimate is average age in this room...

...and you have data on age of all of us.

- 2 Estimate is average age in this room...

...and you have data on age of randomly selected 5 of us.

- 3 Estimate is effect of treatment  $D$  for those in this room...

...and you have data on  $D$  and  $Y$  for all of us.

# Sources of Uncertainty

Abadie, Athey, Imbens and Wooldridge:

*Where is uncertainty about estimate coming from?*

Think about some scenarios:

- 1 Estimate is average age in this room...

...and you have data on age of all of us.

- 2 Estimate is average age in this room...

...and you have data on age of randomly selected 5 of us.

- 3 Estimate is effect of treatment  $D$  for those in this room...

...and you have data on  $D$  and  $Y$  for all of us.

# Sources of Uncertainty

Something is confusing!

- What type of uncertainty do we express with standard standard errors?
- When having a sample of size  $N = 100$ , our standard errors do not take into account whether it is drawn from a population of 1.000, or 1.000.000.
- What on earth are the errors?
- What does it mean that “the Xs are fixed”?
- ...

Guido Imbens talks about the status quo when presenting his current work like Steve Jobs about the blackberry when presenting the iPhone: “Bääää!”

## Abadie, Athey, Imbens and Wooldridge (2020)

Abadie et al. (2020) distinguish between **sampling-based uncertainty** and **design-based uncertainty**.

They propose that it is useful to think (again) about:

- ① the estimand of interest,
- ② the population of interest,
- ③ the sampling process, and
- ④ the assignment process.

# Abadie, Athey, Imbens and Wooldridge (2020)

## The Set-Up

### Set-Up:

- Finite population consisting of  $n$  units.
- Each unit characterized by  $(Y_i, X_i)$ .
- Whether unit  $i$  is in the sample is indicated by  $R_i \in \{0, 1\}$ .

# Sampling-Based Uncertainty

Consider:

- **estimand** which is a function of the full set  $\{(Y_i, X_i)\}_{i=1}^n$ , and
  - **estimator** which is a function of the observed data  $\{(R_i, R_i Y_i, R_i X_i)\}_{i=1}^n$ .
- Uncertainty about estimand arises when we observe the values  $(Y_i, X_i)$  only for sample, i.e. subset of **population!**
- **Sampling-based inference** uses information about the **sampling process** that determines  $\{R_i\}_{i=1}^n$  to assess variability of estimators across different samples.

# Sampling-Based Uncertainty

Table 1: SAMPLING-BASED UNCERTAINTY ( $\checkmark$  IS OBSERVED, ? IS MISSING)

Unit	Actual Sample			Alternative Sample I			Alternative Sample II			...
	$Y_i$	$Z_i$	$R_i$	$Y_i$	$Z_i$	$R_i$	$Y_i$	$Z_i$	$R_i$	...
1	$\checkmark$	$\checkmark$	1	?	?	0	?	?	0	...
2	?	?	0	?	?	0	?	?	0	...
3	?	?	0	$\checkmark$	$\checkmark$	1	$\checkmark$	$\checkmark$	1	...
4	?	?	0	$\checkmark$	$\checkmark$	1	?	?	0	...
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	...
$n$	$\checkmark$	$\checkmark$	1	?	?	0	?	?	0	...

From Abadie et al. (2020).

# Abadie, Athey, Imbens and Wooldridge (2020)

## The Set-Up

### Different Scenario:

- Observe for each unit in the population the value of one of two potential outcomes,  $Y_i^*(1)$  or  $Y_i^*(0)$ , but not both.
- Which potential outcome is observed is indicated by  $X_i \in \{0, 1\}$ .<sup>6</sup>
- Denote the observed outcome as  $Y_i = Y_i^*(X_i)$ .

---

<sup>6</sup>Otherwise we will use  $D_i$  as treatment indicator in this course. But here the main point is that we are talking about an explanatory variable, and those we called  $X_i$  today.

# Design-Based Uncertainty

Consider:

- **estimand** which is a function of the full set  $\{(Y_i^*(1), Y_i^*(0), X_i)\}_{i=1}^n$ , and
  - **estimator** which is a function of the observed data  $\{(Y_i, X_i)\}_{i=1}^n$ .
- Uncertainty about estimand arises because different observations are assigned to treatment across different realisations of the assignment.
- **Design-based inference** uses information about the **assignment process** that determines  $\{X_i\}_{i=1}^n$  to assess the variability of the estimator.

# Design-Based Uncertainty

Table 2: DESIGN-BASED UNCERTAINTY ( $\checkmark$  IS OBSERVED, ? IS MISSING)

Unit	Actual Sample			Alternative Sample I			Alternative Sample II			...
	$Y_i^*(1)$	$Y_i^*(0)$	$X_i$	$Y_i^*(1)$	$Y_i^*(0)$	$X_i$	$Y_i^*(1)$	$Y_i^*(0)$	$X_i$	...
1	$\checkmark$	?	1	$\checkmark$	?	1	?	$\checkmark$	0	...
2	?	$\checkmark$	0	?	$\checkmark$	0	?	$\checkmark$	0	...
3	?	$\checkmark$	0	$\checkmark$	?	1	$\checkmark$	?	1	...
4	?	$\checkmark$	0	?	$\checkmark$	0	$\checkmark$	?	1	...
:	:	:	:	:	:	:	:	:	:	...
$n$	$\checkmark$	?	1	?	$\checkmark$	0	?	$\checkmark$	0	...

From Abadie et al. (2020).

# Abadie, Athey, Imbens and Wooldridge (2020)

## Estimands

$\mathbf{Y}$ ,  $\mathbf{Y}^*(1)$ ,  $\mathbf{Y}^*(0)$ ,  $\mathbf{R}$ ,  $\mathbf{X}$  stacked vectors of corresponding unit-level variables.

### Classification of Estimands:

**Descriptive Estimand:** An estimand which can be written as a function of  $(\mathbf{Y}, \mathbf{X})$ , free of dependence on  $\mathbf{R}$  and on potential outcomes beyond the realized outcome.

**Causal Estimand:** An estimand that depends on potential outcomes  $\mathbf{Y}^*(1)$ ,  $\mathbf{Y}^*(0)$ .

# Abadie, Athey, Imbens and Wooldridge (2020)

Consider three **estimands**:

$$\theta^{\text{sampling}}(\mathbf{Y}, \mathbf{X}) = \frac{1}{n_1} \sum_{i=1}^n X_i Y_i - \frac{1}{n_0} \sum_{i=1}^n (1 - X_i) Y_i$$

$$\theta^{\text{design}}(\mathbf{Y}^*(1), \mathbf{Y}^*(0), \mathbf{R}) = \frac{1}{N} \sum_{i=1}^n R_i (Y_i^*(1) - Y_i^*(0))$$

$$\theta^{\text{causal}}(\mathbf{Y}^*(1), \mathbf{Y}^*(0)) = \frac{1}{n} \sum_{i=1}^n (Y_i^*(1) - Y_i^*(0)),$$

where  $n_0$  and  $n_1$  refer to the **number of units in the population** who are untreated and treated, respectively, and  $N_0$  and  $N_1$  refer to the sample similarly.

Consider the difference-in-sample-means estimator (OLS of  $Y_i$  on  $X_i$  and constant):

$$\hat{\theta} = \frac{1}{N_1} \sum_{i=1}^n R_i X_i Y_i - \frac{1}{N_0} \sum_{i=1}^n R_i (1 - X_i) Y_i.$$

# Abadie, Athey, Imbens and Wooldridge (2020)

## Estimator

Assume random sampling and random assignment.

With appropriate conditioning, the  $\hat{\theta}$  estimator is unbiased for each estimand:

$$\mathbb{E}_{\mathbf{R}}[\hat{\theta} | \mathbf{X}, N_1, N_0] = \theta^{\text{sampling}}$$

$$\mathbb{E}_{\mathbf{X}}[\hat{\theta} | \mathbf{R}, N_1, N_0] = \theta^{\text{design}}$$

$$\mathbb{E}_{\mathbf{X}, \mathbf{R}}[\hat{\theta} | N_1, N_0] = \theta^{\text{total}}$$

Interpretation of conditioning:

- Considering randomness of  $\mathbf{R}$  only gives sampling-based uncertainty.
- Considering randomness of  $\mathbf{X}$  only gives design-based uncertainty.
- Not conditioning accounts for both types of uncertainty.

# Abadie, Athey, Imbens and Wooldridge (2020)

Finally: Variances!

Finally, we can write out the variances of our estimator for each estimand:

$$\begin{aligned} V^{\text{sampling}} &= \mathbb{E}_{\mathbf{X}}[\text{Var}_{\mathbf{R}}(\hat{\theta}|\mathbf{X}, N_1, N_0)|N_1, N_0] &= \frac{S_1^2}{N_1} \left(1 - \frac{N_1}{n_1}\right) + \frac{S_0^2}{N_0} \left(1 - \frac{N_0}{n_0}\right) \\ V^{\text{design}} &= \mathbb{E}_{\mathbf{R}}[\text{Var}_{\mathbf{X}}(\hat{\theta}|\mathbf{R}, N_1, N_0)|N_1, N_0] &= \frac{S_1^2}{N_1} + \frac{S_0^2}{N_0} - \frac{S_{\theta}^2}{N_0 + N_1} \\ V^{\text{total}} &= \text{Var}_{\mathbf{X}, \mathbf{R}}(\hat{\theta}|\mathbf{X}, N_1, N_0) &= \frac{S_1^2}{N_1} + \frac{S_0^2}{N_0} - \frac{S_{\theta}^2}{n_0 + n_1}, \end{aligned}$$

where denote with  $S_0^2$ ,  $S_1^2$ , and  $S_{\theta}^2$  the population variance of  $Y_i^*(0)$ ,  $Y_i^*(1)$  and the treatment effect  $Y_i^*(1) - Y_i^*(0)$ .<sup>7</sup> <sup>8</sup>

---

<sup>7</sup>To arrive at the former two expressions we take expectations over the conditional variances.

<sup>8</sup>For proofs check the supplementary material to the paper, and also Imbens and Rubin (2015), Chapter 6.

# Abadie, Athey, Imbens and Wooldridge (2020)

Finally: Variances!

$$V^{\text{sampling}} = \frac{S_1^2}{N_1} \left(1 - \frac{N_1}{n_1}\right) + \frac{S_0^2}{N_0} \left(1 - \frac{N_0}{n_0}\right)$$

$$V^{\text{design}} = \frac{S_1^2}{N_1} + \frac{S_0^2}{N_0} - \frac{S_\theta^2}{N_0 + N_1}$$

$$V^{\text{total}} = \frac{S_1^2}{N_1} + \frac{S_0^2}{N_0} - \frac{S_\theta^2}{n_0 + n_1}$$

- ① For fixed  $N_0$  and  $N_1$ , if  $n_0, n_1 \rightarrow \infty$ , the total and sampling variance are equal.  
→ All uncertainty comes from randomness in sampling.

# Abadie, Athey, Imbens and Wooldridge (2020)

Finally: Variances!

$$V^{\text{sampling}} = \frac{S_1^2}{N_1} \left(1 - \frac{N_1}{n_1}\right) + \frac{S_0^2}{N_0} \left(1 - \frac{N_0}{n_0}\right)$$

$$V^{\text{design}} = \frac{S_1^2}{N_1} + \frac{S_0^2}{N_0} - \frac{S_\theta^2}{N_0 + N_1}$$

$$V^{\text{total}} = \frac{S_1^2}{N_1} + \frac{S_0^2}{N_0} - \frac{S_\theta^2}{n_0 + n_1}$$

- ② Both when the estimand is  $\theta^{\text{descriptive}}$  or  $\theta^{\text{causal}}$ , ignoring finite population leads to overstatement of variance, but not for  $\theta^{\text{causal, sample}}$ .

Intuition?

# Abadie, Athey, Imbens and Wooldridge (2020)

Finally: Variances!

$$V^{\text{sampling}} = \frac{S_1^2}{N_1} \left(1 - \frac{N_1}{n_1}\right) + \frac{S_0^2}{N_0} \left(1 - \frac{N_0}{n_0}\right)$$

$$V^{\text{design}} = \frac{S_1^2}{N_1} + \frac{S_0^2}{N_0} - \frac{S_\theta^2}{N_0 + N_1}$$

$$V^{\text{total}} = \frac{S_1^2}{N_1} + \frac{S_0^2}{N_0} - \frac{S_\theta^2}{n_0 + n_1}$$

- ③ The expectation of the Eicker-Huber-White estimator is  $\frac{S_1^2}{N_1} + \frac{S_0^2}{N_0}$ .
- Generally over-estimates variance for well-defined estimand.
  - Eicker-Huber-White estimator is assuming infinite super-population!

# Abadie, Athey, Imbens and Wooldridge (2020)

Finally: Variances!

$$V^{\text{sampling}} = \frac{S_1^2}{N_1} \left(1 - \frac{N_1}{n_1}\right) + \frac{S_0^2}{N_0} \left(1 - \frac{N_0}{n_0}\right)$$

$$V^{\text{design}} = \frac{S_1^2}{N_1} + \frac{S_0^2}{N_0} - \frac{S_\theta^2}{N_0 + N_1}$$

$$V^{\text{total}} = \frac{S_1^2}{N_1} + \frac{S_0^2}{N_0} - \frac{S_\theta^2}{n_0 + n_1}$$

- ④ Problem: Unclear how to estimate  $S_\theta^2$ !

→ Eicker-Huber-White estimator implicitly sets it to 0.

→ Check paper for approaches.

# Abadie, Athey, Imbens and Wooldridge

"When Should You Adjust Standard Errors for Clustering?"

What does all of this imply for clustering?

The screenshot shows a dark-themed website for the Chamberlain Seminar. On the left, there's a sidebar with a vertical list of past seminar seasons: Home, Instructions for Attendees, Past Seminars (with a dropdown arrow), Winter 2022, Autumn 2021 (which is highlighted in red), Spring 2021, Winter 2021, Autumn 2020, Spring 2020, and Other Online Seminars.

Friday, October 8, 2021: Guido Imbens (Stanford)

"Clustering Adjustments to Standard Errors" (with Alberto Abadie, Susan Athey, Jeffrey Wooldridge)

Discussant: Colin Cameron (UC Davis)

Moderator: Isaiah Andrews (Harvard)



Friday, October 8, noon ET / 5pm UK

Guido Imbens (Stanford): "Clustering Adjustments to Standard Errors"  
(with Alberto Abadie, Susan Athey, Jeffrey Wooldridge)

Discussant: Colin Cameron  
Moderator: Isaiah Andrews

Watch on [seminar.org #chamberlainseminar](#)

<https://www.chamberlainseminar.org/past-seminars/autumn-2021>

<https://academic.oup.com/qje/article/138/1/1/6750017>

# Plan for Today

- 1 Inference Principles: Introduction
- 2 Classic Approach: Analytic Standard Errors
- 3 Sampling- and Design-Based Uncertainty
- 4 Bootstrap
- 5 Randomisation Inference

# Bootstrap

**Conventional econometrics:** Infer the distribution of a statistic,  $f$  (e.g., t-statistic)

- calculated from a sample with empirical distribution  $F_1$
- drawn from a infinite population with distribution  $F_0$ .

Call this the distribution of  $f(F_1|F_0)$ .

**The Bootstrap:** Estimates the distribution of  $f(F_1|F_0)$

- by drawing random samples  $F_2$  (with replacement) from  $F_1$ ,
- and calculate the statistic  $f$  each time.
- If  $f$  is a smooth function of the data, then  $f(F_2|F_1) \rightarrow_d f(F_1|F_0)$ .

Intuition: treat sample distribution  $F_1$  as though it were the population distribution.

# Bootstrap

**Conventional econometrics:** Infer the distribution of a statistic,  $f$  (e.g., t-statistic)

- calculated from a sample with empirical distribution  $F_1$
- drawn from a infinite population with distribution  $F_0$ .

Call this the distribution of  $f(F_1|F_0)$ .

**The Bootstrap:** Estimates the distribution of  $f(F_1|F_0)$

- by drawing random samples  $F_2$  (**with replacement**) from  $F_1$ ,
- and calculate the statistic  $f$  each time.
- If  $f$  is a smooth function of the data, then  $f(F_2|F_1) \rightarrow_d f(F_1|F_0)$ .

Intuition: treat sample distribution  $F_1$  as though it were the population distribution.

# Bootstrap

## Some Remarks:

- Sometimes analytic errors are not available, or hard to compute.  
(For example when your regression includes “generated regressors”.)
- “Asymptotic refinement”: can sometimes get closer to the true finite-sample distribution than asymptotic approximations.  
→ Requires the bootstrapped statistics to be asymptotically pivotal.
- Bootstrap “feels” like it is addressing sampling uncertainty. But Abadie et al. (2020) clarify in their setting the expectation of the bootstrapped variance equals the Eicker-Huber-White estimator.

# Bootstrap

## Some Remarks:

- Sometimes analytic errors are not available, or hard to compute.  
(For example when your regression includes “generated regressors”.)
- “Asymptotic refinement”: can sometimes get closer to the true finite-sample distribution than asymptotic approximations.  
→ Requires the bootstrapped statistics to be asymptotically pivotal.
- Bootstrap “feels” like it is addressing sampling uncertainty. But Abadie et al. (2020) clarify in their setting the expectation of the bootstrapped variance equals the Eicker-Huber-White estimator.

# Bootstrap

## Some Remarks:

- Sometimes analytic errors are not available, or hard to compute.  
(For example when your regression includes “generated regressors”.)
- “Asymptotic refinement”: can sometimes get closer to the true finite-sample distribution than asymptotic approximations.  
→ Requires the bootstrapped statistics to be asymptotically pivotal.
- Bootstrap “feels” like it is addressing sampling uncertainty. But Abadie et al. (2020) clarify in their setting the expectation of the bootstrapped variance equals the Eicker-Huber-White estimator.

# Bootstrap

Different approaches:

- ① “Pairs bootstrap” or “nonparametric bootstrap”:

Repeatedly sample (with replacement)  $N$  observations from data.

- ② “Parametric bootstrap”:

Keep the  $X$ s fixed, but generate a new dependent variable by resampling from the distribution of residuals  $\hat{\epsilon}$ . (Bad if there is heteroscedasticity).

- ③ “Wild bootstrap”:

Hold  $X$ s fixed, generate new depend. variable  $y_i = X_i' \hat{\beta} \pm \hat{\epsilon}_i$  with probability 1/2.

- ④ “Block bootstrap”:

If there are clusters in the data, you need to resample *whole clusters* (with replacement), to preserve the correlation structure. E.g. for wild bootstrap, all observations within a cluster get  $+\hat{\epsilon}$  or  $-\hat{\epsilon}$ .

X

# Plan for Today

- 1 Inference Principles: Introduction
- 2 Classic Approach: Analytic Standard Errors
- 3 Sampling- and Design-Based Uncertainty
- 4 Bootstrap
- 5 Randomisation Inference

# Randomisation Inference<sup>9</sup>

Long-known approach to **design-based uncertainty**:

- Under **sharp null hypothesis** (e.g.  $\theta = 0 \forall i$ ), we know  $Y_i^*(1)$  and  $Y_i^*(0)$  for all  $i$ .
- Can create many / all alternative assignments, given assignment mechanism, and recalculate  $\hat{\beta}$  or test statistic each time.
- Gives exact, finite sample distribution of  $\hat{\beta}$  or test statistic!
- No assumptions on disturbances!
- Downside: Allows to test sharp hypotheses only.

---

<sup>9</sup>Check Imbens and Rubin (2015), Chapter 5.

Questions?

## References

- 1 Abadie, Alberto, Susan Athey, Guido Imbens, and Jeffrey Wooldridge. 2020. "Sampling-Based versus Design-Based Uncertainty in Regression Analysis". *Econometrica* 88:1, 265–296.
- 2 Abadie, Alberto, Susan Athey, Guido Imbens, and Jeffrey Wooldridge. 2023. "When Should You Adjust Standard Errors for Clustering?". *Quarterly Journal of Economics* 138:1, 1–35.
- 3 Angrist, Joshua and Jörn-Steffen Pischke. 2009. "Mostly Harmless Econometrics". Princeton University Press.
- 4 Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan. 2004. "How Much Should We Trust Differences-In-Differences Estimates?". *The Quarterly Journal of Economics* 119:1, 249–275.
- 5 Hansen, Bruce E. 2022. "Econometrics". Princeton University Press.  
Downloaded at <https://www.ssc.wisc.edu/~bhansen/econometrics/Econometrics.pdf>.
- 6 Young, Alwyn. 2019. "Channelling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results." *The Quarterly Journal of Economics* 134, 557–598.

# Econometrics II

## Lecture 4: Experiments

Konrad Burchardi

Stockholm University

11th of April 2024

# Literature

- ① "Mostly Harmless Econometrics", Angrist and Pischke  
Chapter 2 [introduction]; Chapter 3.2.3 [bad controls ]
- ② "Causal Inference for Statistics, Social and Biomedical Sciences: An Introduction", Imbens and Rubin  
Chapter 4 [introduction]; Chapter 5 [Fisher inference]; Chapter 7.5 [controls]

All mistakes are mine.

# Plan for Today

- 1 Unbiased Estimation
- 2 Balance
- 3 Stratification/Paired Experiments
- 4 Power
- 5 Control/Bad Control
- 6 Attrition
- 7 Canonical Experimental Designs

# Unbiased Estimation

Why randomize?

⇒ Balanced distribution of potential outcomes.

Randomized experiment guarantees, by design, *ex ante*:

$$\mathbb{E}[Y_i(0)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 0] = 0$$

Therefore:

$$\begin{aligned}\mathbb{E}[Y_i(1)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 0] \\ &= \mathbb{E}[Y_i(1)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 1] + \mathbb{E}[Y_i(0)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 0] \\ &= \mathbb{E}[Y_i(1) - Y_i(0)|D_i = 1] \equiv ATT\end{aligned}$$

and in fact:

$$= \mathbb{E}[Y_i(1) - Y_i(0)] \equiv ATE(\text{'average treatment effect'})$$

# How to Randomize?

Why should we think about how to randomise?

Purely random treatment assignment:

- ① Suboptimal fractions of T/C.
  - ① Ex-ante random assignment: unnecessarily low **power**.
- ② Imbalanced distribution of potential outcomes across T/C.
  - ① Ex-post random assignment: problematic **causal inference**.
  - ② Ex-ante random assignment: unnecessarily low **power**.

How should we randomize optimally then?

# Set Fraction of T and C

**Assignment Mechanism:** Randomisation conditional on  $N_1$  and  $N_0$ .

Under this assignment mechanism:  $V(\hat{\beta}|N_0, N_1) = \frac{\sigma_1^2}{N_1} + \frac{\sigma_0^2}{N-N_1}$ .

How to minimise  $V(\hat{\beta})$ ?

With  $\sigma_0^2 = \sigma_1^2$ :  $N_1^* = N/2$ .

With  $\sigma_0^2 \neq \sigma_1^2$ : More observations for noisy outcome.

Budget & Costs: With costs  $c_0$  and  $c_1$ , and a fixed budget  $B$ ,  
then  $\min_{N, N_1} V(\hat{\beta})$  s.s.  $(N - N_1)c_0 + N_1c_1 \leq B$ .  
More observations for cheap outcome.

# How to Randomize?

Why should we think about how to randomise?

Purely random treatment assignment:

- ① Suboptimal fractions of T/C.
  - ① Ex-ante random assignment: unnecessarily low power.
- ② Imbalanced distribution of potential outcomes across T/C.
  - ① Ex-post random assignment: problematic causal inference.
  - ② Ex-ante random assignment: unnecessarily low power.

How should we randomize optimally then?

# Plan for Today

- 1 Unbiased Estimation
- 2 Balance
- 3 Stratification/Paired Experiments
- 4 Power
- 5 Control/Bad Control
- 6 Attrition
- 7 Canonical Experimental Designs

# Balance Tests: How to judge imbalances ex-post?

In practice, treatment arms are unlikely “balanced”.

Imbalances in potential outcomes show up in covariates!

## How to detect imbalances?

- Often t-test for each covariate shown.
  - Conceptually problematic.
  - Statistical significance is not what matters.<sup>1</sup>
- What is useful then?
  - Focus on **size of differences for covariates that impact outcome!**<sup>2</sup>
  - Estimate size of difference using same specification as for differences in outcomes.
  - To check randomisation was implemented correctly, use omnibus test.

---

<sup>1</sup>Altman (1985) notes that such tests amount to assessing the probability of something having occurred by chance when you know that it did occur by chance. “Such a procedure is clearly absurd”.

<sup>2</sup>Imbens and Rubin, 2015

# Forcing Balance

Without further information:

Randomisation conditional on  $N_1$  is best we can do to achieve  $(Y(0), Y(1)) \perp D$  in sample.

But generally have more information:

$$\mathbb{E}[Y_i(D_i)] = f(X_i^+, X_i^-, D_i) \quad (1)$$

where  $X_i^+$  are observable and  $X_i^-$  non-observable covariates.

Cannot force balance of potential outcomes, but...  
...can try to force balance of  $X_i^+$ . Reduce  $V(\hat{\beta})$ .

# Plan for Today

- 1 Unbiased Estimation
- 2 Balance
- 3 Stratification/Paired Experiments
- 4 Power
- 5 Control/Bad Control
- 6 Attrition
- 7 Canonical Experimental Designs

# Stratification: Simple Example

**Idea:** Do not leave imbalance of important covariates to chance.

**Formally:** Restrict assignment mechanisms.

Example 1: Potential outcomes determines as  $Y_i(D_i) = g_i + 1 \times D_i$ , where  $g_i \in \{1, 2\}$  is only covariate,  $N = 4$ ,  $N_1 = 2$  and  $i = 1, 2$  only have  $g_i = 1$ . Note: true  $\beta = 1$ .

Assignment	$d_1$	$d_2$	$d_3$	$d_4$	$\hat{\beta}$
#1	0	1	0	1	1
#2	0	1	1	0	1
#3	1	0	0	1	1
#4	1	0	1	0	1
#5	1	1	0	0	0
#6	0	0	1	1	2

$\Rightarrow V(\hat{\beta})$  lower by excluding last two assignments! **Stratification**.

# Stratification: Practicalities

Extends to several categorical variables. **How?**

⇒ In Problem Set 2 you are asked to simulate effect of stratification on the variance of the estimator.

**Implementation in STATA:**

```
set seed 20230323
gen random = runiform()
sort cat_var1 cat_var2 ... random
gen treatment = mod(_n,2)
```

# Stratification: Practicalities

**Question 1:** Which variables should we stratify on?

General recommendation: covariates strongly related to the outcome.<sup>3</sup> **Why?**

**Question 2:** And when covariates are continuous, or cells sparse?

---

<sup>3</sup>See Bruhn and McKenzie, 2009; Glennerster and Takavarasha, 2013.

# Stratification: Implementations

**What do people do?** (Bruhn and McKenzie, 2009)

- ① Pure Randomisation.
- ② Re-Randomisation.
  - Subjectively decide whether to make another draw; or
  - Re-randomise until some statistic of balance is achieved: or
  - Choose assignment with best balance amongst N draws.
- ③ Matched-Pair ('blocking'): stratified randomisation with 2 units in each stratum.

**Recent insight:**

Matched-Pair designs optimal (under some conditions, in some sense).

# Stratification: Matched-Pair Designs

**Matched-Pair Design** is stratified randomisation with two units in each stratum.

1 Bai (AER, 2022):

*"Optimality of Matched-Pair Designs in Randomized Controlled Trials"*

- Shows optimality (in MSE sense) of a specific matched-pair design: Calculate  $\mathbb{E}[Y_i(1)|X_i] + \mathbb{E}[Y_i(0)|X_i]$ , match adjacent units on that “simple” scalar function.
- Problem: we do not know that sum! Can be estimated using pilot data with large class of estimators, including machine learning techniques.

2 Barrios (2015) – special case:

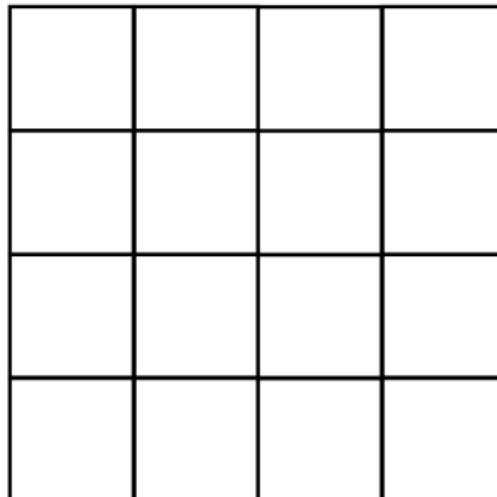
*"Optimal Stratification in Randomized Experiments"*

- With homogeneous treatment effects, best way to choose matched-pairs is to match on  $\mathbb{E}[Y_i(0)|X_i]$ .
- Baseline, but no pilot data needed.

# Stratification: Taking it to the Extreme

**Next level:** No randomization.

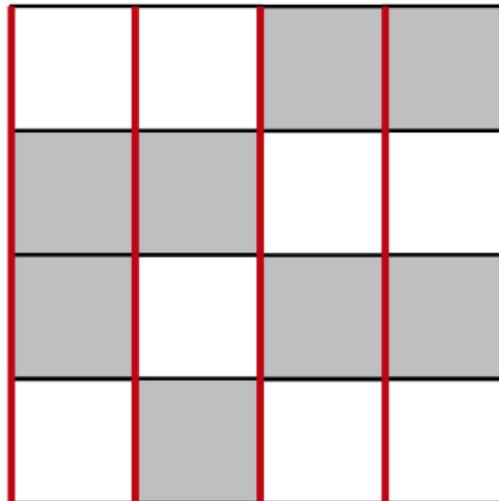
Consider this example, a 'field'. *What is optimal?*



# Stratification: Taking it to the Extreme

**Next level:** No randomization.

Consider this example, a 'field'. *What is optimal?*

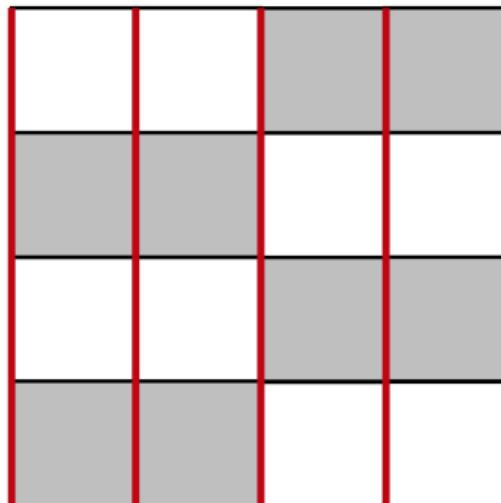


Stratification by  $x_1$ ? Potentially unbalanced marginal distribution of  $x_2$ .

# Stratification: Taking it to the Extreme

**Next level:** No randomization.

Consider this example, a 'field'. *What is optimal?*

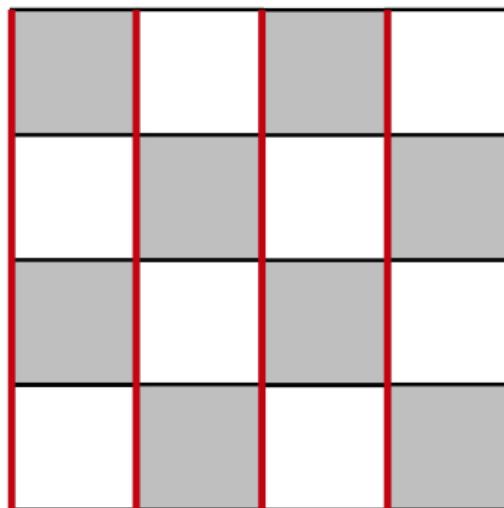


Something like this? Not maximizing power.

# Stratification: Taking it to the Extreme

Next level: No randomization.

Consider this example, a 'field'. *What is optimal?*



Balance joint distribution of  $x_1, x_2$ .

But only two assignments achieve that! Randomization?

# Stratification: Taking it to the Extreme

*Given any prior on how potential outcomes are generated...*

...argument is very general! (Kasy, 2016)

- For any assignment  $\mathbf{D}$ , calculate loss function (MSE, V, ...).
- Expected loss is average across potential assignments.
- Find subset of assignments such that loss is minimised.
- Typically those are two.

Controlled Trial, not Randomized Controlled Trial

- Stratification special case where subset larger than 2.

## Benefits

**Notice:** Stratification has benefits ex-post, and ex-ante!

**After Kasy:** *Why do we randomize then?*

- Optimal solution might be hard to find.
- Might set up the decision problem differently.
- Randomization Inference!

# Plan for Today

- 1 Unbiased Estimation
- 2 Balance
- 3 Stratification/Paired Experiments
- 4 Power
- 5 Control/Bad Control
- 6 Attrition
- 7 Canonical Experimental Designs

# Power: The Concept

Talked about experimental design choices that lead to lower variance of the estimator. Closely related concept: **Power**.

- ① Eventually want to test hypothesis  $H_0$  vs alternative  $H_1$ .
- ② Two types of errors we can make:
  - ① Type I error: reject  $H_0$  when  $H_0$  is true.  
Probability of type I error chosen by setting  $\alpha$ .
  - ② Type II error: fail to reject  $H_0$  when  $H_0$  is false.  
Probability of type II error depends on true effect size, experimental set-up and estimator/test statistics..

# Power: How to increase it?

*How to increase power?*

- ① Reduce the variance of the estimator:
  - Stratification (see before).
  - Baseline Controls (see later).
  - Sample Size.<sup>4</sup>
  - Measurement!
- ② Choice of test statistic.

---

<sup>4</sup>When starting a project, make sure it is ‘powered’ to detect expected effects with reasonably high probability. Funders want to see this in grant applications.

# Plan for Today

- 1 Unbiased Estimation
- 2 Balance
- 3 Stratification/Paired Experiments
- 4 Power
- 5 Control/Bad Control
- 6 Attrition
- 7 Canonical Experimental Designs

## Analysis: Control Variables

- Control variables: pre-determined, observed variables.
- If simple random assignment: no need to include control to guarantee **unbiasedness**...
- ... but might make sense to increase **power**.
- If data is generated by *conditional random assignment*, need to hold variables you conditioned on in randomisation constant in the analysis. (Only *conditional* independence assumption will surely hold, will come back to this.)

## Analysis: Bad Control

Control variables can not be outcomes themselves!

Mathematical reason:

$$\mathbb{E}[Y_i(0)|D_i = 1] = \mathbb{E}[Y_i(0)|D_i = 0]$$

does not imply

$$\mathbb{E}[Y_i(0)|D_i = 1, X_i = x] = \mathbb{E}[Y_i(0)|D_i = 0, X_i = x]$$

where  $X_i$  refers to the (observed) outcomes for the covariate for unit  $i$ . Such an  $X_i$  would be a **bad control**.

## Analysis: Bad Control

Example: Think of  $Y$  as achievement,  $D$  as class size (0 if large/1 if small), and  $X$  as parental help (1 if help/0 if not).  $D_i$  assigned by classical randomized experiment.

Plausibly parents' help responds to variation in  $D$ , so  $X$  is an outcome. Define  $X_i$  as potential outcome analogous to how we defined  $Y_i$ .

Suppose 'control for  $X$ ' in difference in outcomes:

$$\begin{aligned} & \mathbb{E}[Y_i|D_i = 1, X_i = 1] - \mathbb{E}[Y_i|D_i = 0, X_i = 1] \\ &= \mathbb{E}[Y_i(1)|D_i = 1, X_i = 1] - \mathbb{E}[Y_i(0)|D_i = 0, X_i = 1] \\ &= \mathbb{E}[Y_i(1)|X_i(1) = 1] - \mathbb{E}[Y_i(0)|X_i(0) = 1] \quad | \quad \text{by Random Assignment} \\ &= \underbrace{\mathbb{E}[Y_i(1) - Y_i(0)|X_i(1) = 1]}_{\text{'some' causal effect}} + \underbrace{\mathbb{E}[Y_i(0)|X_i(1) = 1] - \mathbb{E}[Y_i(0)|X_i(0) = 1]}_{\text{selection bias}} \end{aligned}$$

## Analysis: Bad Control

$$\underbrace{\mathbb{E}[Y_i(1) - Y_i(0)|X_i(1) = 1]}_{\text{'some' causal effect}} + \underbrace{\mathbb{E}[Y_i(0)|X_i(1) = 1] - \mathbb{E}[Y_i(0)|X_i(0) = 1]}_{\text{selection bias}}$$

**Causal effect:** for population that helps when class size is small.

**Compositional bias:** control sample *should be* those that help when class size is small, but *it is* those that help when class size is large - who are probably more than those who help when size is small. Probably special group!

Same argument applies *generally* when restricting to subgroup with some outcome.  
Example: effect of training on wages – which are only observed for employed.

Extensive margin: easy; intensive margin: infeasible.

## Analysis: Estimate only reduced form of effects

Example: Educational production function  $Y = f(D, X, \theta)$ . Suppose that  $X$  responds to  $D$ .

Comparing outcomes across treatment arms, we estimate:

$$\Delta Y = \frac{\partial f}{\partial D} \Delta D + \frac{\partial f}{\partial X} \frac{\partial X}{\partial D} \Delta D \quad (2)$$

- Sometimes that is what you are interested in.
- But for other questions you might need  $\frac{\partial f}{\partial D}$  or  $\frac{\partial f}{\partial X}$ .

P. Fredriksson: “One reason for the slight dismay in certain quarters over the experimental approach.”

# Plan for Today

- 1 Unbiased Estimation
- 2 Balance
- 3 Stratification/Paired Experiments
- 4 Power
- 5 Control/Bad Control
- 6 Attrition
- 7 Canonical Experimental Designs

# Attrition: Practical Advise

Attrition is common in experiments:

cannot obtain follow-up data for some treated and some un-treated observations.

- Generally good to avoid!
- Unproblematic when attrition is unrelated to potential outcomes.

Unfortunately impossible to know!

- Commonly accepted test: attrition rates unrelated to treatment.

Might still be that attrition is selected differently in T and C.

- When attrition rates differ across T and C:

Some bounding exercise, commonly Lee (2009) bounds.

## Attrition: Lee (2009) Bounds

Suppose the attrition rate is higher in C than in T.

**Worry:** Those with 'best'  $Y_i(0)$  cannot be observed in Control. Comparison of T with C exaggerates ATE.

**Lee Bounds** are an extreme bounding exercise:

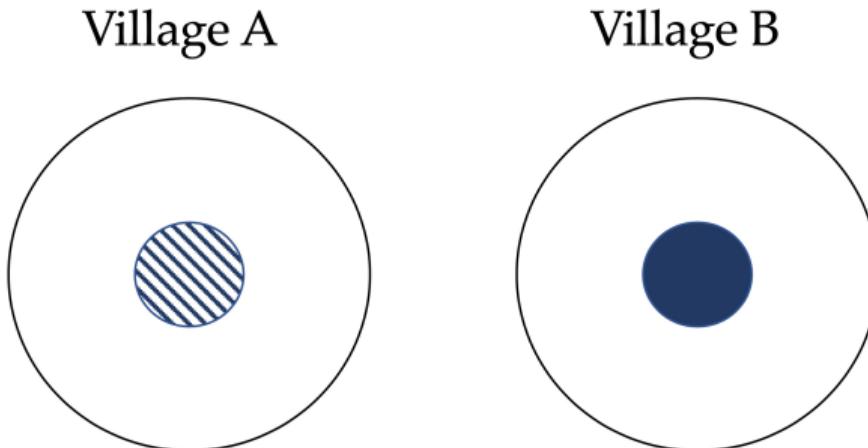
- Drop the 'best' from T, s.t. T and C have same attr. rate, and rerun analysis.
- Drop the 'worst' from T, s.t. T and C have same attr. rate, and rerun analysis.
- Report results from both exercises as bounds.

# Plan for Today

- 1 Unbiased Estimation
- 2 Balance
- 3 Stratification/Paired Experiments
- 4 Power
- 5 Control/Bad Control
- 6 Attrition
- 7 Canonical Experimental Designs

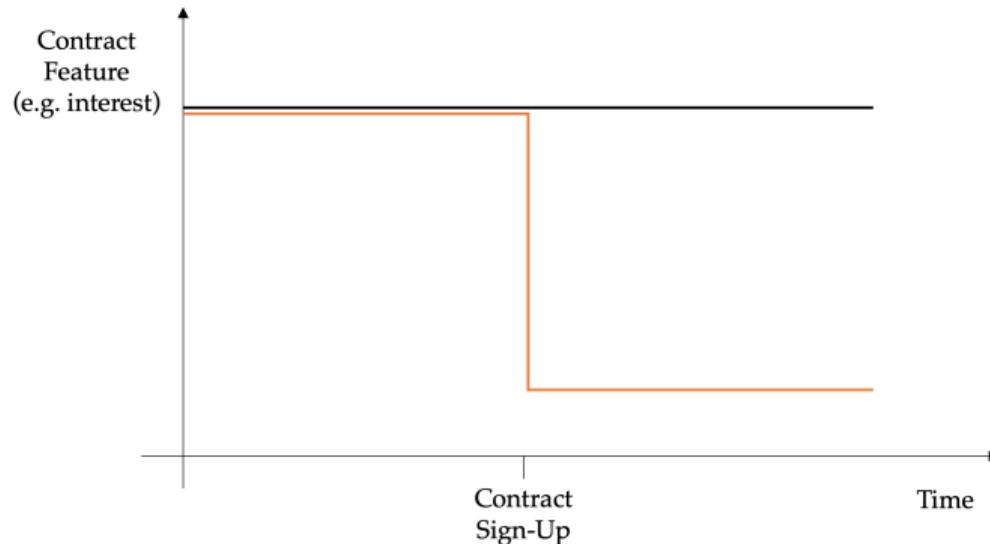
## Spill-Overs: *Miguel and Kremer (2004)*

Suppose you believe your treatment might affect other units, i.e. has **spill-over effects**. How can you estimate those?



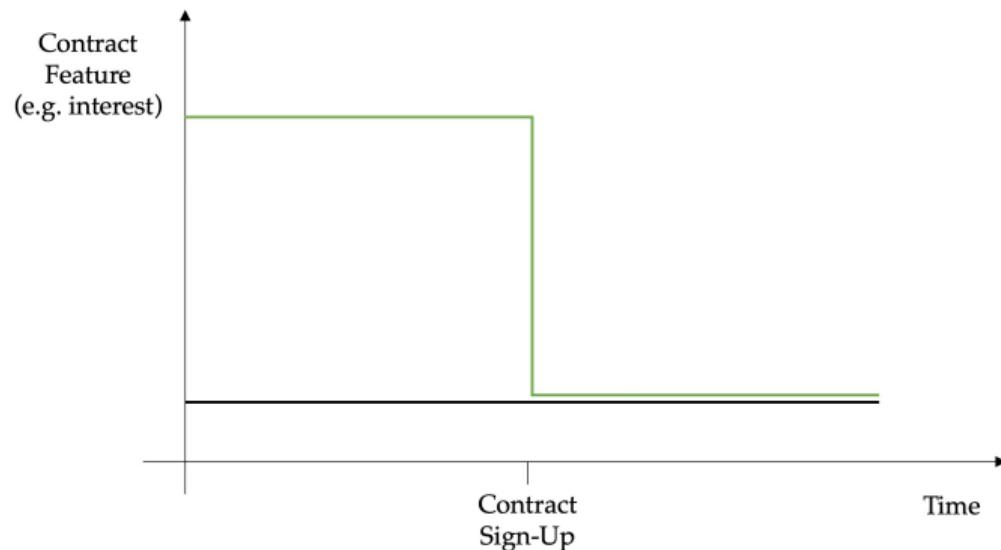
# Moral Hazard and Adverse Selection: *Karlan and Zinman (2009)*

Distinguishing **Moral Hazard** and **Adverse Selection** *empirically* is generally hard.  
Karlan and Zinman figured out one way:



# Moral Hazard and Adverse Selection: *Karlan and Zinman (2009)*

Distinguishing **Moral Hazard** and **Adverse Selection** empirically is generally hard.  
Karlan and Zinman figured out one way:



# Randomized Experiments

- Experiments important to understand on their own right.  
Today just a small introduction.
- Modern econometric approaches approximate that ideal.
- The experimental ideal will often be very useful to think in *observational studies* whether you estimate causal effects.
- Issues we discussed will come up, more or less explicitly: assignment mechanism, conditional independence, balance, biases, reduced form estimation, average treatment effects for subpopulations...

Questions?

# References

- ① Angrist and Pischke (2008): Chapter 2 [introduction]; Chapter 3.2.3 [bad controls]
- ② Imbens and Rubin (2015): Chapter 4 [introduction]; Chapter 5 [Fisher inference], Chapter 7.5 [controls]
- ③ Kasy, Maximilian (2016) "Why Experimenters Might Not Always Want to Randomize, and What They Could Do Instead" *Political Analysis*: 1-15. [stratification]
- ④ Lee, D. S. (2009) "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects". *Review of Economic Studies* 76: 1071–1102. [attrition]
- ⑤ Bai, Yuehao (2022) "Optimality of Matched-Pair Designs in Randomized Controlled Trials". Working Paper. [matched-pair designs]
- ⑥ Barrios, Thomas (2014) "Optimal Stratification in Randomized Experiments". Working Paper. [matched-pair design]
- ⑦ Bruhn, M. and D. McKenzie (2009) "In Pursuit of Balance: Randomization in Practice in Development Field Experiments". *American Economic Journal: Applied Economics* 1:4, 200-232.
- ⑧ Karlan, D. and J. Zinman (2009) "Observing Unobservables: Identifying Information Asymmetries With a Consumer Credit Field Experiment". *Econometrica* 77:6, 1993-2008.
- ⑨ Miguel and Kremer (2004) "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities". *72:1*, 159-217.

# Econometrics II

## Lecture 5: Matching Estimators

Konrad Burchardi

Stockholm University

16th of April 2024

# Literature

- ① "**Causal Inference for Statistics, Social and Biomedical Sciences: An Introduction**", Imbens and Rubin  
Chapters 12.1-12.3
- ② "**Mostly Harmless Econometrics**", Angrist and Pischke  
Chapter 3.3.1-3.3.3

These notes draw on those books. All mistakes are mine.

# Plan for Today

1 Conditional Independence Assumption and Balance

2 Propensity Score

3 Achieving Balance

    Matching

    Propensity Score Matching

    Reweighting

4 Matching and Regressions

5 Practical Issues

# Conditional Independence

In Lecture 1 we discussed that we require

$$\mathbb{E}[Y_i(0)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 0] = 0$$

for causal inference about ATT. This is fundamentally not testable.

Maybe plausible:  $\{Y_i(0), Y_i(1)\} \perp D_i | X_i$ , and  $X_i$  are observable (CIA).<sup>1</sup>

Conditional Independence Assumption implies:

$$\begin{aligned}\mathbb{E}[Y|D = 1, X = x] - \mathbb{E}[Y|D = 0, X = x] \\ = \mathbb{E}[Y(1) - Y(0)|X = x] \equiv \tau_x\end{aligned}$$

Suggests one could condition analysis on  $X$ . But also...

---

<sup>1</sup> Fundamentally not testable either! Requires substantial information.

# Why Covariate Balance?

...when does a simple difference in means comparison give  $\mathbb{E}_X[\tau_x]$ ?

$$\begin{aligned} & \mathbb{E}[Y|D=1] - \mathbb{E}[Y|D=0] \\ = & \int \mathbb{E}[Y|D=1, X=x] f_{X|D=1} dx - \int \mathbb{E}[Y|D=0, X=x] f_{X|D=0} dx \quad [\text{LIE}] \\ = & \int (\mathbb{E}[Y|D=1, X=x] - \mathbb{E}[Y|D=0, X=x]) f_X dx \quad [\text{If } f_{X|D=1} = f_{X|D=0}!] \\ = & \mathbb{E}_X[\tau_x] \end{aligned}$$

Balanced covariate distribution is important!

# Why Covariate Balance?

**Today:**

Check whether (confounding) covariates  $X_i$  are balanced.

Exactly analogous to balance test of randomised experiment.

Make sure the (confounding) covariates  $X_i$  are balanced.

Exactly analogous to stratification in randomised experiment.

# Plan for Today

1 Conditional Independence Assumption and Balance

2 Propensity Score

3 Achieving Balance

    Matching

    Propensity Score Matching

    Reweighting

4 Matching and Regressions

5 Practical Issues

# Propensity Score

Building Intuition, Step 1:

*Why might imbalances in covariates be problematic?*

Imagine for some  $x_1 \neq x_2$

$$\{Y_i(0), Y_i(1)\} | X_i = x_1 \not\sim \{Y_i(0), Y_i(1)\} | X_i = x_2,$$

and we are simply comparing outcomes of treated and control observations.

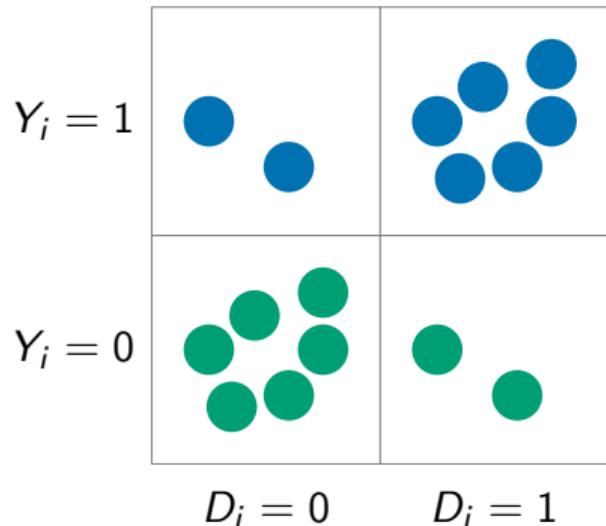
*When will this yield an unbiased estimate of  $\mathbb{E}_x[\tau_x]$ ?*

# Propensity Score

A simple example:

- $x_1 = 0, x_2 = 1$
- $(Y_i(0)|X_i = 0) = (Y_i(1)|X_i = 0) = 0$
- $(Y_i(0)|X_i = 1) = (Y_i(1)|X_i = 1) = 1$

Suppose share treated at  $X_i = 0$  is 0.25 and at  $X_i = 1$  it is 0.75.



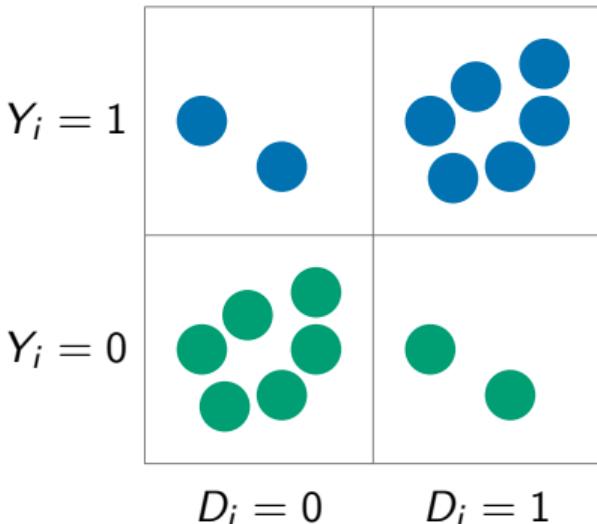
# Propensity Score

A simple example:

- $x_1 = 0, x_2 = 1$
- $(Y_i(0)|X_i = 0) = (Y_i(1)|X_i = 0) = 0$
- $(Y_i(0)|X_i = 1) = (Y_i(1)|X_i = 1) = 1$

Suppose share treated at  $X_i = 0$  is 0.25 and at  $X_i = 1$  it is 0.75.

Find positive treatment effect!



# Propensity Score

A simple example:

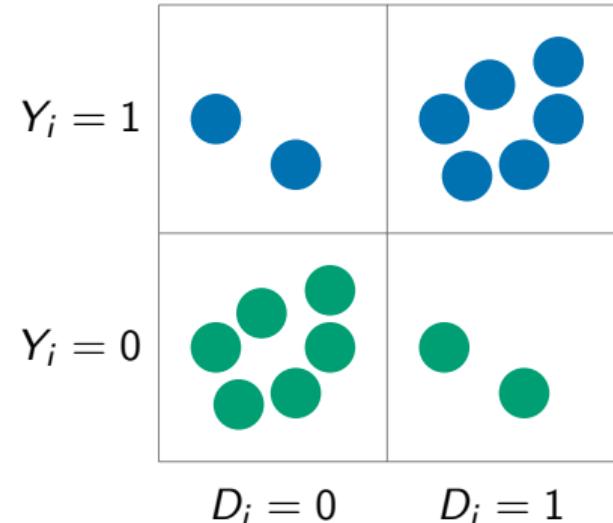
- $x_1 = 0, x_2 = 1$
- $(Y_i(0)|X_i = 0) = (Y_i(1)|X_i = 0) = 0$
- $(Y_i(0)|X_i = 1) = (Y_i(1)|X_i = 1) = 1$

Suppose share treated at  $X_i = 0$  is 0.25 and at  $X_i = 1$  it is 0.75.

Find positive treatment effect!

The problem, somehow:

Observations with different  $(Y_i(0), Y_i(1))$  are assigned to treatment at different frequencies.



# Propensity Score

A simple example:

- $x_1 = 0, x_2 = 1$
- $(Y_i(0)|X_i = 0) = (Y_i(1)|X_i = 0) = 0$
- $(Y_i(0)|X_i = 1) = (Y_i(1)|X_i = 1) = 1$

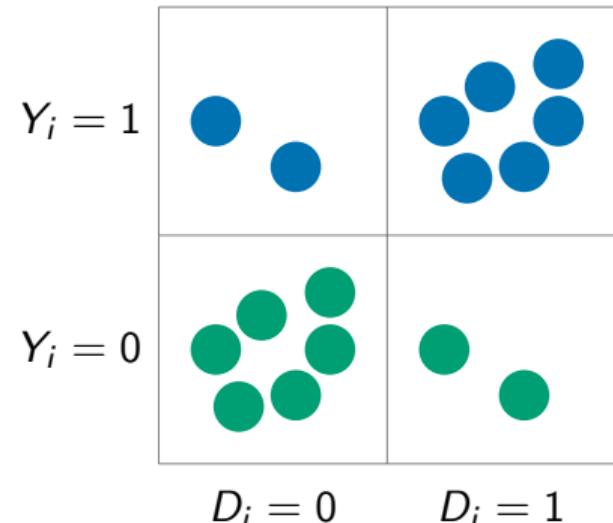
Suppose share treated at  $X_i = 0$  is 0.25 and at  $X_i = 1$  it is 0.75.

Find positive treatment effect!

The problem, somehow:

Observations with different  $(Y_i(0), Y_i(1))$  are assigned to treatment at different frequencies.

Note: Find imbalanced covariate distributions.



