An Introduction to Counterfactual Impact Evaluation

Prof. Dr. Marco Caliendo
(University of Potsdam and IZA Bonn)
caliendo@uni-potsdam.de
www.caliendo.de

Brussels, April 18, 2018
Evaluation Helpdesk of Cohesion Policy 2014-2020
Theory Based and Counterfactual Impact Evaluation

Some relevant literature/links:

Methodological:

- Caliendo, M., and R. Hujer (2006): The Microeconometric Estimation of Treatment Effects - An Overview, *Journal of the German Statistical Society*, 90(1), 197-212.
- Imbens, Guido W., and Jeffrey M. Wooldridge. 2009. Recent Developments in the Econometrics of Program Evaluation. *Journal of Economic Literature*, 47 (1): 5-86.

Practical:

- Bamberger, M., Rugh, J, Church, M., and Fort, L. (2004): Shoestring Evaluation: Designing Impact Evaluations under Budget, Time and Data Constraints, *American Journal of Evaluation*, 25(1): 5-37.
- Caliendo, M. and Kopeinig, S. (2008), Some Practical Guidance for the Implementation of Propensity Score Matching. *Journal of Economic Surveys*, 22: 31-72.
- Gertler, P., Martinez, S.. Premand, P., Rawlings, L and Vermeersch, C. (2011): Impact Evaluation in Practice, 2nd Edition, The World Bank.
- World Bank Repository Impact Evaluation: www.worldbank.org/en/programs/sief-trust-fund/publication/impact-evaluation-in-practice

Counterfactual Impact Evaluation

- Goal: Estimate the causal impact of a certain policy on affected "units"!
- The scope of evaluation topics is virtually unlimited and units can be individuals, firms, regions or even countries.

Some Examples

- Development Policy: Do conditional cash transfers to families increase school attendance rates?
- Labour Market Policy: Do start-up subsidies help unemployed individuals re-integrate into employment?
- Infrastructure policy: Does increased broadband internet access affect employment growth of establishments?
- Tourism policy: Do tax-cuts for the hospitality sector and investment in infrastructure increase regional employment?
- Regional policy: Do structural funds transfers improve regional performance?

Fundamental Evaluation Problem

- All these examples have one thing in common: There is a treatment (intervention, manipulation), there are units (not) affected by it and there is an outcome variable.
- Central question to answer:

  "What would have happened had the affected units not received the treatment?" (counterfactual outcome)
- Causal effect: Comparison of observed outcome with counterfactual situation.
- Fundamental Evaluation Problem: This counterfactual is never observed (for the same unit at the same time)!

$\Rightarrow$ Hence, we need to find a good proxy from a comparison group!

What is (not) a good proxy?

Hypothetical Example

- We want to evaluate a training program for low-skilled individuals (treatment group).
- We have the following data on their employment rates as well as the average employment rates in the population (comparison group):

	Before	After
Treatment group	0.60	0.75
Comparison group	0.74	0.78

- Can we conclude from these figures whether the program was successful?
- Before-after estimator: $0.75 - 0.60 = +0.15$
- Cross-section estimator: $0.75 - 0.78 = -0.03$
- Difference-in-Differences: $(0.75 - 0.60) - (0.78 - 0.74) = +0.11$

Selection Bias

- The major problem with these approaches is that assignment to treatment and comparison group is not random.
- Participants and non-participants might differ even in absence of the program:
  - Individuals may differ in their level of education, labour market experience, ...
  - Firms could differ in terms of productivity, firm size, sector, ...
  - Regions could be different in their population density, age distribution, sectoral composition, ...
- Hence, simple (mean) comparison are not meaningful because of selection bias.

Solving the Selection Problem

- There are a variety of well-established methods to overcome selection bias. Three broad categories:
  - Experimental methods
  - Quasi-experimental methods
  - Non-experimental methods
- Our focus today: Quasi-experimental and non-experimental methods!
- Keep in mind: There is no magic bullet!
  - Each approach has their own strengths and weaknesses and works only if a certain set of assumptions is met.
  - Which one is best for the problem at hand depends on the evaluation question, institutional features, data availability, etc.

