Dakar, Senegal

IsDB-World Bank DIME Impact Evaluation Event

January 29-31, 2019

Transforming Development through Evidence-Based Policy

Non-experimental methods for transport impact evaluation

overview

- In this sessions, we will discuss:
  - the logic of impact evaluation
  - The various methods of non-experimental impact evaluation
  - A case study, to which all major non-experimental methods will be applied

overview

- Today: non-experimental methods for impact evaluation
- Tomorrow: randomized controlled trials (RCTs)
- Wednesday: sampling and power calculations for impact evaluation

Why impact evaluation?

- We want to implement the most effective policies and programs
- So we need a method that enables us to understand what works and what does not work
- The means we need to understand cause and effect
  - We must measure what happened
  - We must also find a way to measure the counterfactual (what *would have* happened if we had not implemented the program)

M&E versus impact evaluation

- Monitoring and evaluation
  - Tracks whether project activities were conducted **(inputs)**
  - Counts the project **outputs** that were delivered/constructed
- Was the road built? Was it on time, and did it meet technical standards?
- Are people/vehicles using the road?

M&E versus impact evaluation

- This standard M&E approach remains very important
- But it **does not** tell us:
  - Did the road lead to increased economic growth/reduced poverty?
  - What is the most precise estimate of this economic growth? How does it compare to alternate uses of scarce resources?
  - Did all groups benefit? Were their winners and losers? To what extent?

Causal inference and counterfactuals

- Fundamentally, to know the effect of our project, we want to observe something that is **fundamentally unobservable**: the counterfactual
- What happened to the village where we built a road, compared to what *would have happened to the same village* if we did not build the road

Causal inference and counterfactuals

- Since we can never observe the same village at the same time both with and without the road, we must use other methods to develop a valid comparison group

impact evaluation

- A historical example:
  - Did the expansion of railways in the late 19th century US contribute to economic growth?

- The Fogel hypothesis:

THE JOURNAL OF ECONOMIC HISTORY

VOL. XXII JUNE 1962 NO. 2

*A Quantitative Approach to the Study of Railroads in American Economic Growth: A Report of Some Preliminary Findings**

Is it legitimate for the historian to consider alternative possibilities to events which have happened? . . . To say that a thing happened the way it did is not at all illuminating. We can understand the significance of what did happen only if we contrast it with what might have happened.

MORRIS RAPHAEL COHEN

I

- Building a road (or a railroad) is probably a good thing. But how good, relative to other investments?

- "the level of per capita income achieved by January 1, 1890 would have been reached by March 31, 1890, if railroads had never been invented."

Source: Lance Davis, https://eh.net/book_reviews/railroads-and-american-economic-growth-essays-in-econometric-history

By contrast, new data sources and methods give a very different answer. (Donaldson and Hornbeck 2016)

RAILROADS AND AMERICAN ECONOMIC GROWTH: A "MARKET ACCESS" APPROACH*

DAVE DONALDSON AND RICHARD HORNBECK

This article examines the historical impact of railroads on the U.S. economy, with a focus on quantifying the aggregate impact on the agricultural sector in 1890. Expansion of the railroad network may have affected all counties directly or indirectly—an econometric challenge that arises in many empirical settings. However, the total impact on each county is captured by changes in that county's "market access," a reduced-form expression derived from general equilibrium trade theory. We measure counties' market access by constructing a network database of railroads and waterways and calculating lowest-cost county-to-county freight routes. We estimate that county agricultural land values increased substantially with increases in county market access, as the railroad network expanded from 1870 to 1890. Removing all railroads in 1890 is estimated to decrease the total value of U.S. agricultural land by 60%, with limited potential for mitigating these losses through feasible extensions to the canal network or improvements to country roads. *JEL* Codes: N01, N51, N71, F1, O1, R1.

New York Times, June 8, 2017: the Fogel hypothesis revisited?