# Propensity Score

Building Intuition, Step 2:

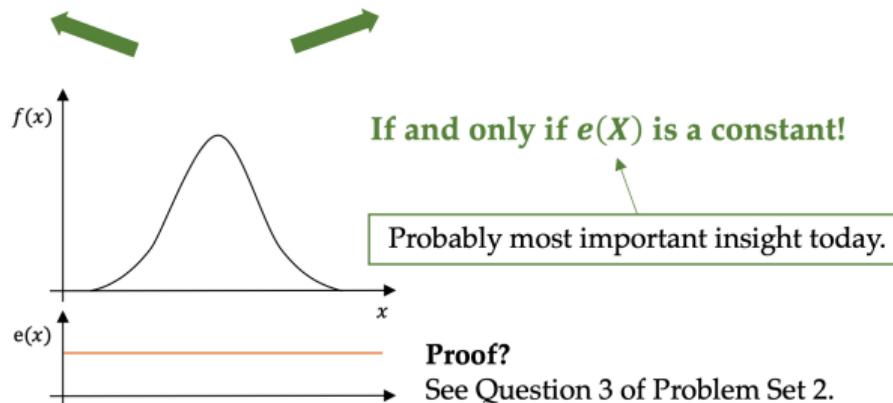
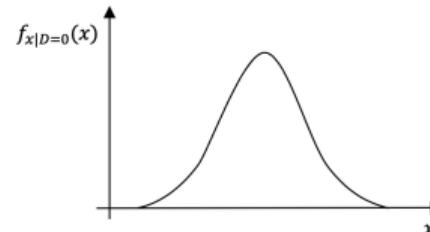
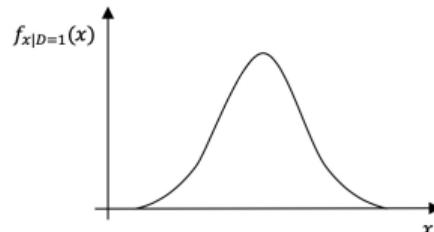
*How might imbalances in covariates occur?*

Denote with  $e(x) \equiv \Pr(D = 1|X = x)$  the probability of treatment given  $x$ .  
We call this the '**propensity score**'.

*Can it be that  $e(x_1) \neq e(x_2)$  for some  $x_1$  and  $x_2$ , yet  $f_{X|D=1}(x) = f_{X|D=0}(x)$  for all  $x$ ?*

# Propensity Score

When will the covariate distribution be balanced across t/c?

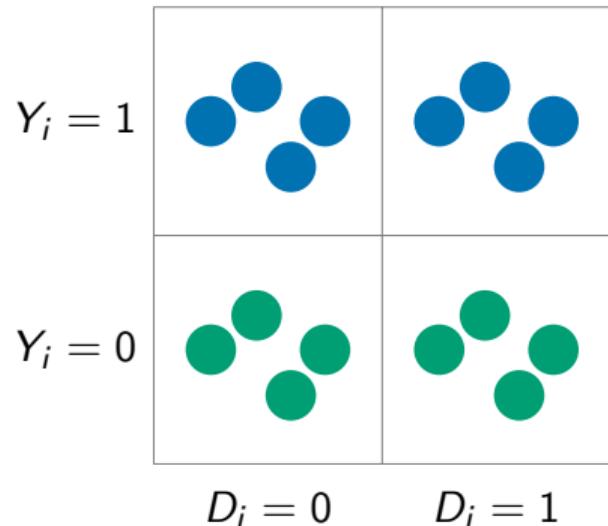


# Propensity Score

A simple example:

- $x_1 = 0, x_2 = 1$
- $(Y_i(0)|X_i = 0) = (Y_i(1)|X_i = 0) = 0$
- $(Y_i(0)|X_i = 1) = (Y_i(1)|X_i = 1) = 1$

Suppose  $e(0) = 0.5$  and  $e(1) = 0.5$ .



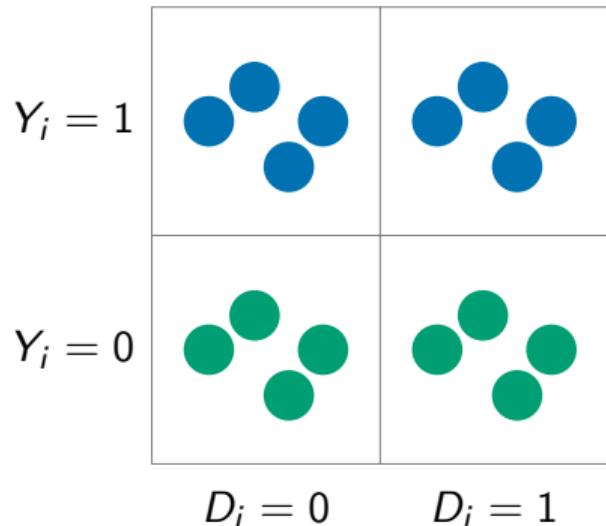
# Propensity Score

A simple example:

- $x_1 = 0, x_2 = 1$
- $(Y_i(0)|X_i = 0) = (Y_i(1)|X_i = 0) = 0$
- $(Y_i(0)|X_i = 1) = (Y_i(1)|X_i = 1) = 1$

Suppose  $e(0) = 0.5$  and  $e(1) = 0.5$ .

Find no treatment effect!



# Propensity Score

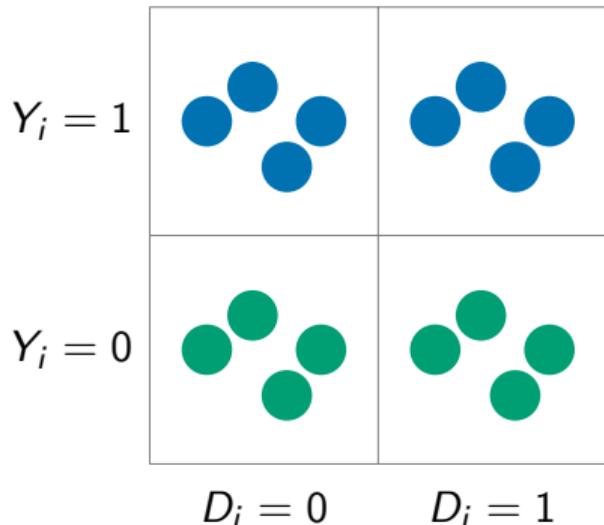
A simple example:

- $x_1 = 0, x_2 = 1$
- $(Y_i(0)|X_i = 0) = (Y_i(1)|X_i = 0) = 0$
- $(Y_i(0)|X_i = 1) = (Y_i(1)|X_i = 1) = 1$

Suppose  $e(0) = 0.5$  and  $e(1) = 0.5$ .

Find no treatment effect!

Find no imbalances in covariate distributions!



# Propensity Score

**Note:** the source of the problem is *not*...

- ... that distribution of outcomes is different at different  $X$ .
- ... that  $e(X_i) \neq 0.5$ .

**Two diagnostics:**

- $X_i = x$  is differently frequent in different treatment arms.

Motivation for standard balance checks.

- $e(X_i)$  not independent of  $x$ .

**Result:** In fact, these two diagnostics are equivalent!<sup>2</sup>

---

<sup>2</sup>Probably most important insight today.

# Plan for Today

1 Conditional Independence Assumption and Balance

2 Propensity Score

3 Achieving Balance

    Matching

    Propensity Score Matching

    Reweighting

4 Matching and Regressions

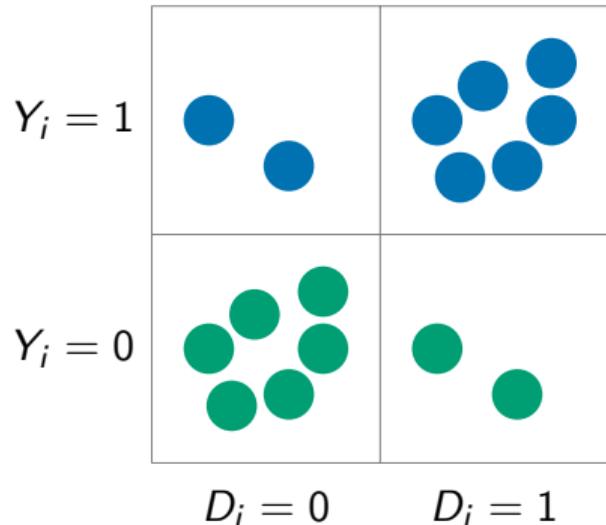
5 Practical Issues

# Achieving Balance

**Solution 1, Matching:**

Control for influence of  $X$  by  
conditioning analysis on  $X$ , find  $\tau_X$ .

You see why that solves the problem?



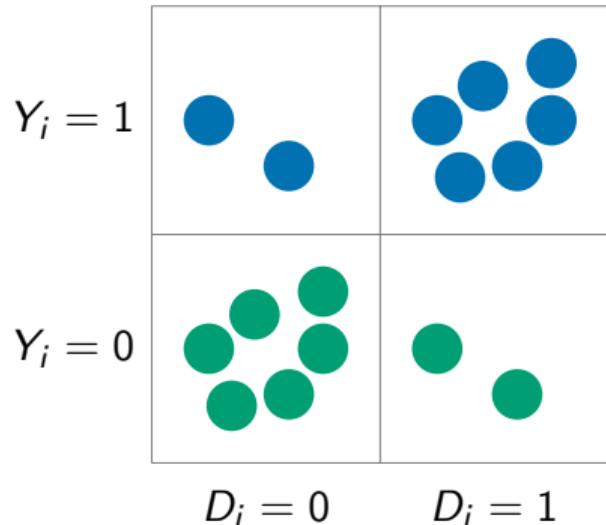
# Achieving Balance

**Solution 1, Matching:**

Control for influence of  $X$  by  
conditioning analysis on  $X$ , find  $\tau_X$ .

You see why that solves the problem?

Fundamental reason why this works:  
Within each  $X$  cell,  $e(X)$  is constant.



# Achieving Balance

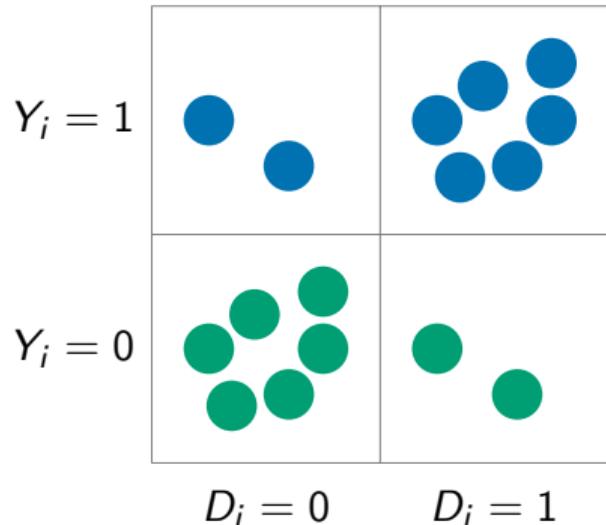
**Solution 1, Matching:**

Control for influence of  $X$  by  
conditioning analysis on  $X$ , find  $\tau_X$ .

You see why that solves the problem?

Fundamental reason why this works:  
Within each  $X$  cell,  $e(X)$  is constant.

(Typically: Average  $\tau_X$ 's by weights of interest.)



# Achieving Balance

Take a just slightly more complex example:

- $X$  has support  $\{x_1, x_2, x_3, x_4\}$ ;
- Distribution of potential outcomes is different;
- $e(x_1) = 0.5, e(x_2) = 0.25, e(x_3) = 0.5, e(x_4) = 0.5$ .

**Solution 2, Balancing Score Matching:**  $D_i \perp X_i | b(X_i)$ .

# Achieving Balance

**Solution 3, Propensity Score Matching:** Match on  $e(X)$ .<sup>3</sup>

Recall from before...

- If  $e(X)$  is constant, distribution of  $X$  in the treatment arms is the same.
- And if distribution of  $X$  is the same, simple difference in outcomes estimates the average treatment effect, by CIA!

Note:

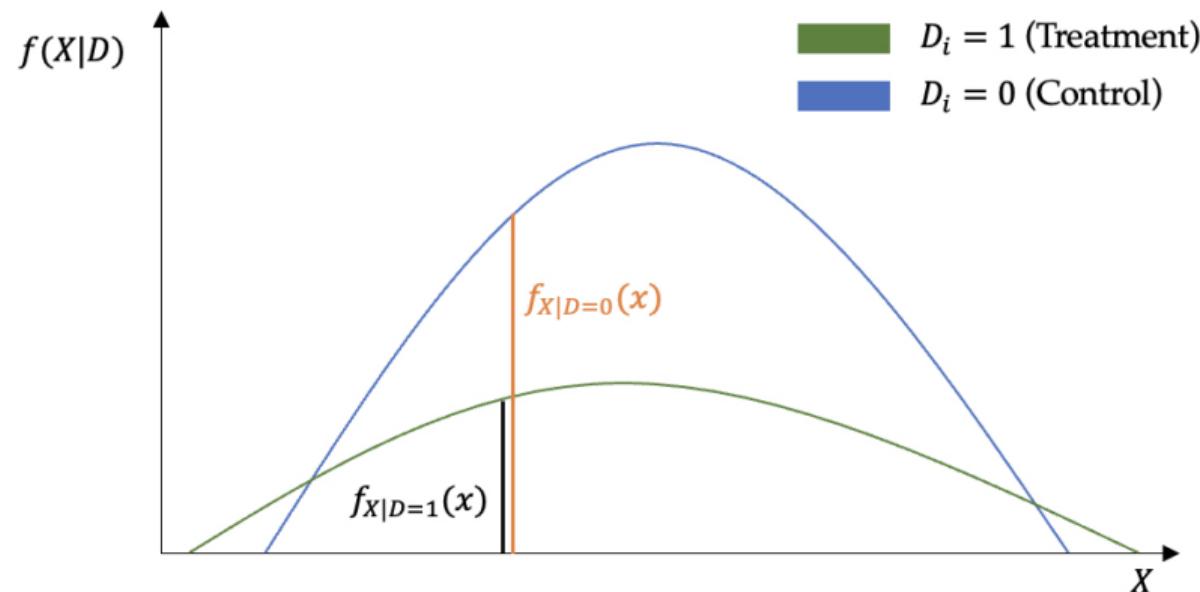
- This is precisely why **randomised trials** are useful, they achieve constant  $e(X)$ .
- Stratification is about forcing realised  $\hat{e}(X)$  to be constant.  
**Like with RCTs, realised  $\hat{e}(X)$  is what matters.** *You see why?*

---

<sup>3</sup>Propensity Score is the coarsest balancing score.

# Achieving Balance

Solution 4, Reweighting (Horvitz and Thomson):

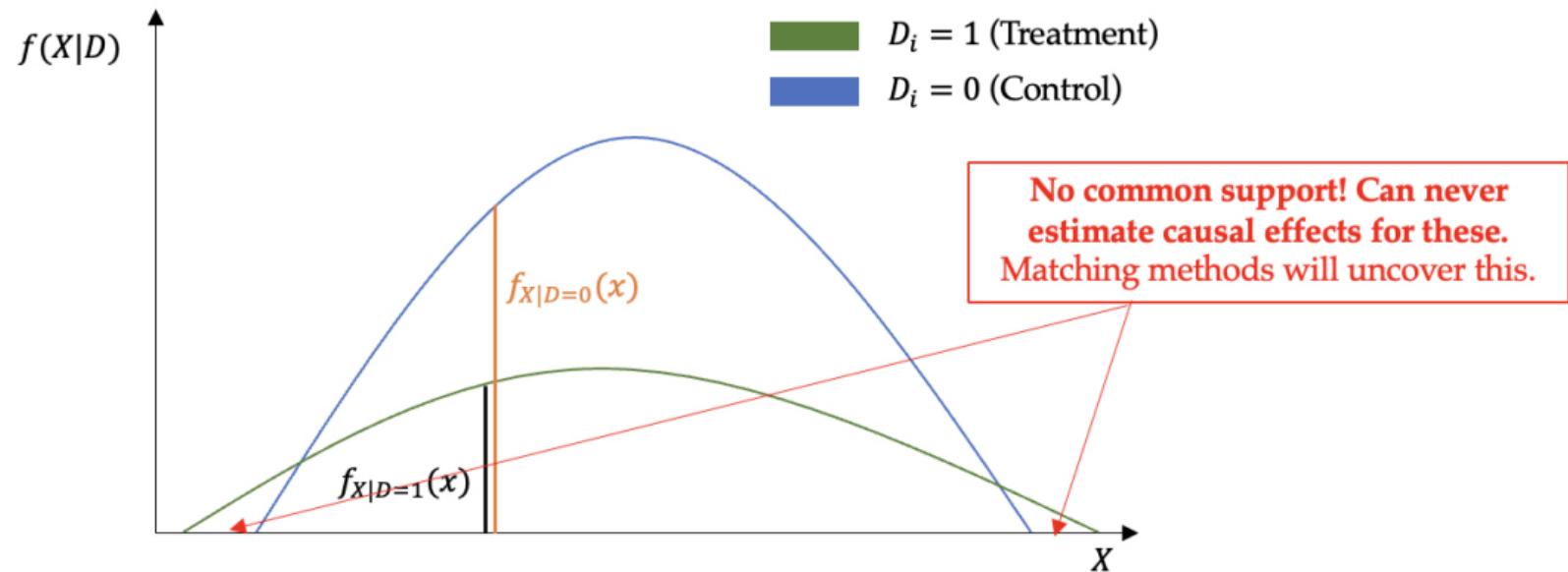


Idea: Find set of weights  $\omega(x)$  for all  $x$  such that

$$\omega(x) f_{X|D=0}(x) = f_{X|D=1}(x).$$

# Achieving Balance

Solution 4, Reweighting (Horvitz and Thomson):



Idea: Find set of weights  $\omega(x)$  for all  $x$  such that

$$\omega(x) f_{X|D=0}(x) = f_{X|D=1}(x).$$

# Achieving Balance

By Bayes' rule (see Angrist and Pischke, ch. 3.3.1):

$$f_{X|D=1}(x) = \frac{Pr(D_i = 1|X_i = x)}{Pr(D_i = 1)} f_X(x) = \frac{e(x)}{e} f_X(x)$$

$$f_{X|D=0}(x) = \frac{Pr(D_i = 0|X_i = x)}{Pr(D_i = 0)} f_X(x) = \frac{1 - e(x)}{1 - e} f_X(x)$$

where  $e(x)$  is the propensity score and  $e = Pr(D_i = 1)$ . Then

$$\omega(x) = \frac{f_{X|D=1}(x)}{f_{X|D=0}(x)} = \frac{e(x)}{1 - e(x)} \frac{1 - e}{e} \quad (1)$$

Then the ATT can be estimated as

$$\hat{\beta}^{ATT} = \frac{1}{N_1} \sum_{i=1}^N D_i Y_i - \frac{1}{N_0} \sum_{i=1}^N (1 - D_i) Y_i \frac{\hat{e}_i(x)}{1 - \hat{e}_i(x)} \frac{1 - \hat{e}}{\hat{e}} \quad (2)$$

According to Imbens and Rubin, sensitive to the estimation of weights.

# Plan for Today

1 Conditional Independence Assumption and Balance

2 Propensity Score

3 Achieving Balance

    Matching

    Propensity Score Matching

    Reweighting

4 Matching and Regressions

5 Practical Issues

# Matching and Regressions

## Solution 5, Regression Control:

*Upside* (see Angrist and Pischke):

Running regression on treatment, fully saturated in  $X$ , have

$$\beta_R = \frac{\mathbb{E}[\sigma_D^2(X_i) \tau_X]}{\mathbb{E}[\sigma_D^2(X_i)]}, \text{ where } \sigma_D^2(X_i) \equiv \mathbb{E}[(D_i - \mathbb{E}[D_i|X_i])^2|X_i]$$

and a standard ATT matching estimator would be

$$\beta_M = \mathbb{E}[\tau_X Pr(X_i|D_i = 1)].$$

- ① Matching estimator weights high  $X$ s with many treatment observations; OLS gives weight to observations with equal treatment shares (max. variance in  $D_i$ ).
- ② If  $\tau_X$  varies little across  $X$  cells, makes little difference.
- ③ None gives any weight to  $X$  cells with no or all treatment.

# Matching and Regressions

## Solution 5, Regression Control:

*Upside* (see Angrist and Pischke):

Running regression on treatment, fully saturated in  $X$ , have

$$\beta_R = \frac{\mathbb{E}[\sigma_D^2(X_i) \tau_X]}{\mathbb{E}[\sigma_D^2(X_i)]}, \text{ where } \sigma_D^2(X_i) \equiv \mathbb{E}[(D_i - \mathbb{E}[D_i|X_i])^2|X_i]$$

and a standard ATT matching estimator would be

$$\beta_M = \mathbb{E}[\tau_X Pr(X_i|D_i = 1)].$$

- ① Matching estimator weights high  $X$ s with many treatment observations; OLS gives weight to observations with equal treatment shares (max. variance in  $D_i$ ).
- ② If  $\tau_X$  varies little across  $X$  cells, makes little difference.
- ③ None gives any weight to  $X$  cells with no or all treatment.

# Matching and Regressions

## Solution 5, Regression Control:

*Upside* (see Angrist and Pischke):

Running regression on treatment, fully saturated in  $X$ , have

$$\beta_R = \frac{\mathbb{E}[\sigma_D^2(X_i) \tau_X]}{\mathbb{E}[\sigma_D^2(X_i)]}, \text{ where } \sigma_D^2(X_i) \equiv \mathbb{E}[(D_i - \mathbb{E}[D_i|X_i])^2|X_i]$$

and a standard ATT matching estimator would be

$$\beta_M = \mathbb{E}[\tau_X Pr(X_i|D_i = 1)].$$

- ① Matching estimator weights high  $X$ s with many treatment observations; OLS gives weight to observations with equal treatment shares (max. variance in  $D_i$ ).
- ② If  $\tau_X$  varies little across  $X$  cells, makes little difference.
- ③ None gives any weight to  $X$  cells with no or all treatment.

# Matching and Regressions

## Solution 5, Regression Control:

*Upside* (see Angrist and Pischke):

Running regression on treatment, fully saturated in  $X$ , have

$$\beta_R = \frac{\mathbb{E}[\sigma_D^2(X_i) \tau_X]}{\mathbb{E}[\sigma_D^2(X_i)]}, \text{ where } \sigma_D^2(X_i) \equiv \mathbb{E}[(D_i - \mathbb{E}[D_i|X_i])^2|X_i]$$

and a standard ATT matching estimator would be

$$\beta_M = \mathbb{E}[\tau_X Pr(X_i|D_i = 1)].$$

- ① Matching estimator weights high  $X$ s with many treatment observations; OLS gives weight to observations with equal treatment shares (max. variance in  $D_i$ ).
- ② If  $\tau_X$  varies little across  $X$  cells, makes little difference.
- ③ None gives any weight to  $X$  cells with no or all treatment.

# Matching and Regressions

*Downside* (see Imbens and Rubin):

Their argument is that regression requires:

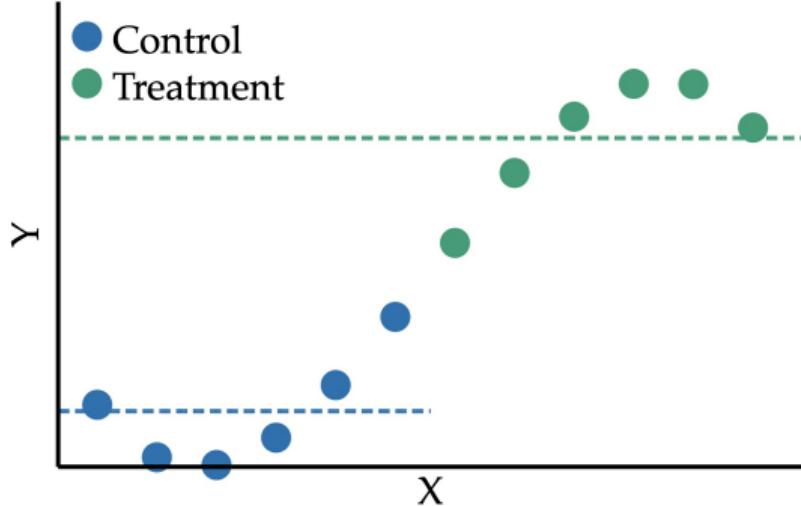
- ① unconfoundedness assumption,

and additionally:

- ② functional form assumptions for  $X$ .

Is **strong** and **unnecessary**.

# Matching and Regressions

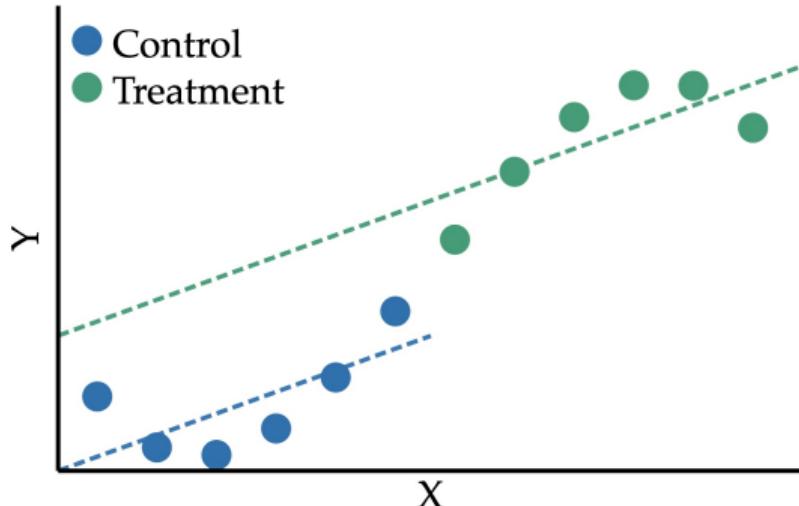


Suppose you run regression of  $Y$  on treatment status  $D$  and...

- **no control for  $X$ ,**
- linear control for  $X$ , or
- quadratic control for  $X$  interacted with  $D$ .

*What is the coefficient estimate on  $D$ ?*

# Matching and Regressions

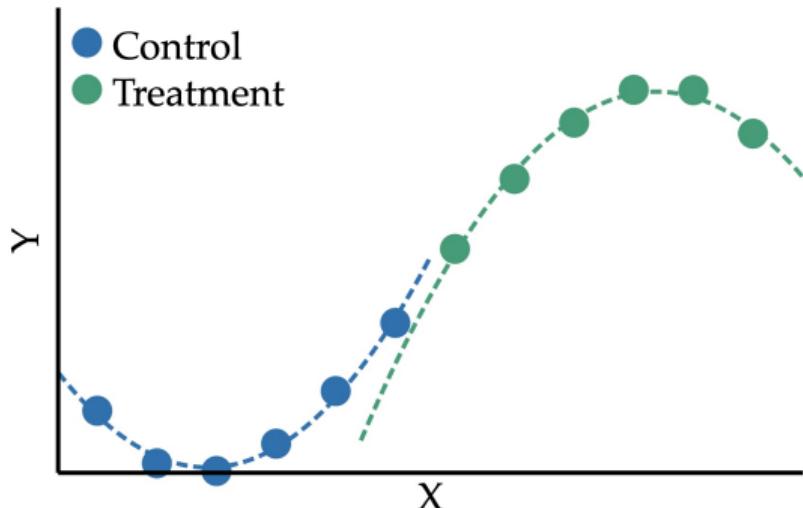


Suppose you run regression of  $Y$  on treatment status  $D$  and...

- no control for  $X$ ,
- **linear control for  $X$ , or**
- quadratic control for  $X$  interacted with  $D$ .

*What is the coefficient estimate on  $D$ ?*

# Matching and Regressions



Suppose you run regression of  $Y$  on treatment status  $D$  and...

- no control for  $X$ ,
- linear control for  $X$ , or
- **quadratic control for  $X$  interacted with  $D$ .**

*What is the coefficient estimate on  $D$ ?*

# Plan for Today

- 1 Conditional Independence Assumption and Balance
- 2 Propensity Score
- 3 Achieving Balance
  - Matching
  - Propensity Score Matching
  - Reweighting
- 4 Matching and Regressions
- 5 Practical Issues

# Practical Issues

*Distinguish:*

## ① Design Stage

### ① Assessing Overlap in Distributions

- How similar are distributions? [Univariate/Multivariate tests.]
- Do similar observations with opposite level of treatment exist?

### ② Estimate Propensity Score

- Estimated propensity score matters; goal is to achieve balance in sample.
- Machine learning methods can potentially be helpful.

### ③ Create Balanced Sample: unbiased/robust inference; power.

- Match estimation sample: 1 to 1; 1 to many; on PS or X...
- Trim sample. Two competing forces for power: sample size vs. match quality.

# Practical Issues

*Distinguish:*

## ① Design Stage

### ① Assessing Overlap in Distributions

- How similar are distributions? [Univariate/Multivariate tests.]
- Do similar observations with opposite level of treatment exist?

### ② Estimate Propensity Score

- Estimated propensity score matters; goal is to achieve balance in sample.
- Machine learning methods can potentially be helpful.

### ③ Create Balanced Sample: unbiased/robust inference; power.

- Match estimation sample: 1 to 1; 1 to many; on PS or  $X$ ...
- Trim sample. Two competing forces for power: sample size vs. match quality.

No need for any  $Y$  here!

Can go back and forth, “play around”, until balanced sample is found...

... just do not condition this on  $Y$  data.

# Practical Issues

## ② Assessment Stage

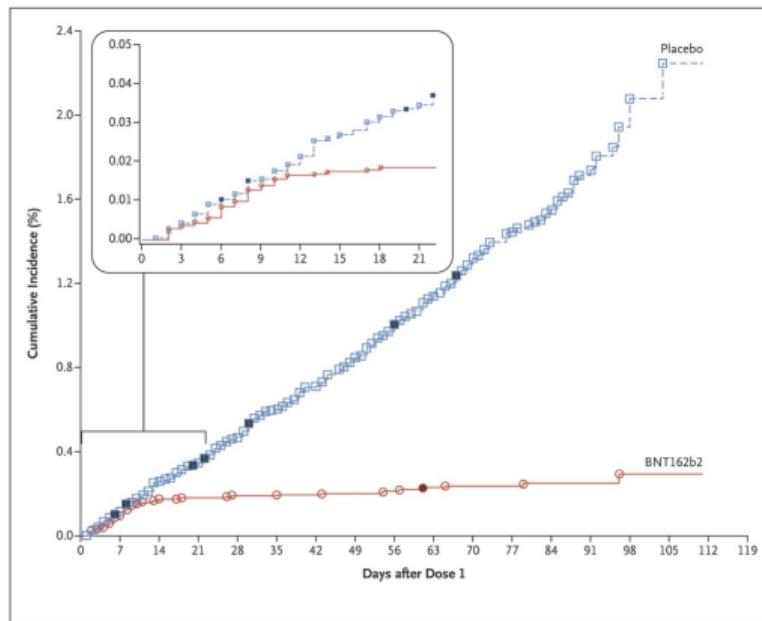
- Use pseudo outcomes to validate the approach (see example on next slide).

## ③ Analysis Stage

- Create PS blocks, or create exact matches, or inexact matches, and estimate mean outcome difference within those.
- Covariate adjustment might increase precision, see Imbens and Rubin, Part III.
- Inference easier with exact matches.

# Pseudo Outcomes: Example

Pfizer/BioNTech released results from an RCT in 12/2020 (NEJM,  $N = 43548$ )



# Pseudo Outcomes: Example

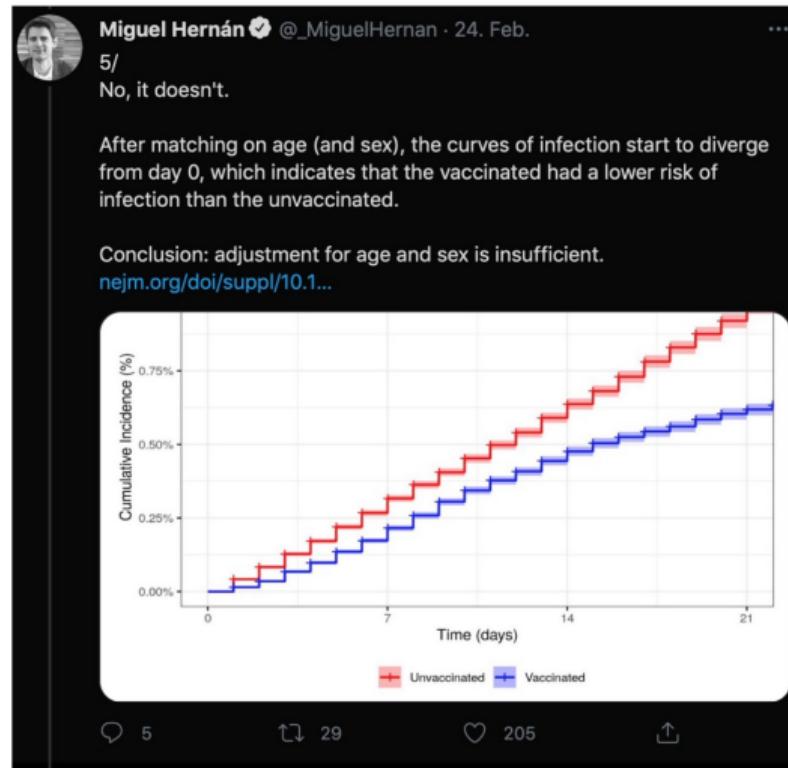
Hernán et al. wanted to run such analysis in non-RCT using large-N registry data from Israel. Here is how they tested their matching algorithm.

Miguel Hernán ✅ @\_MiguelHernan · 24. Feb. ...  
3/  
To adjust for confounding:  
  
We start by identifying potential confounders.  
  
For example: Age  
(vaccination campaigns prioritize older people and older people are more likely to develop severe disease)  
  
Then we choose a valid adjustment method. In our paper, we matched on age.

Miguel Hernán ✅ @\_MiguelHernan · 24. Feb. ...  
4/  
After age adjustment, how do we know if there is residual confounding?  
  
Here is one way to go about that:  
  
We know from the previous randomized trial that the vaccine has no effect in the first few days.  
  
So we check whether matching on age suffices to replicate that finding.

# Pseudo Outcomes: Example

Hernán et al. wanted to run such analysis in non-RCT using large-N registry data from Israel. Here is how they tested their matching algorithm.



# Pseudo Outcomes: Example

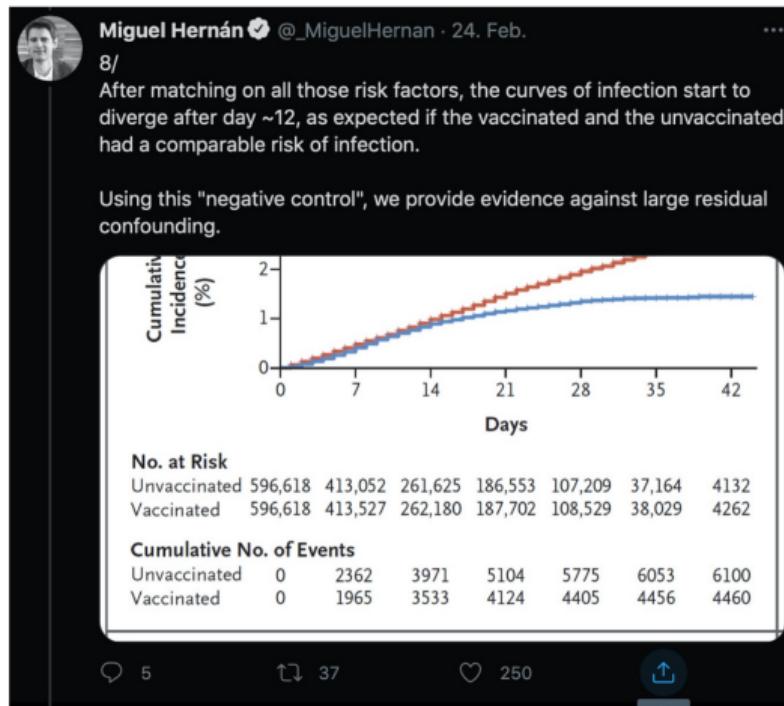
Hernán et al. wanted to run such analysis in non-RCT using large-N registry data from Israel. Here is how they tested their matching algorithm.

Miguel Hernán ✅ @\_MiguelHernan · 24. Feb. ...  
6/  
We learned that we had to match on other #COVID19 risk factors, e.g., location, comorbidities, healthcare use...  
  
And we could do so with high-quality data from @ClalitResearch, part of a health services organization that covers >50% of the Israeli population.  
  
As an example,

Miguel Hernán ✅ @\_MiguelHernan · 24. Feb. ...  
7/  
A vaccinated 76 year-old Arab male from a specific neighborhood who received 4 influenza vaccines in the last 5 years and had 2 comorbidities was matched with an unvaccinated Arab male from the same neighborhood, aged 76-77, with 3-4 influenza vaccines and 2 comorbidities.

# Pseudo Outcomes: Example

Hernán et al. wanted to run such analysis in non-RCT using large-N registry data from Israel. Here is how they tested their matching algorithm.



# Making Matching Persuasive

David McKenzie writes about statistical and "rhetorical" plausibility.<sup>4</sup>

"Rhetorical" plausibility talks explicitly about why some individuals were treated and others were not.

Examples:

- Separate decision-maker with limited information decides on treatment.
- Capacity limits.
- Treatment consequence of randomization.
- Decision maker cares about different outcome than evaluator.

---

<sup>4</sup><https://blogs.worldbank.org/impactevaluations/what-do-you-need-do-make-matching-estimator-convincing-rhetorical-vs-statistical>

# Summary

- Role of balanced covariates distribution.
- Covariates distributions are balanced iff  $e(X)$  is constant.
- Obtain balance forcing  $X$  to be the same (*matching*),  
or forcing  $e(X)$  to be the same (*propensity score matching*).
- Many practical choices how to implement those ideas.

Questions?

# Econometrics II

## Lecture 6: Instrumental Variables

Konrad Burchardi

Stockholm University

18th of April 2024

# Literature

- ① "Mostly Harmless Econometrics", Angrist and Pischke  
Chapter 4.1-4.3, 4.6.1, 4.6.4

These notes draw on those books. All mistakes are mine.

# Plan for Today

1 Introducing IV

2 Understanding IV

3 Common Mistakes

4 Specification Tests

5 Application: Shift-Share Instruments

# Introducing IV

Take standard regression framework<sup>1</sup>:

$$y = \mathbf{X}\beta + \varepsilon,$$

where  $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2]$ . Worried about exogeneity of  $\mathbf{X}_1$ .

Valid Instrumental Variables yield consistent estimates of  $\beta$ .

- ...in the presence of measurement error in  $\mathbf{X}_1$ ;
- ...in case of endogeneity of regressors,  $\mathbb{E}[\varepsilon|\mathbf{X}_1] \neq 0$ .

*How does this work? And what are valid instruments?*

---

<sup>1</sup>Assume constant treatment effect. Later will talk about IV with treatment effect heterogeneity.

# Introducing IV

We require some ‘**instruments**’  $\mathbf{Z}_1$  such that:

- ① **Relevance**:  $\text{plim}_{\bar{N}} \frac{1}{N} (\mathbf{Z}'_1 \mathbf{X}_1) \neq 0$
- ② **Exogeneity**:  $\text{plim}_{\bar{N}} \frac{1}{N} (\mathbf{Z}'_1 \varepsilon) = 0$

Then  $\hat{\beta}_{IV} = (\mathbf{Z}' \mathbf{X})^{-1} \mathbf{Z}' y$  is consistent estimator of  $\beta$ , where  $\mathbf{Z} = [\mathbf{Z}_1, \mathbf{X}_2]$ :

$$\begin{aligned}\text{plim } \hat{\beta}_{IV} &= \text{plim } \left[ \left( \frac{1}{N} \mathbf{Z}' \mathbf{X} \right)^{-1} \frac{1}{N} \mathbf{Z}' (\mathbf{X} \beta + \varepsilon) \right] \\ &= \beta + \left[ \text{plim } \left( \frac{1}{N} \mathbf{Z}' \mathbf{X} \right)^{-1} \times \text{plim } \left( \frac{1}{N} \mathbf{Z}' \varepsilon \right) \right] = \beta,\end{aligned}$$

where we use  $\text{plim}_{\bar{N}} \frac{1}{N} (\mathbf{Z}'_1 \mathbf{X}_1) \neq 0$  [**Relevance**] and  $\text{plim}_{\bar{N}} \frac{1}{N} (\mathbf{Z}'_1 \varepsilon) = 0$  [**Exogeneity**].

# Generalized IV and 2SLS

## 1 Generalized IV

The optimal choice of instruments  $\mathbf{Z}$  is  $\mathbf{P}_Z \mathbf{X}$ .<sup>2</sup>

(Note:  $\mathbf{X}_2$  is optimally instrumented with  $\mathbf{X}_2$ .)

The estimator is called 'generalized IV', defined as:

$$\hat{\beta}_{GIV} = (\mathbf{X}' \mathbf{P}_Z \mathbf{X})^{-1} \mathbf{X}' \mathbf{P}_Z y.$$

## 2 Two-Stage Least-Squares (2SLS)

Given that  $P_Z$  is idempotent and symmetric,  $\hat{\beta}_{GIV}$  is numerically equivalent to:

$$\hat{\beta}_{2SLS} = (\hat{\mathbf{X}}' \hat{\mathbf{X}})^{-1} \hat{\mathbf{X}}' y,$$

where  $\hat{\mathbf{X}} = \mathbf{P}_Z \mathbf{X}$ .<sup>3</sup>

---

<sup>2</sup>In the least asymptotic variance sense.

<sup>3</sup>Proof of consistency works also for  $\hat{\beta}_{GIV}$  and hence  $\hat{\beta}_{2SLS}$ .

# Plan for Today

1 Introducing IV

2 Understanding IV

3 Common Mistakes

4 Specification Tests

5 Application: Shift-Share Instruments

# Anatomy of IV Formula

- The **2SLS** formula shows, we can calculate IV estimator in two steps:
  - ① Regress  $\mathbf{X}$  on  $\mathbf{Z}$  to obtain predicted values  $\hat{\mathbf{X}}$
  - ② Regress  $y$  on  $\hat{\mathbf{X}}$ .

Intuitive interpretation: '**only exploit variation in  $\mathbf{X}$  driven by the instrument**'.
- Meaning **Relevance** Condition:
  - ①  $Z_1$  needs to impact  $X_1$  (conditional on  $X_2$ ).
  - ② At least as many instruments as endogenous variables.  
→ Without it cannot estimate effect of  $X_1$  on  $y$ .
- Meaning **Exogeneity** Condition<sup>4</sup>:
  - ①  $Z_1$  is determined 'like an experiment' (instrument is **external**)...
  - ② and  $Z_1$  affects  $y$  *only* through  $X_1$  (instrument is **excludable**).  
→ Without it do not solve original problem.

---

<sup>4</sup>Sometimes called "Exclusion Restriction" or "Identifying Assumption". **Fundamentally not testable!**

# Intuition for IV

- ① Find variables  $Z_1$  that...

*Relevance:* “shock”  $X_1$ , but...

*Exogeneity:* ...are unrelated to  $y$  otherwise.

Then we see how  $y$  changes when  $X_1$  is shocked!

- ② Only exploit variation in  $X_1$  that we “know to be exogenous”.

- ③ Idea much like an **experiment**:

*Shock the explanatory variable, rather than finding more controls!*

## First Stage, Reduced Form and Second Stage

“**First Stage**” is the (OLS) regression of each element of  $\mathbf{X}_1$  on  $\mathbf{Z}$ .

This tells us how the instruments impact the endogenous variable.

Key to test Condition 1!

“**Reduced Form**” is the (OLS) the regression of  $y$  on  $\mathbf{Z}$ .

This tells us how the instrument is related to outcomes.

(Excludability not necessary.)

“**Second Stage**” is the regression of  $y$  on  $\hat{\mathbf{X}}$ .

This tells us how exogenous changes in  $\mathbf{X}_1$  and  $\mathbf{X}_2$  impact  $y$ .

# IV is Reduced Form over First Stage

In case of one endogenous variable and one regressor, we can write

$$\beta_{IV} = \frac{\text{Cov}(y_i, z_i)}{\text{Cov}(x_i, z_i)} = \frac{\text{Cov}(y_i, z_i) / V(z_i)}{\text{Cov}(x_i, z_i) / V(z_i)}$$

Sample analogue is called **Indirect Least Squares** estimator.

IV estimate is ratio of the reduced form over the first stage coefficient!<sup>5</sup>

**Two-Sample IV** (Angrist and Krueger, 1992):

To calculate IV estimator requires only  $\frac{1}{N_A} \mathbf{Z}' \mathbf{X}$  and  $\frac{1}{N_B} \mathbf{Z}' \mathbf{y}$ . These might come from different samples (from the same population), so  $\mathbf{X}$  and  $\mathbf{y}$  need not be in same data set.

**Split-Sample IV** (Angrist and Krueger, 1995), more efficient:

Find first coefficient in sample A,  $(\mathbf{Z}'_A \mathbf{Z}_A)^{-1} \mathbf{Z}'_A \mathbf{X}_A$  and calculate IV estimate in sample B as regression  $y_B$  on  $\mathbf{Z}_B (\mathbf{Z}'_A \mathbf{Z}_A)^{-1} \mathbf{Z}'_A \mathbf{X}_A$ . Adjust standard errors (Inoue and Solon, 2010).

---

<sup>5</sup> Mathematical fact, also with  $\mathbf{X}_2$ . If your results do not satisfy it, you did something wrong.

## Simple Case: Wald Estimator

Take the case of a single dummy instrument  $z_i$  and one endogenous regressor  $x_i$

$$y_i = \alpha + \beta x_i + \varepsilon_i.$$

Using  $\mathbb{E}[\varepsilon_i | z_i] = 0$ , it follows that  $\mathbb{E}[y_i | z_i] = \alpha + \beta \mathbb{E}[X_i | z_i]$  and:

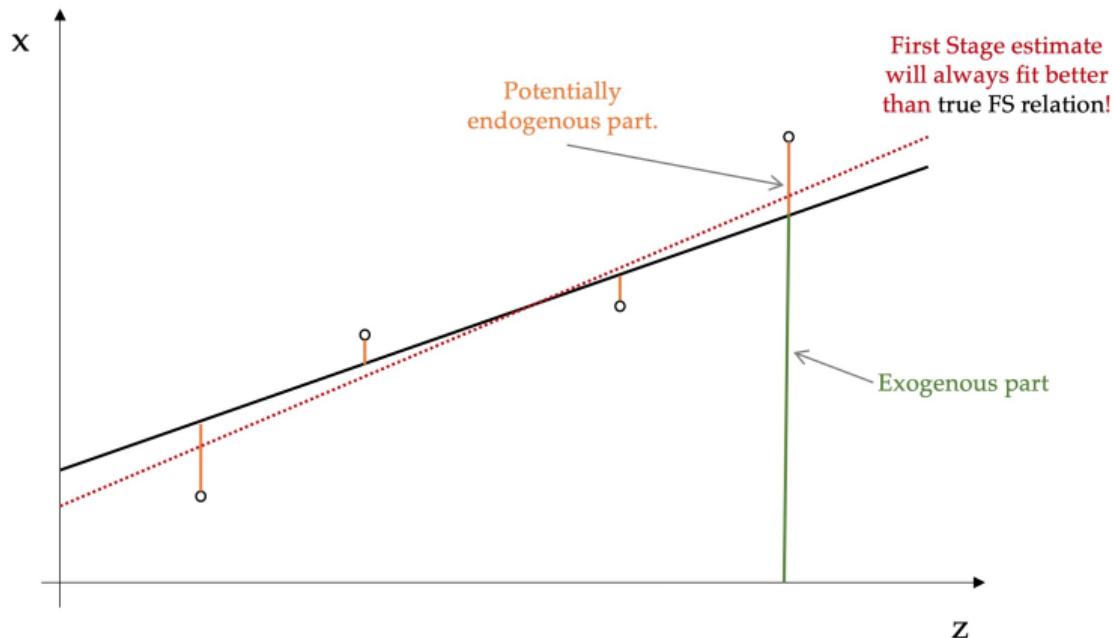
$$\beta = \frac{\mathbb{E}[y_i | z_i = 1] - \mathbb{E}[y_i | z_i = 0]}{\mathbb{E}[x_i | z_i = 1] - \mathbb{E}[x_i | z_i = 0]}$$

Population analogue of **Wald Estimator**.

*Intuition?*

**Experiments:** IV in 'encouragement designs', or with imperfect compliance.

# Why is IV only consistent, but not unbiased?



OLS overfits the First Stage in small samples. [Problem Set 3]

But variance around true First Stage effect decreases with sample size.

# Consistent but not Unbiased

*What can be done about it?*

- ① Test how big problem (likely) is. Test **Relevance** condition!

$$\mathbb{E}[\hat{\beta}_{2SLS} - \beta] \approx \frac{\sigma_{\eta\varepsilon}}{\sigma_\eta^2} \left[ \frac{\mathbb{E}[\pi' \mathbf{Z}_1' \mathbf{Z}_1 \pi] / q}{\sigma_\eta^2} + 1 \right]^{-1}$$

where  $x = \mathbf{Z}_1 \pi + \eta$  is the First Stage, and  $\mathbb{E}[\pi' \mathbf{Z}_1' \mathbf{Z}_1 \pi] / q$  is the First Stage 'population F-statistics' **on the excluded instruments** (not  $\mathbf{X}_2$ ).

- Finite sample bias of IV inversely related to "strength" of instruments; as rule of thumb: with First Stage F-statistics  $< 10$ ,<sup>6</sup> instruments were considered 'weak' (Staiger and Stock, 1997); see also Young (2022).
- If instruments are useless, bias as large as OLS.
- If you add useless instruments, F-statistic falls and bias increases.
- With multiple instruments: KP/AP test of differential variation.

- ② Correct for the degree of bias: FIML estimator (less efficient with strong instr.)

<sup>6</sup>And then? 1. Drop weak instruments; 2. Get better instruments; 3. LIML/JIV; 4. New project.

## IV and Classical Measurement Error

With classical measurement error, where  $x_{1i}^* = x_{1i} + v_i$ :

$$\text{plim } [\hat{\beta}_1] = \beta_1 \frac{\text{Var}(x_{1i})}{\text{Var}(x_{1i}) + \text{Var}(v_i)} \equiv \beta_1 \lambda$$

Now consider you have additionally another measure of  $x_{1i}$ :

$$z_i = x_{1i} + \xi_i, \text{ with } \text{Cov}(v_i, \xi_i) = 0$$

Then the reduced form and first stage identify

$$\gamma_1 = \frac{\text{Cov}(y_i, z_i)}{\text{Var}(z_i)} = \beta_1 \frac{\text{Var}(x_{1i})}{\text{Var}(x_{1i}) + \text{Var}(\xi_i)}; \pi_1 = \frac{\text{Cov}(x_{1i}^*, z_i)}{\text{Var}(z_i)} = \frac{\text{Var}(x_{1i})}{\text{Var}(x_{1i}) + \text{Var}(\xi_i)}$$

Therefore  $\beta = \frac{\gamma_1}{\pi_1}$ , i.e. IV estimator identifies  $\beta_1$ , not  $\beta_1 \lambda$ .

Often if  $\beta_{2SLS} > \beta_{OLS}$  in absolute value - against the readers' expectations-authors conclude: 'IV solved measurement error'.

# Plan for Today

- 1 Introducing IV
- 2 Understanding IV
- 3 Common Mistakes
- 4 Specification Tests
- 5 Application: Shift-Share Instruments

# Getting Standard Errors Right

There is a temptation to calculate the 2SLS estimator by:

- ① running the First Stage as OLS regression of  $\mathbf{X}$  on  $\mathbf{Z}$ ;
- ② calculate the predicted values  $\hat{\mathbf{X}}$ ;
- ③ running the Second Stage as OLS regression of  $y$  on  $\hat{\mathbf{X}}$ .

This will provide you with the correct  $\hat{\beta}_{2SLS}$  (discussed above).

However the standard errors will be wrong! Should be (without proof)

$$y - \mathbf{X}\hat{\beta}_{2SLS},$$

but in the above procedure your statistical package will calculate them as

$$y - \hat{\mathbf{X}}\hat{\beta}_{2SLS}.$$

# Getting First Stage Right

Rewriting the Second Stage we get:

$$y = \hat{\mathbf{X}}_1\beta_1 + \mathbf{X}_2\beta_2 + (\mathbf{X}_1 - \hat{\mathbf{X}}_1)\beta_1 + \varepsilon$$

Note that:

- ①  $\mathbf{X}_2$  is uncorrelated of  $\varepsilon$  (by assumption);
- ②  $\mathbf{X}_2$  is uncorrelated of  $\mathbf{X}_1 - \hat{\mathbf{X}}_1$  (by construction);
- ③  $\hat{\mathbf{X}}_1$  is linear combination of  $[\mathbf{Z}_1, \mathbf{X}_2]$ , asymp. uncorrelated of  $\varepsilon$  (by assumption);
- ④  $\hat{\mathbf{X}}_1$  is uncorrelated of  $\mathbf{X}_1 - \hat{\mathbf{X}}_1$  (by construction).

Together these imply we can consistently estimate  $\beta$ .

Failure to include  $\mathbf{X}_2$  in the First Stage means (2) breaks down.

Failure to run linear First Stage means (2), (4) and (3) might break down.

## Interpreting $R^2$ in Second Stage

The  $R^2$  in Second Stage [when displayed] is not meaningful.

- Residuals are calculated, correctly, as  $y - \mathbf{X}\hat{\beta}_{2SLS}$ . The RSS might be larger than TSS, and hence  $R^2 < 0$ .
- The point of the Second Stage is *not* to fit  $y$  to  $\mathbf{X}$ , but solely to estimate  $\hat{\beta}$ .

What is (somewhat) meaningful is the  $R^2$  in the Reduced Form.

# Basic Mistakes in Typical IV Paper

In my (limited) experience the **most common drawbacks** of IV papers are:

- ① Authors present an instrument that is plausibly external, but might impact  $y$  through multiple channels; authors highlight one channel.  
To save project: **Is Reduced Form interesting?**
- ② Authors do not critically assess plausibility of exclusion restriction.

# Plan for Today

- 1 Introducing IV
- 2 Understanding IV
- 3 Common Mistakes
- 4 Specification Tests
- 5 Application: Shift-Share Instruments

# Discussing Exogeneity

*Relevance* condition can be tested (see above).

*Exogeneity* condition can fundamentally not be tested.

- ① Need to argue, *using understanding of the world*, that it is satisfied.
- ② Might provide 'balance' tests, demonstrating that **Z** is unrelated to baseline variables that might impact *y*.
- ③ Might provide 'placebo' tests, demonstrating that **Z** has no impact on pseudo outcomes, outcomes which it should not impact.

# Order of Identification

With number of instruments in  $Z_1$ ...

- 1 ...greater than number of variables in  $X_1$ , model is '**over-identified**'.  
Efficient to use all instruments, if they are relevant.
- 2 ...equal to number of variables in  $X_1$ , model is '**exactly identified**'.
- 3 ...less than number of variables in  $X_1$ , model is '**under-identified**'.

# Overidentification Tests

In the over-identified case, can calculate Sargan-Hansen/Sargan's J test:

$$J(\hat{\beta}) = N \frac{\hat{\varepsilon}' P_Z \hat{\varepsilon}}{\hat{\varepsilon} \hat{\varepsilon}}$$

Under  $H_0$  that *Exogeneity* is satisfied for all  $Z$ , this is  $\chi^2$ -distributed.