Outline

Outline

2 Evaluation Framework

Program Evaluation - An Ideal World Scenario (1)

- In an ideal world, the evaluator is already involved at early stages of the program design and has influence on the data collected for later evaluation.
- These stages include:
  1. Defining the program's goals
  2. Develop a theory of change
  3. Program design
  4. Implementation and collection of baseline data
  5. Collect final outcome data
  6. Counterfactual impact evaluation
- Process evaluation (focus on program implementation and operation) und impact evaluation should be viewed as complements.
- We can use the information collected in process evaluation to choose amongst alternative evaluation estimators.

Program Evaluation - An Ideal World Scenario (2)

- Important questions which should already be answered at the design stage:
- Aims and measure of success:
  - What are the intended effects of the program?
  - How does one measure the success of the program?
- Theory of change:
  - What is the sequence of events that leads to observed outcomes?
  - Which different channels contribute to the success of the program?
- Empirical strategy:
  - What type of evaluation methodology is to be pursued?
  - How will the necessary data be gathered?
  - How can one distinguish which theoretical mechanisms are most important?

$\Rightarrow$ In an ideal world, the evaluators have sufficient time, budget and high-quality-data at their disposal.

Program Evaluation - The Real World Scenario

– However, in the real world evaluations are often performed under less than optimal circumstances ("shoestring evaluations"):

The Constraints Under which Evaluations must be performed			
Time	Budget	Data	Typical Scenario
×			Evaluator is called in late with tight deadline
	×		Difficulties collecting survey data
		×	No baseline data available, sensitive subject with difficult data collection
×	×		Secondary data is available but little time to analyze it
×		×	Little time and no data has been collected survey design limited due to time constraint
	×	×	Evaluator is called in late, deadline not an issue No access to baseline data, budget is tight
×	×	×	Evaluator is called in late with tight deadline and tight budget, no baseline data and no control group has been identified

Source: Bamberger et. al (2004)

Outline

3 Identifying Causal Effects

Formal Definition of Causal Effects

- Every unit of observation i has two potential outcomes:

$$Y_i = \begin{cases} Y_i^1 & \text{if treated } (D = 1) \\ Y_i^0 & \text{if untreated } (D = 0) \end{cases}$$

- The unit-level causal effect is defined as

$$\Delta_i = Y_i^1 - Y_i^0.$$

- We will never be able to estimate unit-level effects with confidence, hence we focus on population averages.
- The most prominent parameter estimated is the average treatment effect on the treated (ATT):

$$\begin{aligned} \Delta_{ATT} &= E[\Delta \mid D = 1] \\ &= E[Y^1 \mid D = 1] - \underbrace{E[Y^0 \mid D = 1]}_{\textit{unobservable}} \end{aligned}$$

Selection Bias

- Selection bias arises whenever our samples of participants and non-participants are incomparable in some way.
- This means that both groups have different mean outcomes even without treatment:

$$E[Y^0 \mid D = 0] \neq E[Y^0 \mid D = 1]$$

- This incomparability is caused by differences in characteristics that affect selection and our outcome of interest Y.
- These differences may be due to either ...
  - ... observed characteristics or
  - ... unobserved characteristics.
- Depending on the reason for the incomparability, different evaluation methods are needed.