Kenyans Fear Chinese-Backed Railway Is Another 'Lunatic Express'

点击查看本文中文版

By KIMIKO de FREYTAS-TAMURA JUNE 8, 2017

A Kenya Railways employee sending off the train from Nairobi to Mombasa at the new Standard Gauge Railway terminal in Nairobi this week. Adriane Ohanesian for The New York Times

RELATED COVERAGE

8 Taiwanese Are Deported to China A Trial in Kenya APRIL 11, 2016

Taiwan Accuses Kenya of Illegal Deportations as More Are Sent to Chi APRIL 12, 2016

Obama to Push U.S. Trade in Kenya China's Role Grows JULY 24, 2015

Case study

- The Republic of Atlantis is planning a rural road rehabilitation program
  - Agricultural households cannot sell goods at market because poor roads, high transport costs → no profit for cash crop production
  - If you fix the roads, they can produce and sell crops at market -- > higher household incomes and consumption, less poverty

Case study: the program

- First they class all villages into groups:
  - 9,000 villages in the country qualify as high priority for road rehabilitation
  - Given budget limits, the Dept of Transportation opens up the program to 2,000 villages and invites them to apply.
  - Eligible villages must apply by a certain date, otherwise cannot receive program
  - By program deadline, 1,021 villages have applied out of 2,000 → these villages receive the program

Case study: the evaluation

- Minister of Finance:
  - Roads are expensive – if they want this program to be scaled up, he wants evidence on the economic return
  - So the team consults with researchers at Atlantis National University on how to design an evaluation that can inform this decision
- What is the main question that they must answer with this evaluation?

Case study: the evaluation

- Project team has done detailed M&E on previous projects
  - Tracked that roads were actually built up to standard in project villages
  - Also measured that, in project villages:
    - travel time to market centers decreased
    - Vehicle operating costs for car owners decreased
- Did it have any effect incomes or poverty?

Case study: the evaluation

- Based on conversation with evaluation team, they improve methods:
  - Collect information not just about travel times, but collect detailed household consumption data from households
  - They collect this data in **both** the program villages ("treatment") and the comparison villages

Method 1: Single difference

	treated villages	comparison villages	Estimated Impact
Average per capita consumption (Atlantis dollars)	301.6	219.1	82.5*

* = statistically significant at 5% level

Method 1: Single difference

- What does this method tell us about the impact of road upgrading on households' welfare?

Method 1: Single difference

- Is it possible that the villages that are part of phase one are different from those that did not? If so, in which ways?

Method 1: Single difference

	Method 1: Simple Difference		
	Treatment	Comparison	Difference
Number of users	44.26	31.83	12.43*
Pop. density	111.90	109.46	2.44*
Local market [1= Yes]	0.86	0.85	0.01
Number of children per HH	4.83	5.27	-0.44*
Diversification (%)	25.90	25.33	0.57
Sample size	1021	979	

IE CONNECT FOR IMPACT

UKaid i2i DIME WORLD BANK GROUP Transport & ICT sustainable mobility for all

Method 2: matching

- To address these issues, we can us an approach called “matching” (or propensity score matching)
- Use what you know about the villages (observable characteristics) to create treatment and control groups that are similar on these characteristics.

Method 2: matching

- Based on what we know about the villages, (population, distance to market, etc), we estimate a probability that they participated in the program.
  - Example: for each village in treatment group with a (25%/50%/75%) probability of participation, you include one in the control group with (25%/50%/75%) probability of participation

Method 2: matching

- Result:
  - "matched" treatment and control groups which are similar across a broad range of characteristics
  - but which differ on whether or not they took part in the program

Method 2: matching

	Method 1: Simple Difference			Method 2: Propensity Score Matching		
	Treatment	Comparison	Difference	Treatment	Comparison	Difference
Number of users	44.26	31.83	12.43*	43.31	34.18	9.13*
Pop. density	111.90	109.46	2.44*	111.40	110.14	1.26
Local market [1= Yes]	0.86	0.85	0.01	0.86	0.85	0.02
Number of children per HH	4.83	5.27	-0.44*	4.95	5.18	-0.23*
Diversification (%)	25.90	25.33	0.57	26.01	25.41	0.60
Sample size	1021	979		886	751	

IE CONNECT FOR IMPACT

* = statistically significant at 5% level

UKaid i2i DIME WORLD BANK GROUP Transport & ICT sustainable mobility for all

Method 2: matching

- From Table 2, what do you notice about the difference in observable characteristics between the treatment and comparison groups when you switch from using Method 1, Simple Difference, to Method 2, Propensity Score Matching?
- Why do you think that is?