- Empirically straight-forward to implement as  $M$  times  $R^2$  of regression of Second Stage residuals on all elements of  $Z$ .
- Intuition is: in case *Exogeneity* is not satisfied for some instruments, they will be correlated with  $\hat{\varepsilon}$  and  $H_0$  is rejected.
- Omnibus test: Tells you something is wrong, not what.
- With effect heterogeneity, different instruments identify different causal effects, and test is no longer useful.

# Plan for Today

- 1 Introducing IV
- 2 Understanding IV
- 3 Common Mistakes
- 4 Specification Tests
- 5 Application: Shift-Share Instruments

## Shift-Share / Bartik Instruments

$$z_l = \sum_n s_{ln} g_n,$$

a weighted sum of

- common shocks, varying at level  $n = 1, \dots, N$ ,
- weighted by exposure shares, varying at level of outcome  $l = 1, \dots, L$ .

## Shift-Share: Examples

**Bartik (1991) and Blanchard and Katz (1992):**

instrument region  $l$ 's labor demand, where  $g_n$  is the national growth of industry  $n$  and  $s_{ln} \in [0, 1]$  are lagged employment shares; and study impact on wages.

**Autor, Dorn and Hanson (2013):**

instrument Chinese import competition in location  $l$ , where  $g_n$  is growth in Chinese exports in manufacturing industry  $n$  to 8 non-U.S. countries, and  $s_{ln}$  are lagged employment shares; study impact on manufacturing employment / unemployment.

# Shift-Share / Bartik Instruments

## Recent Developments

- ① Borusyak, Kirill, Peter Hull, and Xavier Jaravel (2022) "Quasi- Experimental Shift-Share Research Designs" *Review of Economic Studies*
- ② Goldsmith-Pinkham, Paul, Isaac Sorkin, and Henry Swift (2020) "Bartik Instruments: What, When, Why and How" *American Economic Review*

Questions?

# References

- 1 Angrist and Pischke (2008): Chapter 4.1-4.3, 4.6.1, 4.6.4.
- 2 Angrist, Joshua and Alan B. Krueger (1992) "The Effect of Age at School Entry on Educational Attainment: An Application of Instrumental Variables with Moments from Two Samples" *Journal of the American Statistical Association* 87: 328–336.
- 3 Angrist, Joshua and Alan B. Krueger (1995) "Split-Sample Instrumental Variables Estimates of the Return to Schooling" *Journal of Business and Economic Statistics* 13: 225–235.
- 4 Autor, David, David Dorn, and Gordon Hanson (2013) "The China Syndrome: Local Labor Market Effects of Import Competition in the United States." *American Economic Review* 103 (6): 2121–68.
- 5 Bartik, Timothy (1991) "Who Benefits from State and Local Economic Development Policies?" W.E. Upjohn Institute.
- 6 Blanchard, Olivier, and Lawrence Katz (1992) "Regional Evolutions" *Brookings Papers on Economic Activity* 23(1): 1–76.
- 7 Borusyak, Kirill, Peter Hull and Xavier Jaravel (2022) "Quasi-Experimental Shift-Share Research Designs" *Review of Economic Studies* 89(1): 181–213.

## References

- 8 Goldsmith-Pinkham, Paul, Isaac Sorkin, Henry Swift (2020) "Bartik Instruments: What, When, Why, and How" *American Economic Review* 110(8): 2586–2624.
- 9 Inoue, Atsushi, and Gary Solon (2010) "Two-Sample Instrumental Variables Estimators" *The Review of Economics and Statistics* 92(3): 557–561.
- 10 Staiger, Douglas, and James Stock (1997) "Instrumental Variables Regressions with Weak Instruments" *Econometrica* 65(3): 557–586.
- 11 Young, Alwyn (2022) "Consistency without Inference: Instrumental Variables in Practical Application" *European Economic Review* 147: 104–112.

# Econometrics II

Lecture 7: Instrumental Variables with Heterogeneous Treatment Effects

Konrad Burchardi

Stockholm University

23rd of April 2024

# Literature

- ① "Mostly Harmless Econometrics", Angrist and Pischke  
Chapters 4.4 and 4.5

These notes draw on those books. All mistakes are mine.

# Plan for Today

1 Local Average Treatment Effect

2 Characterizing Compliers

3 Generalisations of LATE

# The LATE

Traditional IV framework is useful:

- Think about estimating constant causal effects.
- Think clearly about source of variation in the  $Z$  used to identify causal effects.

**Today:**  $Y_i(1) - Y_i(0)$  does not have to be same across individuals.

→ Return to the potential outcomes framework.

Focus on simple case where:

- we estimate effect of a *binary treatment*,  $D_i$ , on an outcome  $Y_i$ ;
- the treatment is endogenous, but have *binary instrument*,  $Z_i$ .

(Later we will discuss generalizations.)

# The LATE

## Setup:

- ① Let  $Y_i(d, z)$  denote the potential outcome<sup>1</sup> when  $D_i = d$  and  $Z_i = z$ , and  $D_i(z)$  potential treatment status.
- ② We think of the instrument as causally affecting treatment, but this effect too is allowed to be heterogeneous!

Potential treatment status:  $D_i(0)$  when  $Z_i = 0$ ;  $D_i(1)$  when  $Z_i = 1$

Observed treatment status:  $D_i = D_i(0) + (D_i(1) - D_i(0)) Z_i$

---

<sup>1</sup>Double indexing: candidate instruments might have a direct effect on outcomes.

## LATE: Compliance Types

Therefore there are four 'compliance types':

	$D_i(1)$	$D_i(0)$
Compliers:	1	0
Never-takers:	0	0
Always-takers:	1	1
Defiers:	0	1

Just like with potential outcomes, the compliance type is **unobserved**.

What is **observed**?  $Z_i$  and  $D_i$

	$D_i = 0$	$D_i = 1$
$Z_i = 0$	Compliers, Never-takers	Defier, Always-takers
$Z_i = 1$	Defier, Never-takers	Compliers, Always-takers

## LATE: Assumptions

- ① **Random Assignment** of instrument:  $[\{Y_i(d, z) \forall d, z\}, D_i(1), D_i(0)] \perp Z_i$ .
- ② **Exclusion Restriction:**  $Y_i(d, 1) = Y_i(d, 0)$ ,  $d = 0, 1$ .
- ③ **Relevance:**  $\mathbb{E}[D_i(1) - D_i(0)] \neq 0$ .
- ④ **Monotonicity:**  $D_i(1) \geq D_i(0)$  [no defiers], or  $D_i(1) \leq D_i(0)$  [no compliers].

# LATE Assumptions: Random Assignment

What does random assignment assumption buy us?

$$[\{Y_i(d, z), \forall d, z\}, D_i(1), D_i(0)] \perp Z_i$$

- The treatment being as good as randomly assigned means that  $Z_i$  is independent of potential outcomes and potential treatments.
- Independence implies that **first stage** is average causal effect of  $Z_i$  on  $D_i$ :

$$\begin{aligned}\mathbb{E}[D_i|Z_i = 1] - \mathbb{E}[D_i|Z_i = 0] &= \mathbb{E}[D_i(1)|Z_i = 1] - \mathbb{E}[D_i(0)|Z_i = 0] \\ &= \mathbb{E}[D_i(1) - D_i(0)]\end{aligned}$$

- Independence is sufficient for a causal interpretation of the **reduced form**:

$$\mathbb{E}[Y_i|Z_i = 1] - \mathbb{E}[Y_i|Z_i = 0] = \mathbb{E}[Y_i(D_i(1), 1)] - \mathbb{E}[Y_i(D_i(0), 0)]$$

but this does not link the effect to treatment.

# LATE Assumptions: Exclusion Restriction

What does exclusion restriction assumption buy us?

$$Y_i(d, 1) = Y_i(d, 0) \equiv Y_i(d)$$

The exclusion restriction means that  $Z_i$  affects  $Y_i$  only through  $D_i$ .

Technically, we can write  $Y_i$  as:

$$\begin{aligned} Y_i &= Y_i(0, z) + (Y_i(1, z) - Y_i(0, z))D_i(z) \\ &= Y_i(0) + (Y_i(1) - Y_i(0))D_i(z) \end{aligned}$$

'Random assignment' and the 'exclusion restriction' should look familiar!

# LATE Assumptions: Monotonicity

What does **monotonicity** assumption buy us?

Observed outcome:

$$\begin{aligned}Y_i &= Y_i(0) + (Y_i(1) - Y_i(0))D_i \\&= Y_i(0) + (Y_i(1) - Y_i(0))[D(0) + (D_i(1) - D_i(0))Z_i] \\&= Y_i(0) + (Y_i(1) - Y_i(0))D_i(0) + Z_i(\textcolor{teal}{D}_i(1) - \textcolor{teal}{D}_i(0))(Y_i(1) - Y_i(0))\end{aligned}$$

Then the difference in population means is, using randomization:

$$\mathbb{E}[Y_i|Z_i = 1] - \mathbb{E}[Y_i|Z_i = 0] = \mathbb{E}[(D_i(1) - D_i(0))(Y_i(1) - Y_i(0))]$$

Notice that  $D_i(1) - D_i(0) \in \{-1, 0, 1\}$ .

- For always-takers and never-takers  $D_i(1) - D_i(0) = 0!$   
Do not impact mean difference.
- One of the other two,  $-1$  and  $1$ , is ruled out by monotonicity.

# LATE Assumptions: Monotonicity

What does **monotonicity** assumption buy us? Focus on 'no defiers' case.

- The **reduced form** coefficient estimate is:

$$\begin{aligned}\mathbb{E}[Y_i|Z_i = 1] - \mathbb{E}[Y_i|Z_i = 0] \\ = \mathbb{E}[Y_i(1) - Y_i(0)|D_i(1) - D_i(0) = 1] * Pr(D_i(1) - D_i(0) = 1)\end{aligned}$$

- The **first-stage** regression estimates, using randomisation:

$$\mathbb{E}[D_i|Z_i = 1] - \mathbb{E}[D_i|Z_i = 0] = \mathbb{E}[D_i(1) - D_i(0)] = Pr(D_i(1) - D_i(0) = 1)$$

The ratio of these two is the **IV/Wald-estimator**:

$$\beta^{LATE} = \mathbb{E}[Y_i(1) - Y_i(0)|D_i(1) - D_i(0) = 1]$$

# LATE: Interpretation

$$\beta^{LATE} = \mathbb{E}[Y_i(1) - Y_i(0) | D_i(1) - D_i(0) = 1]$$

IV thus identifies a '**Local Average Treatment Effect**':

- The average effect of treatment in the sub-population whose behaviour was changed because of the value of the instrument, the *compliers*.
- With treatment heterogeneity, this average depends on the instrument.
  - ① With  $> 1$  instrument, use to test whether treatment effects are homogenous.
  - ② Highlights difference between internal and external validity.

# Illustration: Angrist and Evans (AER, 1998)

460

THE AMERICAN ECONOMIC REVIEW

JUNE 1998

TABLE 5—WALD ESTIMATES OF LABOR-SUPPLY MODELS

Variable	1980 PUMS			1990 PUMS			1980 PUMS		
	Mean difference by Same sex	Wald estimate using as covariate:		Mean difference by Same sex	Wald estimate using as covariate:		Mean difference by Twins-2	Wald estimate using as covariate:	
		More than 2 children	Number of children		More than 2 children	Number of children		More than 2 children	Number of children
More than 2 children	0.0600 (0.0016)	—	—	0.0628 (0.0016)	—	—	0.6031 (0.0084)	—	—
Number of children	0.0765 (0.0026)	—	—	0.0836 (0.0025)	—	—	0.8094 (0.0139)	—	—
Worked for pay	-0.0080 (0.0016)	-0.133 (0.026)	-0.104 (0.021)	-0.0053 (0.0015)	-0.084 (0.024)	-0.063 (0.018)	-0.0459 (0.0086)	-0.076 (0.014)	-0.057 (0.011)
Weeks worked	-0.3826 (0.0709)	-6.38 (1.17)	-5.00 (0.92)	-0.3233 (0.0743)	-5.15 (1.17)	-3.87 (0.88)	-1.982 (0.386)	-3.28 (0.63)	-2.45 (0.47)
Hours/week	-0.3110 (0.0602)	-5.18 (1.00)	-4.07 (0.78)	-0.2363 (0.0620)	-3.76 (0.98)	-2.83 (0.73)	-1.979 (0.327)	-3.28 (0.54)	-2.44 (0.40)
Labor income	-132.5 (34.4)	-2208.8 (569.2)	-1732.4 (446.3)	-119.4 (42.4)	-1901.4 (670.3)	-1428.0 (502.6)	-570.8 (186.9)	-946.4 (308.6)	-705.2 (229.8)
In(Family income)	-0.0018 (0.0041)	-0.029 (0.068)	-0.023 (0.054)	-0.0085 (0.0047)	-0.136 (0.074)	-0.102 (0.056)	-0.0341 (0.0223)	-0.057 (0.037)	-0.042 (0.027)

Notes: The samples are the same as in Table 2. Standard errors are reported in parentheses.

## Illustration: Angrist and Evans (AER, 1998)

An aside: **What do you think of the same-sex instrument?**

- ① As good as randomly assigned?

Most likely.

- ② Excludable?

**What do you think?**

- ③ Relevant?

Check the first stage.

- ④ Monotonicity?

**What do you think?**

# LATE Intuition: Attempt 1

What is the intuition for the LATE result?

Observed outcome:

$$Y_i = \underbrace{Y_i(0) + (Y_i(1) - Y_i(0))D_i(0)}_{\text{For always-taker: only this.}} + Z_i(D_i(1) - D_i(0))(Y_i(1) - Y_i(0))$$

For never-taker: only this.

By randomization:

- ① Always- and never-takers are just as frequent and have same outcomes in  $Z_i = 1$  and  $Z_i = 0$  group.
- ② When calculating mean-difference  $\mathbb{E}[Y_i|Z_i = 1] - \mathbb{E}[Y_i|Z_i = 0]$ , they drop out.
- ③ Just need to correct for mean-difference being over all observations, not just compliers: first stage.

# LATE Intuition: Attempt 2

What is the intuition for the LATE result?

Without monotonicity:

	$D_i = 0$	$D_i = 1$
$Z_i = 0$	Compliers, Never-takers	Defier, Always-takers
$Z_i = 1$	Defier, Never-takers	Compliers, Always-takers

With monotonicity:

	$D_i = 0$	$D_i = 1$
$Z_i = 0$	Compliers, Never-takers	Always-takers
$Z_i = 1$	Never-takers	Compliers, Always-takers

By Randomization:

- ① Know fraction of never-takers and their outcome  $\mathbb{E}[Y_i(0)|\text{never taker}]$ .
- ② Know fraction of compliers. This allows us to back out  $\mathbb{E}[Y_i(0)|\text{complier}]$ .
- ③ Same with always-takers and  $\mathbb{E}[Y_i(1)|\text{complier}]$ .

# LATE Intuition: Attempt 3

What is the **intuition** for the LATE result?

- What is the propensity score of always-takers?
- What is the propensity score of never-takers?
- What is the propensity score of compliers?

For whom can you estimate causal effects?

*If you are interested in this interpretation, see  
Heckman and Vytlacil (2005, Econometrica) and Zhou and Xie (2019, JPE).*

## LATE: Monotonicity

Is Monotonicity a sensible assumption? Sometimes.

- A Latent Index model would directly impose monotonicity:

$$D_i(Z_i) = \mathbb{I}(\gamma_0 + \gamma_1 Z_i > v_i). \quad (1)$$

- In many situations it is also a natural assumption:

Think of  $Z$  as assignment to a treatment (e.g. credit access). Few individuals would do the exact opposite **because** they were assigned to a treatment.

## Special Case: Bloom (1984) Estimator

Many experiments randomize the *offer* of a treatment.

- As-treated analysis (OLS) is contaminated by selection bias.
- Intention-to-treat (ITT) analyses diluted by non-compliance.

IV solves problem: IV estimand is average treatment effect of the treated (ATT).<sup>2</sup>

$$\frac{\mathbb{E}[Y_i|Z_i = 1] - \mathbb{E}[Y_i|Z_i = 0]}{\mathbb{E}[D_i|Z_i = 1]} = \frac{\text{ITT Estimate}}{\text{Compliance Rate}} = \mathbb{E}[Y_i(1) - Y_i(0)|D_i = 1]$$

---

<sup>2</sup>Intuitively: *always-takers* are no longer present in  $D_i = 1$  group; the treated group can only consist of compliers.)

# Plan for Today

1 Local Average Treatment Effect

2 Characterizing Compliers

3 Generalisations of LATE

# Counting Compliers

## What fraction of the sample are compliers?

As you saw in Intuition 2, we can count compliers.

- Let  $C_i = \mathbb{I}(D_i(1) > D_i(0))$  indicate whether an individual is a complier.
- We can say how many there are: **simply check the first-stage coefficient.**

$$Pr(C_i = 1) = \mathbb{E}[D_i|Z_i = 1] - \mathbb{E}[D_i|Z_i = 0]$$

Note:

We can pin-point some of the always- and never-takers. **How?**

This is not true for compliers. **Why?**

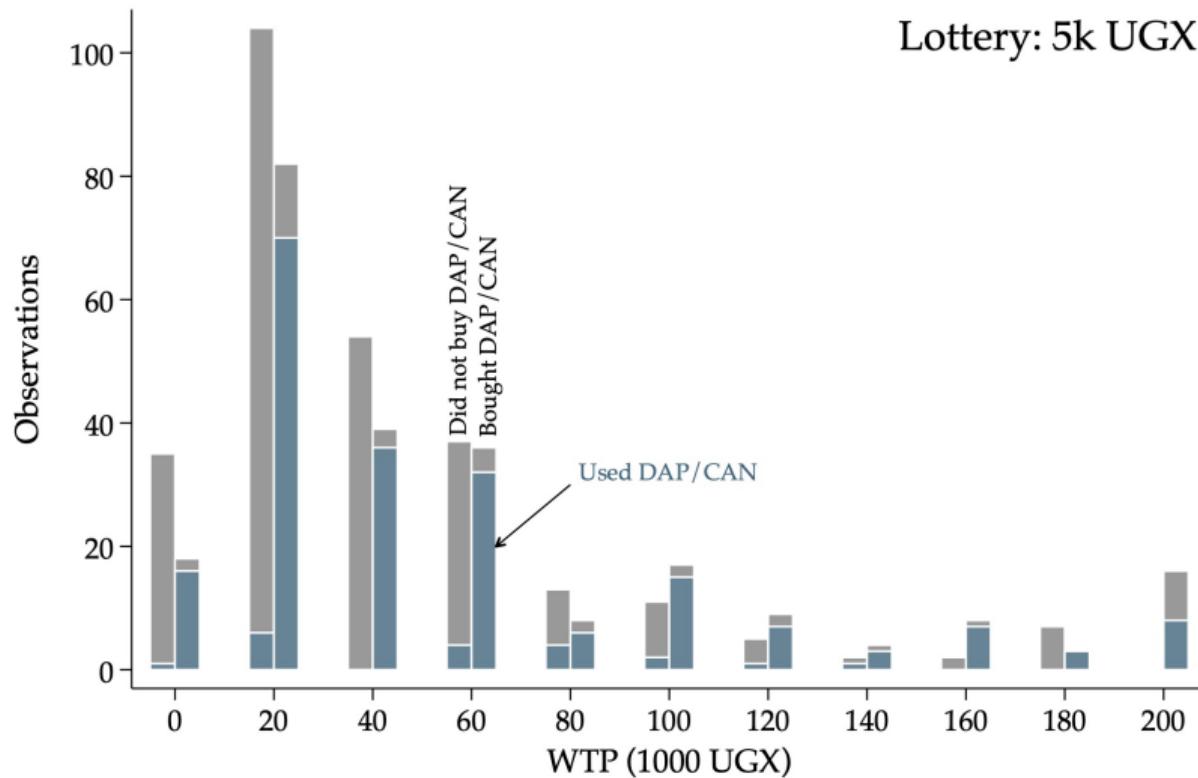
# Counting Compliers: Examples

TABLE 4.4.2  
Probabilities of compliance in instrumental variables studies

Endogenous Variable ( $D$ ) (2)	Instrument ( $z$ ) (3)	Sample (4)	$P[D = 1]$ (5)	First Stage, $P[D_1 > D_0]$ (6)	$P[z = 1]$ (7)	Compliance Probabilities	
						$P[D_1 > D_0   D = 1]$ (8)	$P[D_1 > D_0   D = 0]$ (9)
Veteran status	Draft eligibility	White men born in 1950	.267	.159	.534	.318	.101
		Non-white men born in 1950	.163	.060	.534	.197	.033
More than two children	Twins at second birth	Married women aged 21–35 with two or more children in 1980	.381	.603	.008	.013	.966
		First two children are same sex	.381	.060	.506	.080	.048
High school graduate	Third- or fourth-quarter birth	Men born between 1930 and 1939	.770	.016	.509	.011	.034
	State requires 11 or more years of school attendance	White men aged 40–49	.617	.037	.300	.018	.068

From Mostly Harmless Econometrics.

# Illustration: own work



# Characterizing Compliers

**What are characteristics of compliers relative to sample?**

Take a characteristic  $X_i = x$ .

$$\begin{aligned}\frac{Pr(X_i = x | C_i = 1)}{Pr(X_i = x)} &= \frac{Pr(C_i = 1 | X_i = x)}{Pr(C_i = 1)} \\ &= \frac{\mathbb{E}[D_i | Z_i = 1, X_i = x] - \mathbb{E}[D_i | Z_i = 0, X_i = x]}{\mathbb{E}[D_i | Z_i = 1] - \mathbb{E}[D_i | Z_i = 0]}\end{aligned}$$

This is useful to generalise from sample to other population.

# Illustration: Angrist and Evans (AER, 1998)

Table 21: Complier analysis

A. # compliers				
	$\Pr(D_i = 1)$	$\Pr(Z_i = 1)$	$\Pr(C_i = 1)$	
	0.378 (0.001)	0.506 (0.001)	0.068 (0.002)	0.091 (0.003)
B. Complier characteristics				
	$\Pr(X_i = 1)$	$\Pr(X_i = 1   C_i = 1)$	$\Pr(X_i = 1   C_i = 1) / \Pr(X_i = 1)$	
Age at 1st birth $\geq 25$	0.124 (0.001)	0.096 (0.009)	0.772 (0.073)	
Black	0.0502 (0.0004)	0.038 (0.007)	0.761 (0.135)	
Yrs. of schooling $\geq 12$	0.487 (0.001)	0.513 (0.025)	1.054 (0.051)	
Yrs. of schooling $\geq 16$	0.135 (0.001)	0.0940 (0.010)	0.698 (0.077)	

Notes: Standard errors in parentheses are robust. Data are from the 1980 US Census (PUMS) and correspond (almost) to the sample used in Angrist and Evans (1998). SE:s in Panel A, col. (4), and Panel B, cols. (2) and (3), are calculated using the delta method.

From Peter Fredriksson's old slides.

# Mean Potential Outcomes

Mean complier potential outcomes are also identified. (See Intuition 2.)

		$D_i = 0$	$D_i = 1$
		$Z_i = 0$	$Z_i = 1$
$Z_i = 0$	Compliers, Never-takers	Always-takers	
$Z_i = 1$	Never-takers	Compliers, Always-takers	

Let ,  $\pi^j, j = a, n, c$ , denote shares of always-takers, never-takers, and compliers.

With randomisation of  $Z_i$  they are identified:

- $\pi^a = \mathbb{E}(D_i = 1 | Z_i = 0)$
- $\pi^n = \mathbb{E}(D_i = 0 | Z_i = 1)$
- $\pi^c = \mathbb{E}(D_i = 1 | Z_i = 1) - \mathbb{E}(D_i = 1 | Z_i = 0)$

# Mean Potential Outcomes

- Define the identified quantity

$$\mu_{wz} = \mathbb{E}[Y_i | D_i = d, Z_i = z].$$

- Then

$$\mu_{10} = \mathbb{E}[Y_i(1) | \text{Always-taker}], \quad \mu_{01} = \mathbb{E}[Y_i(0) | \text{Never-taker}]$$

$$\mu_{11} = \frac{\pi^c}{\pi^c + \pi^a} \mathbb{E}[Y_i(1) | \text{Complier}] + \frac{\pi^a}{\pi^c + \pi^a} \mathbb{E}[Y_i(1) | \text{Always-taker}]$$

$$\mu_{00} = \frac{\pi^c}{\pi^c + \pi^n} \mathbb{E}[Y_i(0) | \text{Complier}] + \frac{\pi^n}{\pi^c + \pi^n} \mathbb{E}[Y_i(1) | \text{Never-taker}].$$

- Solve for the mean potential outcomes of compliers:

$$\mathbb{E}[Y_i(0) | \text{Complier}] = \frac{\pi^c + \pi^n}{\pi^c} \mu_{00} - \frac{\pi^n}{\pi^c} \mu_{01}$$

$$\mathbb{E}[Y_i(1) | \text{Complier}] = \frac{\pi^c + \pi^a}{\pi^c} \mu_{11} - \frac{\pi^a}{\pi^c} \mu_{10}$$

# Potential Outcomes Distribution

More generally true for **distribution of potential outcomes of compliers!**

- Let  $g_{c0}(y)$  and  $g_{c1}(y)$  be distributions of  $Y_i(0)$  and  $Y_i(1)$  amongst compliers.
- Let  $f_{dz}(y)$  denote the directly observed distribution when  $D_i = d$  and  $Z_i = z$ .

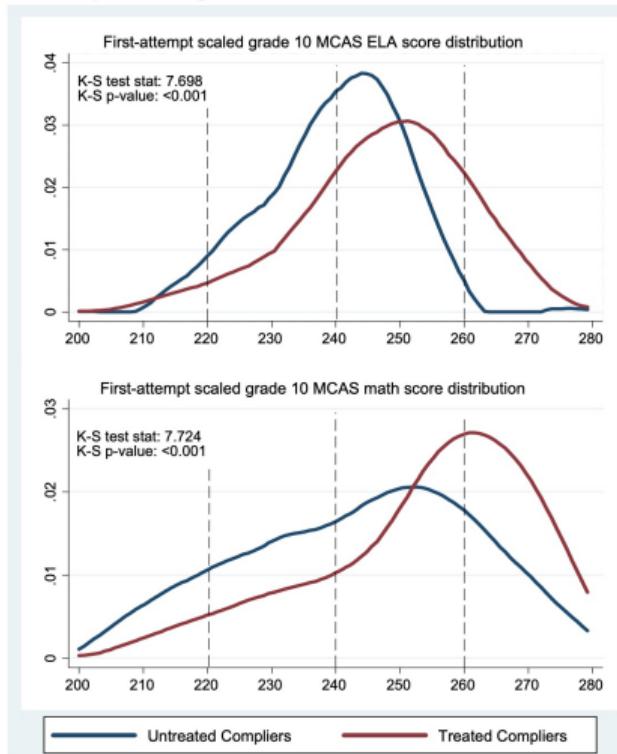
Then

$$g_{c0}(y) = \frac{\pi^c + \pi^n}{\pi^c} f_{00}(y) - \frac{\pi^n}{\pi^c} f_{01}(y)$$
$$g_{c1}(y) = \frac{\pi^c + \pi^a}{\pi^c} f_{11}(y) - \frac{\pi^a}{\pi^c} f_{10}(y)$$

(Imbens and Rubin, 1997; see Abadie, 2002 for convenient alternative.)

# Potential Outcomes Distribution

Figure 1: Complier Distributions for MCAS Scaled Scores



From Angrist's lecture notes (MIT 14.387)

# Plan for Today

1 Local Average Treatment Effect

2 Characterizing Compliers

3 Generalisations of LATE

## LATE with Multiple Instruments

**Scenario:** two (mutually exclusive) binary instruments  $Z_{1i}$  and  $Z_{2i}$ .

- Each instrument can be used to construct separate Wald estimates.  
Both Wald estimates have a LATE interpretation, although the complier population generally differs across the two instruments.
- Alternatively, two instruments can be combined into a single 2SLS estimate.  
→ This is a **weighted average** of the underlying Wald estimates.

## LATE with Multiple Instruments

- Note that LATE defined by each separate instrument is given by

$$\beta_j = \frac{\text{Cov}(Y_i, Z_{ji})}{\text{Cov}(D_i, Z_{ji})}, j = 1, 2$$

when instruments are mutually exclusive.<sup>3</sup>

- Denote the first stage as:

$$\hat{D}_i = \gamma_1 Z_{1i} + \gamma_2 Z_{2i}.$$

- The 2SLS estimand is given by:

$$\beta_{2SLS} = \frac{\text{Cov}(Y_i, \hat{D}_i)}{\text{Var}(\hat{D}_i)} = \gamma_1 \frac{\text{Cov}(Y_i, Z_{1i})}{\text{Cov}(D_i, \hat{D}_i)} + \gamma_2 \frac{\text{Cov}(Y_i, Z_{2i})}{\text{Cov}(D_i, \hat{D}_i)}.$$

---

<sup>3</sup>See Mogstad, Torgovitsky, Walters, AER, 2021.

# LATE with Multiple Instruments

$$\beta_{2SLS} = \frac{\text{Cov}(Y_i, \hat{D}_i)}{\text{Var}(\hat{D}_i)} = \gamma_1 \frac{\text{Cov}(Y_i, Z_{1i})}{\text{Cov}(D_i, \hat{D}_i)} + \gamma_2 \frac{\text{Cov}(Y_i, Z_{2i})}{\text{Cov}(D_i, \hat{D}_i)}$$

Use the definition of  $\beta_j$ :

$$\beta_{2SLS} = \gamma_1 \frac{\text{Cov}(D_i, Z_{1i})}{\text{Cov}(D_i, \hat{D}_i)} \beta_1 + \gamma_2 \frac{\text{Cov}(D_i, Z_{2i})}{\text{Cov}(D_i, \hat{D}_i)} \beta_2$$

or

$$\beta_{2SLS} = \varphi \beta_1 + (1 - \varphi) \beta_2,$$

where

$$\varphi = \gamma_1 \frac{\text{Cov}(D_i, Z_{1i})}{\text{Cov}(D_i, \hat{D}_i)} = \gamma_1 \frac{\text{Cov}(D_i, Z_{1i})}{\gamma_1 \text{Cov}(D_i, Z_{1i}) + \gamma_2 \text{Cov}(D_i, Z_{2i})}$$

Thus 2SLS produces a weighted average of LATE( $Z_1$ ) and LATE( $Z_2$ ).

## LATE with Covariates

A fully saturated first stage, and saturated in covariates second stage produce a weighted average of covariate-specific LATEs:

$$\beta_{2SLS} = \mathbb{E}[\omega(X_i)\beta_{LATE}^X]$$

where

$$\beta_{LATE}^X = \mathbb{E}[Y_i(1) - Y_i(0)|X_i, C_i = 1]$$

and

$$\omega(X_i) = \frac{\text{Var}(\mathbb{E}[D_i|Z_i, X_i]|X_i)}{\mathbb{E}[\text{Var}(\mathbb{E}[D_i|Z_i, X_i]|X_i)]}$$

Remember the regression/matching comparison in Lecture 5?

- Regression: covariate-specific weights were conditional variance of  $D$  given  $X$ .
- Here, variability in variance term comes from  $Z$ . More weight to covariate terms where the instrument creates more variation in fitted values.

# Average Causal Response

Consider a multi-valued treatment.

- Define the average causal response function as

$$Y_i(S) \equiv f_i(S),$$

defines potential outcomes for each individual at any treatment value.

- Take case with binary instrument:  
→ 2SLS provides weighted average of **unit causal effects**.

## Average Causal Response: Assumptions

- ① Randomization and exclusion:  $\{Y_i(0), Y_i(1), \dots, Y_i(\bar{S}), S_i(1), S_i(0)\} \perp Z_i$
- ② Relevance:  $\mathbb{E}[S_i(1) - S_i(0)] \neq 0$
- ③ Monotonicity:  $S_i(1) \geq S_i(0)$  (or  $S_i(1) \leq S_i(0)$ )

Then

$$\begin{aligned}\beta_{2SLS} &= \frac{\mathbb{E}[Y_i|Z_i = 1] - \mathbb{E}[Y_i|Z_i = 0]}{\mathbb{E}[S_i|Z_i = 1] - \mathbb{E}[S_i|Z_i = 0]} \\ &= \sum_{s=1}^S \omega_s \mathbb{E}[Y_i(s) - Y_i(s-1)|S_i(1) \geq s \geq S_i(0)]\end{aligned}$$

where

$$\omega_s = \frac{Pr(S_i < s|Z_i = 0) - Pr(S_i < s|Z_i = 1)}{\mathbb{E}[S_i|Z_i = 1] - \mathbb{E}[S_i|Z_i = 0]}, \text{ and } \sum_s \omega_s = 1.$$

and  $\mathbb{E}[Y_i(s) - Y_i(s-1)|S_i(1) \geq s \geq S_i(0)]$  is the **unit causal effect**:  
average difference in potential outcomes for compliers at point  $s$ .

Questions?

# Econometrics II

## Lecture 8: Fixed Effects and Panel Data

David Schönholzer

Stockholm University

April 25, 2024

# Plan for Today

## 1 Fixed Effects

Definitions

Identification and Estimation

Empirical Bayes

## 2 Panel Data

Structure and Definitions

## 3 Applications

Connected-Set Fixed Effects (AKM)

## 4 Appendix

Group Fixed Effects

# Table of Contents

## 1 Fixed Effects

### Definitions

Identification and Estimation

Empirical Bayes

## 2 Panel Data

Structure and Definitions

## 3 Applications

Connected-Set Fixed Effects (AKM)

## 4 Appendix

Group Fixed Effects

## Dual Indexation

- Consider groups  $\mathbf{J} : \{1, \dots, N\} \rightarrow \{1, \dots, J\}$  and let  $X_{ij} = 1[\mathbf{J}(i) = j]$ , e.g.
  - individuals across villages
  - regions over time
  - race-by-gender-by-grade cells
- Recall model for group means:

$$\begin{aligned} Y_i &= \sum_{j=1}^J X_{ij} \alpha_j + \varepsilon_i \\ &= \alpha_{\mathbf{J}(i)} + \varepsilon_i \end{aligned}$$

- Can re-index with dual indices:  $i = 1, \dots, N_j$  and

$$Y_{ij} = \alpha_j + \varepsilon_{ij}$$

- Two indices is often enough (rarely need e.g.  $Y_{imst}$ )

# Random Effects

- How to interpret  $Y_{ij} = \alpha_j + \varepsilon_{ij}$ ?
- Two ways: *random effects* and *fixed effects*
- Random effects:  $(\alpha_j, \varepsilon_{ij}) \stackrel{\text{iid}}{\sim} F(\mu, \sigma_\alpha^2) \times G(0, \sigma_\varepsilon^2)$  with
  - $\mathbb{E}[Y_{ij}] = \mathbb{E}[\alpha_j] = \mu$  for all  $j$
  - $\text{Var}(\alpha_j) = \sigma_\alpha^2$  and  $\text{Cov}(\alpha_j, \alpha_{j'}) = 0$  for  $j \neq j'$
  - $\text{Var}(\varepsilon_{ij}) = \sigma_\varepsilon^2$  and  $\text{Cov}(\varepsilon_{ij}, \varepsilon_{i'j'}) = 0$  for all  $i, i'$  and  $j, j'$  except if  $i = i'$  and  $j = j'$
  - $\text{Var}(Y_{ij}) = \sigma_\alpha^2 + \sigma_\varepsilon^2$  and  $\text{Cov}(Y_{ij}, Y_{i'j}) = \sigma_\alpha^2$  for  $i \neq i'$
- Parameters of interest:  $\theta = (\mu, \sigma_\alpha^2, \sigma_\varepsilon^2)$
- Primary interest: *variance decomposition*

# Fixed Effects

- Alternative interpretation of  $Y_{ij} = \alpha_j + \varepsilon_{ij}$
- Assume  $\varepsilon_{ij} \stackrel{\text{iid}}{\sim} G(0, \sigma_\varepsilon^2)$  but no distributional assumption on  $\alpha_j$ :
  - $\mathbb{E}[Y_{ij}] = \alpha_j$  for all  $i, j$
  - $\text{Var}(\varepsilon_{ij}) = \sigma_\varepsilon^2$  and  $\text{Cov}(\varepsilon_{ij}, \varepsilon_{i',j'}) = 0$  for  $(i, j) \neq (i', j')$
  - $\text{Var}(Y_{ij}) = \sigma_\varepsilon^2$
- Parameters of interest:  $\theta = (\alpha_1, \dots, \alpha_J, \sigma_\varepsilon^2)$
- Primary interest: estimating *unobserved heterogeneity*
- Why estimate unobserved heterogeneity? Two reasons:
  - ① Defend CIA (or other research design) against omitted variable bias (OVB):

$$\begin{aligned}\tau_j &\equiv \mathbb{E}[Y_{ij}(1) - Y_{ij}(0) | X_{ij}] \\ &= \mathbb{E}[Y_{ij} | D_{ij} = 1, X_{ij}] - \mathbb{E}[Y_{ij} | D_{ij} = 0, X_{ij}]\end{aligned}$$

- ② Direct interest in group means

# Twins Days in Twinsburg, Ohio



## Example of FE to Support CIA: Twins

- Ashenfelter and Krueger (1994): return to schooling using twins
- Earnings  $Y_{ij}$  as function of schooling  $D_{ij}$  for  $i \in \{1, 2\}$ :

$$Y_{ij} = \alpha_j + D_{ij}\beta + \varepsilon_{ij}$$

where  $\alpha_j$  captures unobserved family background

- Let  $D_{jj}^k$  denote schooling reported by  $k \in \{1, 2\}$
- With sample of twins, several ways to estimate  $\beta$ :
  - ① Include  $\alpha_j$  as FE
  - ② Difference between twins:

$$Y_{1j} - Y_{2j} = (D_{1j}^1 - D_{2j}^2) \beta + u_j,$$

which is algebraically equivalent to FE

- ③ Instrument  $(D_{1j}^1 - D_{2j}^2)$  with  $(D_{1j}^2 - D_{2j}^1)$  since  $D_{ij}^i$  is noisy

# Controlling for Unobserved Family Background

TABLE 3—ORDINARY LEAST-SQUARES (OLS), GENERALIZED LEAST-SQUARES (GLS),  
INSTRUMENTAL-VARIABLES (IV), AND FIXED-EFFECTS ESTIMATES OF LOG WAGE  
EQUATIONS FOR IDENTICAL TWINS<sup>a</sup>

Variable	OLS (i)	GLS (ii)	GLS (iii)	IV <sup>a</sup> (iv)	First difference (v)	First difference by IV (vi)
Own education	0.084 (0.014)	0.087 (0.015)	0.088 (0.015)	0.116 (0.030)	0.092 (0.024)	0.167 (0.043)
Sibling's education	—	—	-0.007 (0.015)	-0.037 (0.029)	—	—
Age	0.088 (0.019)	0.090 (0.023)	0.090 (0.023)	0.088 (0.019)	—	—
Age squared (÷ 100)	-0.087 (0.023)	-0.089 (0.028)	-0.090 (0.029)	-0.087 (0.024)	—	—
Male	0.204 (0.063)	0.204 (0.077)	0.206 (0.077)	0.206 (0.064)	—	—
White	-0.410 (0.127)	-0.417 (0.143)	-0.424 (0.144)	-0.428 (0.128)	—	—
Sample size:	298	298	298	298	149	149
R <sup>2</sup> :	0.260	0.219	0.219	—	0.092	—

## Example of Random Effects: Project STAR

- Chetty et al. (2011): effect of kindergarten classrooms on adult earnings  $Y_{ij}$
- Students & teachers randomly assigned to classrooms
- Can study observable classroom characteristics  $D_j$ , e.g. class size:

$$Y_{ij} = \mu + D_j\beta + \varepsilon_{ij}$$

with  $j \perp \varepsilon_{ij}$  due to randomization

- But can also study unobservable classroom characteristics:  $\alpha_j = D_j\beta + \xi_j$
- Two approaches:
  - ①  $\alpha_j$  as FE: F-test of  $H_0 : \alpha_j = \alpha_k$  for all  $j, k$
  - ②  $\alpha_j$  as RE: Estimate  $\sigma_\alpha^2$  using ANOVA (e.g. Searle et al. 2009)

# Testing for Classroom Effects

TABLE VII  
KINDERGARTEN CLASS EFFECTS: ANALYSIS OF VARIANCE

Dependent variable	(1)	(2)	(3)	(4)	(5)	(6)
	Grade K scores	Grade 8 scores			Wage earnings	
p-value of F test on KG class fixed effects	0.000	0.419	0.047	0.026	0.020	0.040
p-value from permutation test	0.000	0.355	0.054	0.029	0.023	0.055
SD of class effects (RE estimate)	8.77%	0.000%	\$1,497	\$1,520	\$1,703	\$1,454
Demographic controls	x	x		x	x	x
Large classes only					x	
Observable class chars.						x
Observations	5,621	4,448	6,025	6,025	4,208	5,983

*Notes.* Each column reports estimates from an OLS regression of the dependent variable on school and class fixed effects, omitting one class fixed effect per school. The p-value in the first row is for an F test of the joint significance of the class fixed effects. The second row reports the p-value from a permutation test, calculated as follows: we randomly permute students between classes within each school, calculate the F-statistic on the class dummies, repeat the previous two steps 1,000 times, and locate the true F-statistic in this distribution. The third row reports the estimated standard deviation of class effects from a model with random class effects and school fixed effects. Grade 8 scores are available for students who remained in Tennessee public schools and took the eighth-grade standardized test any time between 1990 and 1997. Both KG and eighth-grade scores are coded using within-sample percentile ranks. Wage earnings is the individual's mean wage earnings over years 2005–2007 (including 0s for people with no wage earnings). All specifications are estimated on the subsample of students who entered a STAR school in kindergarten. All specifications except (3) control for the vector of demographic characteristics used in Table IV: a quartic in parent's household income interacted with an indicator for whether the filing parent is ever married between 1996 and 2008, mother's age at child's birth, indicators for parent's 401(k) savings and home ownership, and student's race, gender, free-lunch status, and age at kindergarten. Column (5) limits the sample to large classes only; this column identifies pure KG class effects because students who were in large classes were rerandomized into different classrooms after KG. Column (6) replicates column (4), adding controls for the following observable classroom characteristics: indicators for small class, above-median teacher experience, black teacher, and teacher with degree higher than a BA, and classmates' mean predicted score. Classmates' mean predicted score is constructed by regressing test scores on school-by-entry-grade fixed effects and the vector of demographic characteristics listed above and then taking the mean of the predicted scores.

# Table of Contents

## 1 Fixed Effects

Definitions

Identification and Estimation

Empirical Bayes

## 2 Panel Data

Structure and Definitions

## 3 Applications

Connected-Set Fixed Effects (AKM)

## 4 Appendix

Group Fixed Effects

# Identification of Fixed Effects

- Define  $J \times 1$  vector  $\mathbf{x}_i = [X_{i1}, \dots, X_{iJ}]'$
- Single-observation model:  $Y_i = \alpha_{\mathbf{j}(i)} + \varepsilon_i = \mathbf{x}'_i \boldsymbol{\alpha} + \varepsilon_i$
- Interested in  $\boldsymbol{\alpha} = [\alpha_1, \dots, \alpha_J]',$  the  $J \times 1$  vector of FE parameters
- Construct  $N \times 1$  vector  $\mathbf{X}_j = [X_{1j}, \dots, X_{Nj}]'$  for each  $j$
- Stacked model:  $\mathbf{Y} = \mathbf{X}\boldsymbol{\alpha} + \varepsilon$  where
  - $\mathbf{Y} = [Y_1, \dots, Y_N]'$  is  $N \times 1$  outcome vector
  - $\mathbf{X} = [\mathbf{X}_1, \dots, \mathbf{X}_J]$  is  $N \times J$  design matrix of group indicators
  - $\varepsilon = [\varepsilon_1, \dots, \varepsilon_N]'$  is  $N \times 1$  vector of errors
- Pop OLS:  $\boldsymbol{\alpha} = \mathbb{E} [\mathbf{x}_i \mathbf{x}_i']^{-1} \mathbb{E} [\mathbf{x}_i Y_i]$
- OLS requires two conditions for identification of  $\boldsymbol{\alpha}$ :
  - ①  $\mathbb{E} [\mathbf{x}_i \varepsilon_i] = 0$  ( $j$  conditions): exogenous group assignment
  - ②  $\mathbb{E} [\mathbf{x}_i \mathbf{x}_i']$  (and hence  $\mathbf{X}' \mathbf{X}$ ) is full rank, i.e. no multicollinearity

## Endogenous Group Assignment

- First consider violation of exogenous group assignment
- Important distinction: if we run

$$Y_{ij} = \alpha_j + D_{ij}\beta + \varepsilon_{ij}$$

and  $\mathbb{E} [\alpha_j \varepsilon_{ij}] \neq 0$  for some  $j$  (e.g. selection into group)

- This implies the estimates of  $\alpha_j$  are not identified
- And so OLS estimates  $\hat{\alpha}_j$  would be biased
- However, this is *not* necessarily a problem for identifying  $\beta$
- As long as  $\mathbb{E} [D_{ij} \varepsilon_{ij}] = 0$ , we have that  $\hat{\beta} \xrightarrow{P} \beta$
- Follows from linearity of OLS
- For consistent  $\hat{\alpha}_j$ , need CIA (or design) to argue groups form exogenously

# Rank Violations in Fixed Effects

Three prominent causes of multicollinearity:

- ① Including grand mean and  $J$  FE dummies

- Consider  $Y_{ij} = \alpha_j + \varepsilon_{ij}$  with  $J = 3$  and  $N_j = 2$ :

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\alpha} + \boldsymbol{\varepsilon}$$

$$\begin{bmatrix} Y_{11} \\ Y_{21} \\ Y_{12} \\ Y_{22} \\ Y_{13} \\ Y_{23} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{21} \\ \varepsilon_{12} \\ \varepsilon_{22} \\ \varepsilon_{13} \\ \varepsilon_{23} \end{bmatrix}$$

# Over-Parametrization By Including Constant

- However, if we also include constant

$$\begin{bmatrix} Y_{11} \\ Y_{21} \\ Y_{12} \\ Y_{22} \\ Y_{13} \\ Y_{23} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{21} \\ \varepsilon_{12} \\ \varepsilon_{22} \\ \varepsilon_{13} \\ \varepsilon_{23} \end{bmatrix}$$

- Then  $\theta = (\mu, \alpha_1, \alpha_2, \alpha_3)$  is not identified – why?
- Intuition: if you tell me  $(\alpha_1, \alpha_2, \alpha_3)$ , I can compute  $\mu$
- Similarly, if I know  $(\mu, \alpha_1, \alpha_2)$ , I can compute  $\alpha_3$
- Linear algebra:  $\text{rank } (\mathbf{X}'\mathbf{X}) = 3$  but  $\mathbf{X}'\mathbf{X}_{4 \times 4}$ , so below full rank
- This determines whether STATA drops coefficients!