Types of Selection: Examples

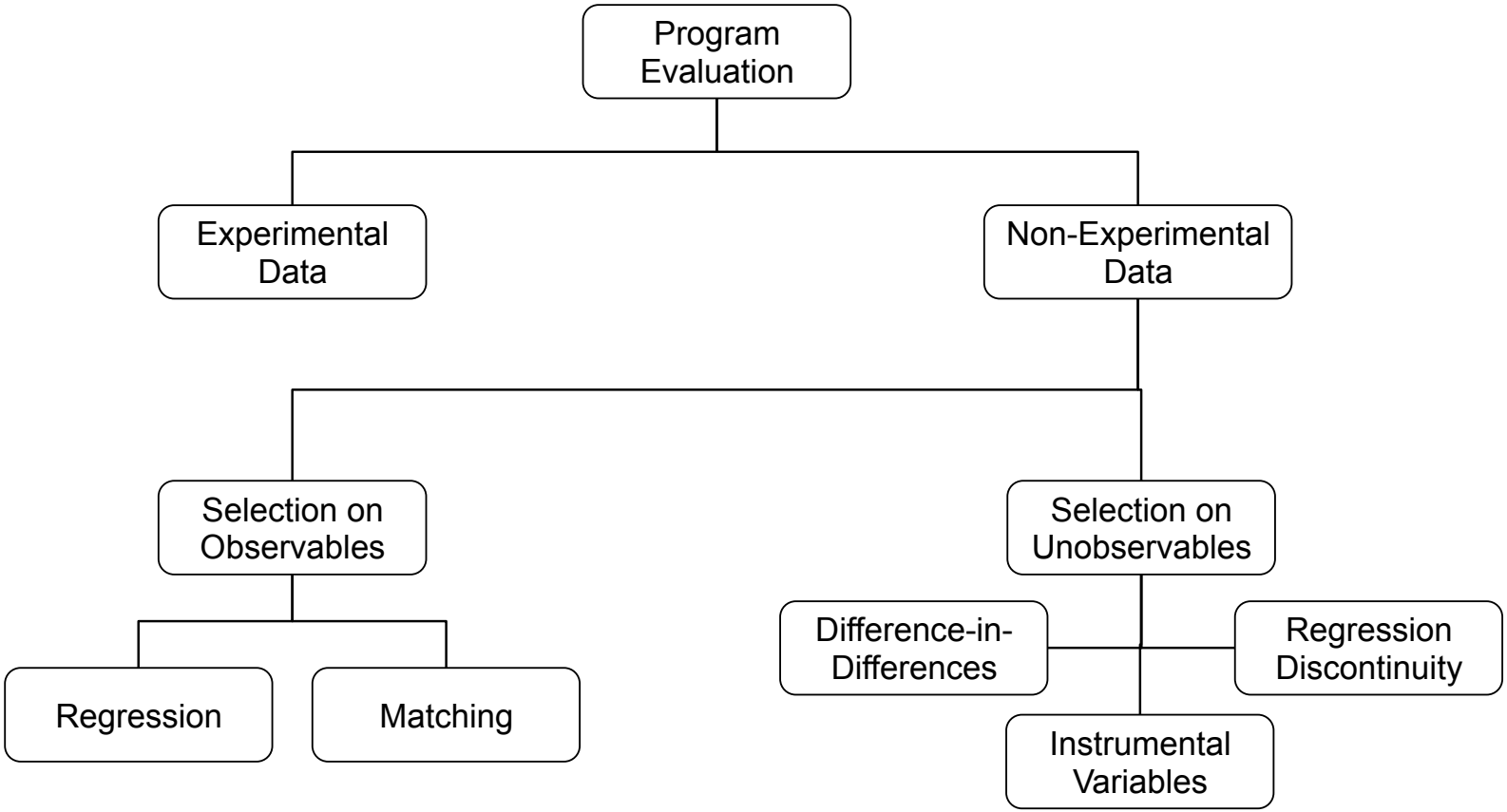
Differences due to observed characteristics

- Participants in active labour market programs have often worse labour market history than non-participants.
- Regions receiving development aid are more likely to have a lower educated work force than other regions.
- Companies that obtain R&D subsidies are often larger and more productive than non-recipients.

Differences due to unobserved characteristics

- Previously unemployed participants in a start-up subsidy may be more motivated than other unemployed individuals.
- Poorer households in developing countries that receive cash transfers may follow a more traditional family values than non-poor households.
- Countries that subsidize loans for start-ups may also have lower beaurocratic burdon to set up a business.