Method 2: matching

	Treated villages	Comparison group	Estimated Impact
Average per capita consumption (Atlantis dollars)	290.23	234.41	55.8*

Method 2: matching

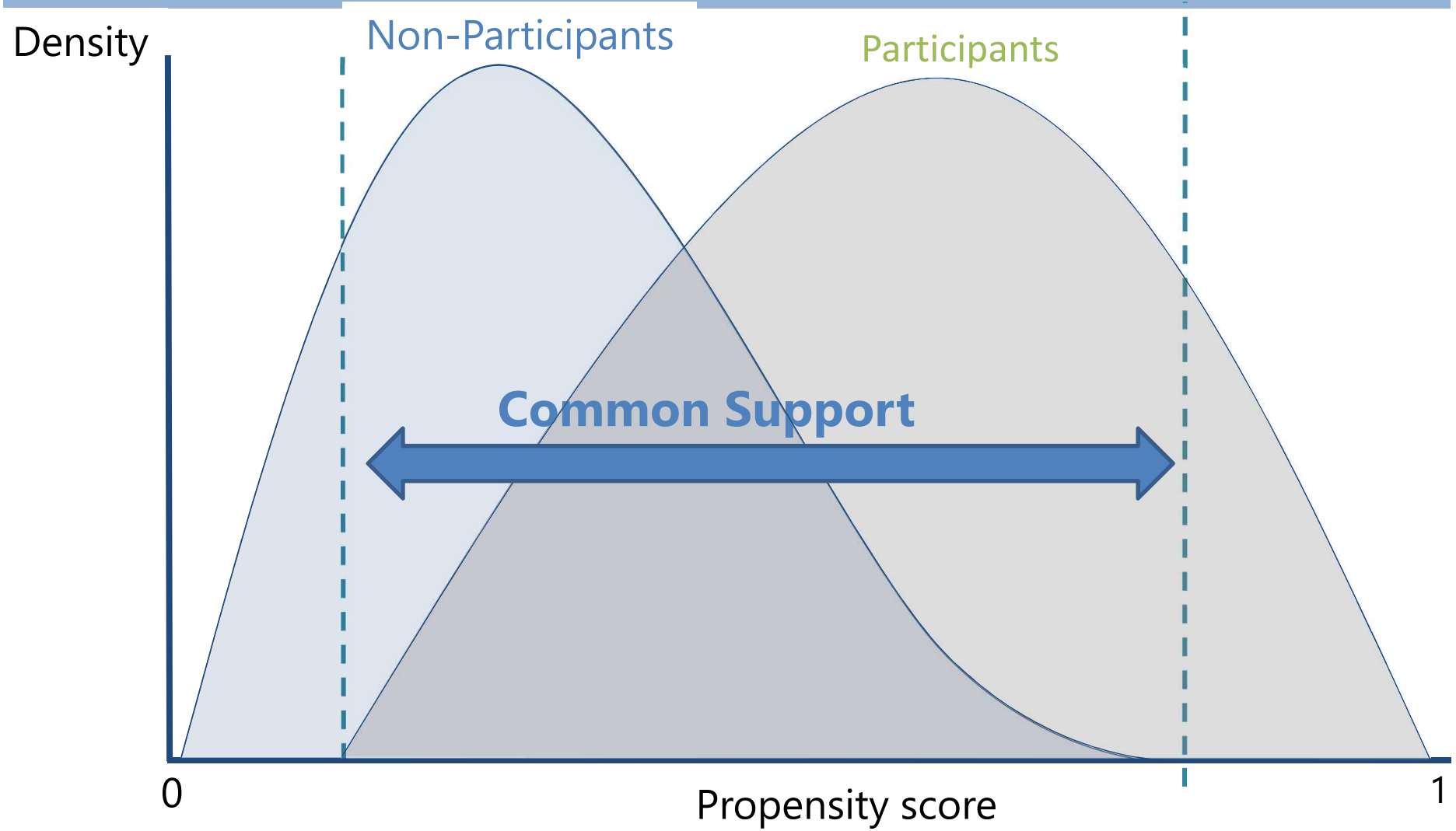
- Why do you think that the estimated impact of the upgrading using Method 2 is smaller than the impact estimated using Method 1?