# Group-Invariant Covariates

## ② Including group-invariant covariates:

- Let  $Y_{ij} = \alpha_j + W_j\gamma + \varepsilon_{ij}$
- Then  $\theta = (\gamma, \alpha_1, \alpha_2, \alpha_3)$  is not identified in:

$$\begin{bmatrix} Y_{11} \\ Y_{21} \\ Y_{12} \\ Y_{22} \\ Y_{13} \\ Y_{23} \end{bmatrix} = \begin{bmatrix} W_1 & 1 & 0 & 0 \\ W_1 & 1 & 0 & 0 \\ W_2 & 0 & 1 & 0 \\ W_2 & 0 & 1 & 0 \\ W_3 & 0 & 0 & 1 \\ W_3 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \gamma \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{21} \\ \varepsilon_{12} \\ \varepsilon_{22} \\ \varepsilon_{13} \\ \varepsilon_{23} \end{bmatrix}$$

- Note that  $W_j$  is just a rescaled version of  $\alpha_j$

## ③ Nested fixed effects:

- For example, consider  $Y_{ijt} = \alpha_j + \xi_{jt} + \varepsilon_{ijt}$
- $\alpha_j$  is group constant;  $\xi_{jt}$  are within-group FE

# Estimation: Least-Square Dummy Variables as Frisch-Waugh-Lovell

- Imagine  $Y_{ij} = \alpha_j + D_{ij}\beta + \varepsilon_{ij}$  and we want  $\hat{\beta}$
- Let  $X_{ij} = 1[i \in j]$  as before
- Consider the following two linear projections:
  - Long regression:  $\mathbb{E}^* [Y_{ij}|D_{ij}, X_{i1}, \dots, X_{iJ}] = D_{ij}\beta + \sum_j X_{ij}\alpha_j$
  - Flipped auxiliary:  $\mathbb{E}^* [D_{ij}|X_{i1}, \dots, X_{iJ}] = \sum_j X_{ij}\alpha_j$
- Then by FWL:  $\beta = \mathbb{E} [\tilde{D}_{ij}\tilde{D}'_{ij}]^{-1} \mathbb{E} [\tilde{D}_{ij}\tilde{Y}_{ij}]$  where
  - $\tilde{D}_{ij} = D_{ij} - \mathbb{E}^* [D_{ij}|X_{i1}, \dots, X_{iJ}]$
  - $\tilde{Y}_{ij} = Y_{ij} - \mathbb{E}^* [Y_{ij}|X_{i1}, \dots, X_{iJ}]$
- But note that these are just group-means regressions:
  - $\mathbb{E}^* [D_{ij}|X_{i1}, \dots, X_{iJ}] = \bar{D}_j = \frac{1}{N_j} \sum_i D_{ij}$
  - $\mathbb{E}^* [Y_{ij}|X_{i1}, \dots, X_{iJ}] = \bar{Y}_j = \frac{1}{N_j} \sum_i Y_{ij}$
- So can do  $Y_{ij} - \bar{Y}_j = (D_{ij} - \bar{D}_j)\beta + \xi_{ij}$  - often much faster!
- This is exactly what `reghdfe` does in STATA

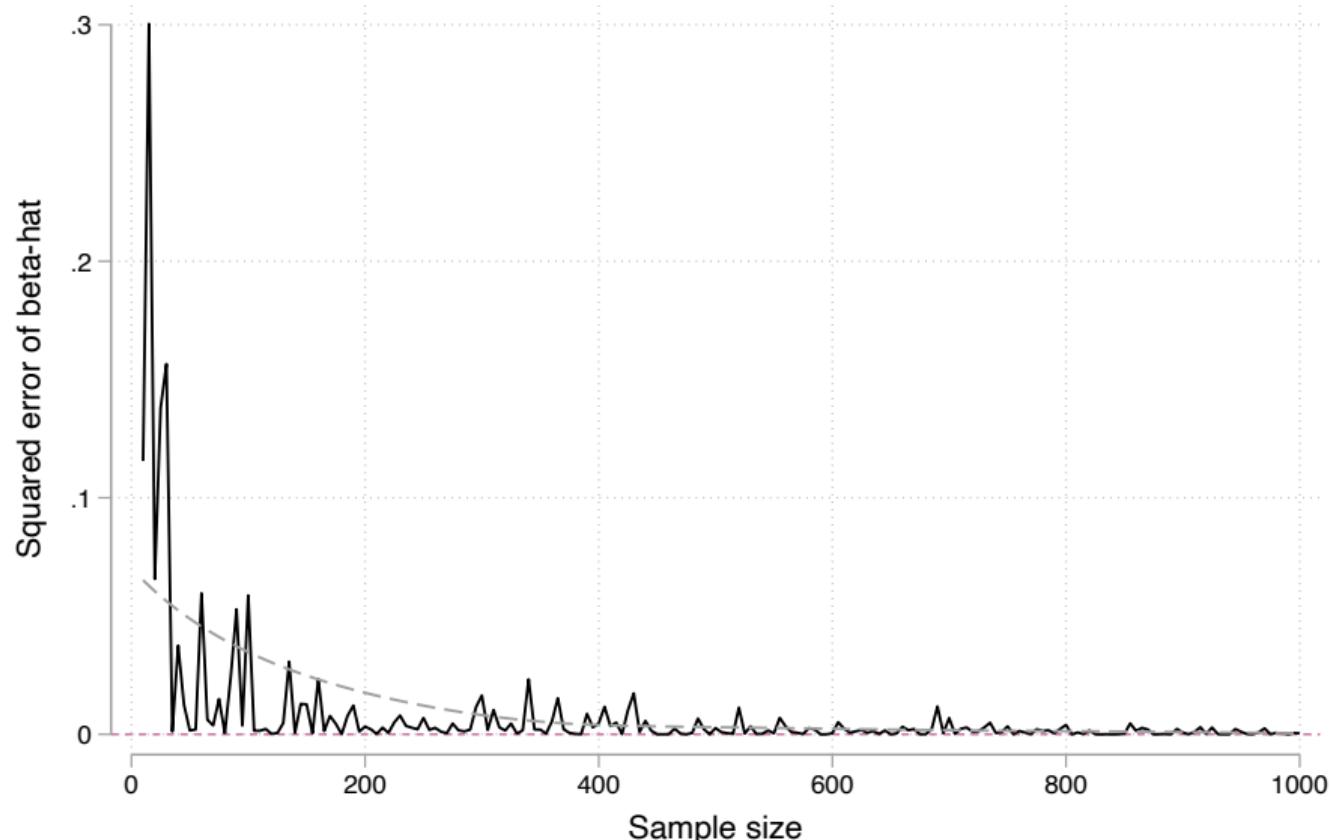
## FE: Unbiased but Not Generally Consistent

- For fixed  $\mathbf{X}_j$  and  $\mathbb{E}[\mathbf{x}_i \varepsilon_i] = 0$ , OLS is unbiased:  $\mathbb{E}[\hat{\alpha}_{OLS}] = \alpha$
- But  $N_j = J/N$  is often constant as  $N$  grows because  $J$  grows as well, e.g.
  - Classrooms:  $N_j \approx 20$  fixed even if  $J$  large
  - Short panels e.g.  $N_j \approx 5$
- Then OLS is *not* consistent:  $\hat{\alpha}_{OLS} \xrightarrow{P} \alpha$
- Illustration of this “incidental parameters problem”:
  - Consider  $Y_{ij} = \alpha_j + D_{ij}\beta + \varepsilon_{ij}$  with  $N_j = 2$
  - Assume  $\alpha_j \sim N(0, 1)$ ,  $\beta = 1$ , and  $\varepsilon_{ij} \sim N(0, 1)$ , all iid
  - Do this for  $J \in [10, 15, 20, \dots, 1000]$
  - Evaluate for each sample size:

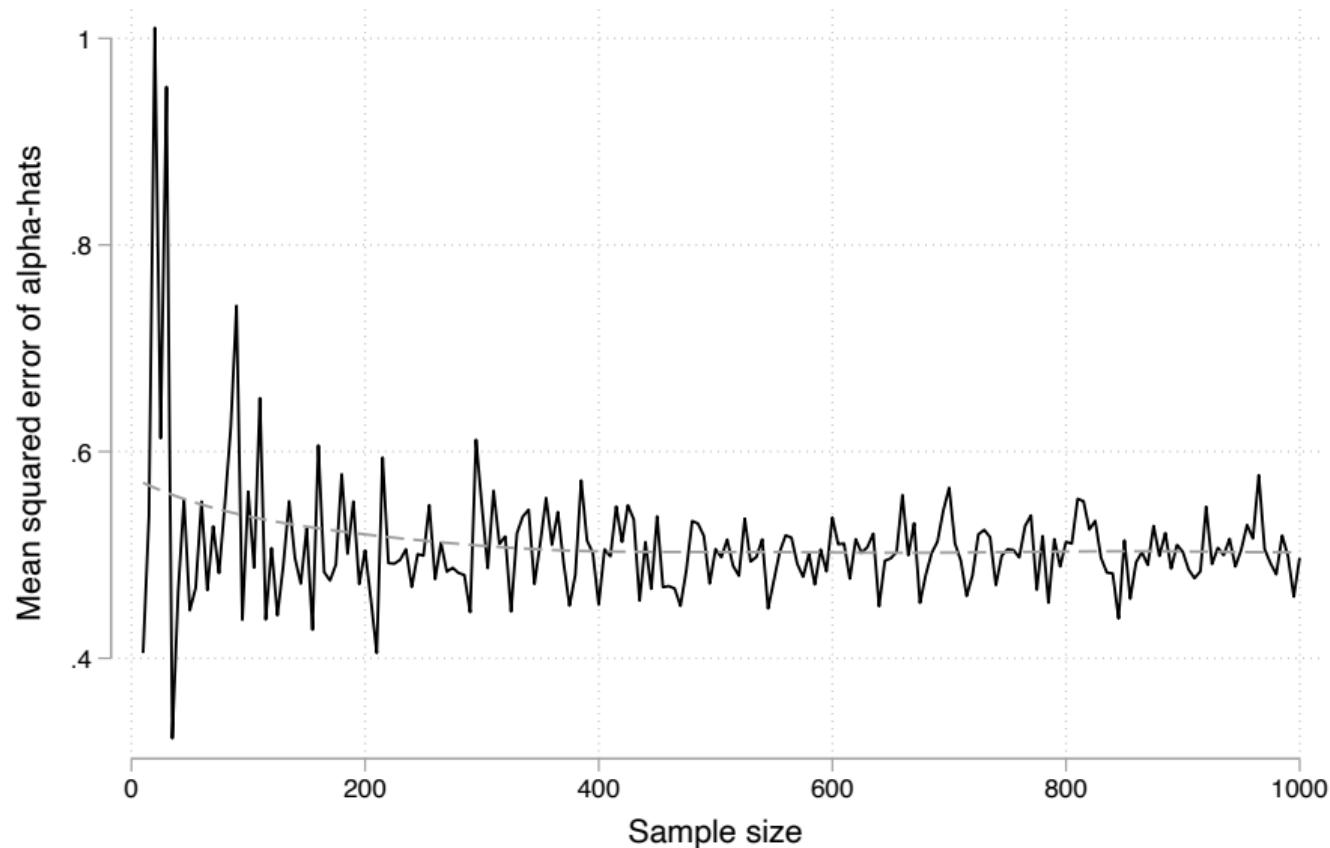
$$\text{Squared error of } \hat{\beta}_{OLS} = (\hat{\beta}_{OLS} - \beta)^2$$

$$\text{Mean squared error of } \hat{\alpha}_{OLS} = \frac{1}{J} \sum_{j=1}^J (\hat{\alpha}_{j,OLS} - \alpha_j)^2$$

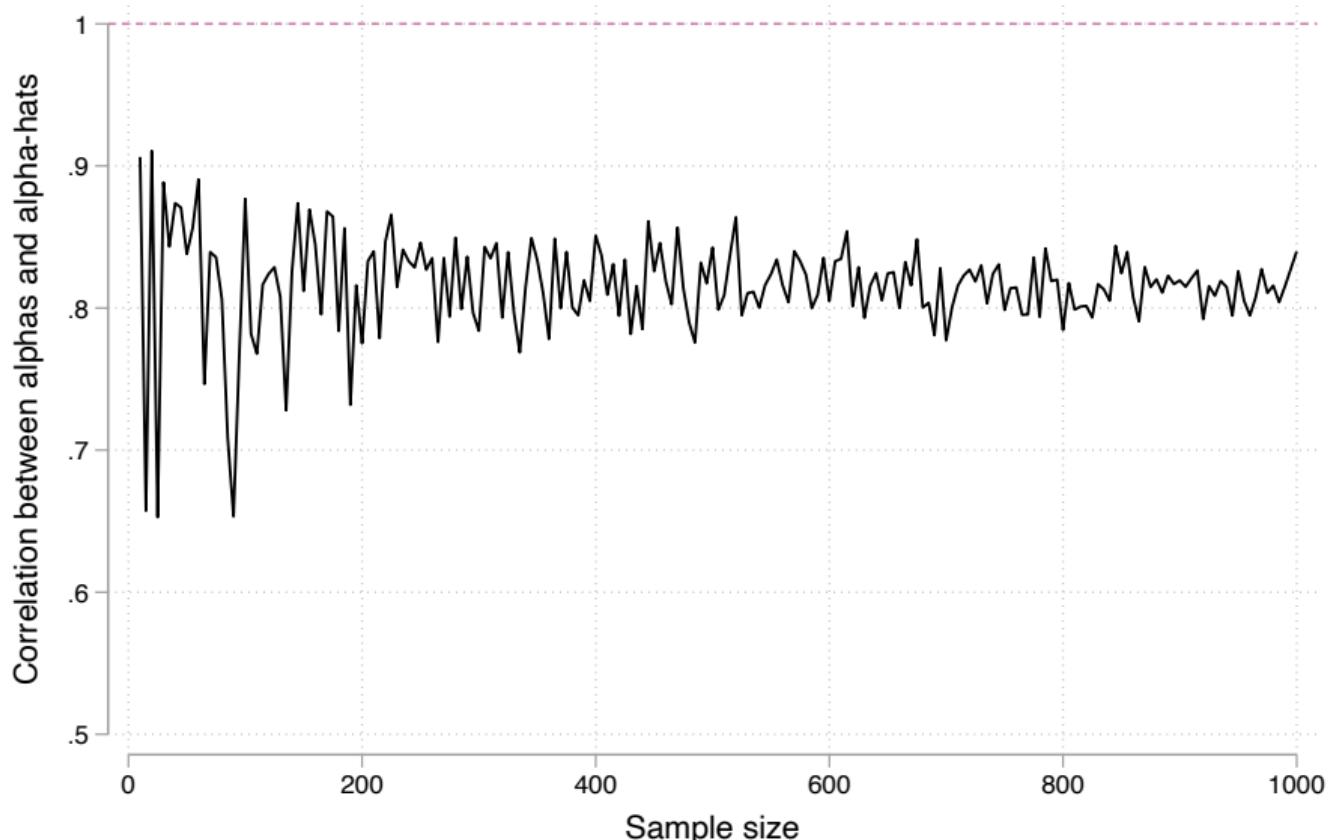
# $\hat{\beta}_{OLS}$ Converges to True Value



But  $\hat{\alpha}_{j,\text{OLS}}$  Does Not



And so  $\text{Corr}(\hat{\alpha}_{j,\text{OLS}}, \alpha_j)$  Never Goes to One



# Table of Contents

## 1 Fixed Effects

Definitions

Identification and Estimation

Empirical Bayes

## 2 Panel Data

Structure and Definitions

## 3 Applications

Connected-Set Fixed Effects (AKM)

## 4 Appendix

Group Fixed Effects

## Shrinking Group Means Towards Grand Mean

- How can we make many group means (or other coefficients) less noisy?
- Consider wages  $Y_{ij}$  for  $N_j = N$  workers in  $J$  regions

$$Y_{ij} = \alpha_j + \varepsilon_{ij}$$

- Let  $\bar{Y}_j = \frac{1}{N} \sum_{i=1}^N Y_{ij}$  be regional average wage
- Want to know  $\alpha_j$  that  $\bar{Y}_j$  would reach when  $N$  large
- Impose parametric restrictions on distribution of  $\alpha_j \rightarrow$  random effects
- Assume

$$\bar{Y}_j | \alpha_j \sim N \left( \alpha_j, \frac{\sigma_\varepsilon^2}{N} \right)$$

where  $\sigma_\varepsilon^2 = \text{Var}(\varepsilon_{ij} | \alpha_j)$  and

$$\alpha_j \sim N(\mu, \sigma_\alpha^2)$$

- Sometimes referred to as mixing distribution

# Posterior of Random Effects

- It can be shown that in this model:

$$\mathbb{E} [\alpha_j | \bar{Y}_j] = \mu + \left( \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \frac{\sigma_\varepsilon^2}{N}} \right) (\bar{Y}_j - \mu)$$

- Intuition:
  - Linear projection of  $\alpha_j$  on  $\bar{Y}_j$  (infeasible)
  - Shrink sample average towards prior mean  $\mu$
  - More weight on  $\bar{Y}_j$  if signal-to-noise ratio  $\lambda = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \frac{\sigma_\varepsilon^2}{N}}$  is high
- Problem: not useful if we don't know  $(\sigma_\alpha^2, \sigma_\varepsilon^2, \mu)$
- Two approaches:
  - ① Bayes: choose  $(\sigma_\alpha^2, \mu)$  based on prior knowledge
  - ② Empirical Bayes: estimate these hyperparameters

# Empirical Bayes Approach

- Estimate  $\mu$  with grand mean:

$$\hat{\mu} = \frac{1}{J} \sum_{j=1}^J \bar{Y}_j$$

- Estimate  $\sigma_\varepsilon^2$  from the “within” variance:

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{J(N-1)} \sum_{j=1}^J \sum_{i=1}^N (Y_{ij} - \bar{Y}_j)^2$$

- Estimate  $\sigma_\alpha^2$  from “overdispersion” in averages:

$$\hat{\sigma}_\alpha^2 = \frac{1}{J-1} \sum_{j=1}^J (\bar{Y}_j - \hat{\mu})^2 - \frac{\hat{\sigma}_\varepsilon^2}{N}$$

- Thus, the empirical Bayes posterior estimate:

$$\widehat{\mathbb{E}} [\alpha_j | \bar{Y}_j] = \hat{\alpha}_j = \hat{\mu} + \frac{\hat{\sigma}_\alpha^2}{\hat{\sigma}_\alpha^2 + \frac{\hat{\sigma}_\varepsilon^2}{N}} (\bar{Y}_j - \hat{\mu})$$

# When is Empirical Bayes Useful?

- Why might  $\hat{\alpha}_j$  be a “better” estimate for  $\alpha_j$  than  $\bar{Y}_j$ ?
- Bias-variance tradeoff: It can be shown for large  $N$  (James-Stein 1961):

$$\mathbb{E} [(\hat{\alpha}_j - \alpha_j)^2 | \alpha_j] \approx \underbrace{(\mu - \alpha_j)^2 (1 - \lambda)^2}_{\text{bias}} + \underbrace{\lambda^2 \frac{\sigma_\varepsilon^2}{N}}_{\text{variance}}$$

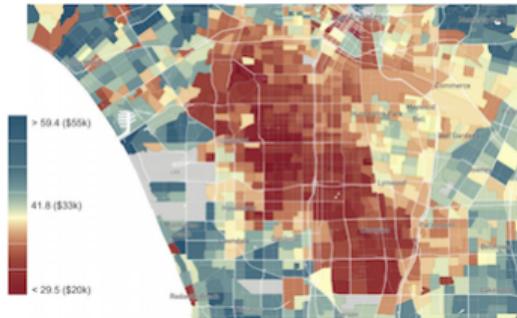
and if  $J \geq 4$ ,  $(\hat{\alpha}_1, \dots, \hat{\alpha}_J)$  has lower MSE than  $(\bar{Y}_1, \dots, \bar{Y}_J)$  for  $(\alpha_1, \dots, \alpha_J)$

- Meaning: worth accepting some bias for lower variance
- Bottom line: shrink when want to estimate many means
- When should I estimate many means? For example:
  - Forecasting / prediction (e.g. map of small area statistics)
  - Input into subsequent regressions
  - Decision making (e.g. which teachers should be fired?)

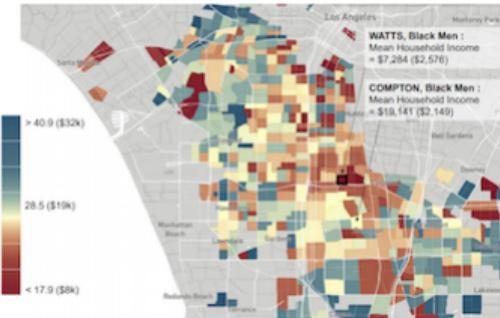
# Prediction: Mapping Small Areas

Chetty et al. (2020)

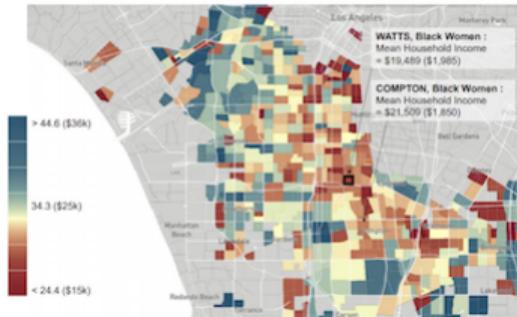
A. All Children: Household Income Given Parents at 25th Percentile



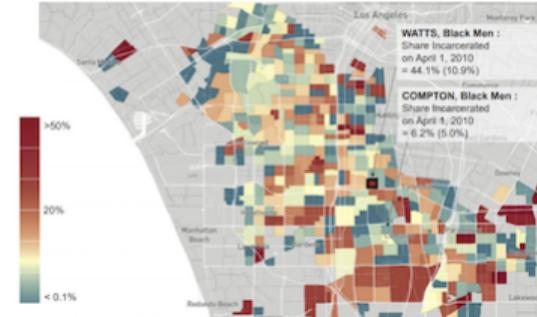
B. Black Men: Household Income Given Parents at 25th Percentile



C. Black Women: Household Income Given Parents at 25th Percentile

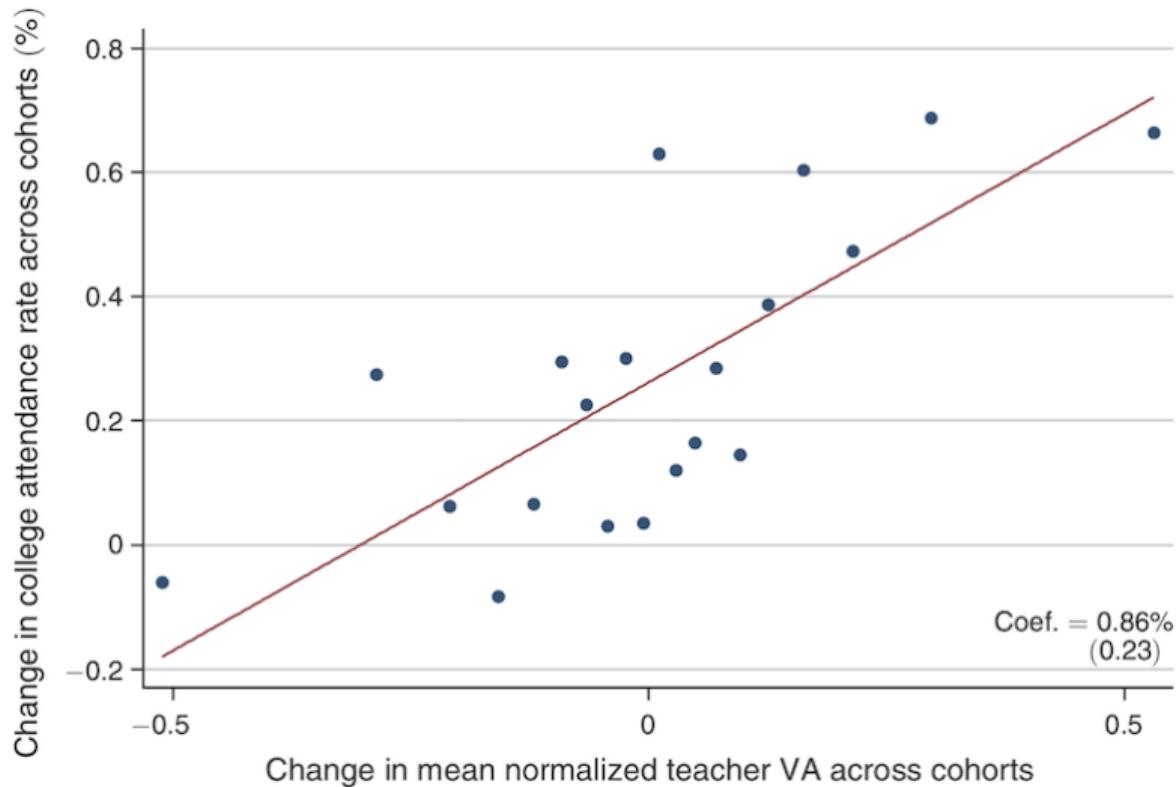


D. Black Men: Incarceration Rates Given Parents at 1st Percentile



# Input into Regressions: Teacher Value Added

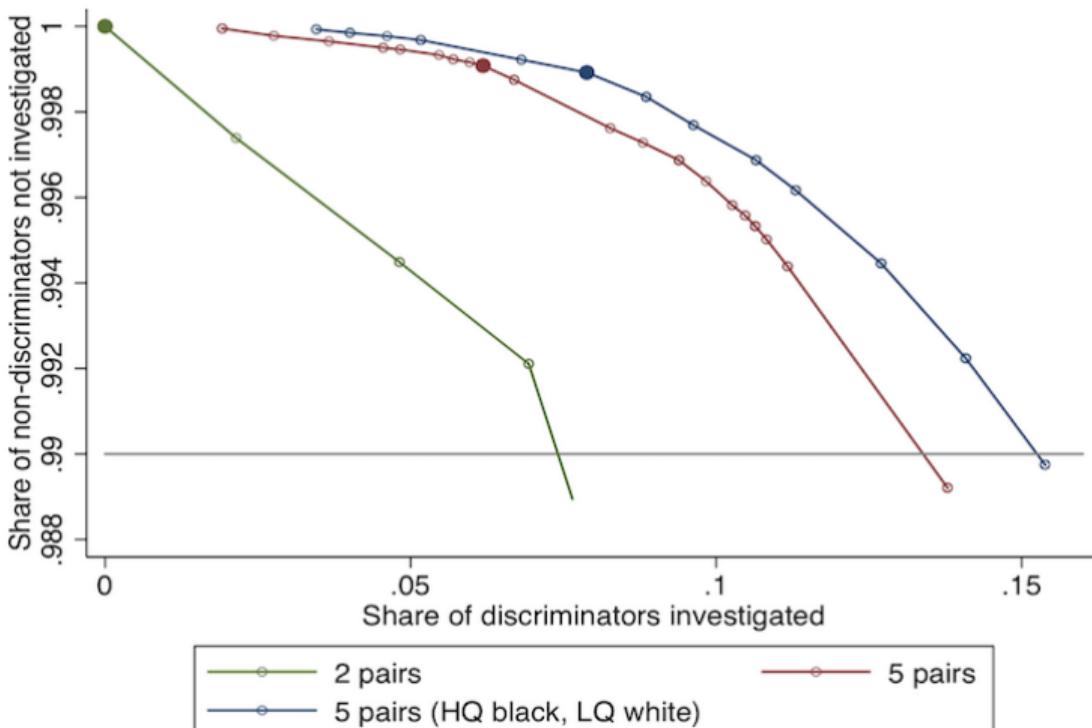
Chetty et al. (2014)



# Decision Rule: Auditing Discriminating Employers

Kline and Walters (2020)

Figure V: Detection/error tradeoffs, NPRS data



# Table of Contents

## 1 Fixed Effects

Definitions

Identification and Estimation

Empirical Bayes

## 2 Panel Data

Structure and Definitions

## 3 Applications

Connected-Set Fixed Effects (AKM)

## 4 Appendix

Group Fixed Effects

# Panel Data Formats

- Careful! We now switch notation:
- Let *units* be indexed by  $i = 1, \dots, N$
- And *periods* indexed by  $t = 1, \dots, T$
- Together, they form *observations*  $(i, t)$
- Two panel data arrangements (“formats”): *Long* and *wide*

# Long Panel

$i$	$t$	$Y_{it}$	$X_{it}$
1	1	5	1
1	2	4	0
1	3	3	0
:	:	:	:
1	$T$	4	1
2	1	3	0
2	2	4	1
:	:	:	:

## Wide Panel

$i$	$Y_{i1}$	$Y_{i2}$	$Y_{i3}$	...	$Y_{iT}$	$X_{i1}$	$X_{i2}$	$X_{i3}$	...
1	5	4	3	...	4	1	0	0	...
2	3	4	...	...	...	0	1	...	...
:	:	:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:	:	:

- Panel data can be *reshaped* between long and wide
- Each format has its uses

## Panel Completeness

- Panel data exist to varying degrees of completeness
- $N \times T$  wide panel  $\mathbf{Y}$  with  $\circ$  observed and  $\bullet$  missing
- In a *strongly balanced* panel, we can observe every entry:

$$\mathbf{Y} = \begin{bmatrix} \circ & \circ & \circ & \circ & \circ & \circ & \circ \\ \circ & \circ & \circ & \circ & \circ & \circ & \circ \\ \circ & \circ & \circ & \circ & \circ & \circ & \circ \\ \circ & \circ & \circ & \circ & \circ & \circ & \circ \end{bmatrix}$$

- *Weakly balanced* panels may or may not have gaps:

$$\begin{bmatrix} \bullet & \bullet & \circ & \circ & \circ & \circ & \bullet \\ \circ & \circ & \circ & \circ & \bullet & \bullet & \bullet \\ \bullet & \circ & \circ & \circ & \circ & \bullet & \bullet \\ \bullet & \bullet & \circ & \circ & \circ & \circ & \bullet \end{bmatrix}$$

and

$$\begin{bmatrix} \bullet & \bullet & \circ & \circ & \circ & \circ & \bullet \\ \bullet & \bullet & \circ & \circ & \bullet & \circ & \circ \\ \circ & \circ & \circ & \bullet & \circ & \bullet & \bullet \\ \bullet & \circ & \circ & \circ & \circ & \bullet & \bullet \end{bmatrix}$$

# Unbalanced Panels and Repeated Cross-Sections

- *Unbalanced panels:*

$$\begin{bmatrix} \bullet & \bullet & \circ & \bullet & \bullet & \circ & \bullet \\ \bullet & \bullet & \bullet & \circ & \bullet & \circ & \circ \\ \circ & \circ & \circ & \bullet & \circ & \bullet & \bullet \\ \circ & \circ & \circ & \circ & \circ & \bullet & \bullet \end{bmatrix}$$

- *Repeated cross-sections:*

$$\begin{bmatrix} \bullet & \bullet & \bullet & \bullet & \bullet & \circ & \bullet \\ \bullet & \bullet & \bullet & \bullet & \bullet & \circ & \bullet \\ \bullet & \bullet & \circ & \bullet & \bullet & \bullet & \bullet \\ \circ & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \end{bmatrix}$$

which we could treat as a *pooled cross-section*

# Cohort Panels

- Can often form balanced panel from repeated cross-section
- How? Define cohorts  $c \in \{1, \dots, C\}$  e.g. birth year

$i$	$t$	$Y_{it}$	$X_{it}$	$c$	$t$	$\bar{Y}_{ct}$
1	1	5	2	1	1	4.2
2	1	4	1	1	2	3.8
3	1	3	1	:	:	:
:	:	:	:	1	$T$	4.1
$i$	2	4	2	2	2	3.4
:	:	3	3	:	:	:
$N$	$T$	4	1	$C$	$T$	3.2

# Table of Contents

## 1 Fixed Effects

Definitions

Identification and Estimation

Empirical Bayes

## 2 Panel Data

Structure and Definitions

## 3 Applications

Connected-Set Fixed Effects (AKM)

## 4 Appendix

Group Fixed Effects

# Units Moving Across Groups over Time

- Extremely useful to observe  $i$  in different groups over  $t$
- Separately identifies unit from group effect
- Consider the following model (Abowd et al. 1999, "AKM")

$$Y_{it} = \alpha_i + \psi_{\mathbf{J}(i,t)} + \mathbf{X}'_{it}\beta + \varepsilon_{it}$$

where  $\mathbf{J} : \mathcal{N} \times \mathcal{T} \rightarrow \mathcal{J}$  assigns  $i$  in  $t$  to  $j \in \mathcal{J} = \{1, \dots, J\}$

- Example: firm of workers in a panel
- Model is isomorphic to standard model but  $J$  treatments
- Interested in  $\psi_j$  or  $Var(\psi_j)$
- E.g. how much variation in wages explained by firms?

# Identification of AKM

- Stacked model:

$$\mathbf{Y} = \mathbf{D}\alpha + \mathbf{F}\psi + \mathbf{X}\xi + \varepsilon$$

- Identification requires

$$\mathbb{E} [\mathbf{D}'\varepsilon] = 0, \quad \mathbb{E} [\mathbf{F}'\varepsilon] = 0, \quad \mathbb{E} [\mathbf{X}'\varepsilon] = 0$$

- Rank condition: set one  $\psi_j = 0$  in each “connected set”
- “Exogenous mobility”:  $\mathbb{E} [\mathbf{F}'\varepsilon] = 0$ 
  - $\Pr (\mathbf{J}(i, t) = j | \varepsilon) = \Pr (\mathbf{J}(i, t) = j) = G_{jt}(\alpha_i, \psi_1, \dots, \psi_J)$
  - Does *not* mean that workers must move randomly across firms
  - It does also *not* require turnover to be random across  $i$
  - But instead, requires no sorting due to wage trends or match quality

## Estimation of AKM

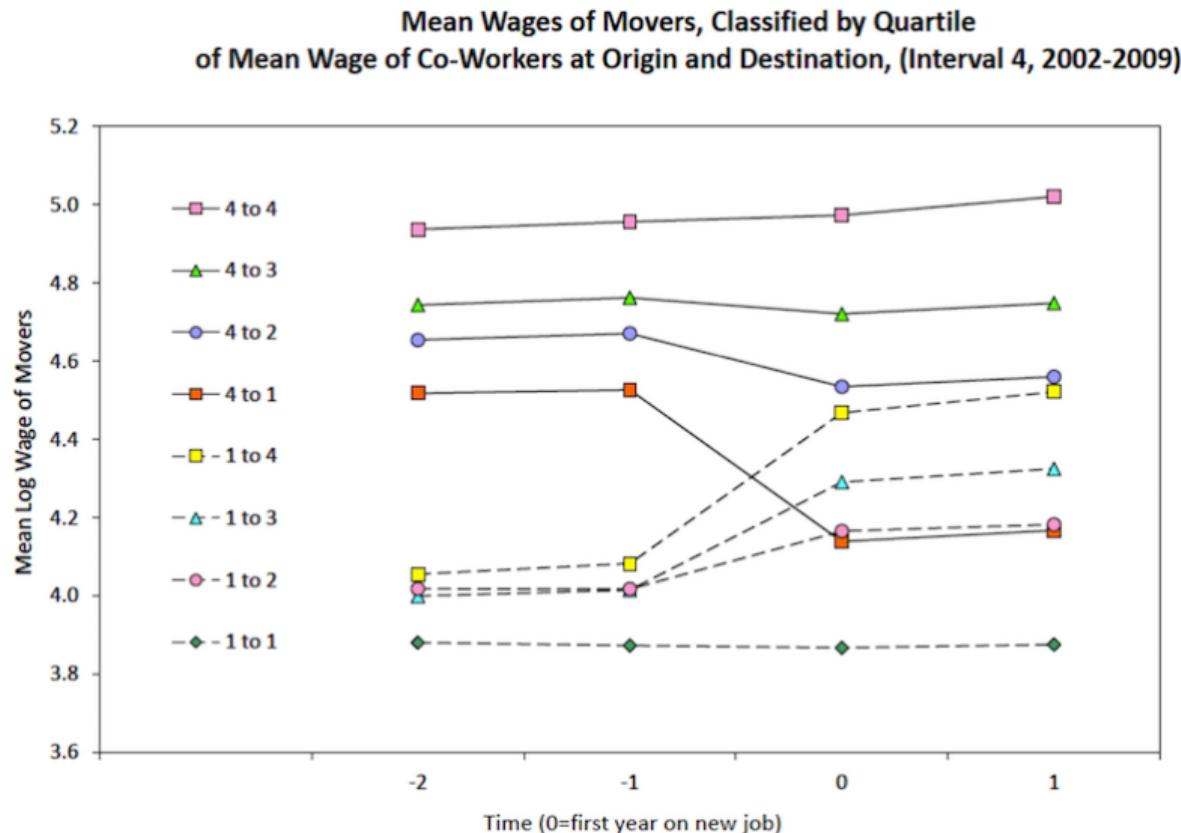
- Inspect stacked model: can estimate via OLS!
  - Worker dummies to estimate  $\alpha_i$ ;
  - Firm dummies to estimate  $\psi_j$
- Recall:  $(\hat{\alpha}, \hat{\psi})$  are inconsistent unless  $T$  gets large
- Can estimate  $\text{Var}(\hat{\psi}_j) = \hat{\sigma}_{\psi}^2$  via Method of Moments:

$$\hat{\sigma}_{\psi}^2 = \frac{1}{NT - 1} \hat{\psi}' \mathbf{F}' \mathbf{Q} \mathbf{F} \hat{\psi}$$

where  $\mathbf{Q}$  is demeaning matrix

- But this estimator is upward biased
- Unbiased estimate via leave-one-out (Kline et al. 2020)

# Ruling Out Moving on Trends or Match Effects: Card et al. (2013)



# Rising Share of Wage Variation due to Firms

TABLE III  
ESTIMATION RESULTS FOR AKM MODEL, FIT BY INTERVAL

	(1) Interval 1 1985–1991	(2) Interval 2 1990–1996	(3) Interval 3 1996–2002	(4) Interval 4 2002–2009
<b>Person and establishment parameters</b>				
Number person effects	16,295,106	17,223,290	16,384,815	15,834,602
Number establishment effects	1,221,098	1,357,824	1,476,705	1,504,095
<b>Summary of parameter estimates</b>				
Std. dev. of person effects (across person-year obs.)	0.289	0.304	0.327	0.357
Std. dev. of establ. Effects (across person-year obs.)	0.159	0.172	0.194	0.230
Std. dev. of Xb (across person-year obs.)	0.121	0.088	0.093	0.084
Correlation of person/establ. Effects (across person-year obs.)	0.034	0.097	0.169	0.249
Correlation of person effects/Xb (across person-year obs.)	-0.051	-0.102	-0.063	0.029
Correlation of establ. effects/Xb (across person-year obs.)	0.057	0.039	0.050	0.112
RMSE of AKM residual	0.119	0.121	0.130	0.135
Adjusted R-squared	0.896	0.901	0.909	0.927
<b>Comparison match model</b>				
RMSE of match model	0.103	0.105	0.108	0.112
Adjusted $R^2$	0.922	0.925	0.937	0.949
Std. dev. of match effect*	0.060	0.060	0.072	0.075
<b>Addendum</b>				
Std. dev. log wages	0.370	0.384	0.432	0.499
Sample size	84,185,730	88,662,398	83,699,582	90,615,841

# Table of Contents

## 1 Fixed Effects

Definitions

Identification and Estimation

Empirical Bayes

## 2 Panel Data

Structure and Definitions

## 3 Applications

Connected-Set Fixed Effects (AKM)

## 4 Appendix

Group Fixed Effects

# Grouping Unobserved Heterogeneity

- Consider model with unobserved groups:

$$Y_{it} = \phi_{g(i),t} + D_{it}\beta + \varepsilon_{it}$$

where  $g : \{1, \dots, N\} \rightarrow \{1, \dots, G\}$  is not known

- If we knew  $\beta$  and  $\boldsymbol{\phi} = [\phi_{1,1}, \dots, \phi_{G,T}]$ , we could estimate:

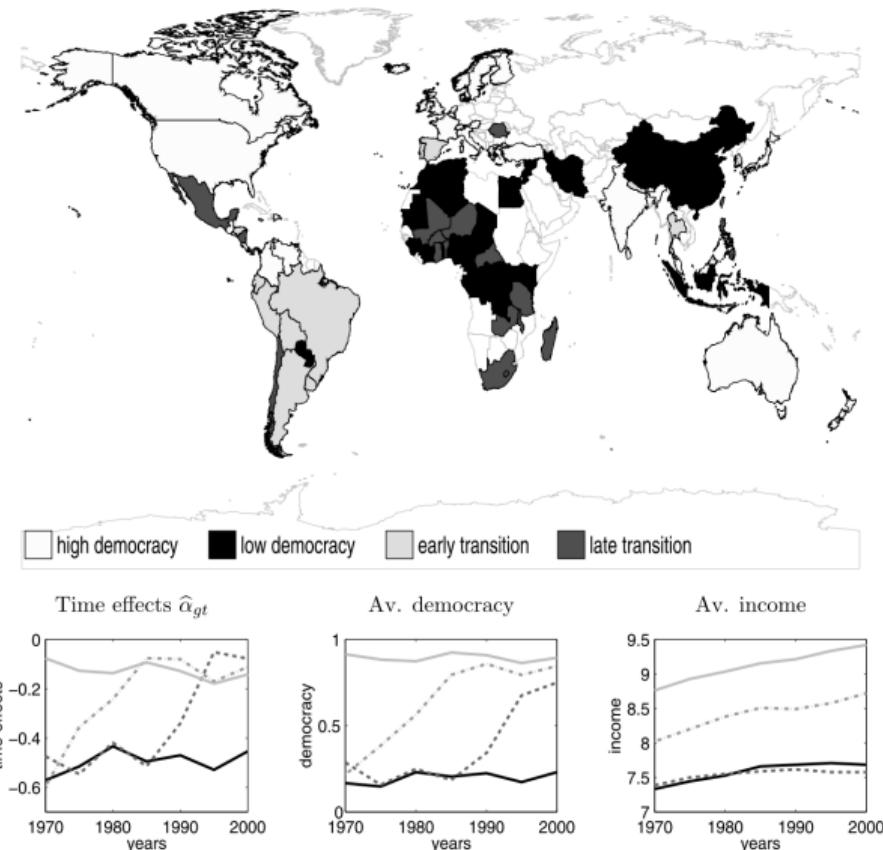
$$\hat{g}(i; \beta, \boldsymbol{\phi}) = \arg \min_{g=\{1, \dots, G\}} \sum_{t=1}^T (Y_{it} - D_{it}\beta - \phi_{g,t})^2$$

- Since we don't know them, we estimate:

$$(\hat{\beta}, \hat{\boldsymbol{\phi}}, \hat{\gamma}) = \arg \min_{(\beta, \boldsymbol{\phi}, \gamma)} \sum_{i=1}^N \sum_{t=1}^T (Y_{it} - D_{it}\beta - \phi_{g,t})^2$$

for  $\gamma = [g(1), \dots, g(N)]$

# Bonhomme and Manresa (2015): Group FE



# Econometrics II

## Lecture 9: Static Difference-in-Differences

David Schönholzer

Stockholm University

April 30, 2024

# Plan for Today

- 1 Causal Effects in Panel Settings
  - Treatment Structures
- 2 The  $2 \times 2$  Difference-in-Differences Design
  - Identification in  $2 \times 2$  DID
  - Estimation of  $2 \times 2$  DID
- 3 Generalized DID Designs
  - Estimation of  $2 \times T$  DID
  - Static Effects in Staggered DID
  - Inference in DID
- 4 Appendix
  - Static Effects in  $2 \times T$  Designs

# Table of Contents

- 1 Causal Effects in Panel Settings
  - Treatment Structures
- 2 The  $2 \times 2$  Difference-in-Differences Design
  - Identification in  $2 \times 2$  DID
  - Estimation of  $2 \times 2$  DID
- 3 Generalized DID Designs
  - Estimation of  $2 \times T$  DID
  - Static Effects in Staggered DID
  - Inference in DID
- 4 Appendix
  - Static Effects in  $2 \times T$  Designs

# Definition of Treatment Structure

- Let  $i \in \mathcal{N} = \{1, \dots, N\}$  and  $t \in \mathcal{T} = \{1, \dots, T\}$
- Consider only balanced panel and  $D_{it} \in \{0, 1\}$  throughout
- A unit's *treatment path* is  $1 \times T$  vector  $\mathbf{D}_i = (D_{i1}, \dots, D_{iT})$ 
  - For example,  $\mathbf{D}_i = (0, 1, 0, 1)$  or  $(1, 1, 0, 0)$
- Say units  $i$  and  $i'$  are in the same *group* if  $\mathbf{D}_i = \mathbf{D}_{i'}$ 
  - Let  $G_i \in \mathcal{G} \subseteq \{1, \dots, T, \infty\}$  denote the group  $i$  is in
  - For example,  $\mathcal{G} = \{1, 2\}$
  - Let  $\mathbf{D}(g)$  be the treatment path of group  $g \in \mathcal{G}$
- Define the  $|\mathcal{G}| \times T$  *grouped treatment structure*  $\mathbf{D}$  as

$$\mathbf{D} = \begin{bmatrix} \mathbf{D}(1) \\ \vdots \\ \mathbf{D}(G) \end{bmatrix} \stackrel{\text{e.g.}}{=} \begin{bmatrix} 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 \end{bmatrix}$$

# Block Structure

- A *block structure* consists of only two groups:
  - ① Treatment group  $G_i = g$ :  $D_{it} = 0$  for  $t < g$ ;  $D_{it} = 1$  for  $t \geq g$
  - ② Control group  $G_i = \infty$ :  $D_{it} = 0$  for all  $t$
- Example:  $2 \times 3$  block structure with  $g = 2$ :

$$\mathbf{D} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 1 \end{bmatrix}$$

- Define
  - Indicator for ever-treated units:  $D_i \equiv \max_t D_{it}$
  - So if  $D_i = 1$ , then  $G_i < \infty$  and if  $D_i = 0$ , then  $G_i = \infty$
  - Indicator for post-treated periods:  $P_t \equiv \max_i D_{it}$
- Typically reverse-sort  $\mathbf{D}(g)$  by  $g$ , e.g.  $\{\infty, 2\}$
- Lemma: Assume there is more than one group.  
Then: Treatment structure is block structure iff  $D_{it} = D_i P_t$

# Staggered Rollout Structure

- *Staggered rollout structure*: at least two groups:
  - Groups  $G_i = g$  have path  $D_{it} = 0$  for  $t < g$ ;  $D_{it} = 1$  for  $t \geq g$
  - There may or may not be a control group  $G_i = \infty$
- Could also be called *absorbing structure*:
  - $D_{is} \leq D_{it}$  for  $s < t$
  - Can define group indices as  $G_i = \arg \min_t D_{it}$
- *Example*:  $3 \times 4$  staggered rollout structure:

$$\mathbf{D} = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

- *Remark 1*: Block structure is special case with  $G_i \in \{g, \infty\}$
- *Remark 2*: In SRS,  $\mathbf{D}$  is just a function of  $\mathcal{G}$  and  $T$ 
  - In the example above,  $T = 4$  and  $G_i \in \{1, 2, 4\}$

# Time Series Structures

- *Time series structure* has only one group

$$G_i = g \text{ with } 1 < g \leq T$$

- For example  $1 \times 4$  with  $G_i = 3$  for all  $i$ :

$$\mathbf{D} = \begin{bmatrix} 0 & 0 & 1 & 1 \end{bmatrix}$$

- Can we ever have cross-sectional variation?
- Sometimes: cohort panel with eligibility varying by cohort
- Example: Petra Persson (2020)
  - Social insurance reform in Sweden in 1988
  - Cohorts based on birth quarter of first child
  - National reform is staggered rollout for cohorts!

# Assumptions for Static Causal Effects

- Start off with *static and homogeneous causal effects*
- Define potential outcomes  $Y_{it}(d)$  with  $d \in \{0, 1\}$
- Built-in assumptions by writing  $Y_{it}(D_{it})$ :
  - SUTVA in panels: no contamination across units
  - No anticipation/memory: no contamination across periods
- Individual causal effect:  $Y_{it}(1) - Y_{it}(0)$
- ATE:  $\mathbb{E}[Y_{it}(1) - Y_{it}(0)]$
- **Estimand is ATT:**  $\tau \equiv \mathbb{E}[Y_{it}(1) - Y_{it}(0)|D_{it} = 1]$

# Table of Contents

## 1 Causal Effects in Panel Settings

Treatment Structures

## 2 The $2 \times 2$ Difference-in-Differences Design

Identification in  $2 \times 2$  DID

Estimation of  $2 \times 2$  DID

## 3 Generalized DID Designs

Estimation of  $2 \times T$  DID

Static Effects in Staggered DID

Inference in DID

## 4 Appendix

Static Effects in  $2 \times T$  Designs

## Basic Idea

- Origin: Snow (1849); prominence: Card and Krueger (1994)
- Effect of minimum wage increase  $D_{it}$  on employment  $Y_{it}$
- New Jersey (NJ) increases minimum wage in mid-1992
  - Call early 1992  $t = 0$  and late 1992  $t = 1$
  - NJ gets  $D_{i1} = 1$ , while Pennsylvania (PA) keeps  $D_{i1} = 0$
- Before enactment: NJ and PA averages:  $\bar{Y}_{\text{NJ},0}$  and  $\bar{Y}_{\text{PA},0}$
- After:  $\bar{Y}_{\text{NJ},1}$  and  $\bar{Y}_{\text{PA},1}$
- Intuitively:

$$\begin{aligned}\text{Treatment effect} &= (\bar{Y}_{\text{NJ},1} - \bar{Y}_{\text{PA},1}) - (\bar{Y}_{\text{NJ},0} - \bar{Y}_{\text{PA},0}) \\ &= (\bar{Y}_{\text{NJ},1} - \bar{Y}_{\text{NJ},0}) - (\bar{Y}_{\text{PA},1} - \bar{Y}_{\text{PA},0})\end{aligned}$$

- Plausible even though  $D_{it}$  is clearly not randomized! Why?
- We now study the mathematics that underlies this intuition

## Setup

- Consider  $2 \times 2$  block structure:

$$\mathbf{D} = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$$

- Let  $t \in \{0, 1\}$  (pre/post) and  $D_i \in \{0, 1\}$  (control/treatment)
- Define the  $2 \times 2$  *CEF matrix*:

$$\begin{bmatrix} \mathbb{E}[Y_{i0}|D_i = 0] & \mathbb{E}[Y_{i1}|D_i = 0] \\ \mathbb{E}[Y_{i0}|D_i = 1] & \mathbb{E}[Y_{i1}|D_i = 1] \end{bmatrix}$$

- The *potential outcomes matrix* (POM) for  $d \in \{0, 1\}$  is

$$\begin{bmatrix} \mathbb{E}[Y_{i0}(d)|D_i = 0] & \mathbb{E}[Y_{i1}(d)|D_i = 0] \\ \mathbb{E}[Y_{i0}(d)|D_i = 1] & \mathbb{E}[Y_{i1}(d)|D_i = 1] \end{bmatrix}$$

# Potential Outcomes Matrices

- Are elements of POM **counterfactual** or “**factual**”?
  - Counterfactuals are **unobserved**, factuals are **observed**
  - We can write  $Y_{it} = [1 - D_i P_t] Y_{it}(0) + D_i P_t Y_{it}(1)$
  - Consider all eight POs in the  $2 \times 2$  block structure POM:

$$Y_{it}(0) : \begin{bmatrix} \mathbb{E}[Y_{i0}(0)|D_i = 0] & \mathbb{E}[Y_{i1}(0)|D_i = 0] \\ \mathbb{E}[Y_{i0}(0)|D_i = 1] & \mathbb{E}[Y_{i1}(0)|D_i = 1] \end{bmatrix}$$

$$Y_{it}(1) : \begin{bmatrix} \mathbb{E}[Y_{i0}(1)|D_i = 0] & \mathbb{E}[Y_{i1}(1)|D_i = 0] \\ \mathbb{E}[Y_{i0}(1)|D_i = 1] & \mathbb{E}[Y_{i1}(1)|D_i = 1] \end{bmatrix}$$

- Notice observability mimics structure of  $\mathbf{D}$
- So the CEF matrix corresponds to following POs:

$$Y_{it} : \begin{bmatrix} \mathbb{E}[Y_{i0}(0)|D_i = 0] & \mathbb{E}[Y_{i1}(0)|D_i = 0] \\ \mathbb{E}[Y_{i0}(0)|D_i = 1] & \mathbb{E}[Y_{i1}(1)|D_i = 1] \end{bmatrix}$$

# Identification of ATT

- We are interested in  $\tau \equiv \mathbb{E} [Y_{it}(1) - Y_{it}(0) | D_{it} = 1]$
- Since  $D_{it} = D_i P_t$  in block structures, we have

$$\begin{aligned}\tau &= \mathbb{E} [Y_{it}(1) - Y_{it}(0) | D_i = 1, P_t = 1] \\ &= \mathbb{E} [Y_{i1}(1) - Y_{i1}(0) | D_i = 1]\end{aligned}$$

and hence

$$\begin{aligned}\tau &= \mathbb{E} [Y_{i1}(1) | D_i = 1] - \mathbb{E} [Y_{i1}(0) | D_i = 1] \\ &= \underbrace{\mathbb{E} [Y_{i1} | D_i = 1]}_{\text{observed}} - \underbrace{\mathbb{E} [Y_{i1}(0) | D_i = 1]}_{\text{need model}}\end{aligned}$$

- We model missing PO using *parallel trends assumption*:

$$\underbrace{\mathbb{E} [Y_{i1}(0) - Y_{i0}(0) | D_i = 1]}_{\text{counterfactual trend in } D_i = 1 \text{ group}} = \underbrace{\mathbb{E} [Y_{i1}(0) - Y_{i0}(0) | D_i = 0]}_{\text{"factual" trend in } D_i = 0 \text{ group}}$$

# Modeling the Counterfactual through Parallel Trends

- Reorganizing parallel trends yields

$$\begin{aligned}\mathbb{E}[Y_{i1}(0) | D_i = 1] &= \mathbb{E}[Y_{i0}(0) | D_i = 1] \\ &\quad + \mathbb{E}[Y_{i1}(0) - Y_{i0}(0) | D_i = 0] \\ &= \underbrace{\mathbb{E}[Y_{i0} | D_i = 1]}_{\text{observed}} + \underbrace{\mathbb{E}[Y_{i1} - Y_{i0} | D_i = 0]}_{\text{observed}}\end{aligned}$$

- From this, it follows that

$$\begin{aligned}\tau &= \mathbb{E}[Y_{i1} | D_i = 1] - \mathbb{E}[Y_{i1}(0) | D_i = 1] \\ &= \underbrace{\mathbb{E}[Y_{i1} - Y_{i0} | D_i = 1]}_{\text{Treated pre/post diff}} - \underbrace{\mathbb{E}[Y_{i1} - Y_{i0} | D_i = 0]}_{\text{Control pre/post diff}}\end{aligned}$$

- So DID simply fills in the missing PO using parallel trends

# Table of Contents

## 1 Causal Effects in Panel Settings

Treatment Structures

## 2 The $2 \times 2$ Difference-in-Differences Design

Identification in  $2 \times 2$  DID

Estimation of  $2 \times 2$  DID

## 3 Generalized DID Designs

Estimation of  $2 \times T$  DID

Static Effects in Staggered DID

Inference in DID

## 4 Appendix

Static Effects in  $2 \times T$  Designs

# Method of Moments in Simple Numerical Example

- Let  $\bar{Y}_{gt} = \frac{1}{N_g} \sum_{i:G_i=g} Y_{it}$  be the group  $g$  mean in  $t$
- Observable group means matrix:

$$\begin{bmatrix} \bar{Y}_{00} & \bar{Y}_{01} \\ \bar{Y}_{10} & \bar{Y}_{11} \end{bmatrix} = \begin{bmatrix} 2 & 3 \\ 3 & 5 \end{bmatrix}$$

- Can use Method of Moments:

$$\begin{aligned}\hat{\tau}_{\text{MM}} &= \widehat{\mathbb{E}} [Y_{i1} - Y_{i0} | D_i = 1] - \widehat{\mathbb{E}} [Y_{i1} - Y_{i0} | D_i = 0] \\ &= (\bar{Y}_{11} - \bar{Y}_{10}) - (\bar{Y}_{01} - \bar{Y}_{00}) \\ &= (5 - 3) - (3 - 2) = 2 - 1 = 1\end{aligned}$$

- Note: equivalently, we estimated  $\widehat{\mathbb{E}} [Y_{i1}(0) | D_i = 1] = 4$

# Additive Separability

- Parallel trends are equivalent to additive separability:

$$\mathbb{E}[Y_{it}(0) | D_i, P_t] = \mu + \alpha D_i + \gamma P_t$$

where  $\mu \equiv \mathbb{E}[Y_{it}(0) | D_i = 0, P_t = 0]$  and

$$\alpha \equiv \mathbb{E}[Y_{it}(0) | D_i = 1, P_t = 0] - \mu$$

$$\gamma \equiv \mathbb{E}[Y_{it}(0) | D_i = 0, P_t = 1] - \mu$$

- Then, use observed outcome as function of POs:

$$Y_{it} = Y_{it}(0) + [Y_{it}(1) - Y_{it}(0)] D_i P_t$$

- Taking conditional expectations, we get

$$\begin{aligned}\mathbb{E}[Y_{it} | D_i, P_t] &= \mathbb{E}[Y_{it}(0) | D_i, P_t] \\ &\quad + \mathbb{E}[Y_{it}(1) - Y_{it}(0) | D_i, P_t] D_i P_t \\ &= \mu + \alpha D_i + \gamma P_t + \tau D_i P_t\end{aligned}$$

# DID Estimation through Regression

- Recall estimating cell means – this is the same:

$$\mathbb{E}[Y_{it}|D_i, P_t] = \mu + \alpha D_i + \gamma P_t + \tau D_i P_t$$

- So this equals saturated OLS model:

$$Y_{it} = \mu + \alpha D_i + \gamma P_t + \tau D_i P_t + \varepsilon_{it}$$

- Hence, let  $\beta = (\mu, \alpha, \gamma, \tau)$  and  $\mathbf{X}_{it} = (1, D_i, P_t, D_i P_t)$ :

$$\hat{\beta}_{OLS} = \left( \sum_{i=1}^N \sum_{t=0}^1 \mathbf{X}_{it} \mathbf{X}'_{it} \right)^{-1} \left( \sum_{i=1}^N \sum_{t=0}^1 \mathbf{X}_{it} Y_{it} \right)$$

and it can be shown that  $\hat{\tau}_{OLS} = \hat{\tau}_{MM}$  if  $N_0 = N_1$

# Card and Krueger (1994) Example

TABLE 3—AVERAGE EMPLOYMENT PER STORE BEFORE AND AFTER THE RISE  
IN NEW JERSEY MINIMUM WAGE

Variable	Stores by state			Stores in New Jersey <sup>a</sup>			Differences within NJ <sup>b</sup>	
	PA (i)	NJ (ii)	Difference, NJ–PA (iii)	Wage = \$4.25 (iv)	Wage = \$4.26–\$4.99 (v)	Wage ≥ \$5.00 (vi)	Low– high (vii)	Midrange– high (viii)
1. FTE employment before, all available observations	23.33 (1.35)	20.44 (0.51)	-2.89 (1.44)	19.56 (0.77)	20.08 (0.84)	22.25 (1.14)	-2.69 (1.37)	-2.17 (1.41)
2. FTE employment after, all available observations	21.17 (0.94)	21.03 (0.52)	-0.14 (1.07)	20.88 (1.01)	20.96 (0.76)	20.21 (1.03)	0.67 (1.44)	0.75 (1.27)
3. Change in mean FTE employment	-2.16 (1.25)	0.59 (0.54)	2.76 (1.36)	1.32 (0.95)	0.87 (0.84)	-2.04 (1.14)	3.36 (1.48)	2.91 (1.41)
4. Change in mean FTE employment, balanced sample of stores <sup>c</sup>	-2.28 (1.25)	0.47 (0.48)	2.75 (1.34)	1.21 (0.82)	0.71 (0.69)	-2.16 (1.01)	3.36 (1.30)	2.87 (1.22)
5. Change in mean FTE employment, setting FTE at temporarily closed stores to 0 <sup>d</sup>	-2.28 (1.25)	0.23 (0.49)	2.51 (1.35)	0.90 (0.87)	0.49 (0.69)	-2.39 (1.02)	3.29 (1.34)	2.88 (1.23)

# DID with Fixed Effects

- What about  $Y_{it} = \alpha_i + \gamma_t + \tau D_i P_t + \varepsilon_{it}$ ? (see TWFE below)
- Concerning  $\gamma_t$ , in  $2 \times 2$ , full rank  $\mathbf{X}_{it} \mathbf{X}'_{it}$  implies
  - Need to normalize one  $\gamma_t$  and include dummy for other one
  - This is equivalent to  $\gamma P_t$ , just notational change
- Concerning  $\alpha_i$ , apply FWL to  $Y_{it}$  and  $D_{it} = D_i P_t$  to get

$$Y_{it} - \bar{Y}_i = \gamma_t + \tau (D_{it} - \bar{D}_i) + \varepsilon_{it} - \bar{\varepsilon}_i$$

- But note that  $D_{it} - \bar{D}_i = 0$  for  $D_i = 0$
- For treatment group,  $D_{it} - \bar{D}_i = D_{it} - \frac{T-g-1}{T}$
- Now let  $\mu = -\tau \times \frac{T-g-1}{T}$  and  $\alpha D_i + \bar{\varepsilon}_i = \bar{Y}_i$
- Then we have recovered  $Y_{it} = \mu + \alpha D_i + \gamma P_t + \tau D_i P_t + \varepsilon_{it}$
- So algebraically equivalent!
- Only true in  $2 \times 2$  block structures, strongly balanced panel

# Table of Contents

## 1 Causal Effects in Panel Settings

Treatment Structures

## 2 The $2 \times 2$ Difference-in-Differences Design

Identification in  $2 \times 2$  DID

Estimation of  $2 \times 2$  DID

## 3 Generalized DID Designs

Estimation of  $2 \times T$  DID

Static Effects in Staggered DID

Inference in DID

## 4 Appendix

Static Effects in  $2 \times T$  Designs

## Estimation $2 \times T$ DID

- Consider now  $2 \times T$  block structures with  $t = 1, \dots, T$ , e.g.

$$\mathbf{D} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix}$$

- $2 \times 2$  pre/post outcome matrix:

$$\begin{bmatrix} \mathbb{E}[Y_{it}|D_i = 0, P_t = 0] & \mathbb{E}[Y_{it}|D_i = 0, P_t = 1] \\ \mathbb{E}[Y_{it}|D_i = 1, P_t = 0] & \mathbb{E}[Y_{it}|D_i = 1, P_t = 1] \end{bmatrix}$$

- Corresponding  $2 \times 2$  sample averages:

$$\begin{bmatrix} \bar{Y}_{0,\text{pre}} & \bar{Y}_{0,\text{post}} \\ \bar{Y}_{1,\text{pre}} & \bar{Y}_{1,\text{post}} \end{bmatrix}$$

where for  $d \in \{0, 1\}$

$$\bar{Y}_{d,\text{pre}} = \frac{1}{N_d(g-1)} \sum_{i:D_i=d, t < g} Y_{it}; \quad \bar{Y}_{d,\text{post}} = \frac{1}{N_d(T-g+1)} \sum_{i:D_i=d, t \geq g} Y_{it}$$

# Method of Moments and Regression Estimators

- MM:  $\hat{\tau}_{\text{MM}} = (\bar{Y}_{1,\text{post}} - \bar{Y}_{1,\text{pre}}) - (\bar{Y}_{0,\text{post}} - \bar{Y}_{0,\text{pre}})$
- Regression:  $\mathbb{E}^* [Y_{it}|1, D_i, P_t, D_i P_t]$  just like in  $2 \times 2$
- $\hat{\tau}_{\text{MM}} = \hat{\tau}_{\text{OLS}}$  if  $N_0 = N_1$  and  $2(g-1) = T$
- Wide format illustration with  $g = 4$ ,  $T = 6$ :

$i$	$D_i$	$Y_{i1}$	$Y_{i2}$	$Y_{i3}$	$Y_{i4}$	$Y_{i5}$	$Y_{i6}$
1	0						
2	0			$\bar{Y}_{0,\text{pre}}$			$\bar{Y}_{0,\text{post}}$
3	0						
4	1						
5	1			$\bar{Y}_{1,\text{pre}}$			$\bar{Y}_{1,\text{post}}$
6	1						

- Run regression in long format

# Table of Contents

- 1 Causal Effects in Panel Settings
  - Treatment Structures
- 2 The  $2 \times 2$  Difference-in-Differences Design
  - Identification in  $2 \times 2$  DID
  - Estimation of  $2 \times 2$  DID
- 3 Generalized DID Designs
  - Estimation of  $2 \times T$  DID
  - Static Effects in Staggered DID
  - Inference in DID
- 4 Appendix
  - Static Effects in  $2 \times T$  Designs

## Treatment-Timing Groups

- Let's turn to staggered rollout structures
- We now have at least two  $G_i < \infty$ , e.g.

$$\mathbf{D} = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}$$

- We might still have a never-treated group, e.g.

$$\mathbf{D} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 \end{bmatrix}$$

- No longer always  $D_{it} = D_i P_t$  (since not block structure)
- Instead, we have  $D_{it} = 1 [t \geq G_i]$  for each  $g \in \mathcal{G}$ 
  - Generalization; collapses to  $D_{it} = D_i P_t$  in block structures

# Identification and Estimation

- Identification of  $\tau = \mathbb{E} [Y_{it}(1) - Y_{it}(0) | D_{it} = 1]$ :
  - Analogous to  $2 \times T$  case (see Appendix)
  - Again requires parallel trends: for  $t \geq 2$  and any  $g, g' \in \mathcal{G}$

$$\mathbb{E} [Y_{it}(0) - Y_{it-1}(0) | G_i = g] = \mathbb{E} [Y_{it}(0) - Y_{it-1}(0) | G_i = g']$$

- Estimation with two-way fixed effects (TWFE) regression:

$$Y_{it} = \alpha_i + \gamma_t + D_{it}\tau_{\text{TWFE}} + \varepsilon_{it}$$

- Called *staggered DID regression* when  $D_{it} = 1 [t \geq G_i]$
- But even used if generic  $D_{it} \in \{0, 1\}$  (i.e.  $D_{it}$  not absorbing)
- Two important recent insights in TWFE:
  - ①  $\tau_{\text{TWFE}} \neq \tau$  if effects are heterogeneous along group/time
  - ② For staggered rollouts,  $\tau_{\text{TWFE}}$  is weighted avg. of  $2 \times T$  DIDs

## Static Group-Time Treatment Effects

- Assume generic  $\mathbf{D}$  with  $N_{gt}$  observations in cell  $(g, t)$ 
  - Consider running  $Y_{it} = \alpha_{G_i} + \gamma_t + \tau_{\text{TWFE}} D_{it} + \varepsilon_{it}$
  - How does  $\tau_{\text{TWFE}}$  relate to  $\tau$ ?
- Write *static group-time treatment effects* as

$$\tau_{gt} = \frac{1}{N_{gt}} \sum_{i \in g} [Y_{it}(1) - Y_{it}(0)]$$

which is called the ATE in cell  $(g, t)$

- Define  $N_1 = \sum_{i,t} D_{it}$  as the number of treated observations
- We can then write the ATT as

$$\tau = \mathbb{E} \left[ \sum_{(g,t): D_{it}=1} \frac{N_{gt}}{N_1} \tau_{gt} \right]$$

which says that  $\tau$  is weighted average of group-time effects

# Decomposing the TWFE Estimator

Theorem (de Chaisemartin & d'Haultfoeuille 2020)

TWFE regression is given by:

$$\tau_{TWFE} = \mathbb{E} \left[ \sum_{(g,t):D_{it}=1} \frac{N_{gt}}{N_1} w_{gt} \tau_{gt} \right]$$

where

$$w_{gt} = \frac{\tilde{D}_{gt}}{\sum_{(g,t):D_{it}=1} \frac{N_{gt}}{N_1} \tilde{D}_{gt}}$$

where  $\tilde{D}_{gt} = D_{it} - \mathbb{E}^* [D_{it} | \alpha_{G_i}, \gamma_t]$  are the residuals of  $D_{it}$  after removing TWFE, which are constant within group (as is  $D_{it}$ ).