Evaluation Methods



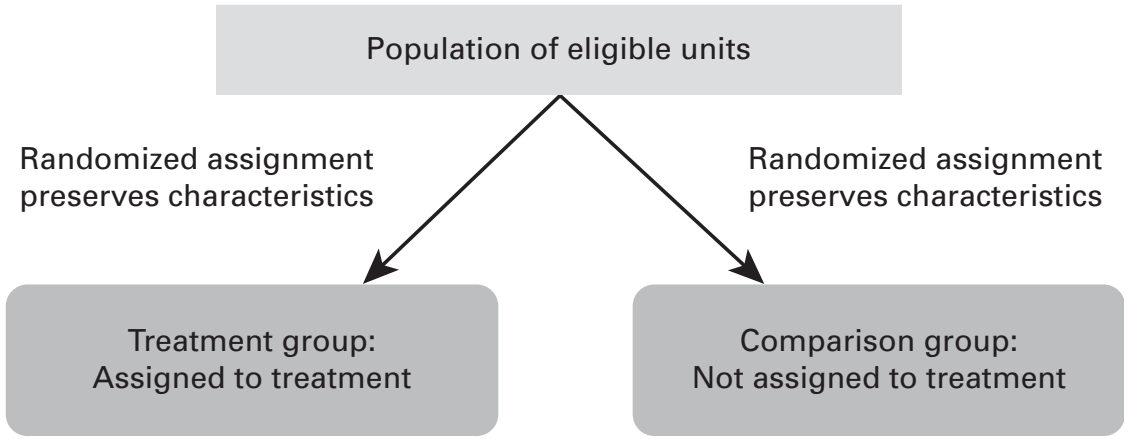
Outline

Evaluation Methods

4.1 Randomised Controlled Trials

Randomised Controlled Trials

– Randomised controlled trials (RCTs) assign units from the eligible population randomly:

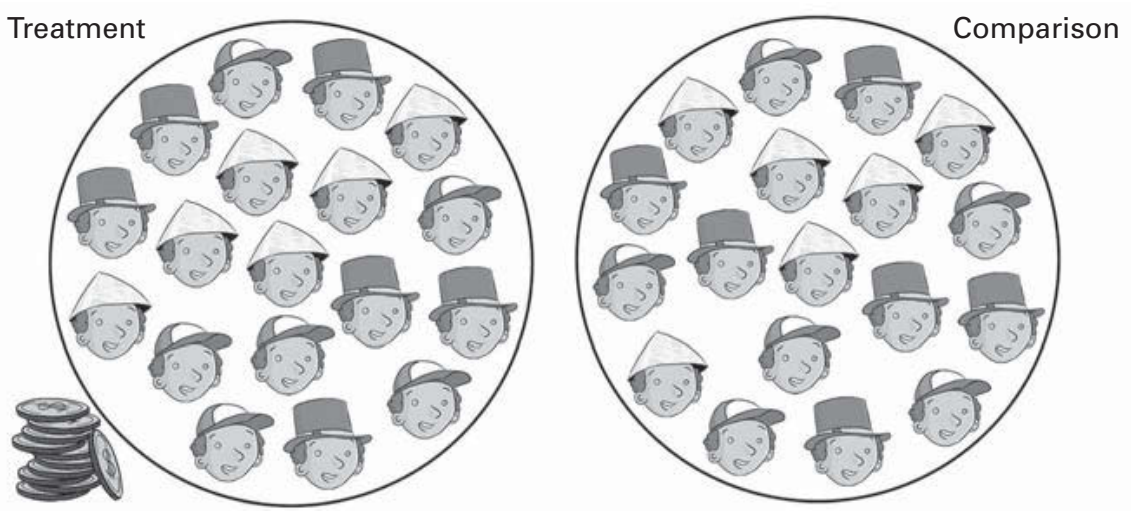


Source: Evaluation in Practice

– This guarantees that participation is unrelated to the units' characteristics.

Randomised Controlled Trials (2)

- Result: RCTs lead to balanced samples in both observed and unobserved characteristics:



Source: Evaluation in Practice

- Therefore, observed outcome differences between the two groups can be solely attributed to the treatment!
- Estimator: Simple cross-sectional mean differences in outcome Y.