Method 2: matching

- Notes:
  - This only accounts for observable traits
  - May lose sample size
    - (e.g if you have villages with 99% probability in treatment group but none in control, these will be dropped, and vice versa)

Method 2: matching

Propensity score for participants and non-participants



Method 3: Difference-in-difference

- There is still the possibility that the two groups are fundamentally different
- The *difference-in-difference* method can help when this is the case.
- We measure household consumption before and after the program, and focus on the change over time, rather than the absolute difference

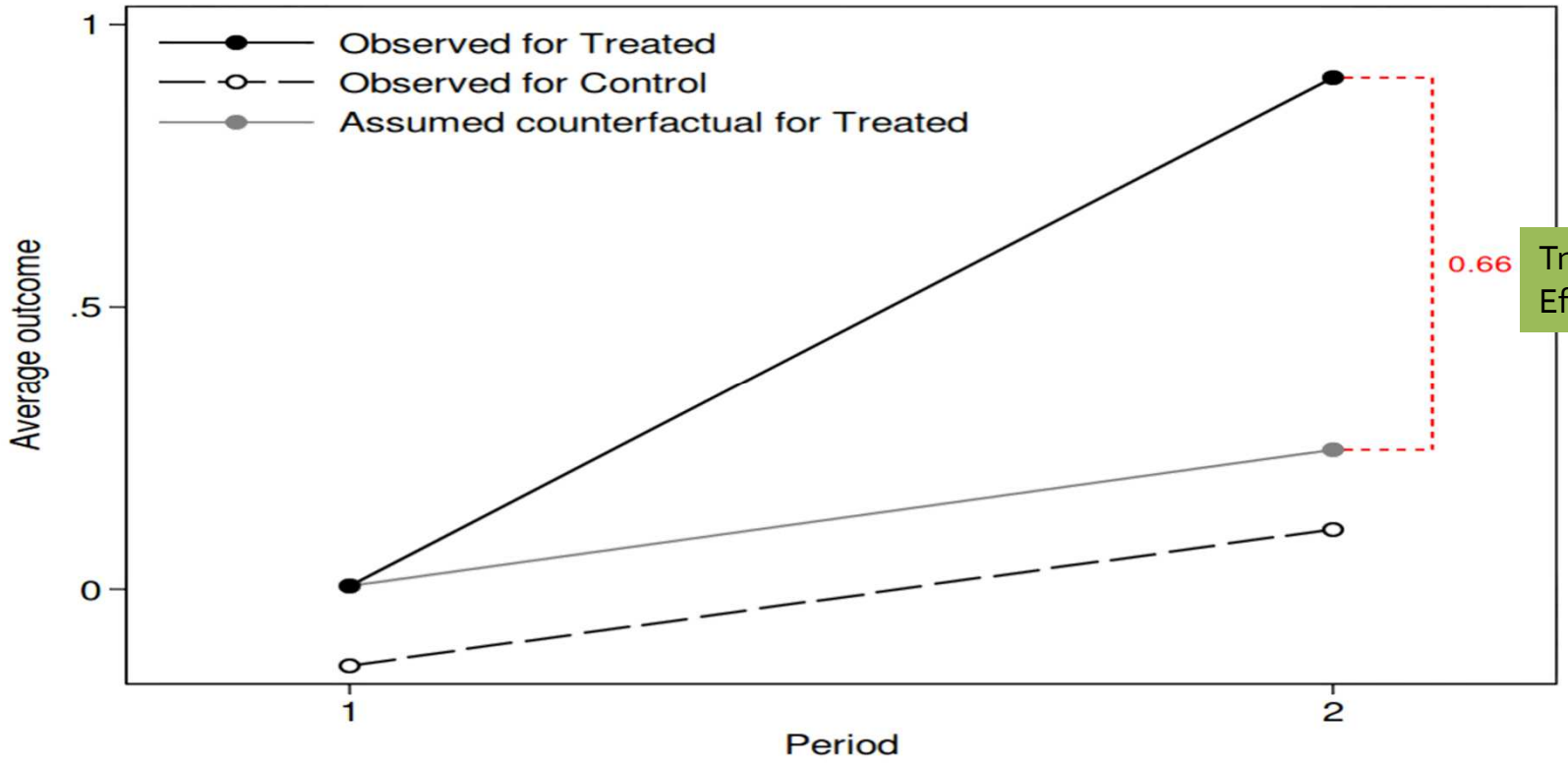
Method 3: Difference-in-difference

- We compare the difference between the change over time for both groups

$$DD = (T_{2016} - T_{2014}) - (C_{2016} - C_{2014}).$$

- This means that even unobservable factors are accounted for, as long as they do not change over time in a way that is correlated with the program

Key assumption: parallel trends



Treatment Effect

Method 3: Difference-in-difference

	TREATMENT upgraded villages	COMPARISON Non upgraded villages	Difference
POST-rural roads upgrading Consumption per capita 2016	301.6	219.1	82.5
PRE- rural roads upgrading Consumption per capita 2014	274.4	219	55.4
PRE- rural roads upgrading Consumption per capita 2012	273.4	218	55.4
Difference in consumption per capita between 2016 and 2014	27.2 (301.6-274.4)	0.1 (219.1-219)	27.1* (301.6-274.4)-(219.1-219) =(Difference-in-Difference)

Method 3: Difference-in-difference

- How could you use this data on consumption per capita in 2012 to improve your analysis? Based on the information in Table 4, what would be your new estimate of the impact of the rural road upgrades on consumption per capita?

- Compare your new estimate to the estimates you obtained with Methods 1 and 2. Is the estimated impact lower or higher? Why do you think this is?

Method 3: Difference-in-difference

- Extra notes
  - Can also use triple difference
  - matching + DD is a common method
  - More powerful when there is significant data before treatment so trends can be examined (and controlled for).
- important weakness:
    - Projects often deliberately targeted based on expectations of differential rates of change
    - “we targeted roads at villages with especially high potential for agricultural growth.”

Method 4: Regression discontinuity (RD)

- Now imagine that instead of allocating the road project to villages that applied on time, the team instead ranked eligible villages
  - based on relevant criteria such as poverty, distance to markets, condition of existing roads
- All 2,000 villages are ranked, and all villages with scores above some threshold receive the program, and those below it do not.

Method 4: Regression discontinuity (RD)

- Treatment and control will be very different in general, *except immediately above and below the threshold*
  - Imagine that the cutoff is 500
  - Wealthy village near the capital = ranked 995
  - Poor remote village 1,000 km from capital = ranked 15
  - **These cannot be meaningfully compared**

Method 4: Regression discontinuity (RD)

- But what about villages 498, 499, 500, 501, 502?
- Presence above or below treatment threshold is essentially arbitrary, (close to) random.
- Villages 499 and 501 are likely very good comparators for each other

Method 4: Regression discontinuity (RD)

- Assignment to the treatment depends on continuous "score" or ranking
  - observations ordered by looking at the score
  - there is a cut-off point for "eligibility" – clearly defined criterion determined *ex ante*
  - cut-off determines the assignment to treatment

conclusions

- High quality non-experimental IE is data-intensive
  - Advances in big data (remote sensing, high frequency/high resolution administrative data, new survey methods) are making this more feasible
  - Many examples in forthcoming presentations

conclusions

- To design conduct high quality impact evaluations of major transport infrastructure projects, we may need the full toolkit of IE methods

conclusions

- In Phase 1 of ieConnect, we have IEs which have:
  - A **non-experimental** component which estimates the impact of transport infrastructure (a road, a corridor, a BRT system)
  - Complementary **experimental** interventions which test key components of program logic