## Understanding the Decomposition Result: Example

- What does this result say?
  - If  $\tau_{gt}$  are constant, then  $\tau_{TWFE} = \tau$  (as  $\sum_{(g,t):D_{it}=1} \frac{N_{gt}}{N_1} w_{gt} = 1$ )
  - But with varying  $\tau_{gt}$ , in general  $\tau_{TWFE} \neq \tau$
  - Hence  $\tau_{TWFE}$  is generally a biased estimator of  $\tau$
- Consider staggered rollout with  $T = 3$  and  $\mathcal{G} = \{2, 3\}$

$$\mathbf{D} = \begin{bmatrix} \mathbf{D}(3) \\ \mathbf{D}(2) \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}$$

- According to FWL, we have  $\tilde{D}_{gt} = D_{it} - \bar{D}_g - \bar{D}_t + \bar{D}$
- This implies:

$$\tilde{D}_{33} = 1 - \frac{1}{3} - 1 + \frac{1}{2} = \frac{1}{6}$$

$$\tilde{D}_{22} = 1 - \frac{2}{3} - \frac{1}{2} + \frac{1}{2} = \frac{1}{3}$$

$$\tilde{D}_{23} = 1 - \frac{2}{3} - 1 + \frac{1}{2} = -\frac{1}{6}$$

## Negative TWFE Although all $(g, t)$ -ATEs are Positive

- Consider, for example,  $\mathbb{E} [\tau_{33}] = \mathbb{E} [\tau_{22}] = 1$  and  $\mathbb{E} [\tau_{23}] = 4$
- It follows then from the Theorem that

$$\tau_{\text{TWFE}} = \frac{1}{2}\mathbb{E} [\tau_{33}] + \mathbb{E} [\tau_{22}] - \frac{1}{2}\mathbb{E} [\tau_{23}] = -\frac{1}{2}$$

- So  $\tau_{\text{TWFE}}$  is negative even though all  $\mathbb{E} [\tau_{gt}]$  are positive!
- However, if e.g.  $\mathbb{E} [\tau_{gt}] = 1$  (i.e. homogeneous), then

$$\tau_{\text{TWFE}} = \tau = 1$$

- The negative weight makes  $\tau_{\text{TWFE}}$  very different from  $\tau$
- But why does this problem happen?

# Where do Negative Weights Come From?

- It arises because it can be shown that in the example:

$$\tau_{\text{TWFE}} = (\text{DID}_1 + \text{DID}_2) / 2$$

where

$$\text{DID}_1 = \mathbb{E} [(\bar{Y}_{22} - \bar{Y}_{21}) - (\bar{Y}_{32} - \bar{Y}_{31})]$$

$$\text{DID}_2 = \mathbb{E} [(\bar{Y}_{33} - \bar{Y}_{32}) - (\bar{Y}_{23} - \bar{Y}_{22})]$$

- So it is the average of all  $2 \times 2$  DIDs:

$$\text{DID}_1 : \left[ \begin{array}{cc|c} 0 & 0 & 1 \\ 0 & 1 & 1 \end{array} \right] \text{ and } \text{DID}_2 : \left[ \begin{array}{c|cc} 0 & 0 & 1 \\ 0 & 1 & 1 \end{array} \right]$$

- $\text{DID}_1 = \mathbb{E} [\tau_{22}]$  as hoped, but  $\text{DID}_2 = \mathbb{E} [\tau_{33}] - (\mathbb{E} [\tau_{23}] - \mathbb{E} [\tau_{22}])$
- So  $\text{DID}_2$  goes awry if  $\tau_{23}$  is very different from  $\tau_{22}$
- Also see Goodman-Bacon (2020) for a similar result

# Robustness to Heterogeneity

- de Chaisemartin and d'Haultfoeuille (2020) propose:
  - Diagnostic to assess sensitivity to heterogeneity
  - Alternative heterogeneity-robust estimator
- The target parameter for estimator is:

$$\tau_S \equiv \mathbb{E} \left[ \frac{1}{N_S} \sum_{(g,t):t \geq 2, D_{it} \neq D_{it-1}} (Y_{it}(1) - Y_{it}(0)) \right]$$

where  $N_S = \sum_{(g,t):t \geq 2, D_{it} \neq D_{it-1}}$  are obs in *switching* cells

- This is the ATE of all switching cells  $D_{it} \neq D_{it-1}$
- In staggered adoption, mean ATE at start of treatment

# Heterogeneity-Robust Estimator

- The estimator is

$$\tau_{\text{dCdH}} = \sum_{t=2}^T (\omega_{+,t} \text{DID}_{+,t} + \omega_{-,t} \text{DID}_{-,t})$$

where  $\omega_{+,t}$  and  $\omega_{-,t}$  are sample weights

- Basic idea:
  - $\text{DID}_{+,t}$  measures “joiner-effect”: [0, 1] against [0, 0]
  - $\text{DID}_{-,t}$  “leaver-effect”: [1, 0] against [1, 1]
  - Ignore the rest – so use only  $\text{DID}_1$  in example
- See `did_multiplegt`

## Real-World Example: Gentzkow et al. (2011)

- $\hat{\beta}_{fe} = \hat{\tau}_{TWFE}$
- $DID_M = \hat{\tau}_{dCdH}$

TABLE 3—ESTIMATES OF THE EFFECT OF ONE ADDITIONAL NEWSPAPER ON TURNOUT

	Estimate	Standard error	Observations
$\hat{\beta}_{fd}$	0.0026	0.0009	15,627
$\hat{\beta}_{fe}$	-0.0011	0.0011	16,872
$DID_M$	0.0043	0.0014	16,872
$DID_M^{pl}$	-0.0009	0.0016	13,221
$DID_M$ , on placebo subsample	0.0045	0.0019	13,221

*Notes:* This table reports estimates of the effect of one additional newspaper on turnout, as well as a placebo estimate of the common trends assumption underlying  $DID_M$ . Estimators are computed using the data of Gentzkow, Shapiro, and Sinkinson (2011), with state-year fixed effects as controls. Standard errors are clustered by county. To compute the  $DID_M$  estimators, the number of newspapers is grouped into 4 categories: 0, 1, 2, and more than 3.

# Table of Contents

## 1 Causal Effects in Panel Settings

Treatment Structures

## 2 The $2 \times 2$ Difference-in-Differences Design

Identification in  $2 \times 2$  DID

Estimation of  $2 \times 2$  DID

## 3 Generalized DID Designs

Estimation of  $2 \times T$  DID

Static Effects in Staggered DID

Inference in DID

## 4 Appendix

Static Effects in  $2 \times T$  Designs

## Serial Correlation in $Y_{it}$ in Long Panels

- Consider staggered rollouts with moderately large  $T$  (say 30)
- $\text{Cov}(Y_{it}, Y_{is})$  and  $\text{Cov}(D_{it}, D_{is})$  are often high in this case
  - Makes it likely that  $\text{Cov}(\varepsilon_{it}, \varepsilon_{is})$  is high as well
  - Ignoring serial correlation overstates precision of  $\hat{\beta}_{\text{OLS}}$
- Bertrand et al. (2004):
  - Robust standard errors reject too often
  - Treatment-unit clustered SEs work well → do this
  - Some “fancier” standard errors work, others do not
- Takeaway: always use  $G_i$ -clustered SEs as default

# Table of Contents

- 1 Causal Effects in Panel Settings
  - Treatment Structures
- 2 The  $2 \times 2$  Difference-in-Differences Design
  - Identification in  $2 \times 2$  DID
  - Estimation of  $2 \times 2$  DID
- 3 Generalized DID Designs
  - Estimation of  $2 \times T$  DID
  - Static Effects in Staggered DID
  - Inference in DID
- 4 Appendix
  - Static Effects in  $2 \times T$  Designs

## $2 \times T$ Design Setup

- Consider now  $2 \times T$  block structures with  $t = 1, \dots, T$ , e.g.

$$\mathbf{D} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix}$$

- Two groups:  $G_i \in \{g, \infty\}$  with  $1 < g \leq T$
- Parameter of interest still:  $\tau = \mathbb{E}[Y_{it}(1) - Y_{it}(0)|D_{it} = 1]$
- Now write  $P_t \equiv \max_i D_{it} = 1 [t \geq g]$
- Just like before, can split estimand:

$$\begin{aligned}\tau &= \mathbb{E}[Y_{it}(1)|D_i = 1, P_t = 1] - \mathbb{E}[Y_{it}(0)|D_i = 1, P_t = 1] \\ &= \underbrace{\mathbb{E}[Y_{it}|D_i = 1, P_t = 1]}_{\text{observed}} - \underbrace{\mathbb{E}[Y_{it}(0)|D_i = 1, P_t = 1]}_{\text{need model}}\end{aligned}$$

## Parallel Trends in $2 \times T$

- Parallel trends: for  $t \geq 2$

$$\mathbb{E}[Y_{it}(0) - Y_{it-1}(0) | D_i = 1] = \mathbb{E}[Y_{it}(0) - Y_{it-1}(0) | D_i = 0]$$

so that specifically for  $t \geq g$

$$\begin{aligned}\mathbb{E}[Y_{it}(0) | D_i = 1] &= \mathbb{E}[Y_{it}(0) | D_i = 1, P_t = 1] \\ &= \mathbb{E}[Y_{it-1}(0) | D_i = 1] \\ &\quad + \mathbb{E}[Y_{it}(0) - Y_{it-1}(0) | D_i = 0]\end{aligned}$$

- So we are already identified if  $g = T$  (same as  $2 \times 2$ )

# Multiple Post-Periods

- Recursively use  $\mathbb{E}[Y_{it-1}(0) | D_i = 1]$  until it “turns blue”
  - If  $t = g$ :  $\mathbb{E}[Y_{it-1}(0) | D_i = 1] = \mathbb{E}[Y_{it-1} | D_i = 1, P_{t-1} = 0]$
  - If  $t = g + 1$  (i.e. second post-period), then

$$\begin{aligned}\mathbb{E}[Y_{it-1}(0) | D_i = 1] &= \mathbb{E}[Y_{it-2} | D_i = 1, P_{t-2} = 0] \\ &\quad + \mathbb{E}[Y_{it-1}(0) - Y_{it-2}(0) | D_i = 0]\end{aligned}$$

- And so forth, with  $s \geq 0$  for  $t + s = g + 1$
- Hence we can show that  $\tau$  is identified:

$$\begin{aligned}\tau &= (\mathbb{E}[Y_{it} | D_i = 1, P_t = 1] - \mathbb{E}[Y_{it} | D_i = 1, P_t = 0]) \\ &\quad - (\mathbb{E}[Y_{it} | D_i = 0, P_t = 1] - \mathbb{E}[Y_{it} | D_i = 0, P_t = 0])\end{aligned}$$

# Econometrics II

## Lecture 10: Dynamic Difference-in-Differences

David Schönholzer

Stockholm University

May 2, 2024

# Plan for Today

## 1 Evaluating DID Designs

Parallel Trends

## 2 Dynamic DID

Definitions: Dynamic Heterogeneous PO and ATT

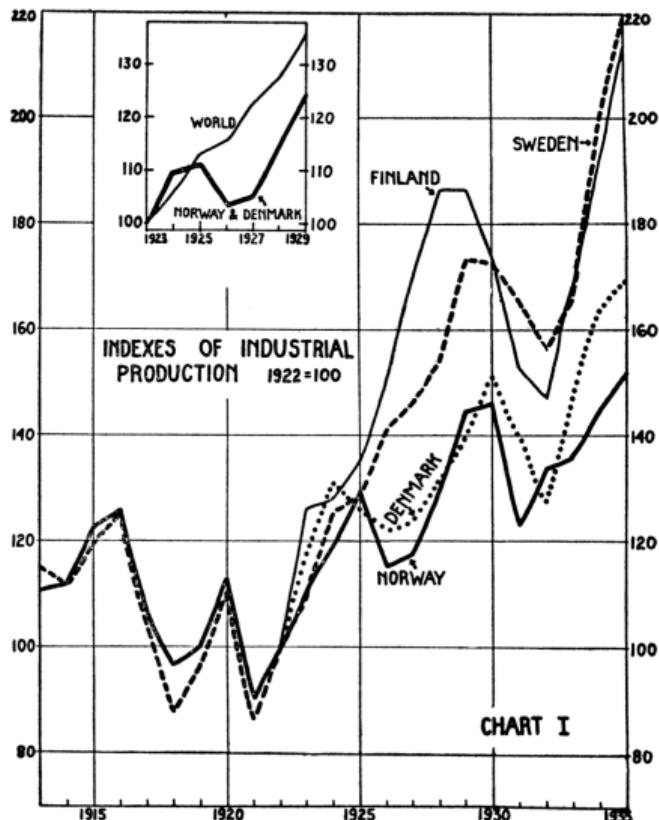
Dynamic Effects

Group-Time Heterogeneity

## 3 Appendix

Functional Form

# Example: Lester (1937)



- Great identification in macro exists!
  - DK and NO reintroduced gold standard
  - FI and SE did not
  - Large difference in industrial production
- Issues we study today:
  - ① Parallel trends:
    - Are FI and SE good counterfactuals?
    - Did they evolve similarly before 1925?
  - ② Time heterogeneity / dynamics:
    - How did effect change over time?
  - ③ Unit heterogeneity:
    - Did DK and NO respond differently?
    - Different in short-run or long-run?

# Table of Contents

## 1 Evaluating DID Designs

Parallel Trends

## 2 Dynamic DID

Definitions: Dynamic Heterogeneous PO and ATT

Dynamic Effects

Group-Time Heterogeneity

## 3 Appendix

Functional Form

# Parallel Trends Assumption and Pre-Trends

- PTA is untestable! We can never observe  $\mathbb{E} [Y_{it} (0) | D_{it} = 1]$
- However, we can test for parallel *pre-trends*
  - Absence of pre-trends provides evidence for PT
  - But common shocks remain a threat
- Need at least one  $G_i \in [3, \infty)$  to test for pre-trends, e.g.

$$\mathbf{D} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

- *Placebo test:*  $\mathbb{E} [Y_{i2} - Y_{i1} | D_i = 1] = \mathbb{E} [Y_{i2} - Y_{i1} | D_i = 0]$
- Note that with  $g \geq 3$ , this corresponds to

$$\mathbb{E} [Y_{i2} (0) - Y_{i1} (0) | D_i = 1] = \mathbb{E} [Y_{i2} (0) - Y_{i1} (0) | D_i = 0]$$

- And if it holds, we might believe

$$\mathbb{E} [Y_{i3} (0) - Y_{i2} (0) | D_i = 1] = \mathbb{E} [Y_{i3} (0) - Y_{i2} (0) | D_i = 0]$$

# Testing for Parallel Trends in Staggered Rollout

- Consider staggered rollout with at least one  $G_i \in [3, \infty)$
- We run (see tsvarlist)

$$Y_{it} = \alpha_i + \gamma_t + \sum_{\ell \in \mathcal{L}} 1[t = G_i + \ell] \beta_\ell + \tau D_{it} + \varepsilon_{it}$$

where  $\mathcal{L} = \{-g+1, -g+2, \dots, -2\}$  is set of lags

- Test  $\beta_\ell = 0$  jointly or individually
- Note we set  $\beta_{-1} = 0$  at the outset
  - Avoids multicollinearity (due to  $\gamma_t, \beta_\ell, \tau$ )
  - Expresses  $\beta_\ell$  (and  $\tau$ ) relative to the period before treatment
- E.g. in  $2 \times 3$  block structure with  $g = 3$  such that  $\mathcal{L} = \{-2\}$ :

$$Y_{it} = \alpha_i + \gamma_t + D_i 1[t = 1] \beta_{-2} + \tau D_i P_t + \varepsilon_{it}$$

and test  $\beta_{-2} = 0$ ; if reject, then there are pre-trends

# Pre-Trend Test Example: Duflo (2001)

TABLE 3—MEANS OF EDUCATION AND LOG(WAGE) BY COHORT AND LEVEL OF PROGRAM CELLS

	Years of education			Log(wages)		
	Level of program in region of birth			Level of program in region of birth		
	High (1)	Low (2)	Difference (3)	High (4)	Low (5)	Difference (6)
<i>Panel A: Experiment of Interest</i>						
Aged 2 to 6 in 1974	8.49 (0.043)	9.76 (0.037)	-1.27 (0.057)	6.61 (0.0078)	6.73 (0.0064)	-0.12 (0.010)
Aged 12 to 17 in 1974	8.02 (0.053)	9.40 (0.042)	-1.39 (0.067)	6.87 (0.0085)	7.02 (0.0069)	-0.15 (0.011)
Difference	0.47 (0.070)	0.36 (0.038)	0.12 (0.089)	-0.26 (0.011)	-0.29 (0.0096)	0.026 (0.015)
<i>Panel B: Control Experiment</i>						
Aged 12 to 17 in 1974	8.02 (0.053)	9.40 (0.042)	-1.39 (0.067)	6.87 (0.0085)	7.02 (0.0069)	-0.15 (0.011)
Aged 18 to 24 in 1974	7.70 (0.059)	9.12 (0.044)	-1.42 (0.072)	6.92 (0.0097)	7.08 (0.0076)	-0.16 (0.012)
Difference	0.32 (0.080)	0.28 (0.061)	0.034 (0.098)	0.056 (0.013)	0.063 (0.010)	0.0070 (0.016)

*Notes:* The sample is made of the individuals who earn a wage. Standard errors are in parentheses.

# Limitations of Pre-Trend Testing

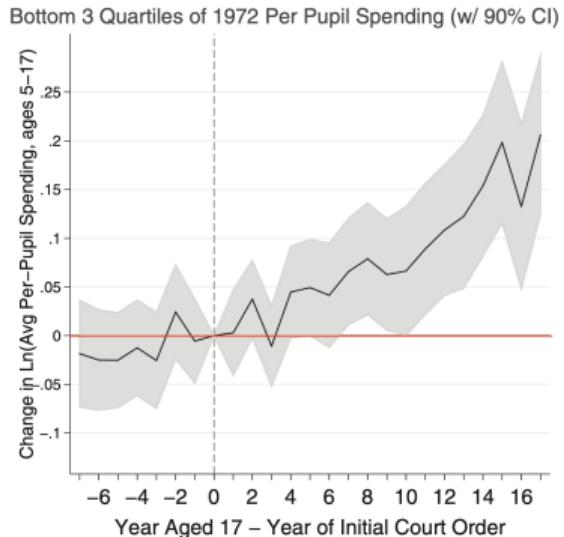


FIGURE I

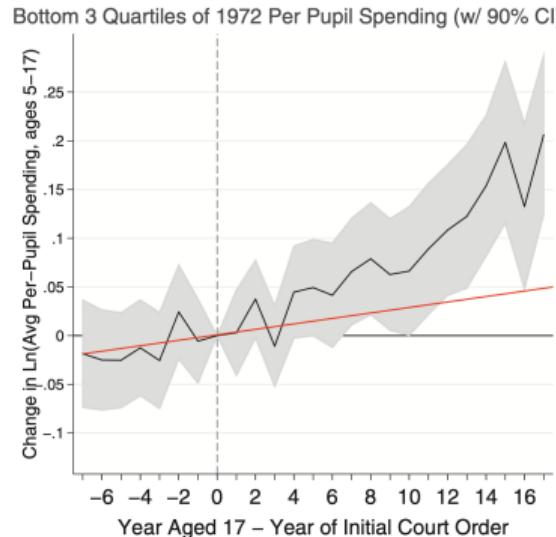


FIGURE I

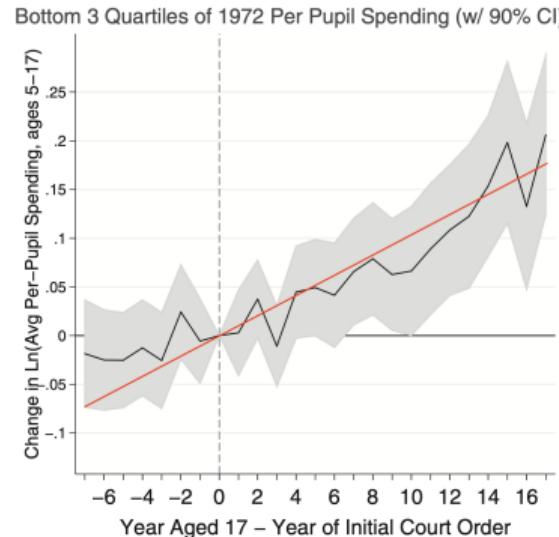
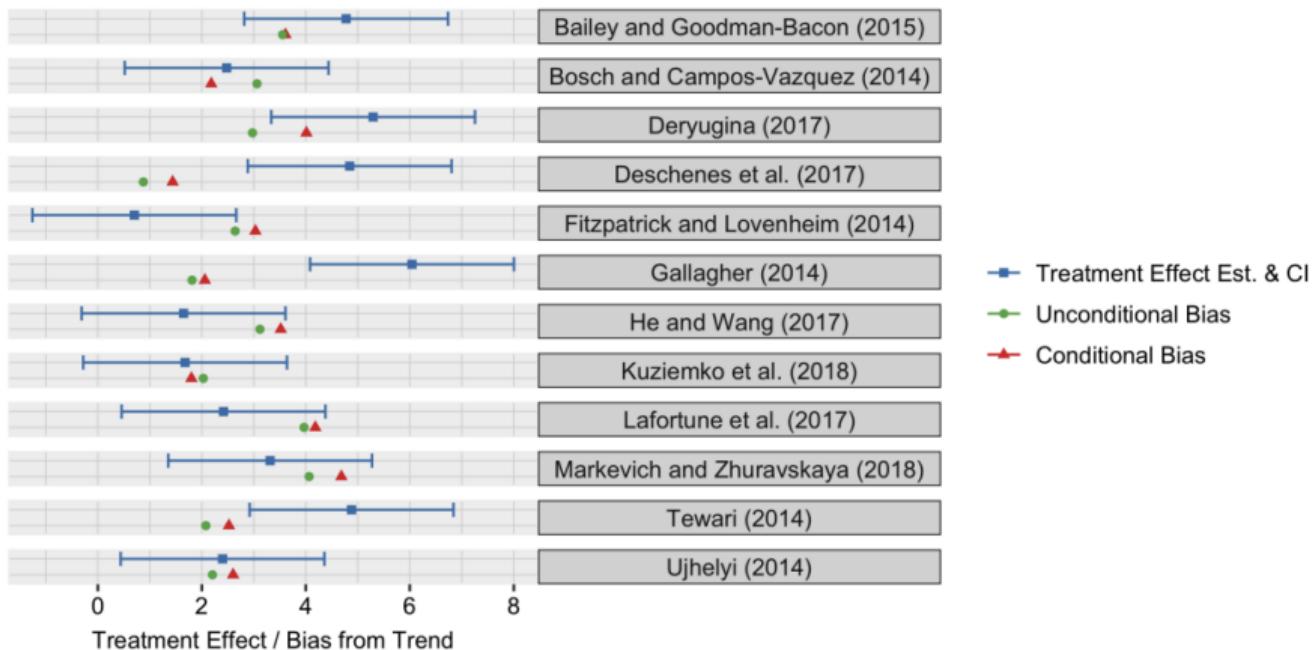


FIGURE I

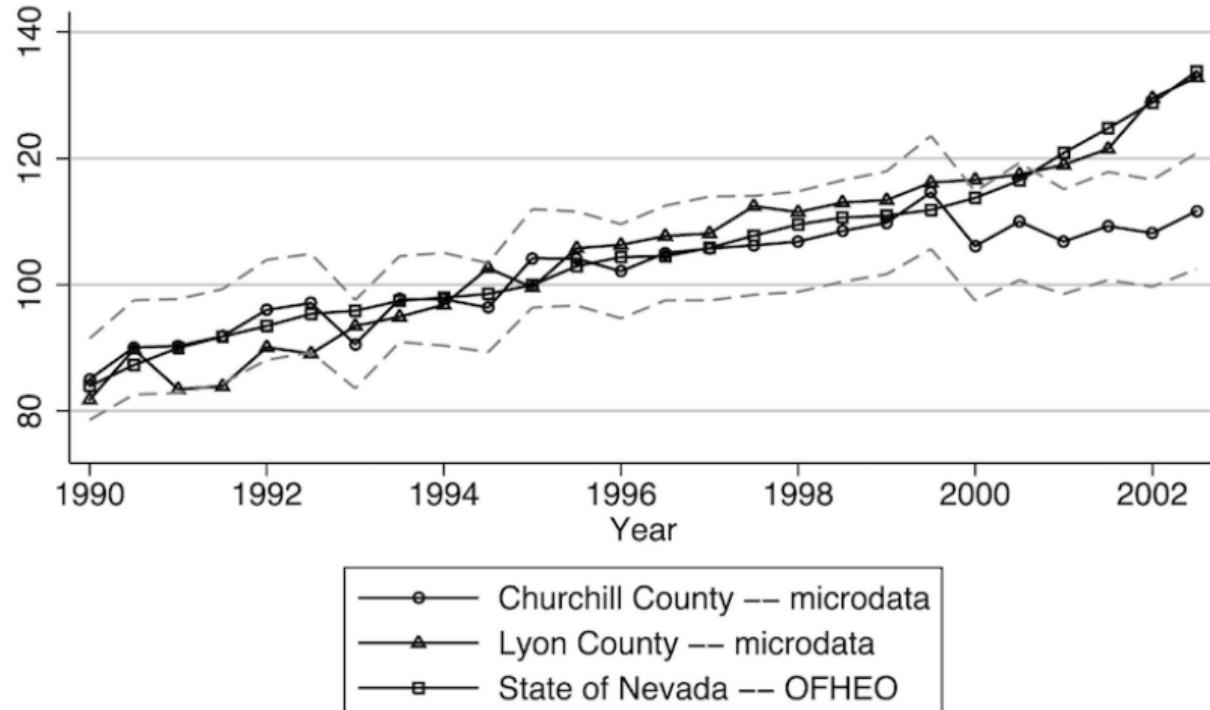
- Example: Jackson et al (2016)
  - Cannot reject zero pre-trend
  - But also cannot reject moderate or strong pre-trend
- How good does pre-trend testing work to detect potential biases?

# Roth (2022): Pre-Trend Testing Issues

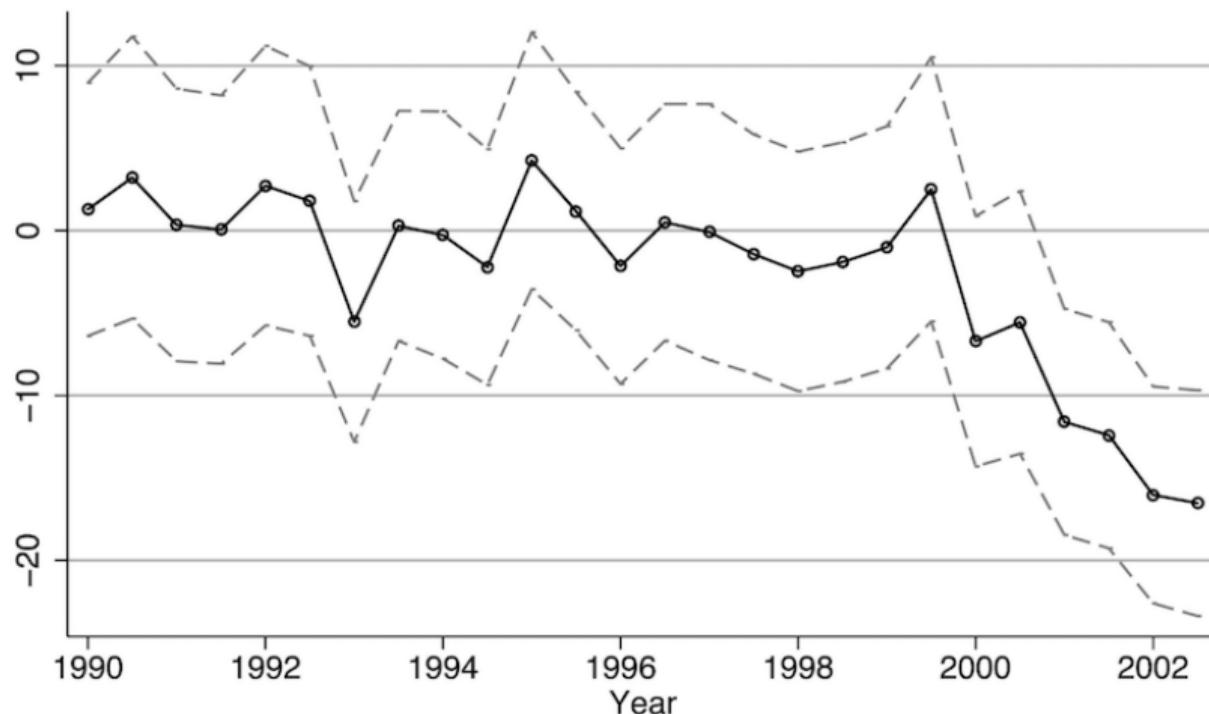
- 1 Low power: pre-tests struggle to detect meaningful biases from pre-trends
- 2 Pre-test bias: passing a pre-test can increase bias



## Davis (2004) Example: Raw Data in Block Structure

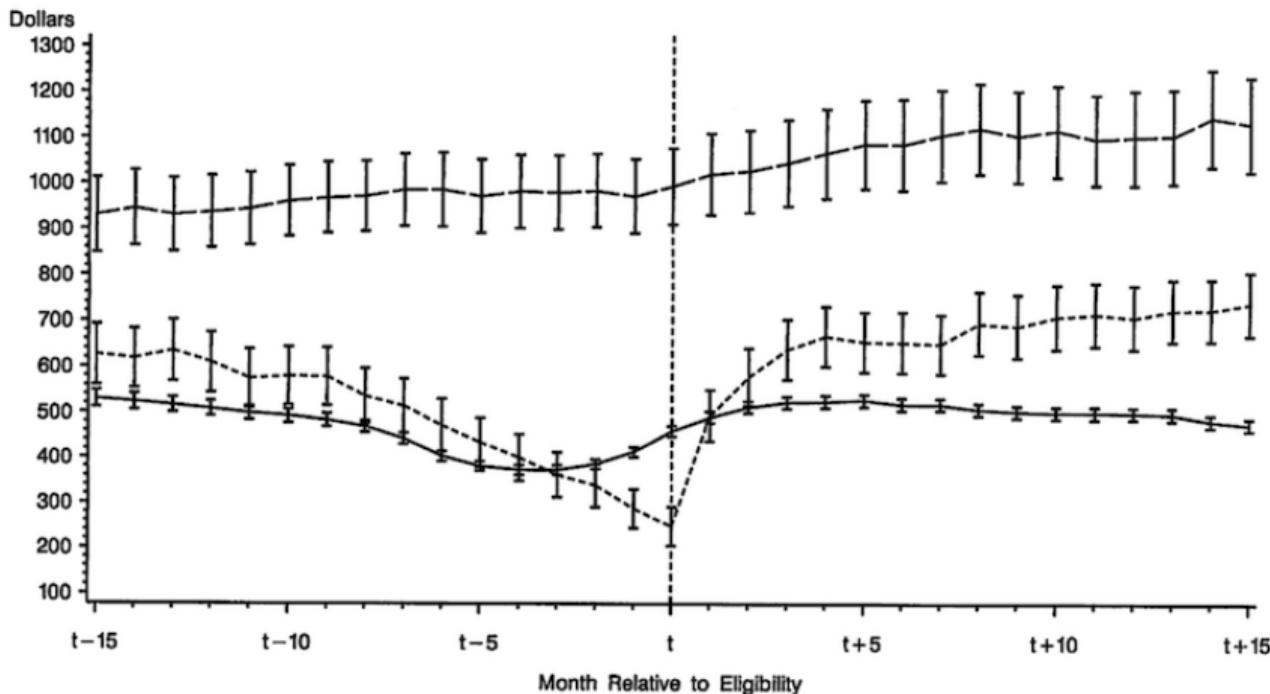


## Davis (2004) Example: Dynamic DID Estimates



# Example of Failed Pre-Trend: Ashenfelter's Dip

Mean Self-Reported Monthly Earnings  
SIPP Eligibles and JTPA Controls and ENPs  
Male Adults



# What if There Are Pre-Trends?

- Parallel trends unlikely to hold in the presence of pre-trends
- Several ways to proceed:
  - ① Linear time trends  $\kappa_g \times t$ , but may absorb part of  $\tau$  and generally unreliable
  - ② Richer fixed effects: e.g. state-by-year FE
  - ③ IV to correct for pre-trend (Freyaldenhoven et al. 2019)
  - ④ Set-identify treatment effects (Roth and Rambachan 2022)
  - ⑤ Find different design
- Next lecture: new methods for when parallel trends are unlikely to hold
  - ① Matched DID
  - ② Synthetic control methods
- Making causal claims with pre-trends will meet resistance
- But DID with pre-trend can be descriptively interesting

# Famous Despite Pre-Trend: Mass Layoffs

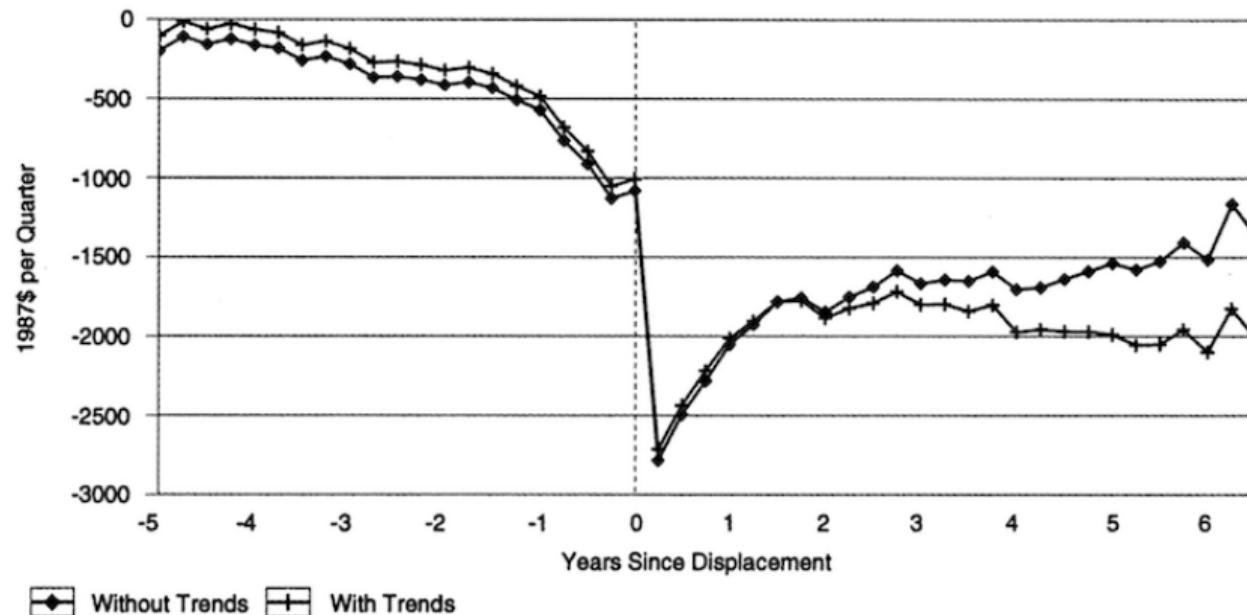


FIGURE 2. EARNINGS LOSSES FOR SEPARATORS IN MASS-LAYOFF SAMPLE

# Table of Contents

## 1 Evaluating DID Designs

Parallel Trends

## 2 Dynamic DID

Definitions: Dynamic Heterogeneous PO and ATT

Dynamic Effects

Group-Time Heterogeneity

## 3 Appendix

Functional Form

# Group-Dependent Potential Outcomes

- Now interested in *heterogeneous, dynamic ATT*
- Need to generalize our PO for treatment timing
  - Define  $Y_{it}(g)$  as PO if treated starting in  $g \geq 2$  (i.e. ignore always-treated case)
  - E.g. for  $T = 3$ , there are three counterfactuals for each  $(i, t)$ :

$$Y_{it}(2), Y_{it}(3), Y_{it}(\infty)$$

where  $Y_{it}(\infty)$  is the PO for the never-treated

- Corresponding treatment paths:

$$g = 2 : [0 \ 1 \ 1], \quad g = 3 : [0 \ 0 \ 1], \quad \text{and } g = \infty : [0 \ 0 \ 0]$$

- How many POs for unit  $i$  are there in this example?  $T \times |\mathcal{G}| = 3 \times 3 = 9$
- So even for small structures, there are *many* POs

# Observed Outcome as Function of Potential Outcomes

- Compare to  $Y_{it}(d)$  with  $d \in \{0, 1\}$ 
  - Were only able to consider static counterfactuals:

*What if  $(i, t)$  had not been treated?*

- But now, can consider dynamic counterfactuals:

*What if  $(i, t)$  had been treated earlier/later?*

- How do we relate POs to observed outcome?

- Let  $D_i^g = 1 [G_i = g]$
- Then the observed outcome is given by

$$Y_{it} = \sum_{g \in \mathcal{G}} D_i^g Y_{it}(g)$$

- Assume *no anticipation*:  $Y_{it}(g) = Y_{it}(g')$  for all  $i; g, g' > t$

# Dynamic Group-Time ATT

- We now define the *dynamic group-time ATT* as

$$\tau_t(g) \equiv \mathbb{E}[Y_{it}(g) - Y_{it}(\infty) | G_i = g]$$

- $\tau_t(g)$  formalize heterogeneity along two dimensions:

- 1 *Period-specific* heterogeneity:  $\tau_t(g) \neq \tau_{t'}(g)$
- 2 *Group-specific* heterogeneity:  $\tau_t(g) \neq \tau_t(g')$

- Consider simple example:  $T = 3$  and  $G_i \in \{2, 3\}$

- Period-specific effects might differ:  $\tau_2(2) \neq \tau_3(2)$

$$\mathbf{D} = \begin{bmatrix} 0 & 0 & 1 \\ 0 & \textcolor{red}{1} & \textcolor{red}{1} \end{bmatrix}$$

- Group-specific effects might differ:  $\tau_3(2) \neq \tau_3(3)$

$$\mathbf{D} = \begin{bmatrix} 0 & 0 & \textcolor{red}{1} \\ 0 & 1 & \textcolor{red}{1} \end{bmatrix}$$

- Effects could differ along both dimensions

# Table of Contents

## 1 Evaluating DID Designs

Parallel Trends

## 2 Dynamic DID

Definitions: Dynamic Heterogeneous PO and ATT

**Dynamic Effects**

Group-Time Heterogeneity

## 3 Appendix

Functional Form

## Event Study Parameters

- We focus first on *event-specific* heterogeneity
- To this end, define *event study* parameters:

$$\tau_{g+\ell}(g) = \mathbb{E} [Y_{ig+\ell}(g) - Y_{ig+\ell}(\infty) | G_i = g]$$

where  $\ell$  is the number of periods ("lags") since treatment

- $\ell < 0$ : periods before first treatment period
- $\ell \geq 0$ : periods after treatment "switched on"
- In block structures (i.e. single  $G_i < \infty$ ),  $\tau_{g+\ell}(g)$  collapses to

$$\tau_\ell \equiv \mathbb{E} [Y_{ig+\ell}(g) - Y_{ig+\ell}(\infty) | D_i = 1]$$

- For  $T = 5$  and  $g = 3$ , we would have  $\tau_{-2}, \tau_{-1}, \tau_0, \tau_1, \tau_2$
- Due to *no anticipation*:  $\tau_\ell = 0$  for  $\ell < 0$
- So we are after  $\tau_0, \tau_1$  and  $\tau_2$ : *dynamic effects*
- E.g. policy effects in year of enactment; one, two years later

# Identification of $\tau_\ell$ in Staggered Rollouts

- No surprise: requires *parallel trends assumption* – assume

$$\mathbb{E} [Y_{it}(\infty) - Y_{it-1}(\infty) | G_i = g]$$

is the same for all  $g \in \mathcal{G}$

- Logic:
  - First identify  $\tau_0$  using PTA (same as static case)
  - Then identify  $\tau_1$  using PTA and  $\tau_0$
  - And so forth, up to  $\tau_{T-g+1}$
- Relies on *no anticipation*:  $Y_{it}(g) = Y_{it}(g')$  for  $g, g' > t$ 
  - Specifically,  $Y_{it}(\infty) = Y_{it}(g)$  for  $g > t$
  - Which is why this works even if no  $G_i = \infty$

# Estimation of $\tau_\ell$ in Block Structures

- Construct event-time indicators:
  - Define  $P_t^\ell = 1 [t = g + \ell]$  for  $\ell \in \{-g + 1, \dots, T - g\}$ , e.g.  $\{-2, -1, 0, 1, 2\}$
  - Set  $P_t^{-1} = 0$  as reference time (and avoid multicollinearity)
  - So relevant set of event-time indicators is e.g.  $\{-2, 0, 1, 2\}$
- Then run *dynamic DID regression*:

$$Y_{it} = \alpha_i + \gamma_t + \sum_{\ell \in \mathcal{L}} D_i P_t^\ell \tau_\ell + \varepsilon_{it}$$

- By including  $\ell < 0$  in  $\mathcal{L}$ , estimate pre-trend
  - Normalization  $P_t^{-1} = 0$  equivalent to  $\tau_{-1} = 0$
  - $\tau_\ell = 0$  for  $\ell < 0$  if *no anticipation* holds
  - Violated by e.g. Ashenfelter's Dip

# Matrix Visualization of Dynamic DID Estimation

- What are post indicators picking up? Assume:

$$\mathbf{D} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 \end{bmatrix}$$

and note that (demonstrating on three out of four  $P_t^\ell$ )

$$\begin{array}{c} P_t^{-2} : \\ \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 \end{bmatrix} \end{array} \quad \begin{array}{c} P_t^0 : \\ \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 \end{bmatrix} \end{array} \quad \begin{array}{c} P_t^1 : \\ \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 \end{bmatrix} \end{array}$$

- Thus, the interactions  $D_i P_t^\ell$  capture (for each  $\tau_\ell$ ):

$$\begin{array}{c} D_i P_t^{-2} : \\ \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 \end{bmatrix} \end{array} \quad \begin{array}{c} D_i P_t^0 : \\ \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 \end{bmatrix} \end{array} \quad \begin{array}{c} D_i P_t^1 : \\ \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 \end{bmatrix} \end{array}$$

i.e. measure treatment mean in  $\ell$  relative to  $-1$

## Estimation of $\tau_\ell$ in Staggered Rollout

- Treatment timing differs across groups  $G_i$ 
  - Varying number of pre- and post-periods per group
- Data determines bounds  $(\ell_L, \ell_U)$ : largest number of pre/post periods
  - Display only subset of pre/post estimates with sufficient precision: judgement call
- Then define event-time treatments:
  - Let  $D_{it}^\ell = 1 [t = G_i + \ell]$  for  $\ell_L < \ell < \ell_U$
  - Again set  $D_{it}^{-1} = 0$  as reference and avoid multicollinearity
- Then run *event study regression*:

$$Y_{it} = \alpha_i + \gamma_t + \sum_{\ell \in \mathcal{L}} D_{it}^\ell \tau_\ell + \varepsilon_{it}$$

- Note: imposing  $\tau_\ell(g) = \tau_\ell$  (homogeneity) for all  $g$ 
  - Leads to similar issues as  $\tau_{TWFE}$  if violated
  - In fact, it is also TWFE!