Randomised Controlled Trials (3)

Hypothetical Example

- Let's revisit our hypothetical example on the training program for low-skilled individuals.
- Assume we have access to experimental data:

	Before	After	% low-skilled
Treatment group	0.60	0.75	100
Experimental controls	0.60	0.67	100
Comparison group	0.74	0.78	30
Low-skilled	0.60	0.67	100
High-skilled	0.80	0.83	0

- Random assignment guarantees balanced characteristics in treated and experimental control sample.
- Experimental estimator: $0.75 - 0.67 = +0.08$
- The non-experimental comparison group also consists of high-skilled individuals with high employment rates.

Example RCT: Progresa

Schultz (2004): School Subsidies for the Poor: Evaluating the Mexican Progresa Poverty Program

- Research Question: Do conditional cash transfers to poor mothers in rural Mexico raise their children's school enrolment rates?
- Treatment: Mothers receive monthly transfers if their children attend school.
- Data: Survey data, gathered in 1997/1998. $N \approx 39,000$.
- Method: Randomised controlled fiel experiment. Poor households are randomly assigned to treatment or control group.
- Results: Progresa significantly increased enrolment rates and educational attainment of program participants!

RCT: Pros, Cons, Pitfalls and Requirements

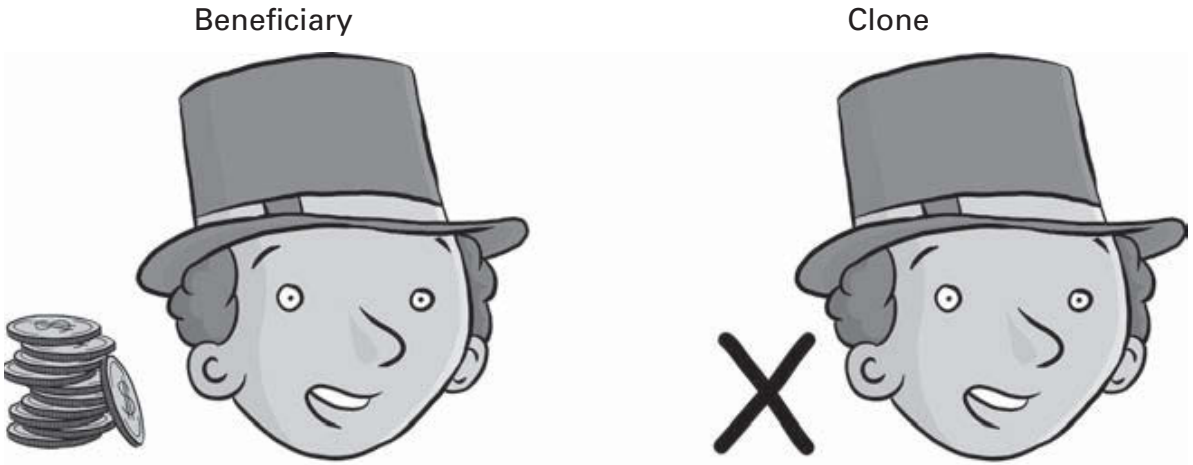
- Pros and Cons:
  - (+) Credible, intuituve estimates of causal effects (high internal validty)
  - (–) Costly, ethical concerns.
- Although social experiments seem to be very appealing in providing a simple solution to the fundamental evaluation problem, there are potential threats undermining their internal and external validity.
- Pitfalls:
  - Randomization may sometimes fail to produce balanced samples.
  - Subjects knowing they take part in an experiment may behave differently ("hawthorne effect").
  - Individuals willing to take part in an experiment may be systematically different from the population of interest (randomization bias $\Rightarrow$ low external validity).
- Requirements:
  - Close cooperation between researchers and policymakers.
  - Sufficient number of units to be randomised.
- In many situations RCTs will not be feasible and we need to think about identifying causal impacts with non-experimental data.