# Matrix Visualization of Event Study Estimation

- What are event-study indicators picking up?

$$\mathbf{D} = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 \end{bmatrix}$$

- Here  $(\ell_L, \ell_U) = (-4, 4)$ , hence we want  $\tau_{-4}, \tau_{-3}, \tau_{-2}, \tau_0, \tau_1, \tau_2, \tau_3, \tau_4$ 
  - But might want to display only e.g.  $\tau_{-2}, \dots, \tau_2$  for clarity
  - $\tau_{-4}, \tau_{-3}, \tau_{-2}$  are placebo tests;  $\tau_0, \dots, \tau_4$  are treatment effect estimates
  - $\tau_{-1} = 0$  is reference category
  - So for example, looking at 3 (of 8)  $D_{it}^{\ell}$ :

$$\begin{array}{c} D_{it}^{-2} : \\ \begin{bmatrix} 0 & 0 & \color{blue}{0} & \color{green}{0} & 1 \\ \color{blue}{0} & \color{green}{0} & 1 & 1 & 1 \\ \color{green}{0} & 1 & 1 & 1 & 1 \end{bmatrix} \end{array} \quad \begin{array}{c} D_{it}^0 : \\ \begin{bmatrix} 0 & 0 & 0 & \color{green}{0} & \color{blue}{1} \\ 0 & \color{green}{0} & \color{blue}{1} & 1 & 1 \\ \color{green}{0} & 1 & 1 & 1 & 1 \end{bmatrix} \end{array} \quad \begin{array}{c} D_{it}^1 : \\ \begin{bmatrix} 0 & 0 & 0 & \color{green}{0} & 1 \\ 0 & \color{green}{0} & 1 & \color{blue}{1} & 1 \\ \color{green}{0} & 1 & \color{blue}{1} & 1 & 1 \end{bmatrix} \end{array}$$

again measure treatment mean in  $\ell$  relative to  $-1$

# Event Time Visualization

- Event study is equivalent to the following procedure:

- ① Residualize  $Y_{it}$  with respect to  $\alpha_i$  and  $\gamma_t$
- ② Move all treatment paths into *event time* indexed by  $\ell$ :

$$\begin{bmatrix} 0 & 0 & 0 & \textcolor{green}{0} & 1 & \cdot & \cdot & \cdot \\ \cdot & \cdot & 0 & \textcolor{green}{0} & 1 & 1 & 1 & \cdot \\ \cdot & \cdot & \cdot & \textcolor{green}{0} & 1 & 1 & 1 & 1 \end{bmatrix}$$

- ③ Compute means of  $\tilde{Y}_{it}$  per event time, e.g. for  $\tau_1$  (i.e.  $D_{it}^1$ ):

$$\begin{bmatrix} 0 & 0 & 0 & \textcolor{green}{0} & 1 & \cdot & \cdot & \cdot \\ \cdot & \cdot & 0 & \textcolor{green}{0} & 1 & \textcolor{blue}{1} & 1 & \cdot \\ \cdot & \cdot & \cdot & \textcolor{green}{0} & 1 & \textcolor{blue}{1} & 1 & 1 \end{bmatrix}$$

relative to  $-1$

# Jensen (2007) Example: Raw Data in Staggered Rollout

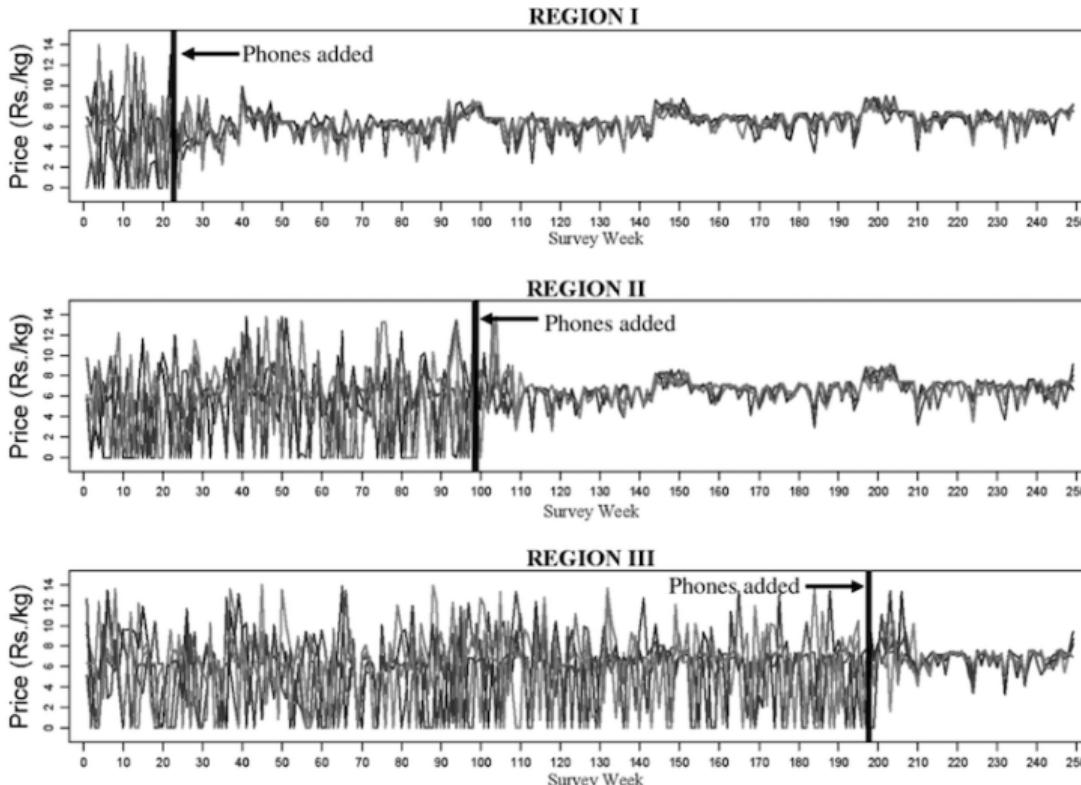


FIGURE IV  
Prices and Mobile Phone Service in Kerala

# Table of Contents

## 1 Evaluating DID Designs

Parallel Trends

## 2 Dynamic DID

Definitions: Dynamic Heterogeneous PO and ATT

Dynamic Effects

**Group-Time Heterogeneity**

## 3 Appendix

Functional Form

# Elementary Target Parameter

- Recall dynamic group-time ATT as

$$\tau_t(g) = \mathbb{E}[Y_{it}(g) - Y_{it}(\infty) | G_i = g]$$

- We have already seen special case  $\tau_\ell$ 
  - Here we assumed  $\tau_\ell(g) = \tau_\ell$  (i.e. homogeneity across groups)
  - But TWFE nature of estimator may lead to bias
- It turns out  $\tau_t(g)$  itself is non-parametrically identified
  - Shown in Callaway and Sant'Anna (published in 2021 and 4,538 citations)
- Once we have an estimand  $\tau_t(g)$ , can aggregate:
  - ① By period:  $\tau_t = \frac{1}{\sum_{g:g \leq t} N_g} \sum_{g:g \leq t} \tau_t(g)$
  - ② By group:  $\tau(g) = \frac{1}{T-g+1} \sum_{t:g \leq t} \tau_t(g)$
  - ③ By event time:  $\tau_{\ell,CS} = \frac{1}{\sum_{g:g+\ell \leq T} N_g} \sum_{g:g+\ell \leq T} \tau_{g+\ell}(g)$

# Aggregate Target Parameter

- Imagine want to study minimum wage changes that occurred over 2004-2007
- Want to know effect of MW increases on employment
- These estimators allow us to make comparisons like:
  - ①  $\tau_t$ : 2006 vs 2007 effect of MW increase occurring in 2004
  - ②  $\tau(g)$ : average post-effect of 2004 vs 2006 MW increase
  - ③  $\tau_\ell(g)$ : effect after 3 years of 2004 vs 2006 MW increase
- Can further aggregate  $\tau_t(g)$ :
  - Mean period effect:  $\tau_{\text{period}} = \frac{1}{T} \sum_t \tau_t$
  - Mean group effect:  $\tau_{\text{group}} = \frac{1}{|\mathcal{G} \setminus \infty|} \sum_{g: g \neq \infty} \tau(g)$
  - “Simple” effect:  $\tau_{\text{simple}} = \frac{1}{\sum_t \sum_{g: g \leq t} N_g} \sum_t \sum_{g: g \leq t} \tau_t(g)$
- Good news!  $\tau_{\text{simple}}$  is unbiased even under heterogeneity

# Callaway & Sant'Anna Estimator

- Basic idea of (one of many) estimators:
  - Consider staggered rollout with  $\infty \in \mathcal{G}$
  - Then let

$$\hat{\tau}_{t,cs}(g) \equiv (\bar{Y}_{g,t} - \bar{Y}_{\infty,t}) - (\bar{Y}_{g,g-1} - \bar{Y}_{\infty,g-1})$$

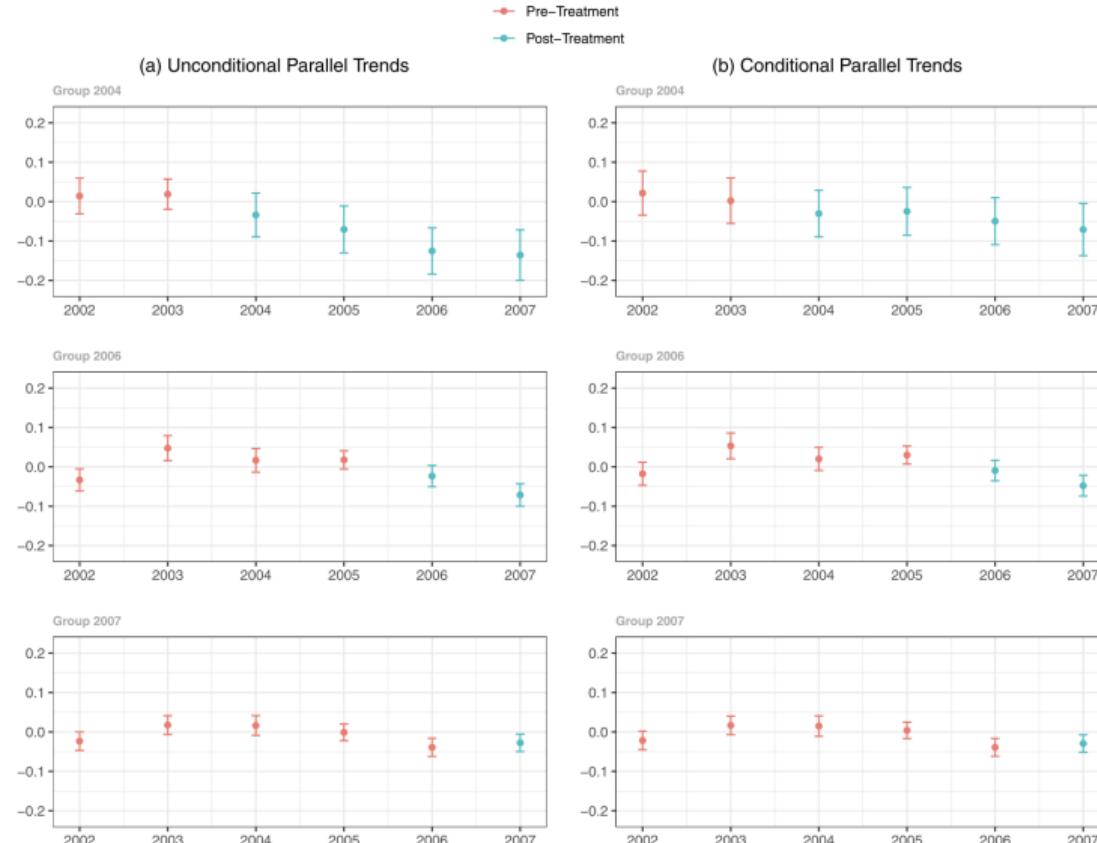
$$\text{where } \bar{Y}_{gt} = \frac{1}{N_{gt}} \sum_{i: G_i=g} Y_{it}$$

- This is the DID that compares outcomes ...
  - ... between  $t$  and period right before treatment  $g-1$  ...
  - ... for the group first treated in period  $g$  ...
  - ... relative to the never-treated group  $\infty$
- For example, for  $T=5, \mathcal{G}=\{3, 4, \infty\}$ :

$$\hat{\tau}_4(4) : \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \end{bmatrix}; \hat{\tau}_4(3) : \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \end{bmatrix}; \hat{\tau}_3(3) : \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \end{bmatrix}$$

- Implementation: Run dynamic DID group-by-group, then aggregate

# Minimum Wage Example: Estimates of $\tau_t(g)$



## Alternative: Imputation Estimator (Borusyak et al 2024)

- Recall parallel trends  $\leftrightarrow$  additive separability:

$$Y_{it}(\infty) = \alpha_i + \gamma_t + \varepsilon_{it}$$

using **only not-yet-treated** observations, i.e. those with  $g > t$

- Can use this to infer  $\hat{Y}_{it}(\infty)$  and construct estimate

$$\hat{\tau}_{it} = Y_{it} - \hat{Y}_{it}(\infty)$$

where  $Y_{it} = Y_{it}(g)$  for  $G_i = g$

- In the example with  $T = 5$  and  $\mathcal{G} = \{3, 4, \infty\}$ :

Use  $\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \cdot & \cdot \\ 0 & 0 & \cdot & \cdot & \cdot \end{bmatrix}$  to infer  $\begin{bmatrix} \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & 0 & 0 \\ \cdot & \cdot & 0 & 0 & 0 \end{bmatrix}$  compared to  $\begin{bmatrix} \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & 1 & 1 \\ \cdot & \cdot & 1 & 1 & 1 \end{bmatrix}$

# Imputation Estimator for Block Structures

- Can aggregate  $\hat{\tau}_{it}$  similarly as with the CS estimator
- In general hard to characterize estimator in closed form
- For block structures with  $G_i = g$  for all  $i$ , it turns out we can write

$$\hat{\tau}_{t,\text{IE}} = (\bar{Y}_{g,t} - \bar{Y}_{g,1:g-1}) - (\bar{Y}_{\infty,t} - \bar{Y}_{\infty,1:g-1})$$

where  $\bar{Y}_{g,1:g'-1} = \frac{1}{N_{g,1:g'-1}} \sum_{t=1}^{g'-1} Y_{it}$  is mean in  $g$  for  $t = 1, \dots, g' - 1$

- For example if  $T = 5$  and  $\mathcal{G} = \{3, \infty\}$ :

$$\hat{\tau}_3 : \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 \end{bmatrix}; \hat{\tau}_4 : \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 \end{bmatrix}$$

- We can see the following tradeoff:
  - 1  $\hat{\tau}_{t,\text{IE}}(g)$  uses more data than  $\hat{\tau}_{t,\text{CS}}(g)$ : smaller SEs
  - 2 However, requires stronger parallel trend assumption: for periods  $t = 1, \dots, g - 1$

# How to Choose an Estimator: Two Decisions

- 1 Can I still use a classical dynamic DID or event study estimator?
  - Dynamic DID in block structures are fine: no group het, allowing for time het
  - With staggered rollout, TWFE leads to forbidden comparisons
  - These are generally problematic when lots of large negative weights
- 2 Which heterogeneous-robust estimator should I use?
  - Besides CS and EI, there are others (Sun and Abraham 2021, Wooldridge 2021)
  - Tradeoff: efficiency versus strength of parallel trends assumption (see also [here](#))
  - Packages exist to implement all these methods (see [here](#))
  - Interpretation in these new estimators can be misleading (see [here](#) and [here](#))
  - Typically, these estimators produce similar results (often similar to event studies)
  - Can be useful to build them yourself to have full control

# Table of Contents

## 1 Evaluating DID Designs

Parallel Trends

## 2 Dynamic DID

Definitions: Dynamic Heterogeneous PO and ATT

Dynamic Effects

Group-Time Heterogeneity

## 3 Appendix

Functional Form

# What is the DID Functional Form (FF) Issue?

- For  $2 \times 2$ , parallel trends is *invariant to transformations* if:

$$\mathbb{E}[h(Y_{i1}(0)) - h(Y_{i0}(0)) | D_i = d]$$

is the same for  $d \in \{0, 1\}$  for all strictly monotonic  $h(\cdot)$

- Says: validity of DID assumption does not depend on units
- For example, does PT hold in logs or levels?
  - Consider  $i \in \{0, 1\}$  (only two units)

$$\begin{bmatrix} Y_{00}(0) & Y_{01}(0) \\ Y_{10}(0) & Y_{11}(0) \end{bmatrix} = \begin{bmatrix} 1 & e \\ e & e^2 \end{bmatrix} \text{ and log transform: } \begin{bmatrix} 0 & 1 \\ 1 & 2 \end{bmatrix}$$

where  $e \approx 2.72$  is Euler's number

- Note that  $2 - 1 = 1 - 0$  (PT holds in logs)
- However,  $e^2 - e \neq e - 1$  (PT does *not* hold in levels)

# When is Parallel Trends Invariant to FF?

Requires parallel trends in CDFs of untreated POs: Define

- $F_{D=d,t=s}^{Y(0)}(y)$ : the CDF of  $Y_{i,t=s}(0) | D_i = d$  (group & time)
- $F_{t=s}^{Y(0)}(y)$ : CDF of  $Y_{is}(0)$  depending only on time
- $F_{D=d}^{Y(0)}(y)$ : CDF of  $Y_{it}(0) | D_i = d$  depending only on group

Proposition (Roth and Sant'Anna 2021)

PT is invariant to transformation iff

$$F_{D=1,t=1}^{Y(0)}(y) - F_{D=1,t=0}^{Y(0)}(y) = F_{D=0,t=1}^{Y(0)}(y) - F_{D=0,t=0}^{Y(0)}(y)$$

This holds iff

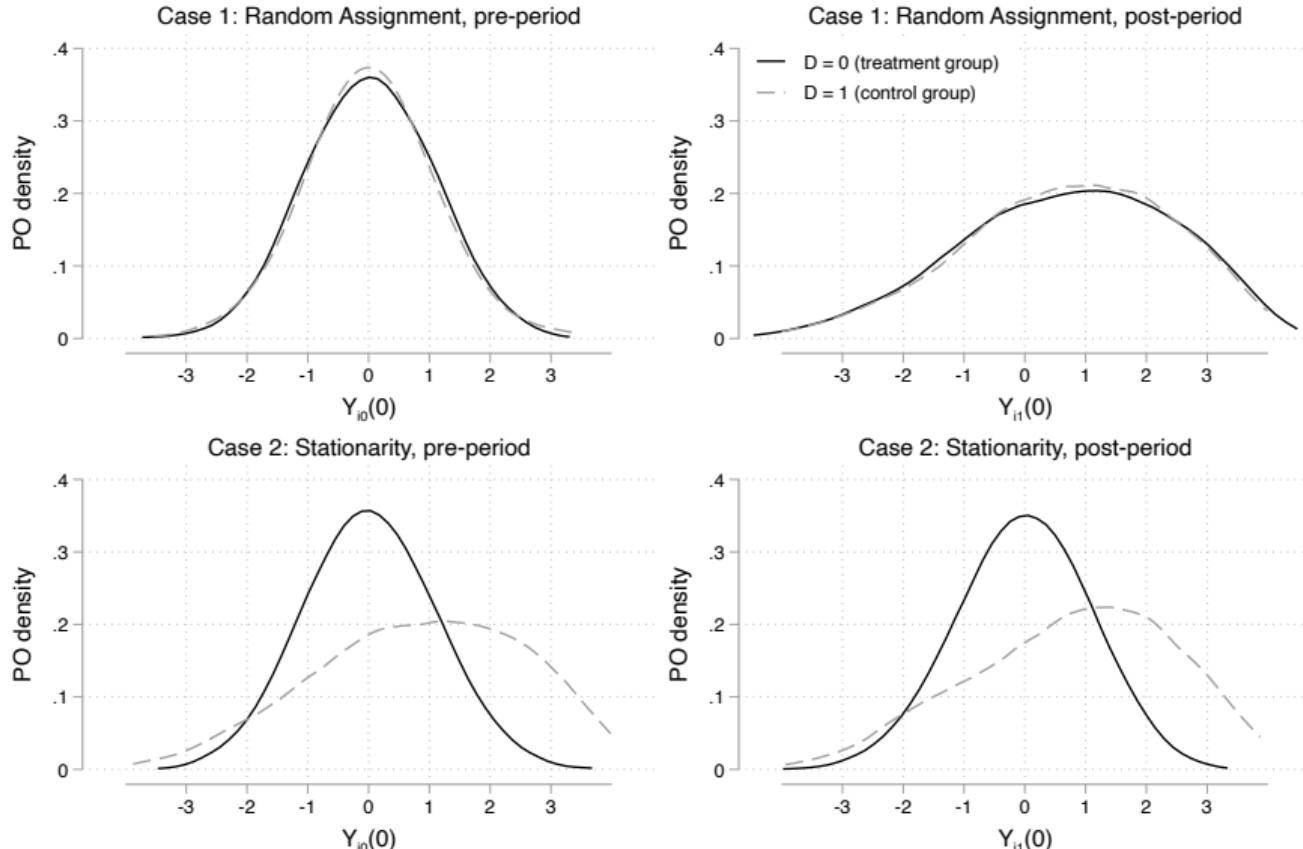
$$F_{D=d,t=s}^{Y(0)}(y) = \theta F_{t=s}^{Y(0)}(y) + (1 - \theta) F_{D=d}^{Y(0)}(y)$$

for all  $y \in \mathbb{R}$  and  $d, s \in \{0, 1\}$ , where  $\theta \in [0, 1]$

# When is Parallel Trends in CDFs Satisfied?

- What does this proposition say?
  - Need strong distributional assumptions for FF not to matter
  - These assumptions fall into three specific cases
- The three cases for which PT in CDFs can hold:
  - ① Random assignment ( $\theta = 1$ ):  $F_{D=1,t}^{Y(0)}(y) = F_{D=0,t}^{Y(0)}(y)$
  - ② Stationary  $\mathbf{Y}_i(0)$  ( $\theta = 0$ ):  $F_{D=d,t=1}^{Y(0)}(y) = F_{D=d,t=0}^{Y(0)}(y)$
  - ③ Partially random and partially stationary ( $\theta \in (0, 1)$ )
- Remarks:
  - Binary outcomes  $Y_{it} \in \{0, 1\}$ : always invariant to FF
  - PT can hold in logs and levels even if  $F_{D=0,t=0}^{Y(0)} \neq F_{D=1,t=0}^{Y(0)}$
  - Can hold even if  $\mathbb{E}[Y_{i0}(0) | D_i = 0] \neq \mathbb{E}[Y_{i0}(0) | D_i = 1]$

# Visualization of Cases: Monte Carlo Sample of POs



# What Are Practical Implications of this Result?

Three important take-aways:

- ① If treatment is (as-if) random, PT is invariant of FF
- ② If not, then DID is sensitive to FF, unless CDFs are parallel
  - Three cases: as-if random; stationarity; or combination
- ③ If PT likely to be sensitive to FF, then justify a specific FF
  - For example, context may say  $\mathbb{E} [\log Y_{it} (0) | \alpha_i, \gamma_t] = \alpha_i + \gamma_t$
  - Then not necessary to be invariant to FF!

# Econometrics II

## Lecture 11: Matched DID and Synthetic Controls

David Schönholzer

Stockholm University

May 7, 2024

# What if Parallel Trends Assumption (PTA) Is Unlikely to Hold?

- Review of DID so far: if PTA holds, then can identify treatment effects:
  - Static and dynamic: lectures 9 and 10
  - Homogeneous or heterogeneous: first parts of lectures versus second parts
- But what if PTA unlikely to hold?
  - Briefly discussed some solutions, e.g. more fixed effects
  - Today: what to do if evidence against PTA or a priori unlikely
  - These methods are typically dynamic and robust to heterogeneity
  - Also allow us to estimate effects when only few units are treated
- Many practical applications!

# Plan for Today

## 1 Matched Difference-in-Differences

Matching with Panel Data

Example

## 2 Synthetic Control Methods

Basic Idea

Examples

## 3 Appendix

Synthetic Controls with Many Treated Units

Matrix Completion Methods

Setup

Examples

# Table of Contents

## 1 Matched Difference-in-Differences

Matching with Panel Data

Example

## 2 Synthetic Control Methods

Basic Idea

Examples

## 3 Appendix

Synthetic Controls with Many Treated Units

Matrix Completion Methods

Setup

Examples

## Reminder: Counterfactuals and Within-Cell Comparisons

- **Covariate imbalance:** control group not a good counterfactual

$$\mathbb{E}[Y_i|D_i = 0] = \mathbb{E}[Y_i(0)|D_i = 0] \neq \mathbb{E}[Y_i(0)|D_i = 1]$$

- Potential solution: control for  $\mathbf{X}_i$  such that

$$\mathbb{E}[Y_i(0)|D_i = 0, \mathbf{X}_i = x] = \mathbb{E}[Y_i(0)|D_i = 1, \mathbf{X}_i = x]$$

and compare treated and control only within cells defined by  $x$

- **PTA violation:** control group *trend* not a good counterfactual

$$\mathbb{E}[Y_{i1} - Y_{i0}|D_i = 0] = \mathbb{E}[Y_{i1}(0) - Y_{i0}(0)|D_i = 0] \neq \mathbb{E}[Y_{i1}(0) - Y_{i0}(0)|D_i = 1]$$

- Potential solution: control for  $\mathbf{X}_i$  such that

$$\mathbb{E}[Y_{i1}(0) - Y_{i0}(0)|D_i = 0, \mathbf{X}_i = x] = \mathbb{E}[Y_{i1}(0) - Y_{i0}(0)|D_i = 1, \mathbf{X}_i = x]$$

and compare treated and control *trends* only within cells defined by  $x$

## Conditional Parallel Trends

- Consider the  $2 \times 2$  DID with  $t \in \{0, 1\}$  with pre-treatment covariate vector  $\mathbf{X}_i$ ;
- We now make the **conditional parallel trends assumption (CPTA)**:

$$\mathbb{E}[Y_{i1}(0) - Y_{i0}(0)|D_i = 0, \mathbf{X}_i] = \mathbb{E}[Y_{i1}(0) - Y_{i0}(0)|D_i = 1, \mathbf{X}_i] \quad \text{a.s.}$$

- We need **sufficient overlap**:  $0 < \Pr(D_i = 1|\mathbf{X}_i) < 1$  almost surely
- We can then identify the ATT at a given  $\mathbf{X}_i = x$  as

$$\tau(x) = \underbrace{\mathbb{E}[Y_{i1} - Y_{i0}|D_i = 1, \mathbf{X}_i = x]}_{\Delta \text{ over time in } D_i=1, \mathbf{X}_i=x} - \underbrace{\mathbb{E}[Y_{i1} - Y_{i0}|D_i = 0, \mathbf{X}_i = x]}_{\Delta \text{ over time in } D_i=0, \mathbf{X}_i=x}$$

i.e. the DID within sub-population  $x$  identifies the ATT for  $x$

- Can then identify unconditional  $\tau$  by averaging:

$$\tau = \mathbb{E}[Y_{i1}(1) - Y_{i1}(0)|D_i = 1] = \mathbb{E}[\tau(\mathbf{X}_i)|D_i = 1]$$

# OLS Approach to CPTA

- May consider running

$$Y_{it} = \alpha_i + \gamma_t + \beta D_i P_t + (\mathbf{X}_i P_t)' \delta + \varepsilon_{it}$$

- However,  $\beta \neq \tau$  unless the ATT is homogeneous w.r.t.  $\mathbf{X}_i$
- Intuition: This OLS setup implicitly models CEF of  $Y_{i1} - Y_{i0}$  as follows:
  - CEF depends on  $\mathbf{X}_i$  with a constant slope  $\delta$ , regardless of  $D_i$
  - If ATT varies e.g. by age then change in CEF may depend on  $D_i$
- See Abadie (2005) for details on this issue
- Luckily, there are semi-/non-parametric solutions to this problem

# Regression Adjustment to DID with CPTA

- Heckman et al (1997) exploit:

$$\begin{aligned}\tau &= \mathbb{E} [\tau(\mathbf{X}_i) | D_i = 1] \\ &= \mathbb{E} [Y_{i1} - Y_{i0} | D_i = 1] - \mathbb{E} [\mathbb{E} [Y_{i1} - Y_{i0} | D_i = 0, \mathbf{X}_i] | D_i = 1]\end{aligned}$$

where the first line follows from the LIE

- Estimate CEF for controls by  $x$  and average using treated  $\mathbf{X}_i$ -distribution:

$$\hat{\tau}_{\text{RA}} = \frac{1}{N_1} \sum_{i:D_i=1} \left\{ (Y_{i1} - Y_{i0}) - \hat{\mathbb{E}} [Y_{i1} - Y_{i0} | D_i = 0, \mathbf{X}_i] \right\}$$

- $\hat{\mathbb{E}} [Y_{i1} - Y_{i0} | D_i = 0, \mathbf{X}_i]$  is estimated CEF evaluated for treated  $\mathbf{X}_i$ ;
- How do we construct estimates for it?
  - Many semi-/non-parametric solutions
  - Often used: matching

# DID Regression Adjustment Through Matching

- Goal: find control units with similar (or identical)  $\mathbf{X}_i$  as treated
- Example of matching estimation procedure:
  - ① For each treated  $i$ , choose unit  $m(i)$  such that  $m(i) = \arg \min_{m: D_i=0} \|\mathbf{X}_i - \mathbf{X}_m\|$
  - ② Estimate  $\widehat{\mathbb{E}} [Y_{i1} - Y_{i0} | D_i = 0, \mathbf{X}_i]$  using  $\widehat{\mathbb{E}} [Y_{m(i),1} - Y_{m(i),0}]$
  - ③ Assign  $G_i$  to  $m(i)$  and estimate e.g. matched event study as

$$Y_{it} = \alpha_i + \gamma_t + \underbrace{\sum_{\ell \in \mathcal{L}} 1[t = g + \ell] \delta_\ell}_{\text{avg. trend in treated and matched controls}} + \underbrace{\sum_{\ell \in \mathcal{L}} D_i P_t^\ell \beta_\ell}_{\text{deviation in treated}} + \varepsilon_{it}$$

- Intuitively appealing, often with very compelling results, plus het-robust!
- However, inference has complications (often ignored), Abadie Imbens (2006)

# Table of Contents

## 1 Matched Difference-in-Differences

Matching with Panel Data

Example

## 2 Synthetic Control Methods

Basic Idea

Examples

## 3 Appendix

Synthetic Controls with Many Treated Units

Matrix Completion Methods

Setup

Examples

## Goldschmidt and Schmieder (2017)

- Interested in effect of outsourcing on daily wages
- Use matched administrative worker-firm data from Germany
- Outsourcing: workers with same job and site but switch to subsidiary firm
- Concern: control group should capture time and life cycle effects
- Matched DID approach:
  - ① Let  $D_i$  be whether a worker is ever affected by an outsourcing event
  - ② Donor pool of control workers: same industry/occupation but  $D_i = 0$
  - ③  $\mathbf{X}_i$ : tenure, establishment size, wages at  $\ell = -2$  and  $-3$
  - ④ Run probit of  $D_i$  on  $\mathbf{X}_i$  to and compute fitted values  $\hat{e}(\mathbf{X}_i)$
  - ⑤ Match each  $i$  to  $m(i) = \arg \min_{m: D_i=0} |\hat{e}(\mathbf{X}_i) - \hat{e}(\mathbf{X}_m)|$

# Match Evaluation

TABLE I  
CHARACTERISTICS OF OUTSOURCED AND NONOUTSOURCED WORKERS

	Outsourced at $t = -1$	Matched non-OS at $t = -1$	FCSL at BSF/temp	FCSL not at BSF/temp
Avg establishment daily wage in euro	78.83 (20.16)	77.42 (20.32)	53.65 (19.59)	74.49 (17.94)
Establishment effect*	0.03 (0.14)	0.03 (0.15)	-0.14 (0.18)	0.02 (0.15)
Establishment size	1,120.63 (2,416.86)	1,107.55 (3,207.42)	265.41 (385.18)	1,683.45 (5,204.99)
Real daily wage in euro	69.93 (29.47)	69.96 (30.73)	51.07 (24.80)	63.71 (25.36)
Age in years	42.29 (7.98)	43.63 (9.75)	40.25 (8.49)	41.87 (8.43)
Female	0.45	0.46	0.40	0.40
Years of education	10.16 (1.17)	10.23 (1.34)	9.93 (1.06)	10.06 (0.89)
College degree	0.02	0.03	0.01	0.01
Living in West Germany	0.86	0.88	0.85	0.94
Working full-time	0.78	0.76	0.70	0.78
Tenure in years	8.58 (5.80)	8.51 (6.32)	3.91 (3.83)	6.16 (5.29)
Food occupation	0.21	0.21	0.05	0.14
Cleaning occupation	0.11	0.11	0.41	0.24
Security occupation	0.03	0.03	0.11	0.08
Logistics occupation	0.34	0.34	0.42	0.53
Observations	21,195	21,195	6,412,854	35,201,181

Notes. Mean of each variable with standard deviation in parentheses. Columns (1) and (2) include on-site outsourced and matched nonoutsourced workers age 25–55 with at least 2 years of tenure in year before outsourcing. Statistics calculated in year before outsourcing. Columns (3) and (4) include workers in food, cleaning, security, and logistics occupations who are age 25–55 and employed at an establishment with 50 or more workers. Column (3) includes these workers who are employed at business service firms (BSF) or temp firms, and column (4) includes these workers who are not employed at BSF or temp firms. All columns exclude East Germany prior to 1997.

\*The establishment effects are the predicted fixed effects from the AKM model described in Section IV.A. The establishment effects are normalized to be equal to 0 in the sample of all workers from 1979 to 2009 (the period we use for the AKM model).

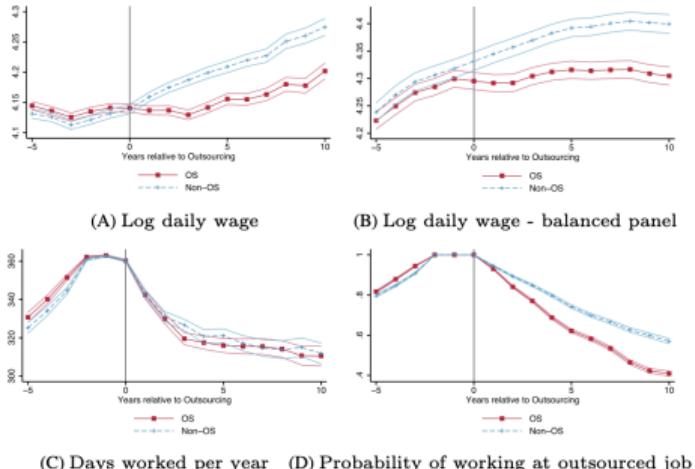
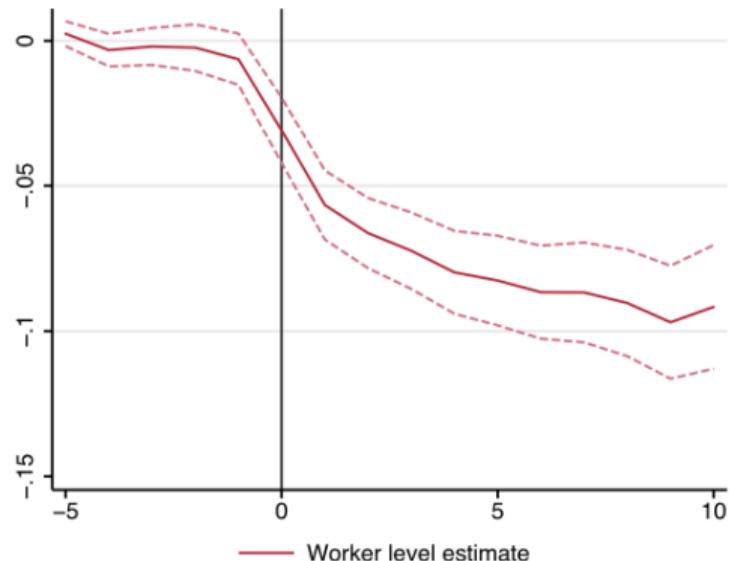


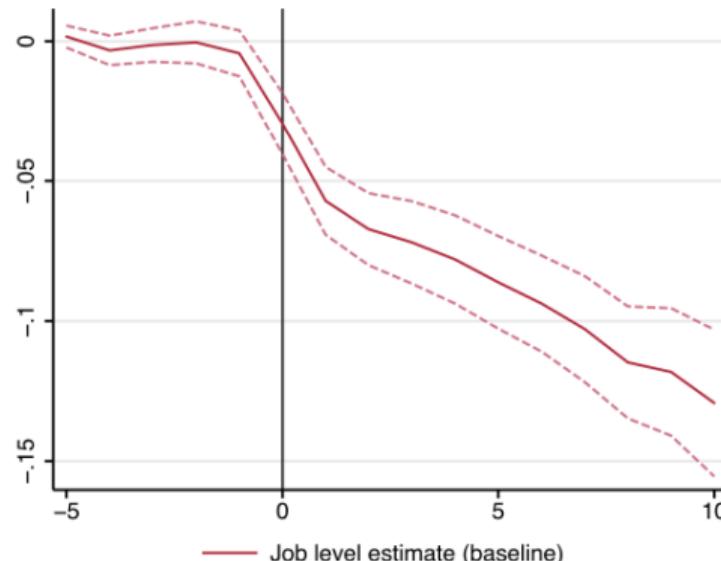
FIGURE IV  
Employment Outcomes of Outsourced and Nonoutsourced Workers before and after On-site Outsourcing

The figures follow two groups of workers: the first is a group of workers who are outsourced between year  $t = -1$  and  $t = 0$  (the first year at the new establishment), while the second group is a control group of nonoutsourced workers. The control group was chosen by finding workers employed in the same industry and occupation with similar tenure and establishment size in the year prior to outsourcing, who have similar wages two and three years prior to outsourcing as the outsourced workers. The figures show average characteristics of the workers in the two groups before and after the outsourcing event. Panels A, C, and D show data from the unbalanced panels of workers in the outsourced and control group. Panel B restricts the data to a balanced panel of individuals observed in each year from 5 years before to 10 years after the outsourcing event.

# Goldschmidt and Schmieder (2017): Event Studies on Daily Wages



(A) All worker observations before and after outsourcing



(B) Sample restricted to observations remaining at the same job

# Table of Contents

## 1 Matched Difference-in-Differences

Matching with Panel Data

Example

## 2 Synthetic Control Methods

Basic Idea

Examples

## 3 Appendix

Synthetic Controls with Many Treated Units

Matrix Completion Methods

Setup

Examples

## Return to NJ and PA

- Recall that we said

$$\begin{aligned}\text{Treatment effect} &= (\bar{Y}_{\text{NJ},1} - \bar{Y}_{\text{PA},1}) - (\bar{Y}_{\text{NJ},0} - \bar{Y}_{\text{PA},0}) \\ &= (\bar{Y}_{\text{NJ},1} - \bar{Y}_{\text{NJ},0}) - (\bar{Y}_{\text{PA},1} - \bar{Y}_{\text{PA},0})\end{aligned}$$

- We showed that DID uses

$$\hat{\tau}_{\text{DID}} = \underbrace{(\bar{Y}_{\text{NJ},1} - \bar{Y}_{\text{NJ},0})}_{\Delta \text{ over time in } D_i = 1} - \underbrace{(\bar{Y}_{\text{PA},1} - \bar{Y}_{\text{PA},0})}_{\Delta \text{ over time in } D_i = 0}$$

- Alternatively, could do

$$\hat{\tau}_{\text{SCM}} = \underbrace{(\bar{Y}_{\text{NJ},1} - \bar{Y}_{\text{PA},1})}_{\Delta \text{ over group in } t = 1} - \underbrace{(\bar{Y}_{\text{NJ},0} - \bar{Y}_{\text{PA},0})}_{\Delta \text{ over group in } t = 0}$$

- This is approach of *synthetic control method* (SCM)

## A Simple $2 \times T$ Case

- Consider some  $2 \times T$  block structure with  $g = T$ , e.g.

$$\mathbf{D} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

- Imagine only one unit  $i = N$  in treatment group
- We are interested in

$$\begin{aligned}\tau &= \mathbb{E}[Y_{iT}(1) - Y_{iT}(0) | D_i = 1] \\ &= \textcolor{blue}{Y_{NT}(1)} - \textcolor{red}{Y_{NT}(0)} = Y_{NT} - \underbrace{Y_{NT}(0)}_{\text{model}}\end{aligned}$$

- Imagine cross-sectional OLS with all  $i = 1, \dots, N - 1$ :

$$Y_{iT} = \phi_0 + \sum_{t=1}^{T-1} \phi_t Y_{it} + \varepsilon_{iT}$$

which Athey et al. (2018) call *horizontal regression*

# Horizontal Regression

- Why is it called horizontal regression?

$$\begin{bmatrix} Y_{i1} & \cdots & Y_{iT-1} & \leftarrow Y_{iT} \\ Y_{N1} & \cdots & Y_{NT-1} & Y_{NT} \end{bmatrix}$$

i.e. regressing post-control on pre-control

- We can then construct the fitted value

$$\hat{Y}_{NT} = \hat{\phi}_0 + \sum_{t=1}^{T-1} \hat{\phi}_t Y_{Nt}$$

as our estimate of  $Y_{NT}(0)$ , so that  $\hat{\tau} = Y_{NT} - \hat{Y}_{NT}$

- This is very close what DID using PT does! It sets:

$$\hat{\phi}_0 = \bar{Y}_{1,\text{pre}} - \bar{Y}_{0,\text{pre}}$$

$$\hat{\phi}_t = \frac{1}{T-1}$$

# Vertical Regression

- Consider instead a *vertical regression*:

$$\begin{bmatrix} Y_{i1} & \cdots & Y_{iT-1} & Y_{iT} \\ \uparrow & \uparrow & \uparrow \\ Y_{N1} & \cdots & Y_{NT-1} & Y_{NT} \end{bmatrix}$$

i.e. regressing pre-treatment on pre-control

- Construct the fitted value

$$\hat{Y}_{NT} = \hat{\lambda}_0 + \sum_{i=1}^{N-1} \hat{\lambda}_i Y_{iT}$$

- This is almost what SCM does, with restrictions:

- 1  $\hat{\lambda}_0 = 0$
- 2  $\hat{\lambda}_i \geq 0$  for all  $i \leq N - 1$
- 3  $\sum_{i=1}^{N-1} \hat{\lambda}_i = 1$

- Unlike DID, PT does *not* have to hold unconditionally for SCM

# Implementation

- Let

$$\begin{bmatrix} \mathbf{Y}_{c,\text{pre}} & \mathbf{Y}_{c,T} \\ \mathbf{Y}'_{t,\text{pre}} & Y_{t,T} \end{bmatrix}$$

where

- $\mathbf{Y}_{c,\text{pre}} = (\mathbf{Y}_{c,1}, \dots, \mathbf{Y}_{c,T-1})$  is  $(N - 1) \times (T - 1)$  pre-control matrix
- $\mathbf{Y}_{t,\text{pre}} = (Y_{N1}, \dots, Y_{NT-1})'$  is  $(T - 1) \times 1$  pre-treatment vector
- $\mathbf{Y}_{c,T} = (Y_{1T}, \dots, Y_{N-1,T})'$  is  $(N - 1) \times 1$  post-control vector
- SCM is typically implemented as CMD:

$$\min_{\lambda=(\lambda_1, \dots, \lambda_{N-1})} (\mathbf{Y}_{t,\text{pre}} - \mathbf{Y}'_{c,\text{pre}} \lambda)' \mathbf{W} (\mathbf{Y}_{t,\text{pre}} - \mathbf{Y}'_{c,\text{pre}} \lambda)$$

with  $\mathbf{W}$  a symmetric weighting matrix

- Yields weights  $(\hat{\lambda}_1, \dots, \hat{\lambda}_{N-1})$  that form *synthetic control group*

# Table of Contents

## 1 Matched Difference-in-Differences

Matching with Panel Data

Example

## 2 Synthetic Control Methods

Basic Idea

Examples

## 3 Appendix

Synthetic Controls with Many Treated Units

Matrix Completion Methods

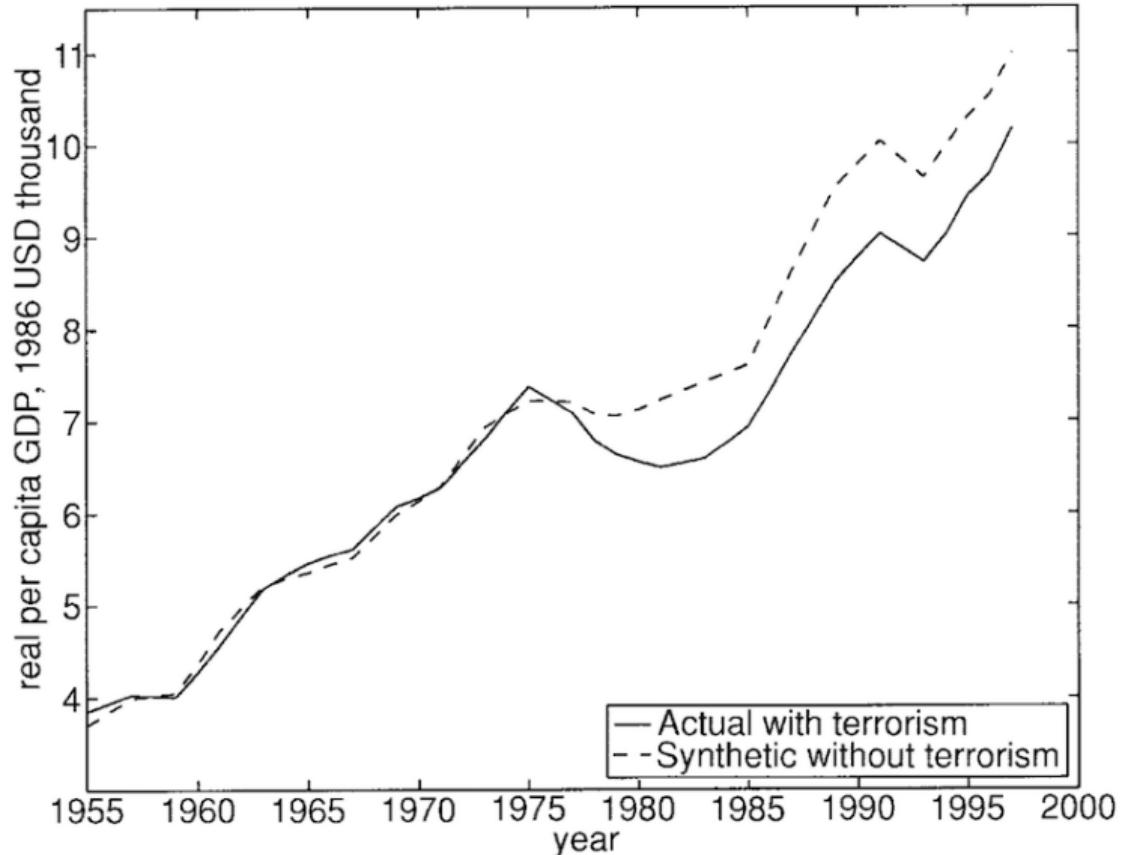
Setup

Examples

## Abadie & Gardeazabal (2003)

- Question: what is economic effect of terrorism?
- Authors study the Basque country (treatment unit)
- Want to use other Spanish regions as control units
  - But no parallel pre-trends in average of others
  - Construct weighted average using SCM – birth of SCM
- What are the weights?
  - $\hat{\lambda}_i = 0$  for most regions
  - $\hat{\lambda}_{\text{Catalonia}} = 0.85$  and  $\hat{\lambda}_{\text{Madrid}} = 0.15$
- So does this synthetic control group “work”?

## Actual and Synthetic Control Trend



# Andersson (2019): OECD Average (Left); Synthetic Sweden (Right)

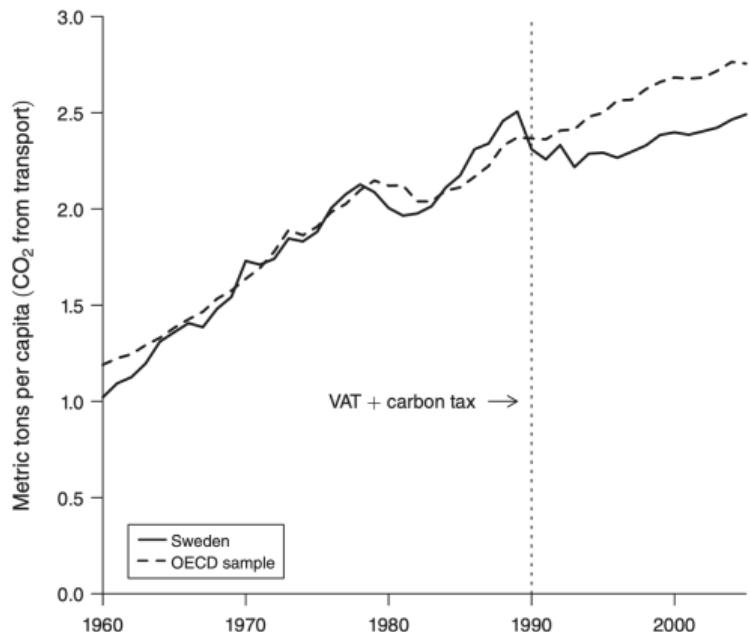


FIGURE 3. PATH PLOT OF PER CAPITA CO<sub>2</sub> EMISSIONS FROM TRANSPORT DURING 1960–2005:  
SWEDEN VERSUS THE OECD AVERAGE OF MY 14 DONOR COUNTRIES

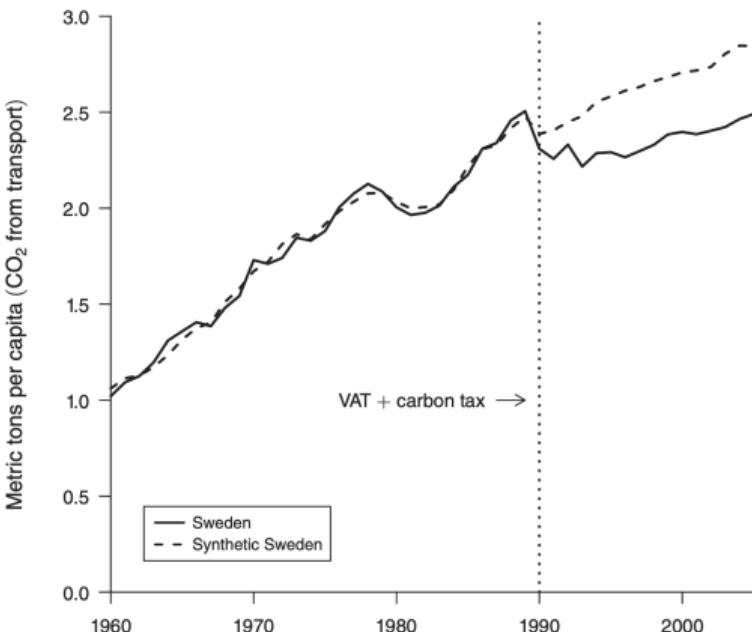


FIGURE 4. PATH PLOT OF PER CAPITA CO<sub>2</sub> EMISSIONS FROM TRANSPORT DURING 1960–2005:  
SWEDEN VERSUS SYNTHETIC SWEDEN

# Comparison and Country Weights

TABLE 1—CO<sub>2</sub> EMISSIONS FROM TRANSPORT PREDICTOR MEANS BEFORE TAX REFORM

Variables	Sweden	Synth. Sweden	OECD sample
GDP per capita	20,121.5	20,121.2	21,277.8
Motor vehicles (per 1,000 people)	405.6	406.2	517.5
Gasoline consumption per capita	456.2	406.8	678.9
Urban population	83.1	83.1	74.1
CO <sub>2</sub> from transport per capita 1989	2.5	2.5	3.5
CO <sub>2</sub> from transport per capita 1980	2.0	2.0	3.2
CO <sub>2</sub> from transport per capita 1970	1.7	1.7	2.8

*Notes:* All variables except lagged CO<sub>2</sub> are averaged for the period 1980–1989. GDP per capita is purchasing power parity (PPP)—adjusted and measured in 2005 US dollars. Gasoline consumption is measured in kilograms of oil equivalent. Urban population is measured as percentage of total population. CO<sub>2</sub> emissions are measured in metric tons. The last column reports the population-weighted averages of the 14 OECD countries in the donor pool.

TABLE 2—COUNTRY WEIGHTS IN SYNTHETIC SWEDEN

Country	Weight	Country	Weight
Australia	0.001	Japan	0
Belgium	0.195	New Zealand	0.177
Canada	0	Poland	0.001
Denmark	0.384	Portugal	0
France	0	Spain	0
Greece	0.090	Switzerland	0.061
Iceland	0.001	United States	0.088

*Note:* With the synthetic control method, extrapolation is not allowed so all weights are between  $0 \leq w_j \leq 1$  and  $\sum w_j = 1$ .

# Treatment Effect Estimate

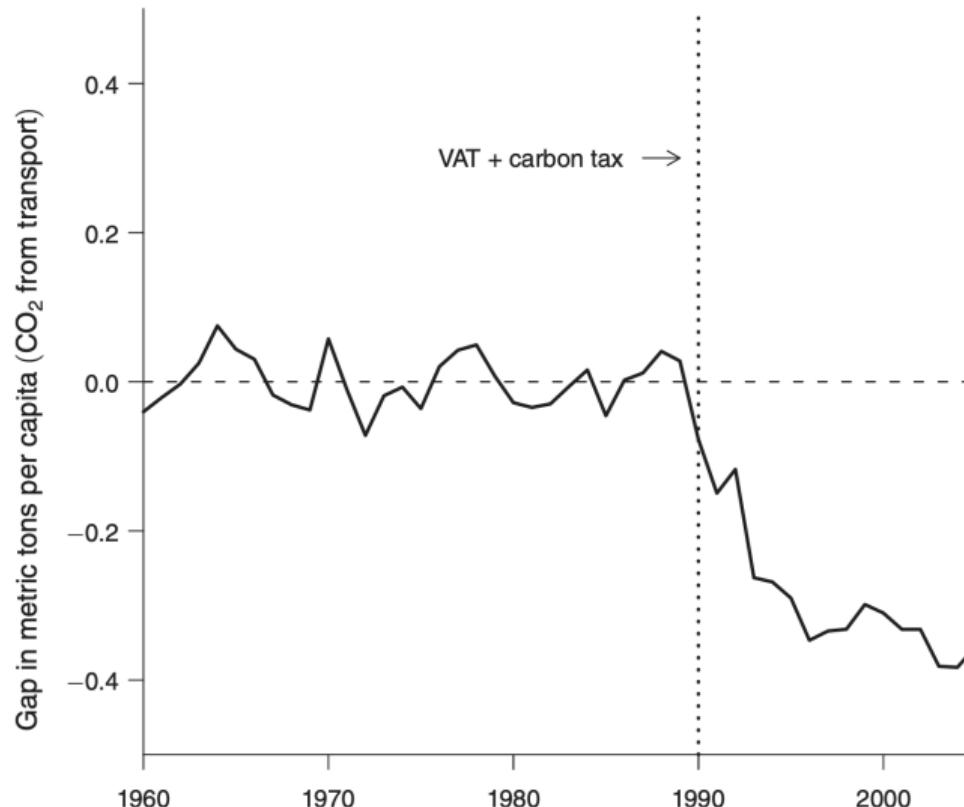


FIGURE 5. GAP IN PER CAPITA  $\text{CO}_2$  EMISSIONS FROM TRANSPORT BETWEEN SWEDEN AND SYNTHETIC SWEDEN

# Placebo in Time

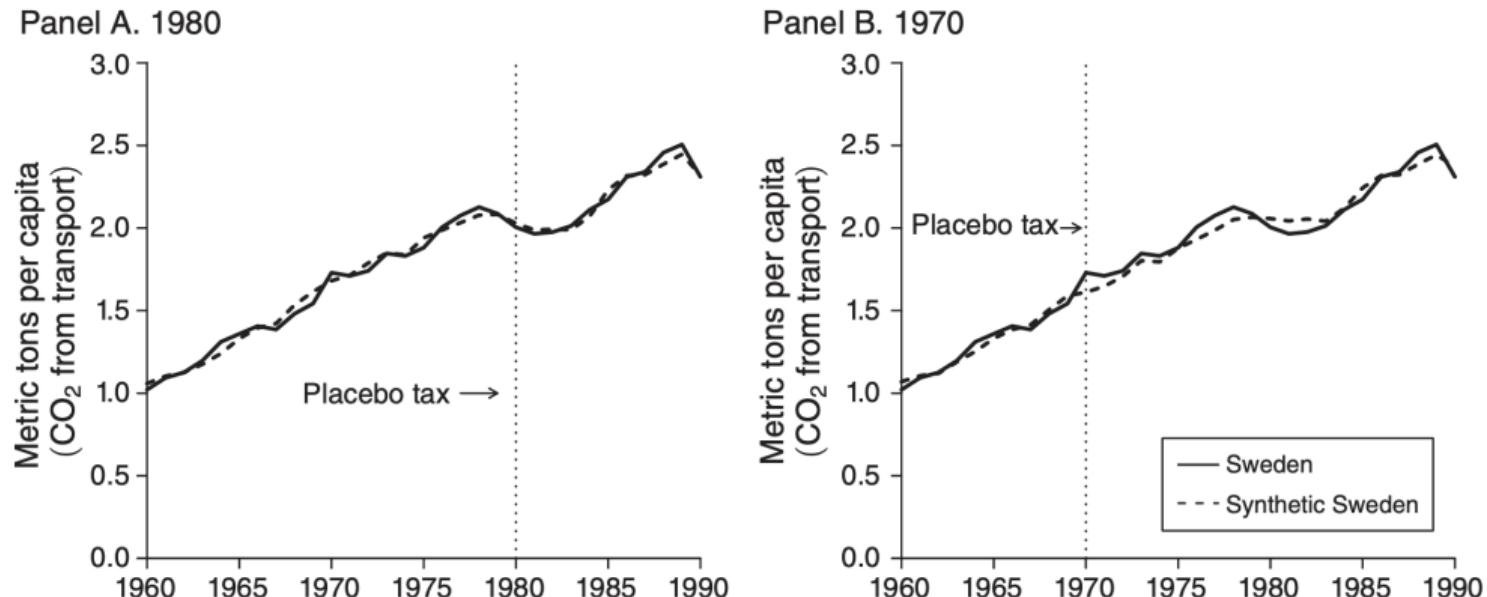
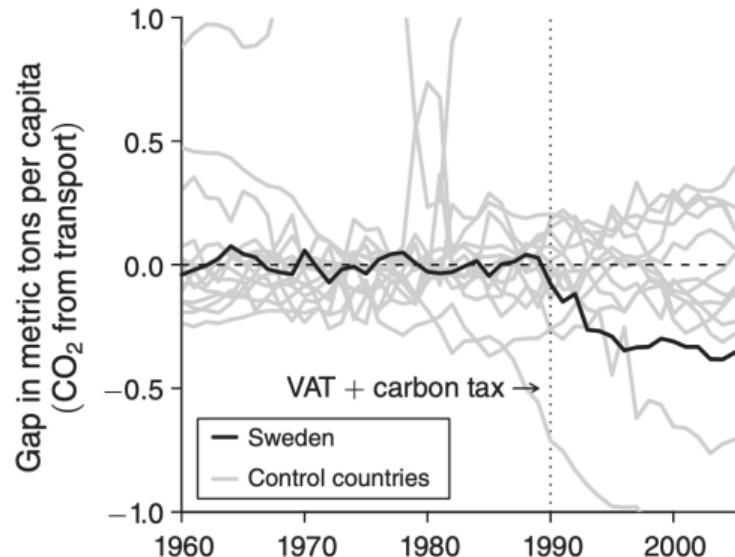


FIGURE 6. PLACEBO IN-TIME TESTS

*Notes:* In panel A, the placebo tax is introduced in 1980, ten years prior to the actual policy changes. In panel B, the placebo tax is introduced in 1970.

# Placebo Across Countries and Inference

Panel A



Panel B

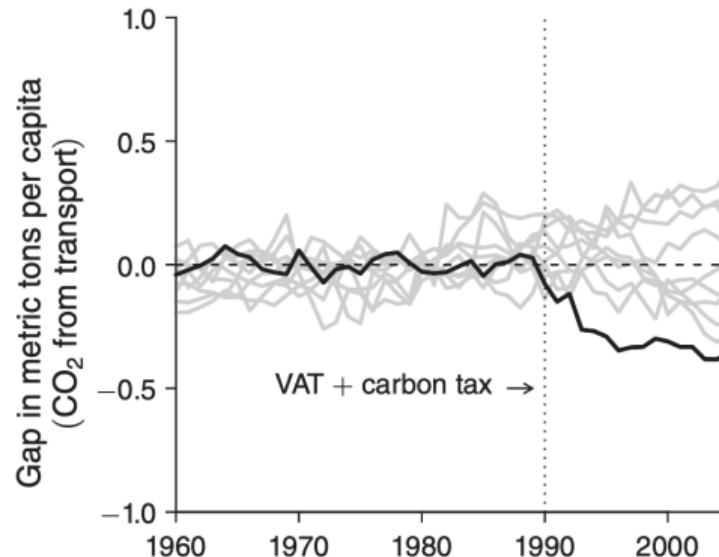


FIGURE 7. PERMUTATION TEST: PER CAPITA  $\text{CO}_2$  EMISSIONS GAP IN SWEDEN AND PLACEBO GAPS FOR THE CONTROL COUNTRIES

Notes: Panel A shows per capita  $\text{CO}_2$  emissions gap in Sweden and placebo gaps in all 14 OECD control countries. Panel B shows per capita gap in Sweden and placebo gaps in nine OECD control countries (countries with a pretreatment MSPE 20 times higher than Sweden's are excluded).

# Table of Contents

## 1 Matched Difference-in-Differences

Matching with Panel Data

Example

## 2 Synthetic Control Methods

Basic Idea

Examples

## 3 Appendix

### Synthetic Controls with Many Treated Units

Matrix Completion Methods

Setup

Examples

## Multiple Treated Units

- Ben-Michael et al (2022) develop SCM for staggered rollout: multiple units
- Same setup as before, but  $N_0$  units never-treated and  $N_1 = N - N_0$  treated
- Treatment times given by  $G_i$  with  $G_i = \infty$  for never-treated
- Continue to assume **no anticipation**:  $Y_{it}(g) = Y_{it}(\infty)$  for  $g > t$ , implying:

$$Y_{it} = 1[t < G_i] Y_{it}(\infty) + 1[t \geq G_i] Y_{it}(G_i)$$

- **Target parameter**: unit-specific dynamic treatment effect for each  $i : G_i < \infty$ :

$$\tau_{il} = Y_{ig+\ell}(G_i) - Y_{ig+\ell}(\infty)$$

and by no anticipation we have  $\tau_{il} = 0$  for any  $\ell < 0$

- Can use these to compute **dynamic ATT estimates**:

$$\tau_\ell = \frac{1}{N_1} \sum_{i: G_i < \infty} \tau_{il}$$

# SCM Estimator

- Define donor set  $\mathcal{D}_{i\ell}$ : all never treated or not-yet treated  $j \neq i$
- The SCM weights for treated unit  $i$  are estimated via

$$\min_{\lambda_i \in \Lambda_i} \underbrace{\frac{1}{G_i - 1} \sum_{\ell < 0, i: G_i < \infty} \left( Y_{ig+\ell} - \sum_{j=1}^N \lambda_{ji} Y_{jg+\ell} \right)^2}_{\text{SCM objective}} + \underbrace{\rho \sum_{i=1}^N \lambda_{ji}^2}_{\text{Regularization}}$$

with  $\Lambda_i$  such that  $\lambda_{ji} = 0$  when  $j \notin \mathcal{D}_{i\ell}$

- Regularization tunes objective towards including all units
- Missing potential outcome is then estimated as

$$\hat{Y}_{ig+\ell}(\infty) = \sum_{j=1}^N \lambda_{ji} Y_{jg+\ell}$$

- And treatment effect estimates are  $\hat{\tau}_{i\ell} = Y_{ig+\ell} - \hat{Y}_{ig+\ell}(\infty)$

# Mean Squared Placebo Treatment Effect

- Let  $\ell_U$  be number of pre-treatment estimates
- Then  $\hat{\tau}_i^{\text{pre}} = (\hat{\tau}_{i,\ell_U}, \dots, \hat{\tau}_{i,-1}) \in \mathbb{R}^{\ell_U}$  is vector of placebo effects
- Hence the SCM objective is the mean squared placebo effect:

$$q_i(\hat{\lambda}_i)^2 \equiv \frac{1}{G_i - 1} \|\hat{\tau}_i^{\text{pre}}\|_2^2 = \frac{1}{G_i - 1} \sum_{\ell < 0, i: G_i < \infty} \left( Y_{ig+\ell} - \sum_{j=1}^N \lambda_{ji} Y_{jg+\ell} \right)^2$$

- Want to pick weights  $\hat{\lambda}_i$  to minimize this error

## Combining Multiple Treated Units

- We now want  $\lambda_i$  for all  $i$  with  $G_i < \infty$
- Collect these into  $N \times N_1$  matrix  $\Lambda = [\lambda_{N_0+1}, \dots, \lambda_N]$
- Can write  $\hat{\tau}_\ell$  in two equivalent ways:

$$\hat{\tau}_\ell = \frac{1}{N_1} \sum_{i:G_i < \infty} \hat{\tau}_{i\ell}$$

$$(\text{average of unit-specific SCM-estimates}) = \frac{1}{N_1} \sum_{i:G_i < \infty} \left[ Y_{ig+\ell} - \sum_{j=1}^N \hat{\lambda}_{ji} Y_{jg+\ell} \right]$$

$$(\text{SCM-estimate for average treated unit}) = \frac{1}{N_1} \sum_{i:G_i < \infty} Y_{ig+\ell} - \sum_{i:G_i < \infty} \sum_{j=1}^N \frac{\hat{\lambda}_{ji}}{N_j} Y_{jg+\ell}$$

# Two Goodness-of-Fit Measures

- Depending on interpretation of  $\hat{\tau}_\ell$ , can evaluate pre-fit via:
  - ① Root mean square of pre-treatment fits across treated units:

$$q^{\text{sep}}(\hat{\Lambda}) = \sqrt{\frac{1}{N_1} \sum_{i:G_i < \infty} q_i^2(\lambda_i)} = \sqrt{\frac{1}{N_1} \sum_{i:G_i < \infty} \frac{1}{G_i - 1} \|\hat{\tau}_i^{\text{pre}}\|_2^2}$$

- ② Root mean square of pre-treatment fit of average unit:

$$q^{\text{pool}}(\hat{\Lambda}) = \frac{1}{\sqrt{\ell_U}} \left\| \frac{1}{N_1} \sum_{i:G_i < \infty} \hat{\tau}_i^{\text{pre}} \right\|_2$$

- So two different ways to pick weights! Separately or pooled

# Lower Bound of Error in Pre-Treatment Fit

- Focus on effect of first period of treatment  $\tau_0$
- Can show that lower bound of estimation error is weighted average:

Theorem (Estimation error as weighted average)

*Under assumptions on the DGP it can be shown that the error is bounded by:*

$$|\hat{\tau}_0 - \tau_0| \leq \rho^{pool} q^{pool}(\hat{\Lambda}) + \rho^{sep} q^{sep}(\hat{\Lambda}) + \zeta$$

*where  $\rho^{pool}$  and  $\rho^{sep}$  depend on treatment structure and data, and  $\zeta$  is noise*

→ Optimal SCM for staggered rollout is partially pooled!

# Table of Contents

## 1 Matched Difference-in-Differences

Matching with Panel Data

Example

## 2 Synthetic Control Methods

Basic Idea

Examples

## 3 Appendix

Synthetic Controls with Many Treated Units

Matrix Completion Methods

Setup

Examples

# Matrix Completion Methods

- Goal: estimate causal effect of  $D_i \in \{0, 1\}$  in panel data
- We have seen two approaches:
  - ① DID: estimate stable pattern over  $t$  across  $i$
  - ② Synthetic control: estimate stable pattern over  $i$  across  $t$
- New approach: Matrix completion (Athey et al. 2018)
  - Find stable pattern over  $(i, t)$
  - Data-driven mix of patterns across units and periods
  - Combines machine learning/computer science with metrics

## Setup: Control Counterfactual $\mathbf{Y}$

- Consider  $N \times T$  matrix  $\mathbf{Y}$  of outcomes (wide format)
- Only observe  $Y_{it}$  for some  $i$  and  $t$ 
  - Let  $(i, t) \in \mathcal{M}$  be missing entries
  - $(i, t) \in \mathcal{O}$  are observed entries
- Potential outcomes: observe only  $Y_{it}(0)$  or  $Y_{it}(1)$
- Causal inference approach:
  - Impute  $\mathbf{Y}(0) \equiv \mathbf{Y}$  matrix for  $i$  with  $D_i = 1$
  - Use those to construct  $\mathbb{E}[Y_{it}(1) - Y_{it}(0)|D_i = 1]$

# Data Structure

- 1 Complete treatment matrix:

$$\mathbf{D} = \begin{bmatrix} 0 & 0 & 1 & \cdots & 0 \\ 1 & 1 & 0 & \cdots & 1 \\ 0 & 1 & 0 & \cdots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & 0 & \cdots & 1 \end{bmatrix}$$

where  $D_{it} = 1$  if  $(i, t) \in \mathcal{M}$  and zero otherwise

- 2 Incomplete counterfactual outcome matrix:

$$\mathbf{Y} = \begin{bmatrix} Y_{11} & Y_{12} & ? & \cdots & Y_{iT} \\ ? & ? & Y_{23} & \cdots & ? \\ Y_{31} & ? & Y_{33} & \cdots & ? \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Y_{N1} & ? & Y_{N3} & \cdots & ? \end{bmatrix}$$

# Patterns of Missing Data

- Block structure: subset treated (i.e. missing) for all  $t \geq e$

$$\mathbf{Y} = \begin{bmatrix} \circ & \circ & \circ & \circ & \cdots & \circ \\ \circ & \circ & \circ & \circ & \cdots & \circ \\ \circ & \circ & \circ & \circ & \cdots & \circ \\ \circ & \circ & \bullet & \bullet & \cdots & \bullet \\ \circ & \circ & \bullet & \bullet & \cdots & \bullet \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \circ & \circ & \bullet & \bullet & \cdots & \bullet \end{bmatrix}$$

where  $\bullet$  denotes  $(i, t) \in \mathcal{M}$  and  $\circ$  mean  $(i, t) \in \mathcal{O}$

# Special Block Structures

- Two special cases of block structures:

$$\mathbf{Y} = \begin{bmatrix} \circ & \circ & \cdots & \circ & \circ \\ \circ & \circ & \cdots & \circ & \circ \\ \circ & \circ & \cdots & \circ & \bullet \\ \circ & \circ & \cdots & \circ & \bullet \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \circ & \circ & \cdots & \circ & \bullet \end{bmatrix} \quad \text{and} \quad \mathbf{Y} = \begin{bmatrix} \circ & \circ & \circ & \cdots & \circ & \circ \\ \circ & \circ & \circ & \cdots & \circ & \circ \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \circ & \circ & \circ & \cdots & \circ & \circ \\ \circ & \circ & \bullet & \cdots & \bullet & \bullet \end{bmatrix}$$

- Left: ATT estimable by DID
- Right: ATT estimable by synthetic control approach

# Staggered Adoption

- Staggered adoption

$$\mathbf{Y} = \begin{bmatrix} \circ & \circ & \circ & \circ & \cdots & \circ \\ \circ & \circ & \circ & \circ & \cdots & \bullet \\ \circ & \circ & \circ & \bullet & \cdots & \bullet \\ \circ & \circ & \bullet & \bullet & \cdots & \bullet \\ \circ & \circ & \bullet & \bullet & \cdots & \bullet \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \circ & \bullet & \bullet & \bullet & \cdots & \bullet \end{bmatrix}$$

- ATT estimable by event study

# Thin and Fat Matrices

- Thin and fat matrices:

$$\mathbf{Y} = \begin{bmatrix} \circ & \circ & \circ \\ \circ & \circ & \circ \\ \circ & \circ & \bullet \\ \circ & \circ & \bullet \\ \vdots & \vdots & \vdots \\ \circ & \circ & \bullet \end{bmatrix} \quad \text{and} \quad \mathbf{Y} = \begin{bmatrix} \circ & \circ & \circ & \cdots & \circ & \circ \\ \circ & \circ & \circ & \cdots & \circ & \circ \\ \circ & \circ & \bullet & \cdots & \bullet & \bullet \end{bmatrix}$$

- Thin matrix: e.g. classical DID with  $T = 3$
- Fat matrix: e.g. synthetic control for single treated unit

# Nuclear Norm Matrix Completion Estimator

- Idea: data-driven mix of horizontal and vertical prediction
- Model counterfactual as  $\mathbf{Y} = \mathbf{L}^* + \boldsymbol{\varepsilon}$  with  $\mathbb{E} [\boldsymbol{\varepsilon} | \mathbf{L}^*] = 0$
- How about just minimizing sum of squared errors?

$$\min_{\mathbf{L}} \frac{1}{|\mathcal{O}|} \sum_{(i,t) \in \mathcal{O}} (Y_{it} - L_{it})^2$$

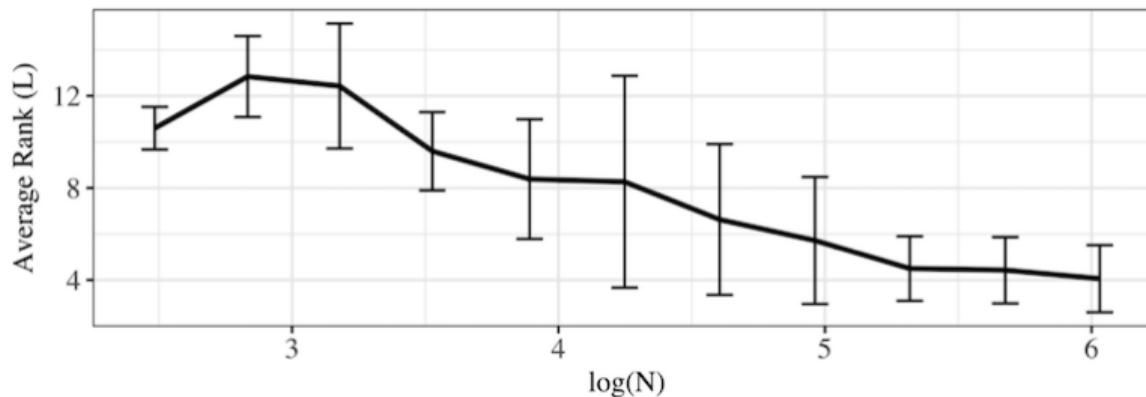
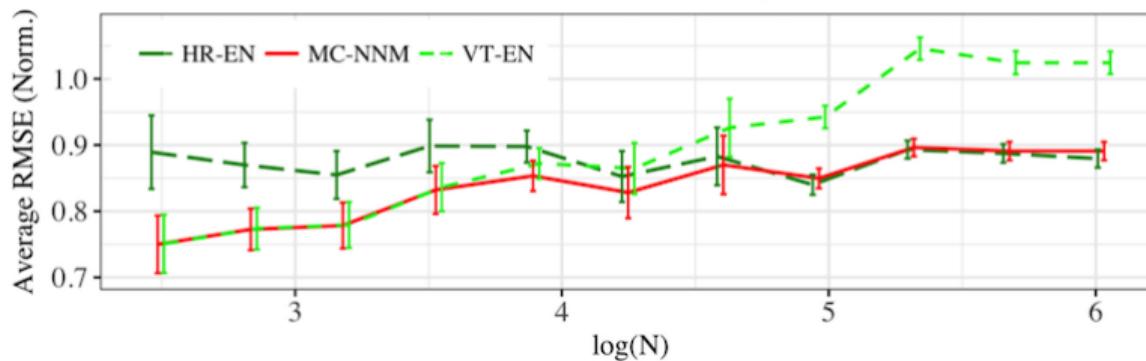
- Does not work! Does not depend on  $L_{it}$  for  $(i, t) \in \mathcal{M}$
- Instead:

$$\widehat{\mathbf{L}} = \arg \min_{\mathbf{L}} \left\{ \frac{1}{|\mathcal{O}|} \sum_{(i,t) \in \mathcal{O}} (Y_{it} - L_{it})^2 + \lambda \|\mathbf{L}\|_* \right\}$$

with penalty factor  $\lambda$  and nuclear matrix norm  $\|\cdot\|_*$

# Athey et al. (2018): Impute Missing Stock Data

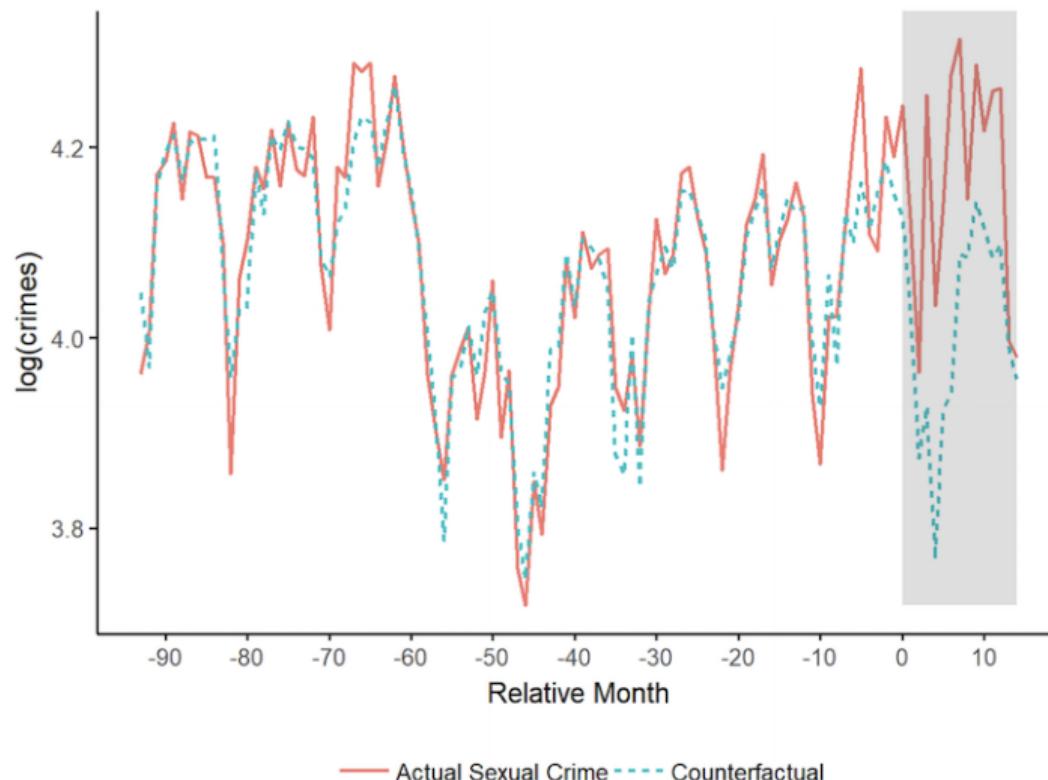
$N \times T = 4900$  Fraction Missing = 0.25



# Levy and Mattsson (2020): Impute Counterfactual Sex Crimes

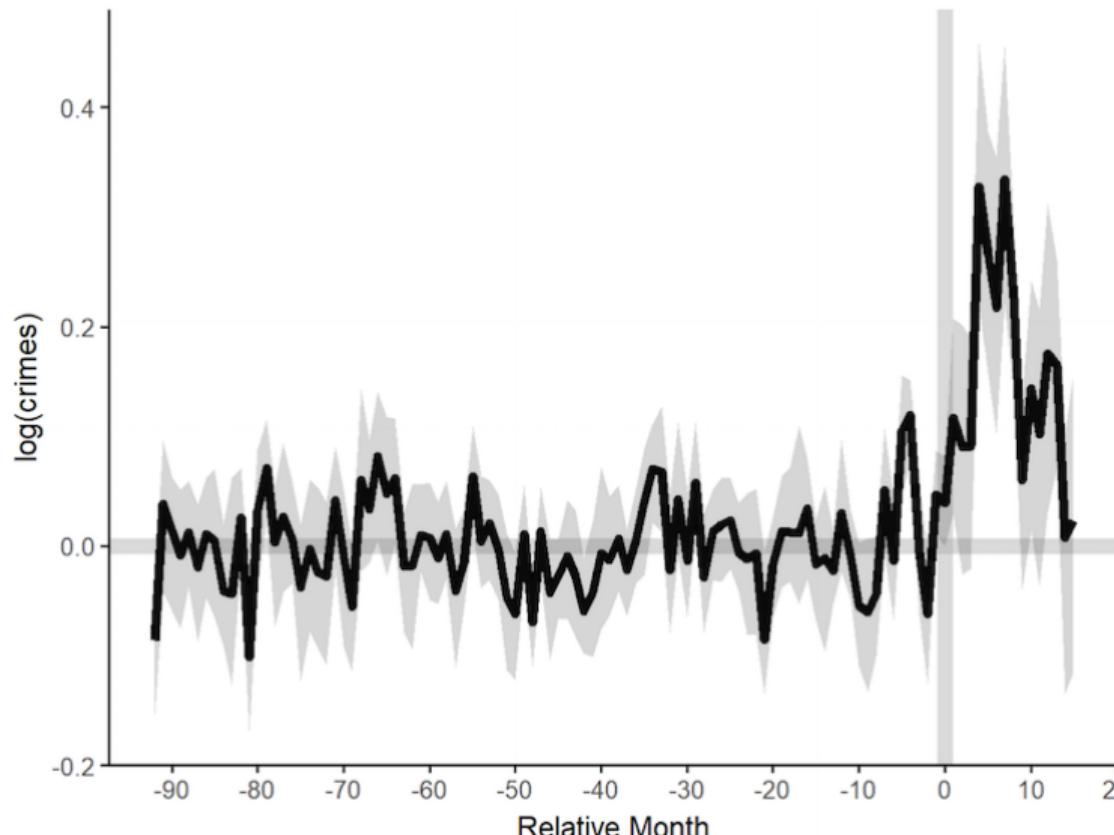
Figure 5: Matrix Completion Results

(a) Counterfactual Versus Actual Outcomes



# Implied Treatment Effect of #MeToo

(b) Average Treatment Effect



# Econometrics II

## Lecture 12: Regression Discontinuity Designs

David Schönholzer

Stockholm University

May 10, 2024

# Table of Contents

## 1 RD Basics

Identification and Interpretation

Estimation and Visualization

## 2 Tuning and Diagnostics

Tuning: Global Versus Local

Diagnostics: Balance and Bunching

Optimal Bandwidth and Inference

## 3 Special Cases

Fuzzy RD Design

Multiple Discontinuities and Multidimensional RDD

Spatial/Boundary Discontinuity Design

## 4 Appendix

Discrete Running Variable

Regression Kink Design

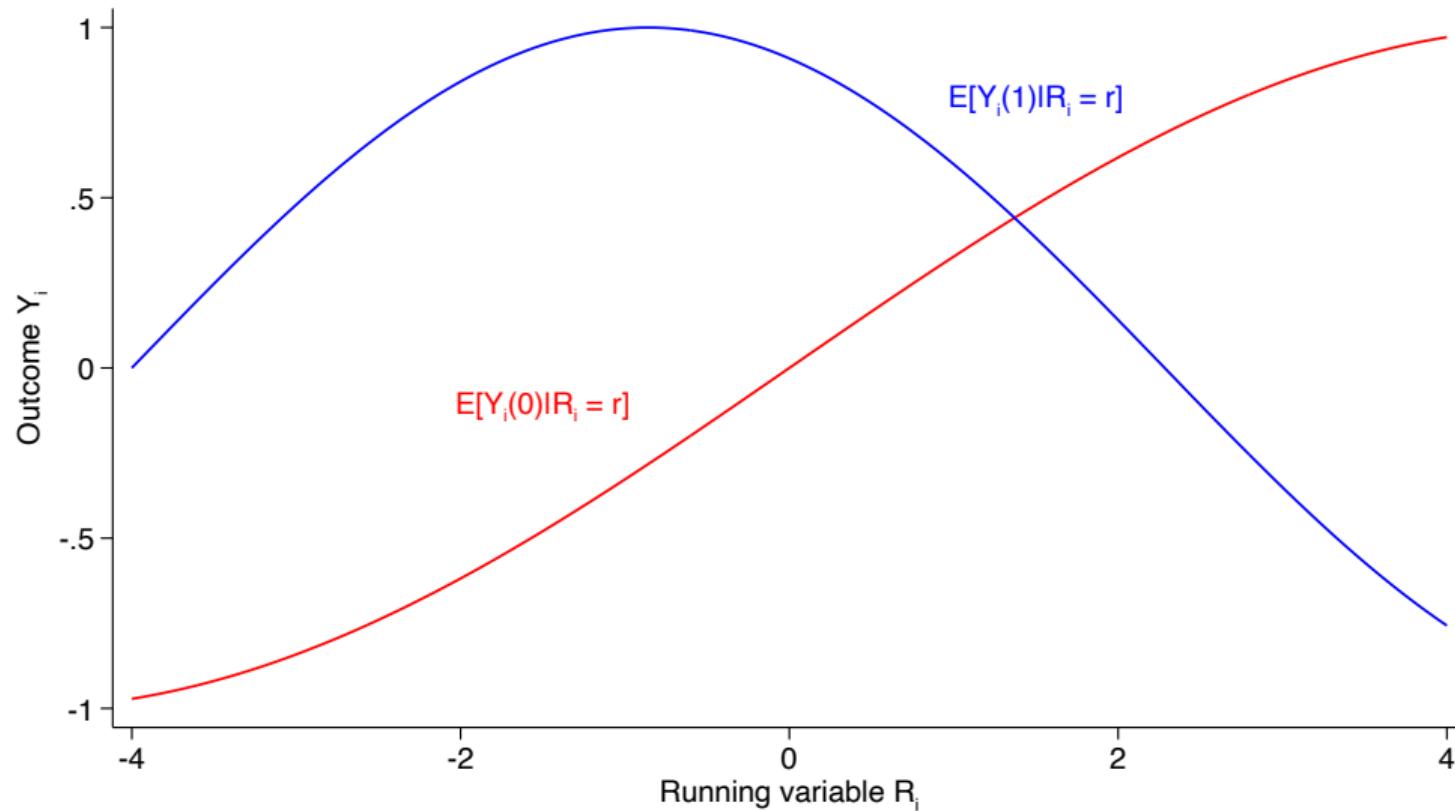
## Basic Setup

- Consider setting with treatment  $D_i \in \{0, 1\}$  and POs  $Y_i(1)$  and  $Y_i(0)$
- Suppose that  $D_i$  is *deterministic* function of observed covariate  $R_i$ :

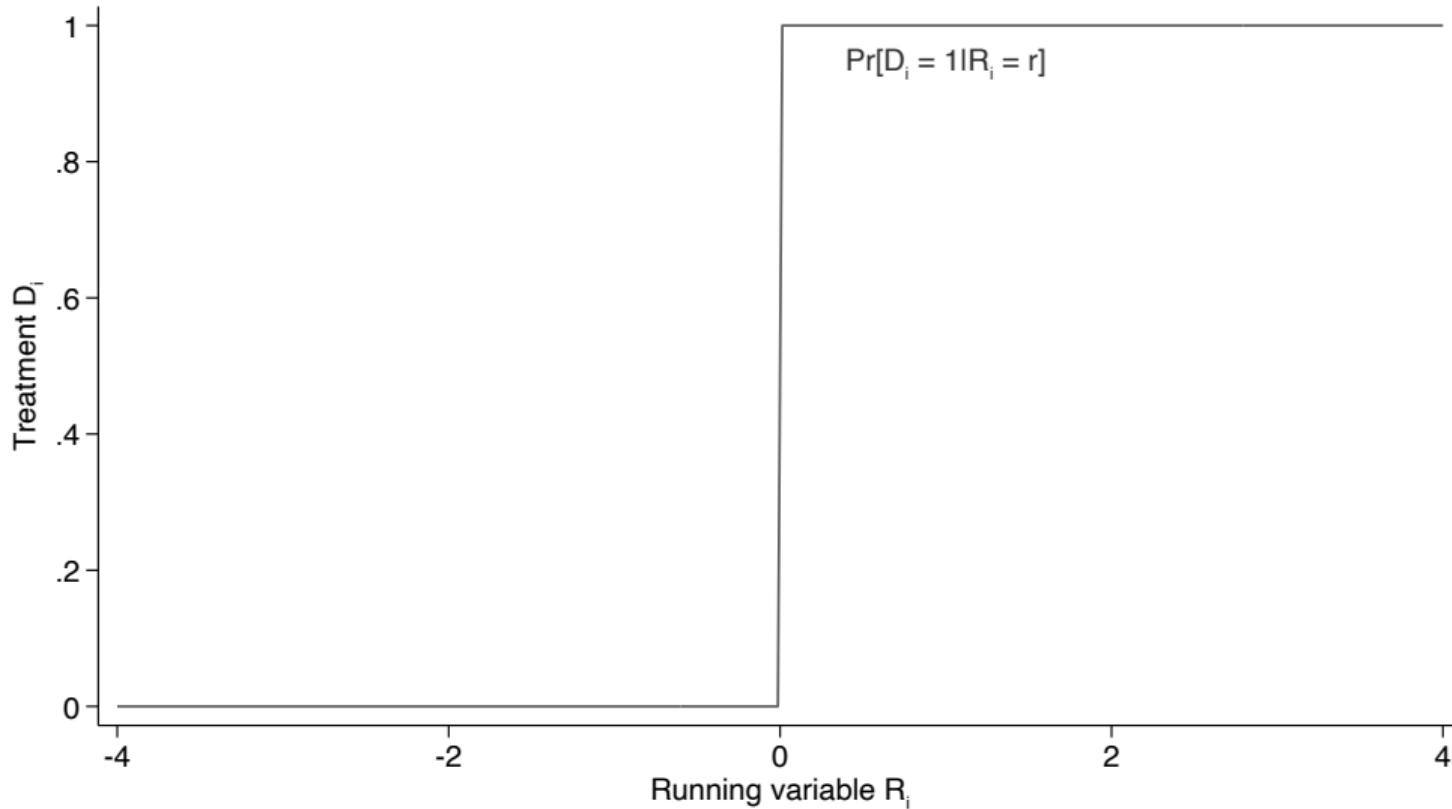
$$D_i = 1[R_i > c]$$

- $R_i$  is called the **running variable**
- This is a **sharp RD** because treatment switches from 0 to 1 at threshold
- We observe  $Y_i(1)$  when  $R_i > c$  and  $Y_i(0)$  when  $R_i \leq c$
- Example: Scholarship awarded to students above test score threshold
- Basic idea:
  - Compare observations just above and below  $c$
  - Intuitively, treatment may be as good as randomly assigned near  $R_i = c$

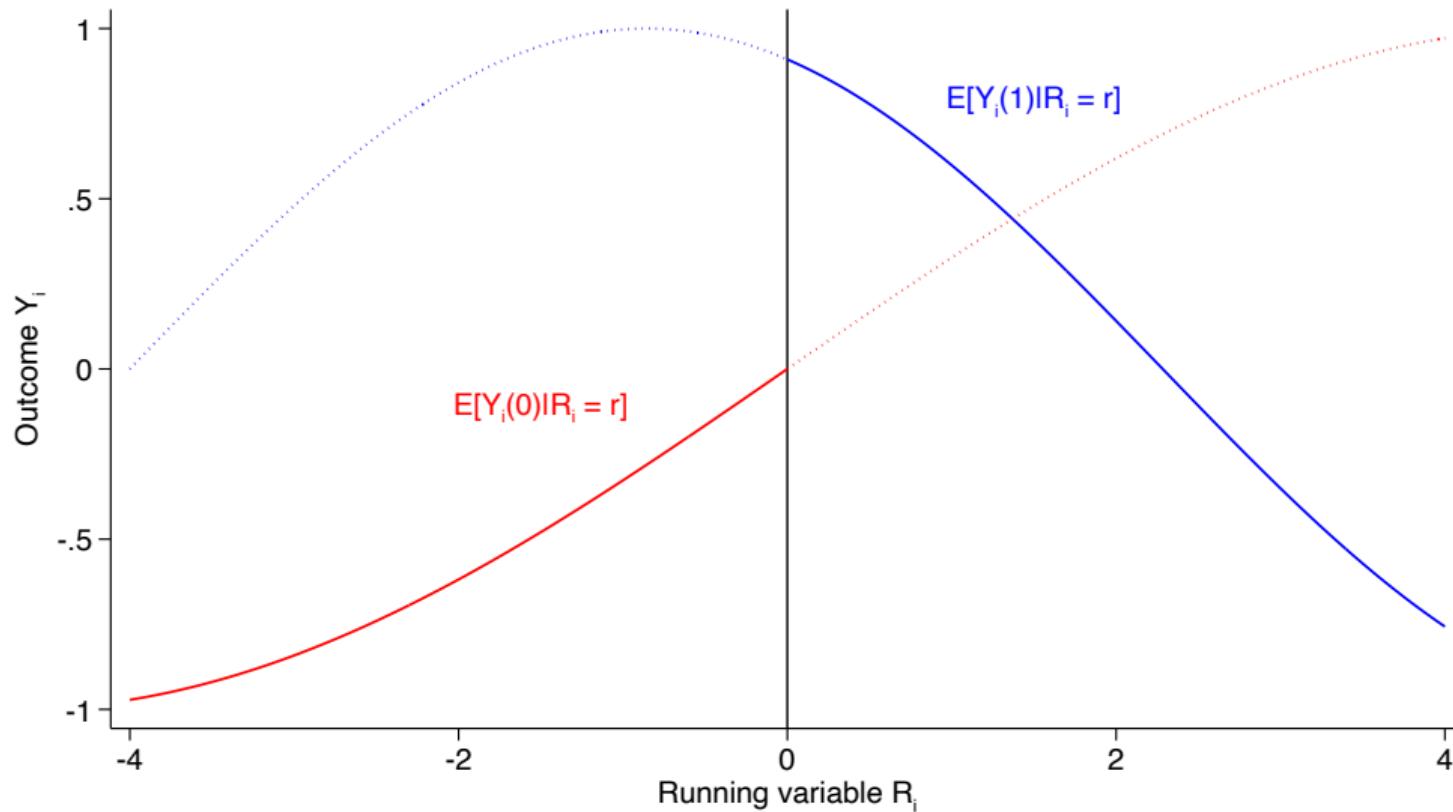
# Potential Outcomes



# Assignment Probabilities



# Observed Outcomes



# Sharp RD Identification

- Key identifying assumption: POs are smooth at the threshold  $c$ :

$$\lim_{r \uparrow c} \mathbb{E}[Y_i(d) | R_i = r] = \lim_{r \downarrow c} \mathbb{E}[Y_i(d) | R_i = r], \quad d \in \{0, 1\}$$

- Potential outcome CEFs need to be **continuous at the threshold**
- The population just below must not be discretely different than above
- If the assumption holds, then

$$\begin{aligned}\lim_{r \downarrow c} \mathbb{E}[Y_i | R_i = r] - \lim_{r \uparrow c} \mathbb{E}[Y_i | R_i = r] &= \lim_{r \downarrow c} \mathbb{E}[Y_i(1) | R_i = r] - \lim_{r \uparrow c} \mathbb{E}[Y_i(0) | R_i = r] \\ &= \mathbb{E}[Y_i(1) | R_i = c] - \mathbb{E}[Y_i(0) | R_i = c] \\ &= \mathbb{E}[Y_i(1) - Y_i(0) | R_i = c] \equiv \tau(c)\end{aligned}$$

# Interpretation

- Note that  $\tau(c)$  is the treatment effect **only for individuals with  $R_i = c \rightarrow \text{LATE}$**
- But identification is **nonparametric**:
  - No assumptions about distribution of  $Y_i(d)$
  - Other than continuity of CEFs
- We have a **CIA condition**:

$$(Y_i(1), Y_i(0)) \perp D_i | R_i$$

- But no common support: conditional on  $R_i$ , always  $D_i = 0$  or  $D_i = 1$
- RD is local extrapolation outside support to predict  $\mathbb{E}[Y_i(d)|R_i = c]$

# Table of Contents

## 1 RD Basics

Identification and Interpretation

Estimation and Visualization

## 2 Tuning and Diagnostics

Tuning: Global Versus Local

Diagnostics: Balance and Bunching

Optimal Bandwidth and Inference

## 3 Special Cases

Fuzzy RD Design

Multiple Discontinuities and Multidimensional RDD

Spatial/Boundary Discontinuity Design

## 4 Appendix

Discrete Running Variable

Regression Kink Design

# Basic Estimation Problem

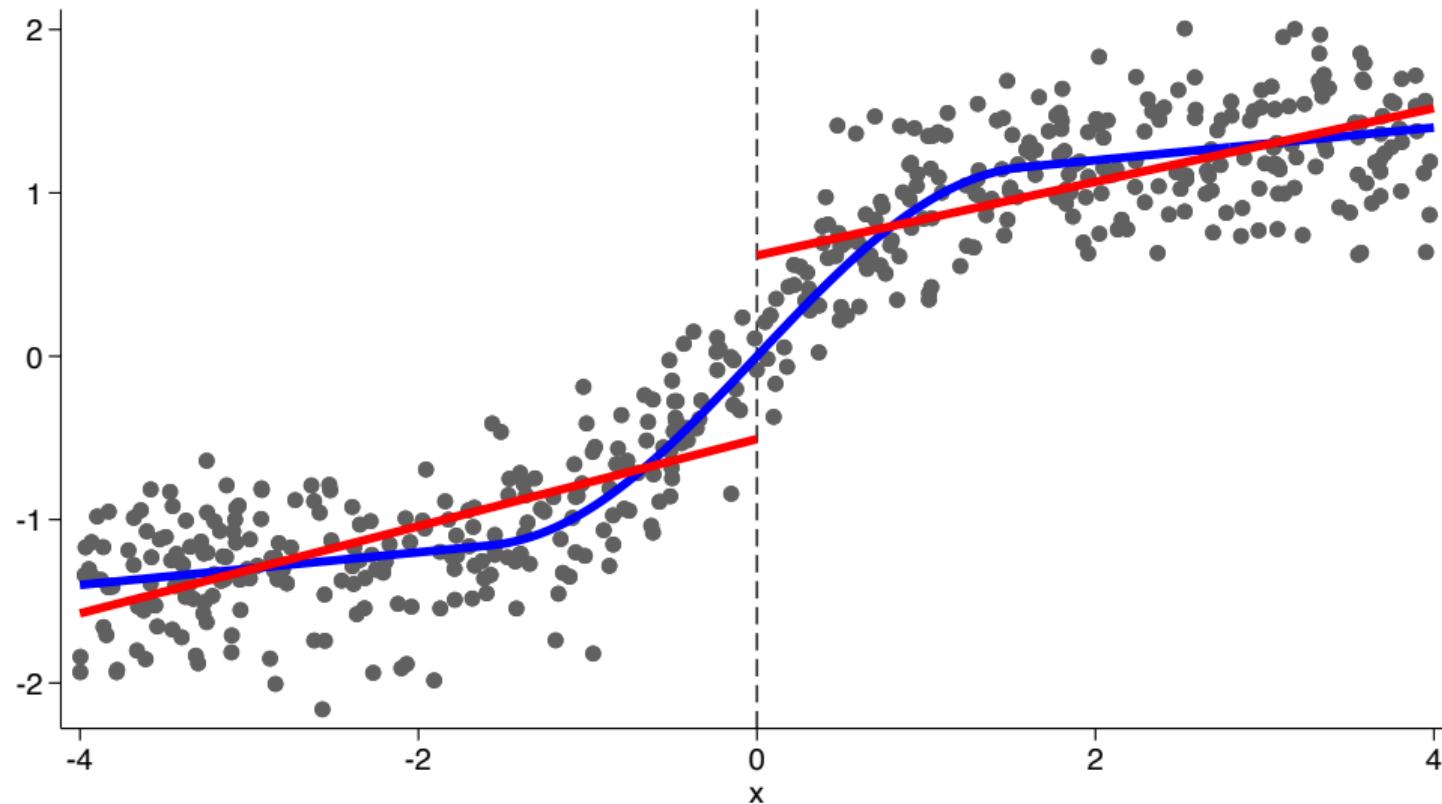
- Implementing RD requires estimating right- and left-side limits:

$$\lim_{r \uparrow c} \mathbb{E}[Y_i | R_i = r]$$

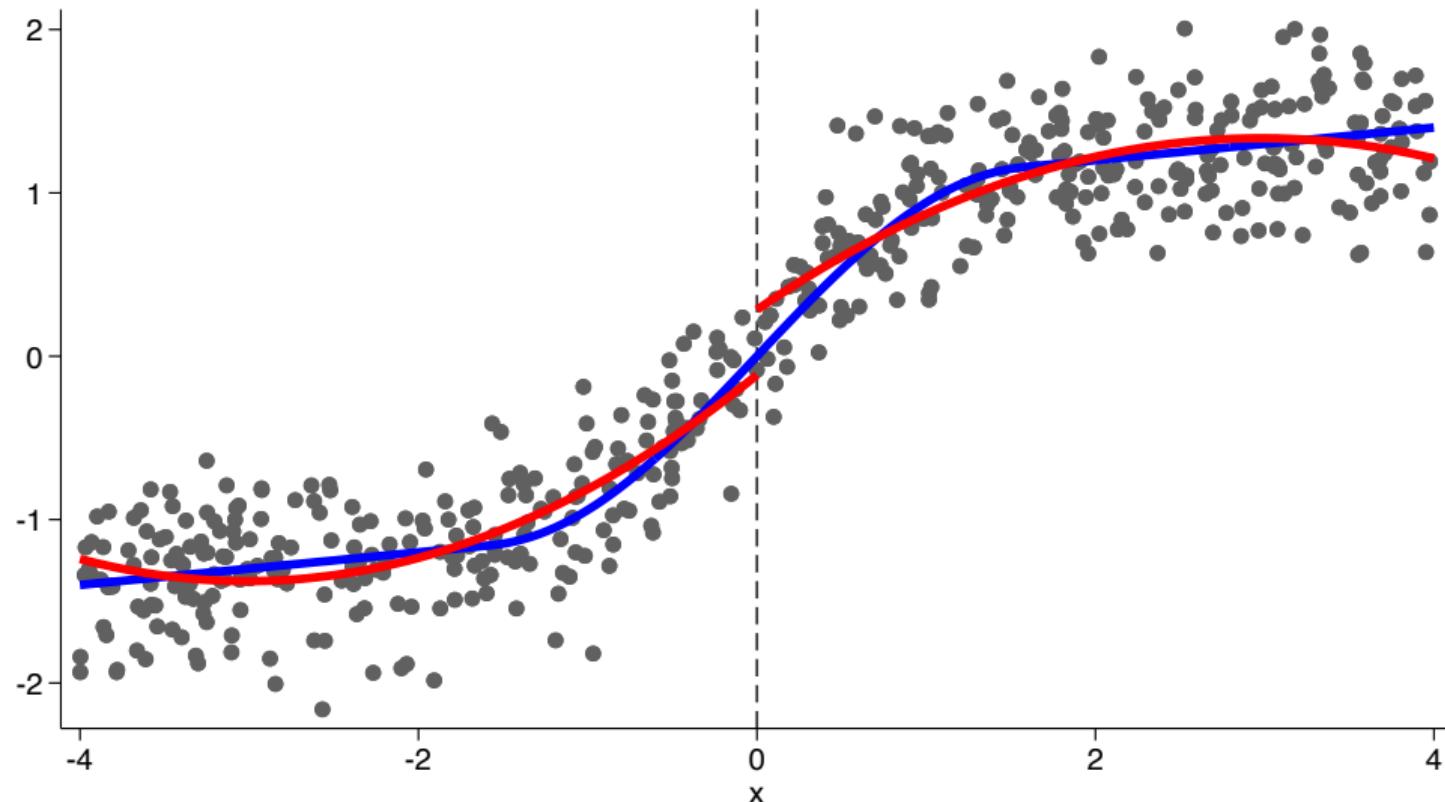
$$\lim_{r \downarrow c} \mathbb{E}[Y_i | R_i = r]$$

- Since we extrapolate, functional form used to approximate CEF important
  - Not enough to rely on OLS approximation theorems
  - With insufficiently flexible specification, might mistake nonlinearity for effect
  - But too flexible specification reduces precision and may overfit
- How do we balance this tradeoff?

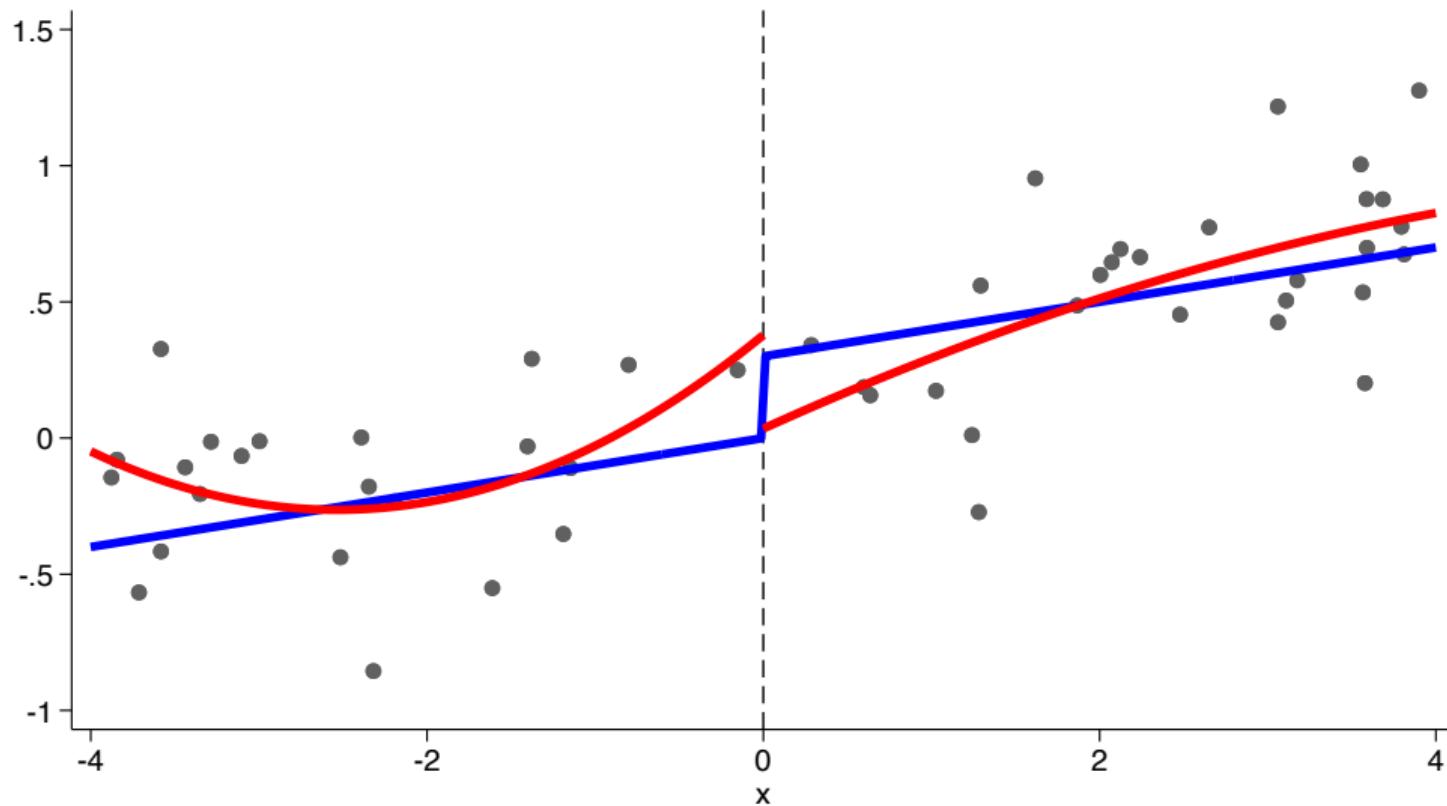
# False Positive: Mistaking Nonlinearity as Discontinuity



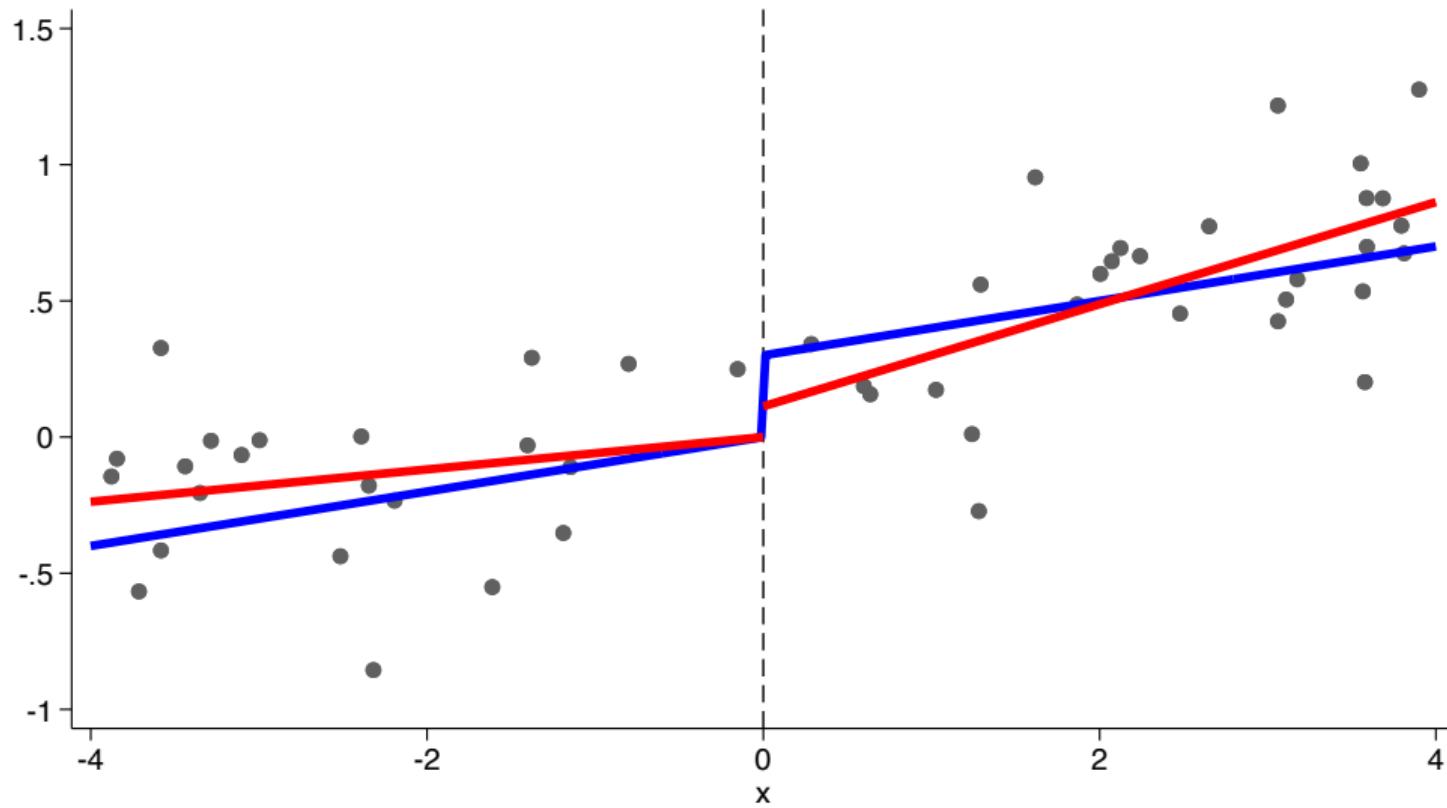
## Increased Flexibility May Avoid Detecting False Positive



# False Negative: Overfitting



# Less Flexibility May Avoid Overfitting



# Estimation: Global Parametric versus Local Nonparametric

- 1 **Global parametric:** approximate  $\mathbb{E}[Y_i|R_i]$  using polynomials
  - Traditional approach
  - Can work with moderately large datasets ( $N \approx 1000$ )
  - Usually uses all data, even when far from cutoff
- 2 **Local nonparametric:** approximate  $\mathbb{E}[Y_i|R_i]$  using nonparametric techniques
  - More modern approach
  - Typically requires large datasets ( $N > 1000$ )
  - Uses only small subset of data close to the cutoff
  - Often useful to compare across methods – stability suggests robust finding

# Global Parametric Specification

- Estimating equation:

$$Y_i = \alpha + \underbrace{\beta 1 [R_i > c]}_{\text{Discontinuity}} + \underbrace{\sum_{k=1}^K \gamma_{0k} (R_i - c)^k}_{\text{Polynomial for all data}} + \underbrace{\sum_{k=1}^K \gamma_{1k} 1 [R_i > c] (R_i - c)^k}_{\text{Polynomial only on right side}} + \varepsilon_i$$

- $K$ -th order polynomials to capture CEFs on either side
- Using separate coefficients  $\gamma_{0k}$  and  $\gamma_{1k}$  on either side
- Key parameter:  $\beta$ , the size of the jump at the threshold
- Treatment effect estimate:  $\widehat{\beta}$

## Local (i.e. Nonparametric) Linear Specification

- Estimating equation for **left-hand side CEF** using **local linear**:

$$(\hat{\alpha}_0, \hat{\delta}_0) = \arg \min_{\alpha_0, \delta_0} \sum_{i=1}^N 1[R_i \leq c] K\left(\frac{R_i - c}{h}\right) (Y_i - \alpha_0 - \delta_0(R_i - c))^2$$

- Similarly, for the **right-hand side CEF** and **local linear**, we run:

$$(\hat{\alpha}_1, \hat{\delta}_1) = \arg \min_{\alpha_1, \delta_1} \sum_{i=1}^N 1[R_i > c] K\left(\frac{R_i - c}{h}\right) (Y_i - \alpha_1 - \delta_1(R_i - c))^2$$

- Here,  $\hat{\alpha}_0$  is an estimate of  $\lim_{r \uparrow c} \mathbb{E}[Y_i | R_i = r]$  and  $\hat{\alpha}_1$  of  $\lim_{r \downarrow c} \mathbb{E}[Y_i | R_i = r]$
- Hence the **treatment effect estimate** is  $\hat{\alpha}_1 - \hat{\alpha}_0$
- Note that  $K(\cdot)$  is a Kernel with bandwidth  $h$
- Linear term eliminates “boundary bias” (Fan and Gijbels 1992)

# Visualization

- RD is appealing because the design is relatively simple and transparent
  - Key relationships: outcome & running variable; treatment & running variable
  - These relationships can be plotted → visual evidence of discontinuity
  - In practice, people won't believe an RD if the result isn't visible in a plot
- Plotting tips:
  - Automated plotting procedure `rdplot` (Calonico et al, 2014b)
  - May use own plotter using e.g. scatter of bin means and `lpoly`
  - `binscatter` also has an `, rd` option

# Table of Contents

## 1 RD Basics

- Identification and Interpretation
- Estimation and Visualization

## 2 Tuning and Diagnostics

- Tuning: Global Versus Local
- Diagnostics: Balance and Bunching
- Optimal Bandwidth and Inference

## 3 Special Cases

- Fuzzy RD Design
- Multiple Discontinuities and Multidimensional RDD
- Spatial/Boundary Discontinuity Design

## 4 Appendix

- Discrete Running Variable
- Regression Kink Design

# Choosing Global Versus Local

- Remaining issues:
  - 1 Global versus local?
  - 2 How to choose order of polynomial?
  - 3 How to choose bandwidth?
- No real conceptual distinction: need to choose bandwidth & polynomial order
  - 1 Bandwidth: how much data to include on either side of cutoff
  - 2 Polynomial order: how much flexibility to allow
- Typical choices:
  - 1 Global parametric: infinite or large bandwidth, higher-order polynomial
  - 2 Local nonparametric: small bandwidth, first-order polynomial (linear)
- But under uniform kernel, given bandwidth, and poly order they are the same!