Evaluation Methods

4.2 Matching

Matching (1)

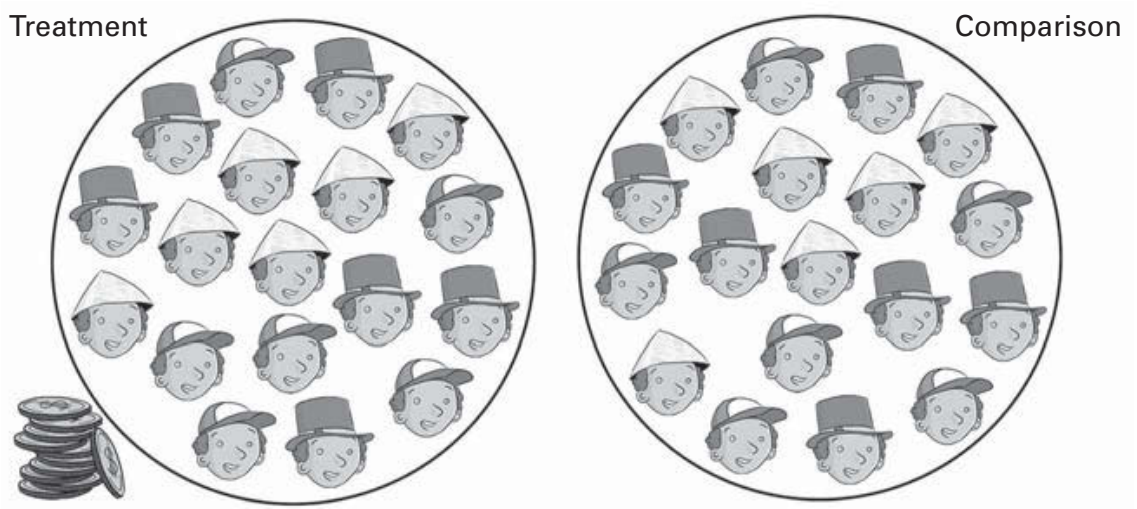
- Matching methods aim to mimic an RCT with observational data.
- Idea: Choose for each participant, one (or many) statistical twins from the sample of non-participants.
- They should be identical in all relevant characteristics! This is a very strong requirement and requires informative data.



Source: Evaluation in Practice

Matching (2)

– Similar to an RCT, this leads to a balanced sample:



Source: Evaluation in Practice

– Estimator: Simple cross-sectional mean differences in outcome Y on the matched sample.

Matching (3)

Hypothetical Example

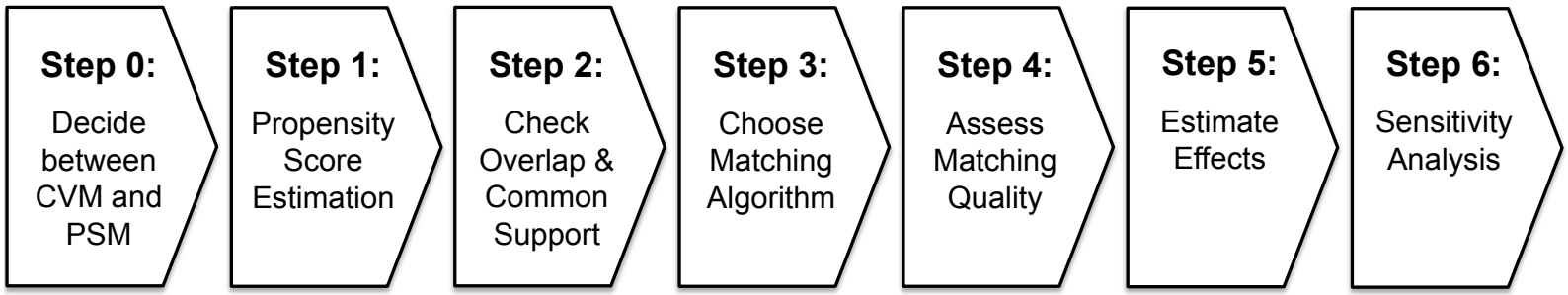
- Let's return to our hypothetical example on the training program for low-skilled individuals.
- The matching procedure picks the statistical twins (low-skilled) from the comparison group.

	Before	After	% low-skilled
Treatment group	0.60	0.75	100
Experimental controls	0.60	0.67	100
Comparison group	0.74	0.78	30
Low-skilled	0.60	0.67	100
High-skilled	0.80	0.83	0

- Matching estimator: $0.75 - 0.67 = +0.08$
- Matching re-creates the experimental estimates when all relevant characteristics are observed.