## Global versus Local: Polynomial Order

- Global estimators can rely heavily on data far from cutoff
  - This might lead to very sensitive estimates (Gelman and Imbens 2014)
  - But if CEF is well-approximated by polynomial, legitimate and efficient
- Choosing optimal polynomial order:
  - In contrast to optimal bandwidth, not much known about optimal poly order
  - Generally, higher orders may be suboptimal (Pei et al. 2021)
- In practice: typically want to show robustness to tuning choices
- If you have the power, local linear generally preferred

# Table of Contents

## 1 RD Basics

- Identification and Interpretation
- Estimation and Visualization

## 2 Tuning and Diagnostics

- Tuning: Global Versus Local
- Diagnostics: Balance and Bunching**
- Optimal Bandwidth and Inference

## 3 Special Cases

- Fuzzy RD Design
- Multiple Discontinuities and Multidimensional RDD
- Spatial/Boundary Discontinuity Design

## 4 Appendix

- Discrete Running Variable
- Regression Kink Design

# Diagnostics Issues

- RD identifying assumption: PO distributions smooth around threshold
  - $1[R_i > c]$  must be as good as randomly assigned near  $R_i = c$
  - May be violated if individuals can exactly control  $R_i$
  - Example: self-reported income, people misreport for program eligibility
  - Just like in DID, this is **untestable**, but diagnostics can assess plausibility
- Two diagnostics:
  - 1 Covariate balance: check that no sorting on characteristics
  - 2 Bunching: check that no sorting by frequency

## RD Diagnostic I: Covariate Balance

- If people sort in the neighborhood of  $R_i = c$ , expect imbalance in covariates
- Motivates a check for whether there is a discontinuity in  $\mathbb{E}[X_i|R_i = c]$  at  $R_i = c$
- Analogous to balance check in baseline variables in an RCT
- Implement by using  $X_i$  as the RD dependent variable
  - Often done visually for primary variables of concern
  - Balance table for comparison of larger number of characteristics

## RD Diagnostic II: Bunching

- Instead of the mean **value** of  $X_i$ , the **frequency** could exhibit a jump
- McCrary (2008) proposes a test for whether the RD is compromised:
  - Logic: Imagine individuals strategically locate above/below threshold
  - Would then expect “bunching” on the side of threshold that is more preferable
- More generally, sorting may generate anomalies in the distribution of  $R_i$
- McCrary suggests looking for a discontinuity in the density of  $R_i$  near  $c$
- Typically implemented visually
- Sometimes, a “donut” strategy can ameliorate sorting concerns
  - Basic idea: leave out observations very close to threshold (Barreca et al 2011)
- Bunching is bad for credibility of RD, but may in itself be interesting:
  - Bunching estimators in public economics

# Table of Contents

## 1 RD Basics

- Identification and Interpretation
- Estimation and Visualization

## 2 Tuning and Diagnostics

- Tuning: Global Versus Local
- Diagnostics: Balance and Bunching
- Optimal Bandwidth and Inference**

## 3 Special Cases

- Fuzzy RD Design
- Multiple Discontinuities and Multidimensional RDD
- Spatial/Boundary Discontinuity Design

## 4 Appendix

- Discrete Running Variable
- Regression Kink Design

# Kernels and Bandwidths

- What kernel to use? Many plausible choices
- Popular choice: “edge” (or triangle) kernel:

$$K(u) = 1[|u| \leq 1] \times (1 - |u|)$$

- Has optimality properties in boundary estimation problems (Cheng et al 1997)
- It is also intuitively appealing: generates weighting function:

$$K\left(\frac{R_i - c}{h}\right) = 1[|R_i - c| \leq h] \times \left(1 - \frac{|R_i - c|}{h}\right)$$

- $h$  can be interpreted as cutoff distance for data inclusion
- Weights fall linearly from 0 to 1 in included sample
- Should not make big difference, but variation across kernels rarely shown

# Optimal Bandwidth

- Recent literature on optimal choice of bandwidth  $h$
- Bias/variance tradeoff: smaller bandwidth reduces bias but reduces precision
- Intuitively, if little curvature in CEF, bias from large  $h$  is small
- Imbens and Kalyanaraman (IK, 2012):
  - Asymptotic approximation of MSE of RD estimator
  - Derive MSE-minimizing bandwidth
  - Optimal bandwidth depends on curvature of CEF near discontinuity
  - Use plug-in estimators of parameters governing curvature

# Robust Confidence Intervals

- The IK bandwidth is well-suited for estimation
- But what about inference?
- Calonico, Cattaneo and Titiunik (CCT, 2014) show it works poorly for inference
  - Non-negligible bias term in the estimate
  - Naive inference can lead to misleading confidence intervals
- Alternative by CCT:
  - CCT advocate using a second, smaller bandwidth to eliminate this bias
  - “Undersmooth”, i.e. prioritize small bias
  - Report robust standard error using this smaller bandwidth

# Optimal Bandwidth in Practice

- Unfortunately, CCT bandwidth often much smaller than IK bandwidth
- CCT generates large confidence intervals
- May limit feasibility of many potential RD projects:
  - Lots of data required to detect an effect with CCT robust CIs
  - Noisy effects are very hard to write up and sell
- In practice, there is some slack with using alternatives to CCT
  - Primacy of visualization: compelling figures guide robustness requirements
  - Plotting effect and CIs for different bandwidths can be useful
- Implementation of IK and CCT: see `rdrobust`

# Table of Contents

## 1 RD Basics

- Identification and Interpretation
- Estimation and Visualization

## 2 Tuning and Diagnostics

- Tuning: Global Versus Local
- Diagnostics: Balance and Bunching
- Optimal Bandwidth and Inference

## 3 Special Cases

- Fuzzy RD Design
- Multiple Discontinuities and Multidimensional RDD
- Spatial/Boundary Discontinuity Design

## 4 Appendix

- Discrete Running Variable
- Regression Kink Design

## Fuzzy RD Setup

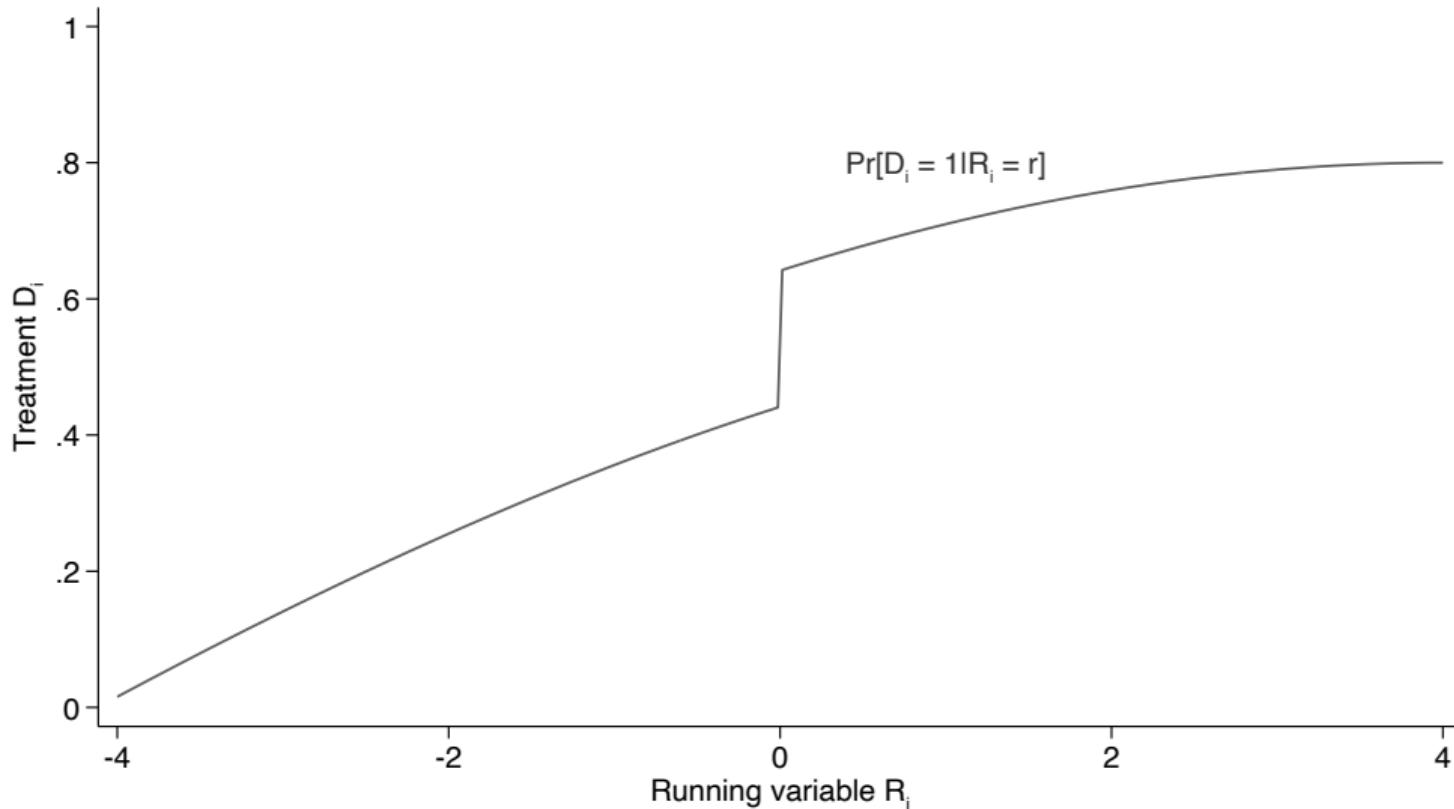
- In many cases, assignment probability does not jump from zero to one at  $c$ 
  - Instead of treatment being deterministic with  $R_i$ , its probability jumps
  - Other factors besides running variable may affect treatment assignment
- Suppose that

$$\lim_{r \uparrow c} \Pr(D_i = 1 | R_i = r) \neq \lim_{r \downarrow c} \Pr(D_i = 1 | R_i = r)$$

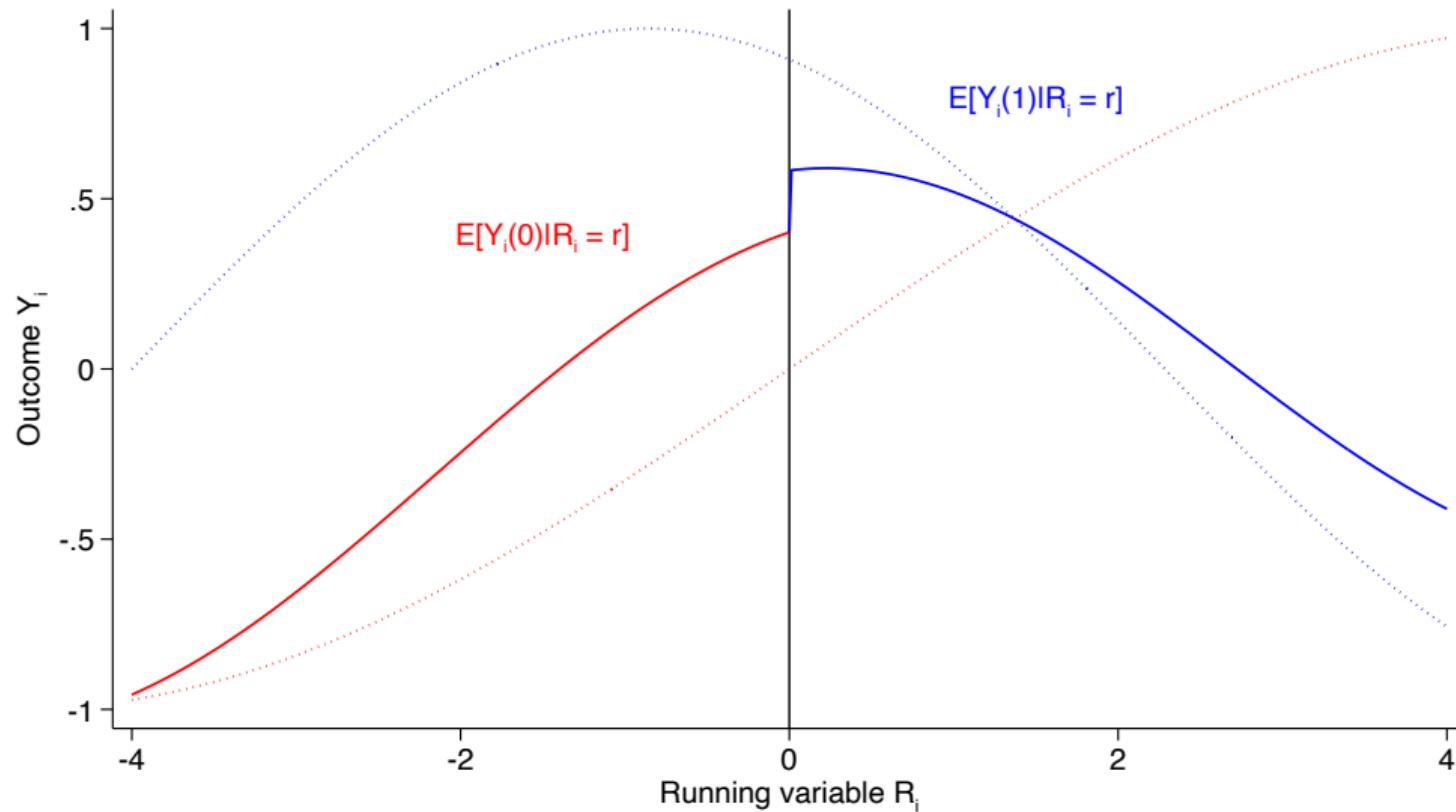
i.e. treatment probability jumps at  $R_i = c$ , but not necessarily from zero to one

- This is the **fuzzy RD scenario**
- Example:
  - Income threshold determines program eligibility
  - But not all households participate

# Assignment Function in Fuzzy RD



# Observed Outcome in Fuzzy RD



# Fuzzy RD Assumptions

- As before, assume  $Y_i(1)$  and  $Y_i(0)$  are smooth around the threshold
- Let  $D_i(1)$  and  $D_i(0)$  denote treatment status above/below  $R_i = c$
- **New assumptions:**
  - 1 **Potential treatment continuity:**  $D_i(1)$  and  $D_i(0)$  are smooth around the threshold
  - 2 **Monotonicity:** crossing threshold weakly increases treatment probability:

$$D_i(1) \geq D_i(0) \text{ for all } i$$

- Can again investigate comparison of individuals above and below threshold:

$$\lim_{r \downarrow c} \mathbb{E}[Y_i | R_i = r] - \lim_{r \uparrow c} \mathbb{E}[Y_i | R_i = r]$$

with our usual relationship that  $Y_i = Y_i(0) + [Y_i(1) - Y_i(0)] D_i$

## Difference in Outcomes and Treatment near Threshold

- We get:  $\lim_{r \downarrow c} \mathbb{E}[Y_i | R_i = r] - \lim_{r \uparrow c} \mathbb{E}[Y_i | R_i = r] =$ 
$$= \lim_{r \downarrow c} \mathbb{E}[Y_i(0) + (Y_i(1) - Y_i(0)) D_i | R_i = r]$$
$$- \lim_{r \uparrow c} \mathbb{E}[Y_i(0) + (Y_i(1) - Y_i(0)) D_i | R_i = r]$$
$$= \lim_{r \downarrow c} \mathbb{E}[(Y_i(1) - Y_i(0)) D_i(1) | R_i = r] - \lim_{r \uparrow c} \mathbb{E}[(Y_i(1) - Y_i(0)) D_i(0) | R_i = r]$$
$$= \mathbb{E}[(Y_i(1) - Y_i(0))(D_i(1) - D_i(0)) | R_i = c]$$
$$= \mathbb{E}[Y_i(1) - Y_i(0) | D_i(1) > D_i(0), R_i = c] \times \Pr(D_i(1) > D_i(0) | R_i = c)$$

where the last step uses monotonicity

- Similarly, comparing compliance probabilities at the threshold;

$$\lim_{r \downarrow c} \mathbb{E}[D_i | R = c] - \lim_{r \uparrow c} \mathbb{E}[D_i | R = c] = \Pr(D_i(1) > D_i(0) | R_i = c)$$

## Fuzzy RD is IV

- Then we can take the ratio:

$$\frac{\lim_{r \downarrow c} \mathbb{E}[Y_i | R_i = r] - \lim_{r \uparrow c} \mathbb{E}[Y_i | R_i = r]}{\lim_{r \downarrow c} \mathbb{E}[D_i | R = c] - \lim_{r \uparrow c} \mathbb{E}[D_i | R = c]} = \mathbb{E}[Y_i(1) - Y_i(0) | D_i(1) > D_i(0), R_i = c]$$

- i.e. the jump in the outcome CEF, over the jump in the treatment probability
- this identifies the LATE of:
  - individuals who are at the threshold... (as in strict RD)
  - individuals who comply with the treatment (i.e. not always/never takers)
- Look familiar?

→ Fuzzy RD is IV!

- Using a threshold indicator  $Z_i = 1[R_i > c]$  as instrument for treatment
- ... in the neighborhood of the threshold

# Fuzzy RD Implementation

- Can implement fuzzy RD with global 2SLS approach:

$$D_i = \lambda + \pi 1[R_i > c] + \sum_{k=1}^K \theta_{0k} (R_i - c)^k + \sum_{k=1}^K \theta_{1k} 1[R_i > c] (R_i - c)^k + \xi_i$$

$$Y_i = \alpha + \beta \hat{D}_i + \sum_{k=1}^K \gamma_{0k} (R_i - c)^k + \sum_{k=1}^K \gamma_{1k} 1[R_i > c] (R_i - c)^k + \varepsilon_i$$

- $Z_i = 1[R_i > c]$  is the excluded instrument
- Alternatively, can estimate each of the four limits using local linear approach
  - The two limits of the outcome above/below threshold
  - The two limits of the treatment probability

# Table of Contents

## 1 RD Basics

- Identification and Interpretation
- Estimation and Visualization

## 2 Tuning and Diagnostics

- Tuning: Global Versus Local
- Diagnostics: Balance and Bunching
- Optimal Bandwidth and Inference

## 3 Special Cases

- Fuzzy RD Design
- Multiple Discontinuities and Multidimensional RDD**
- Spatial/Boundary Discontinuity Design

## 4 Appendix

- Discrete Running Variable
- Regression Kink Design

# Multiple Discontinuities with Scalar Running Variable

- There may be more than one discontinuity along running variable
- Three approaches:
  - ① Nonparametric solution: estimate each RD separately, then average across
    - Only feasible if discontinuities are far enough from one another
    - Requires sufficient precision on each RD separately
  - ② Impose stationarity:
    - Concretely, running variable is distance to **nearest** discontinuity
    - Can then again use local linear methods
  - ③ Fully parametric:
    - Global approach with polynomials
    - Discontinuities may or may not be restricted to be the same

## Multiple Discontinuity Example: Fredriksson et al (2013)

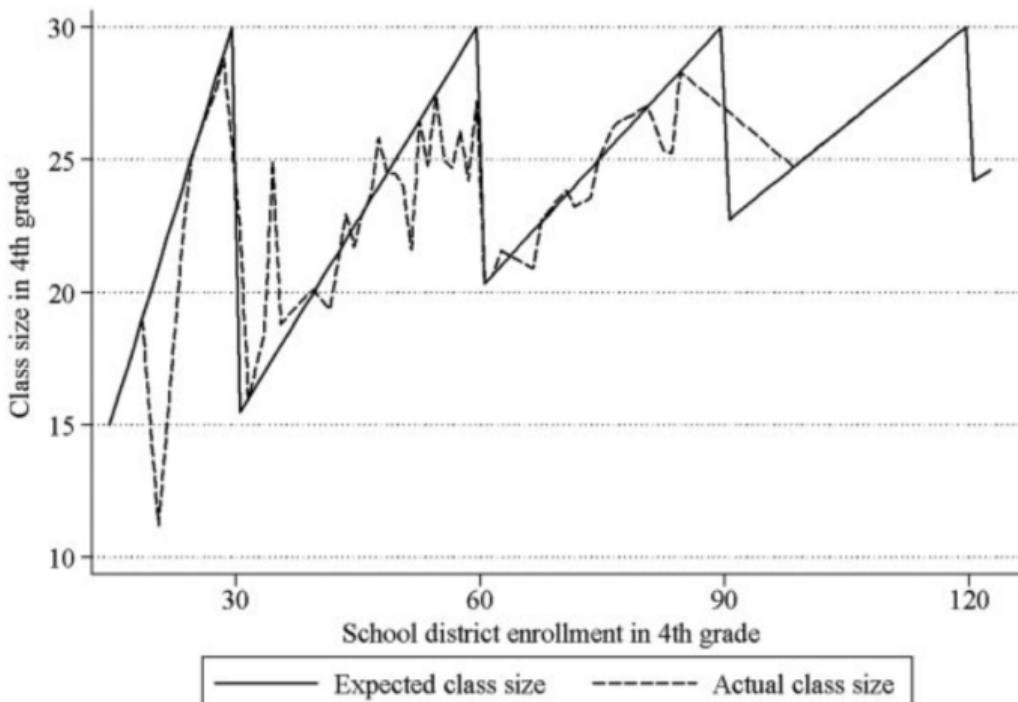


FIGURE II

Expected and Actual Class Size in Grade 4 by Enrollment in Grade 4

# Residualized Outcome Around Standardized Running Variable

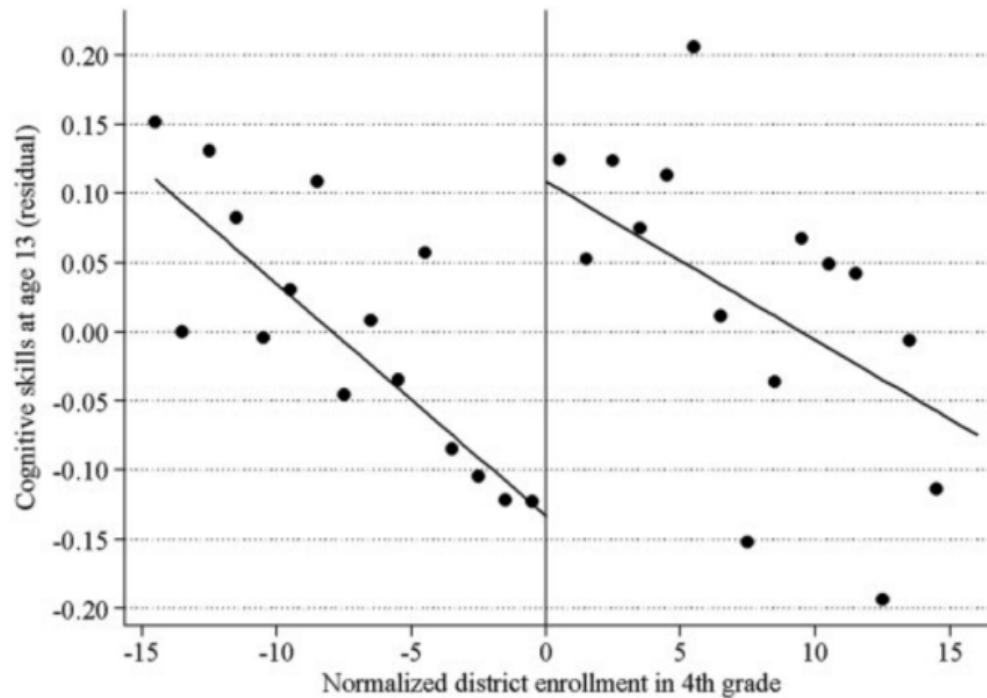
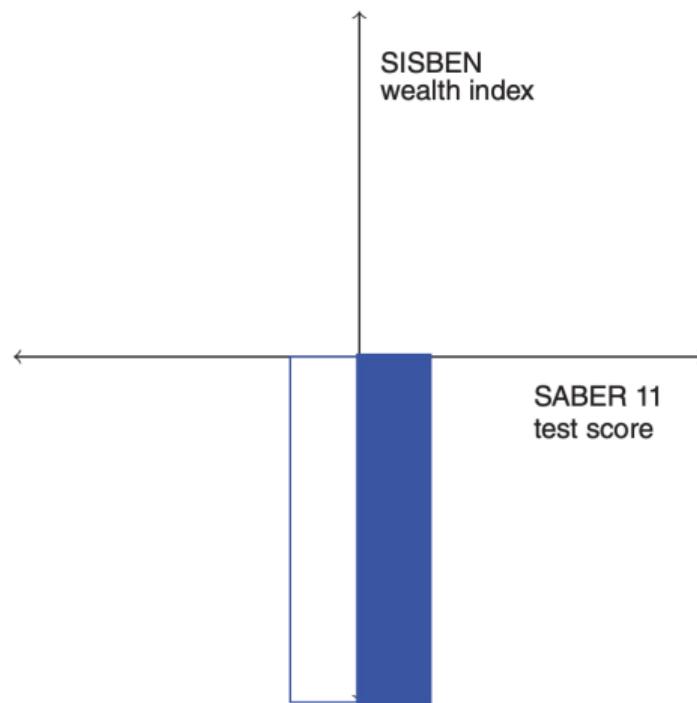


FIGURE VI

Cognitive Ability at Age 13 by Enrollment in Grade 4

# Fuzzy Multidimensional RD Example: Londoño-Vélez et al (2020)

Panel A. SABER 11 as  $R_i$



Panel B. SISBEN as  $R_i$

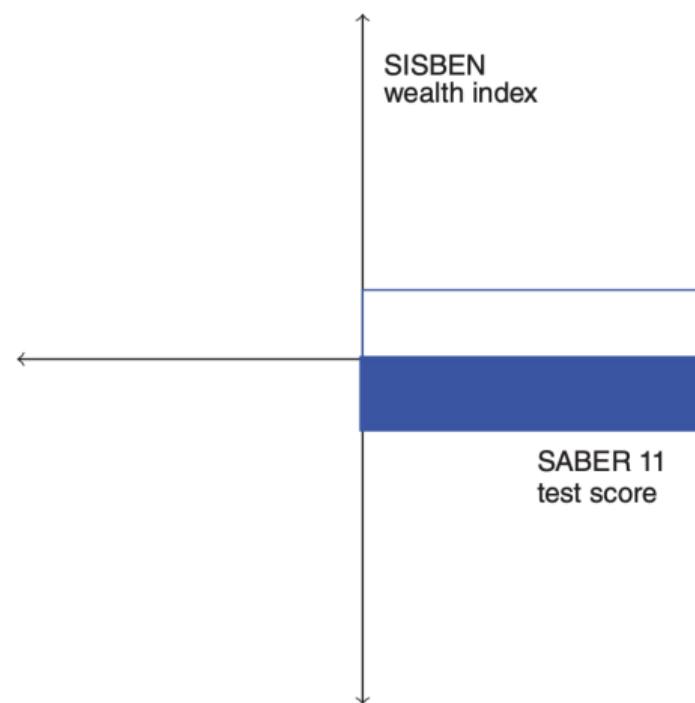
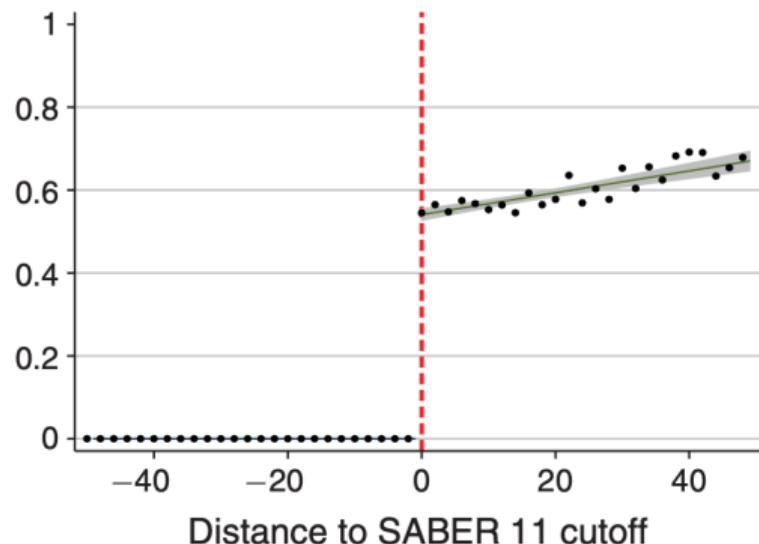


FIGURE 2. ILLUSTRATION OF THE TWO TYPES OF COMPLIERS

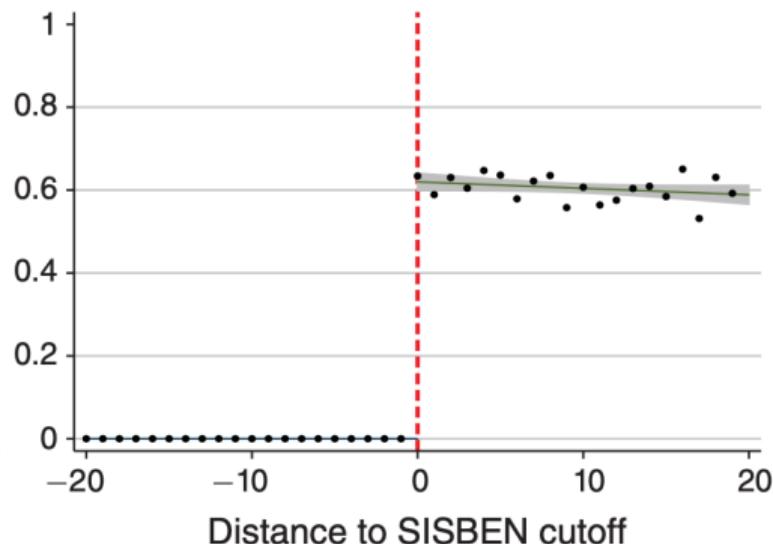
# Assignment Probabilities Along Each Running Variable

Panel A.  $R_i = \text{SABER 11 test score}$



Sample restricted to SISBEN-eligible individuals

Panel B.  $R_i = \text{SISBEN wealth index}$

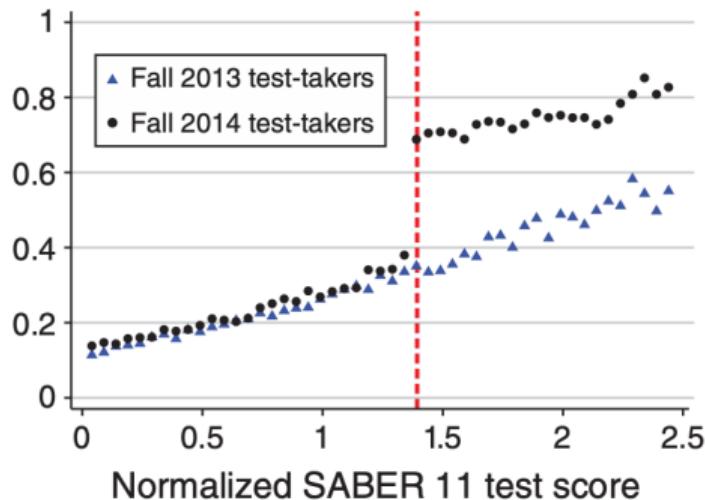


Sample restricted to SABER 11-eligible individuals

FIGURE 3. DISCONTINUITY IN THE PROBABILITY OF RECEIVING SPP FINANCIAL AID

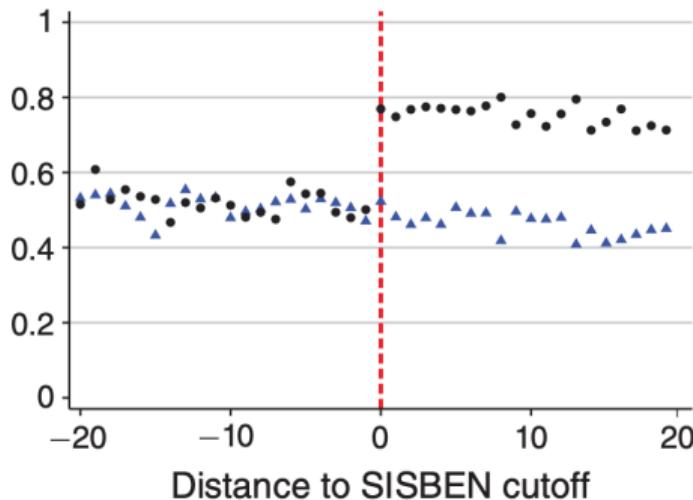
# Actual and Placebo Outcomes

Panel A.  $R_i = \text{SABER 11 test score}$



Sample restricted to SISBEN-eligible individuals

Panel B.  $R_i = \text{SISBEN wealth index}$



Sample restricted to SABER 11-eligible individuals

FIGURE 5. PLACEBO TEST USING PRE-TREATMENT PERIOD

# Table of Contents

## 1 RD Basics

Identification and Interpretation  
Estimation and Visualization

## 2 Tuning and Diagnostics

Tuning: Global Versus Local  
Diagnostics: Balance and Bunching  
Optimal Bandwidth and Inference

## 3 Special Cases

Fuzzy RD Design  
Multiple Discontinuities and Multidimensional RDD  
**Spatial/Boundary Discontinuity Design**

## 4 Appendix

Discrete Running Variable  
Regression Kink Design

# Boundary Discontinuity Design

- Many researchers used institutional boundaries as discontinuities
- Popular in urban, development, and economic history
- Challenges:
  - ① Non-trivial two-dimensional boundary: it may arbitrarily move through space
  - ② Individuals usually actively sort near boundaries
- Two general approaches:
  - ① Nonparametric local RD
  - ② Parametric RD with spatial polynomial in latitude and longitude
- Caveats:
  - Even passing diagnostics, concerns about sorting in unobservables
    - Exclude areas with geographic discontinuities (e.g. highways)
    - Include boundary segment fixed effects
  - May require model to convince people that RD is interesting
  - Even without sorting, spatial spillovers may be important

# He et al (2020): Design

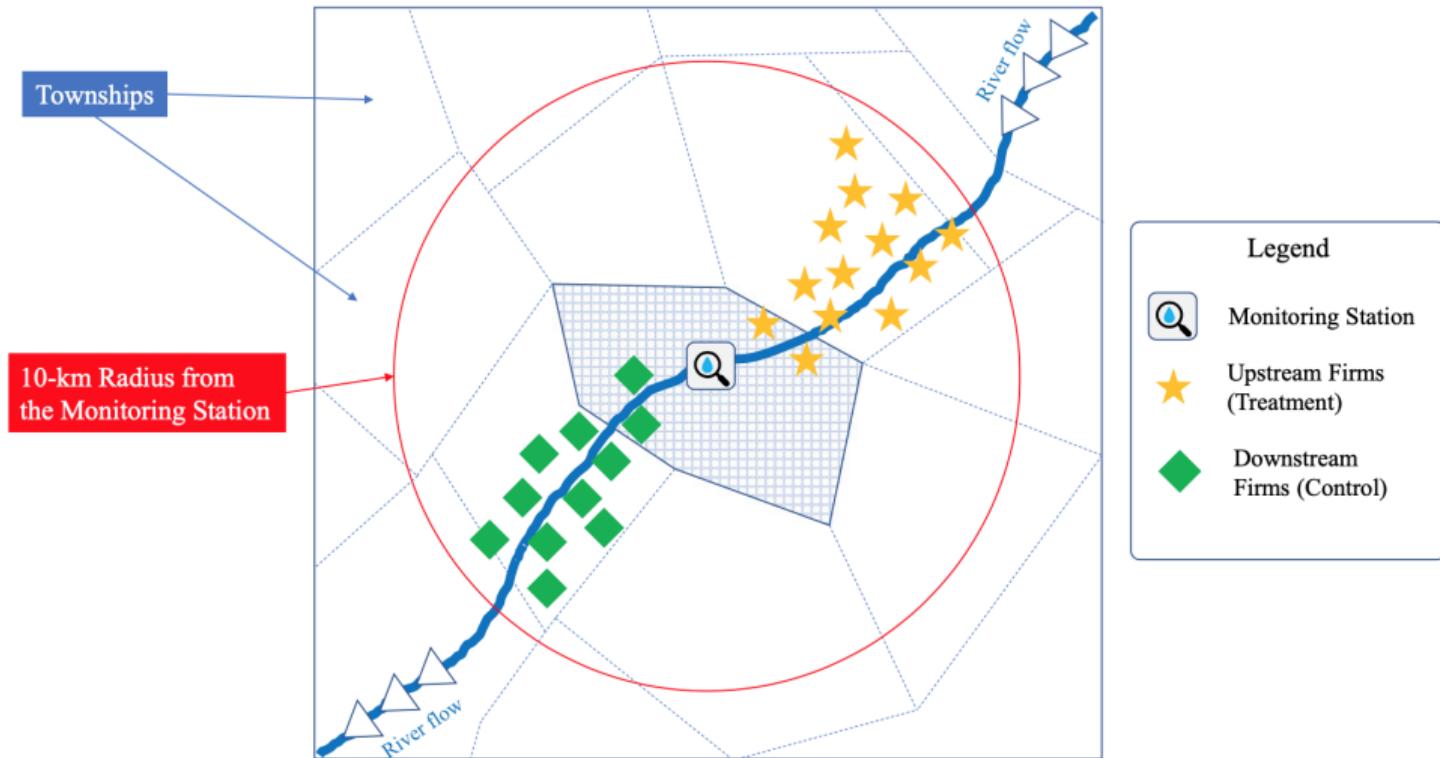
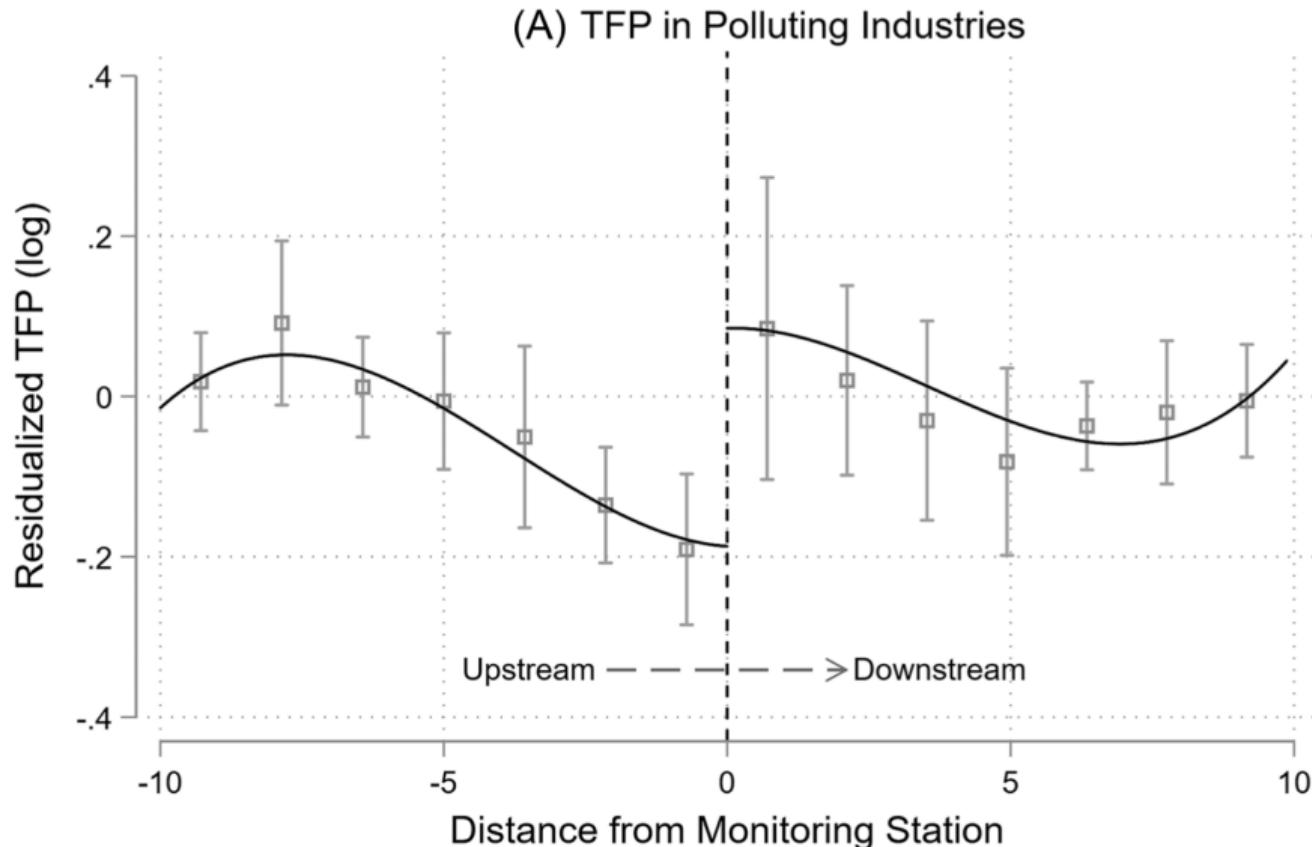


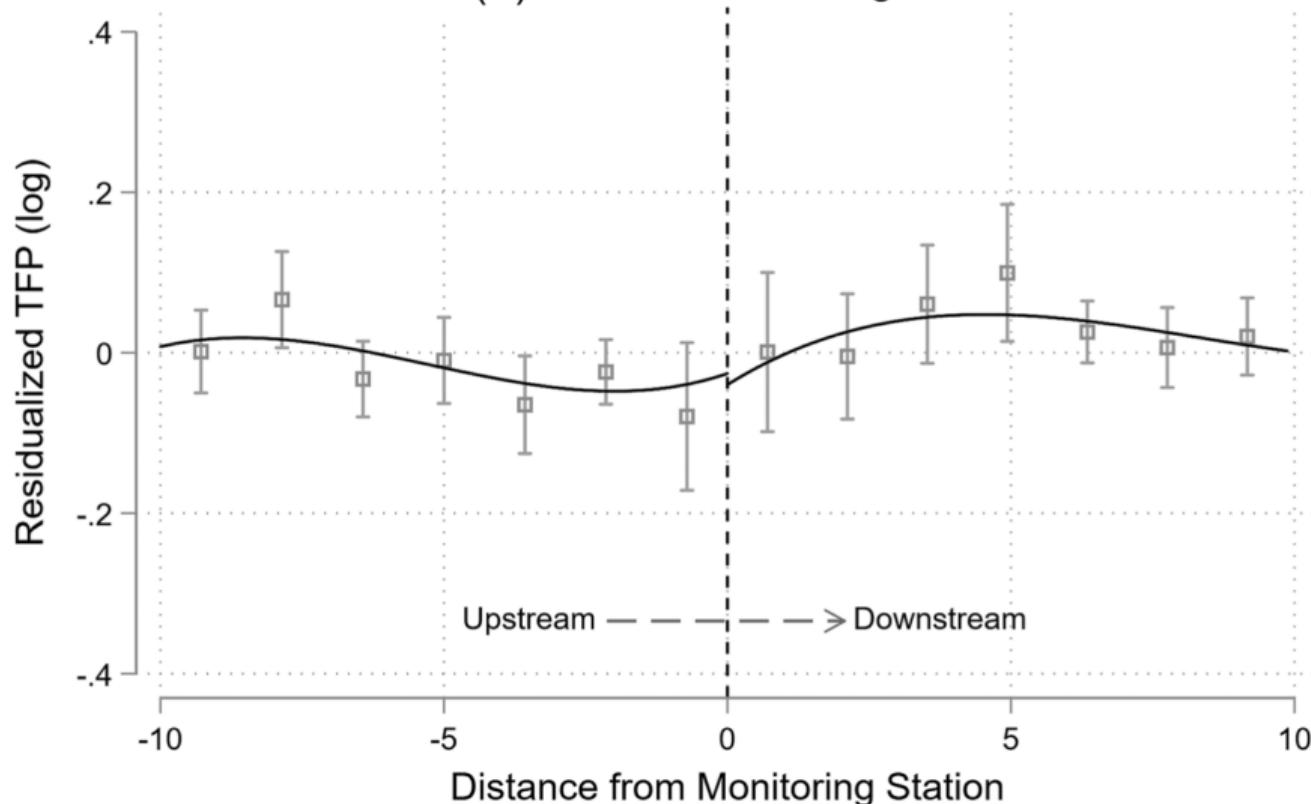
FIGURE II

# Main Effects of Spatial Discontinuity



# Placebo of Non-Polluting Firms

(B) TFP in Non-Polluting Industries



# RD Estimates

TABLE I  
THE UPSTREAM–DOWNSTREAM TFP GAP

	Polluting industries			Nonpolluting industries		
	(1)	(2)	(3)	(4)	(5)	(6)
<b>Panel A: No control</b>						
RD in TFP (log)	0.34	0.37	0.32	-0.03	0.04	0.01
(downstream – upstream)	(0.57)	(0.59)	(0.56)	(0.15)	(0.18)	(0.18)
Bandwidth (km)	4.203	3.889	3.622	5.887	5.168	4.522
<b>Panel B: Station FE + industry FE absorbed</b>						
RD in TFP (log)	0.36**	0.38**	0.34**	0.03	0.04	-0.02
(downstream – upstream)	(0.17)	(0.17)	(0.15)	(0.09)	(0.09)	(0.09)
Bandwidth (km)	5.723	5.523	5.144	5.890	5.479	6.091
<b>Panel C: Station by industry FE absorbed</b>						
RD in TFP (log)	0.27*	0.29**	0.29**	0.02	0.04	0.03
(downstream – upstream)	(0.15)	(0.15)	(0.14)	(0.06)	(0.06)	(0.07)
Bandwidth (km)	4.496	4.333	4.689	5.692	5.204	4.430
Obs.	6,224	6,224	6,224	11,502	11,502	11,502
Kernel	Triangle	Epanech.	Uniform	Triangle	Epanech.	Uniform

*Notes.* Each cell in the table represents a separate RD regression. The running variable is the distance between a firm and a monitoring station, where negative (positive) distance means firms are located to the upstream (downstream) of the monitoring stations. The positive coefficients indicate that downstream firms have higher TFP than upstream firms. TFP is estimated using the [Olley and Pakes \(1996\)](#) method, with “upstream polluting” added as an additional state variable. The discontinuities at monitoring stations are estimated using local linear regressions and MSE-optimal bandwidth proposed by [Calonico, Cattaneo, and Titiunik \(2014\)](#) for different kernel weighting methods. Standard errors clustered at the monitoring station level are reported below the estimates. \* significant at 10%, \*\* significant at 5%, \*\*\* significant at 1%.

# Difference-in-Discontinuity

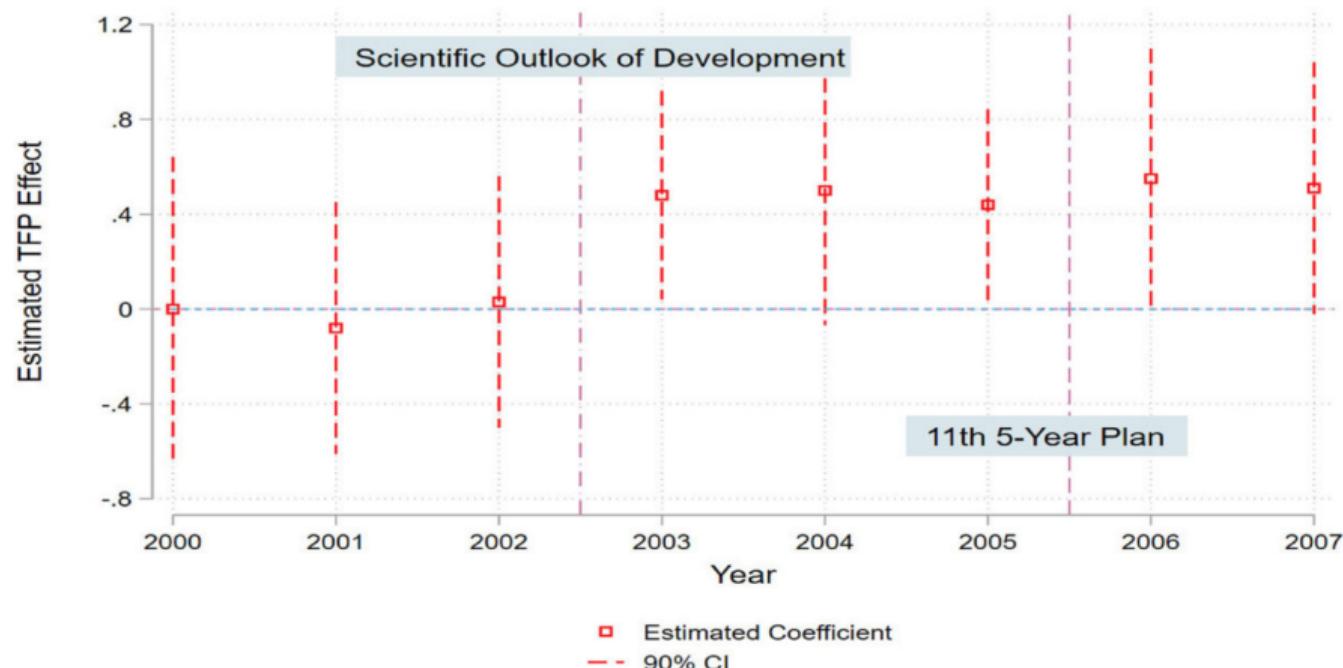


FIGURE V  
RD Estimates by Year

# Table of Contents

## 1 RD Basics

- Identification and Interpretation
- Estimation and Visualization

## 2 Tuning and Diagnostics

- Tuning: Global Versus Local
- Diagnostics: Balance and Bunching
- Optimal Bandwidth and Inference

## 3 Special Cases

- Fuzzy RD Design
- Multiple Discontinuities and Multidimensional RDD
- Spatial/Boundary Discontinuity Design

## 4 Appendix

- Discrete Running Variable
- Regression Kink Design

## RD with a Discrete Running Variable

- In some cases, the running variable is discrete
- Example: treatment assignment based on test score with few questions
- Estimation techniques generally work, despite discreteness
- However, inference is problematic:
  - Card and Lee (2008) suggest clustering on values of running variable
  - But Kolesar and Rothe (2018) show that these CI have poor coverage
  - Propose alternative with guaranteed coverage under restrictions on CEF
- Conceptually distinct issue:
  - True running variable is continuous
  - But only observe running variable rounded at discrete values
  - See Dong (2015) for this issue

# Table of Contents

## 1 RD Basics

- Identification and Interpretation
- Estimation and Visualization

## 2 Tuning and Diagnostics

- Tuning: Global Versus Local
- Diagnostics: Balance and Bunching
- Optimal Bandwidth and Inference

## 3 Special Cases

- Fuzzy RD Design
- Multiple Discontinuities and Multidimensional RDD
- Spatial/Boundary Discontinuity Design

## 4 Appendix

- Discrete Running Variable
- Regression Kink Design

# Regression Kink Design

- Extension of RDD: the **regression kink design** (RKD) by Card et al (2015)
- Recall RDD logic:
  - Binary treatment  $D_i(R_i)$ , e.g.  $D_i = 1[R_i > c]$  in strict RDD
  - Exploit discontinuous **level change** in treatment (probability)
- Basic RKD logic:
  - Continuous treatment  $S_i(R_i)$ , such as monthly unemployment payment
  - Exploit discontinuous **slope change** ("kink") in treatment intensity
- Suppose  $S_i = b(R_i)$ , e.g.  $S_i = \gamma_0 R_i + 1[R_i > c] \times \gamma_1 R_i$  with  $\gamma_1 > 0$
- Specifically,  $b(\cdot)$  is a continuous function with kink at  $c$
- Let  $f_i(s)$  denote  $i$ 's potential outcome if treated with intensity  $s = S_i$

# RKD Logic and Estimation

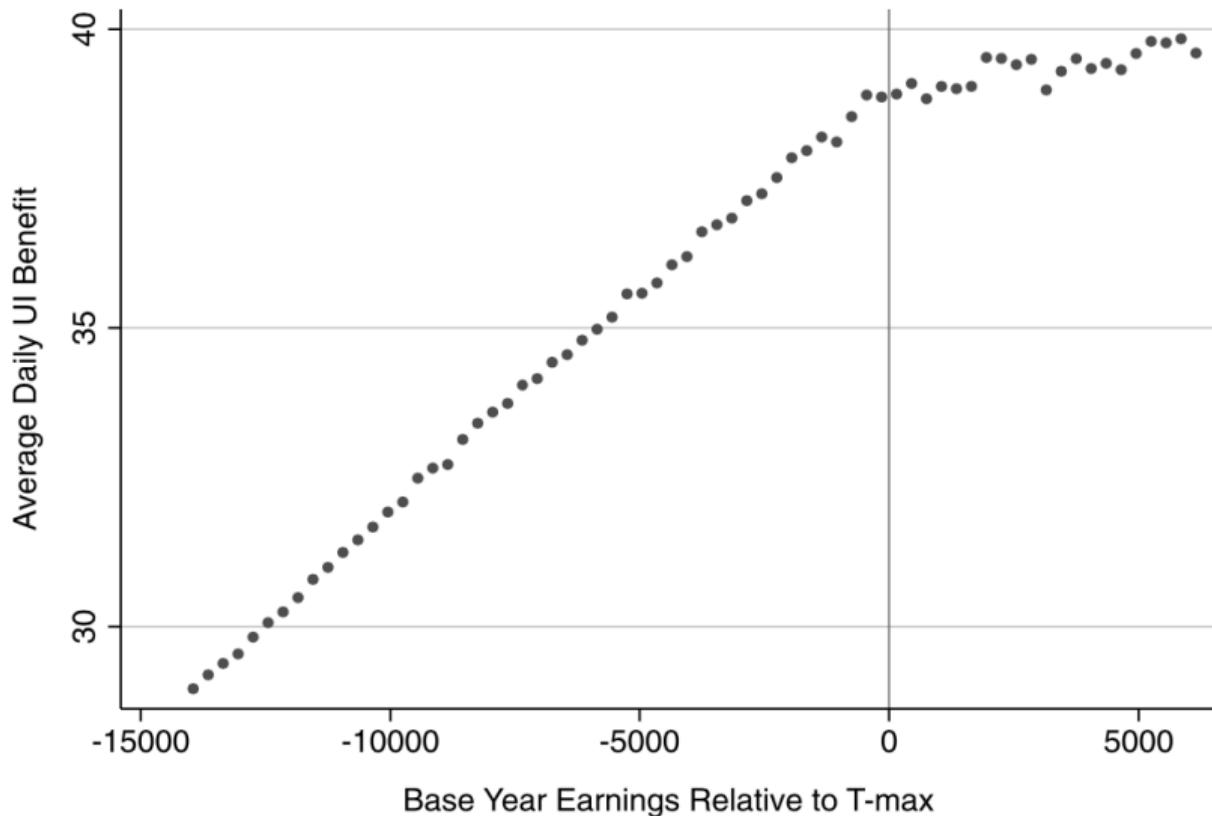
- Then under mild regularity conditions:

$$\frac{\lim_{r \downarrow c} \frac{d\mathbb{E}[Y_i|R_i=r]}{dr} - \lim_{r \uparrow c} \frac{d\mathbb{E}[Y_i|R_i=r]}{dr}}{\lim_{r \downarrow c} b'(r) - \lim_{r \uparrow c} b'(r)} = \mathbb{E} [f'_i(S_i)|R_i = c]$$

which says the ratio of ...

- ... discontinuity in outcome derivative ...
  - ... over discontinuity in the treatment derivative ...
  - ... identifies the average marginal effect of treatment ...
  - ... for individuals at the threshold
- 
- As before, assumption is that POs are smooth around threshold
    - Any kink in outcome CEF must then be due to treatment
    - Diagnostics: check for kinks in covariates or bunching near  $c$
  - Estimation: analogous to RDD; see `rdrobust` for RKD implementation
  - Even more data-intensive than RDD!**

## RKD Example: Card et al (2015)



## RKD Example: Outcome