Propensity Score Matching

- Curse of dimensionality: If the number of relevant characteristics is large, it may be very difficult to find an exact match!
- One solution: Propensity-score matching summarizes all information in one index and choose the closest non-participant in terms of that index.
- Implementation:



Source: Caliendo/Kopeinig (2008)

Matching: Pros, Cons, Pitfalls and Requirements

- Pros and Cons:
  - (+) Intuitive by mimicking an RCT
  - (+) Can be applied in many settings
  - (−) Only balances observed characteristics
- Pitfalls:
  - Some matching methods may not balance samples satisfactorily (alternatives: automatic balancing through algorithms).
  - If groups are very different, not all participants may be matched with a non-participant and effects can only be estimated for a subset of the treated units.
  - Estimator fails if there are differences in unobserved characteristics that affect the outcome of interest.
- Requirements:
  - Very good and rich data.
  - Good knowledge of the institutional setting and selection process.

Better Data Helps A Lot!

- Implementing a matching approach in a credible way is not easy. Better data helps a lot!
- Often, the estimates can be improved by combining several data sources:
  - Individual- and firm-level data are often available from administrative records at low cost (e.g. through national employment agencies).
  - Regional/country-level data are provided by (inter-) national statistics agencies.
- New trends:
  - Augment individual or firm data with regional data to make sure units operate in the same kind of economic environment.
  - Merging admin data with survey data allows the evaluator to enrich the admin data with information on "usually unobserved" characteristics (personality, preferences, expectations, etc.).

Example Matching: Start-Up Subsidies (1)

Caliendo/Künn/Weißenberger (2016): Personality traits and the evaluation of start-up subsidies

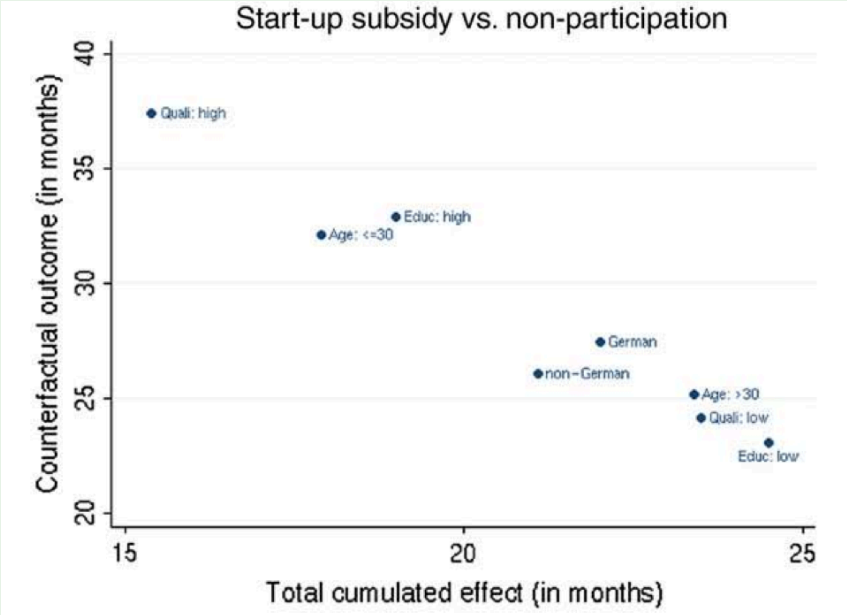
- Research Question: Are start-up subsidies for the unemployed an effective active labour market program? And do omitted personality traits pose a threat to the reliability of the matching estimates?
- Treatment: Unemployed individuals willing to set-up a business obtain monthly transfers for up to 15 months.
- Data: Combination of administrative and survey data. $N \approx 1,300$.
- Source of selection bias: Participants self-select into the program; participants differ in their characteristics from non-participants!
- Method: Matching participants and non-participants based on a large set of characteristics and pre-treatment outcomes.
- Results:
  - Positive effects on employment probabilities and income.
  - Results are robust to the inclusion of usually unobserved personality traits!

Example Matching: Start-Up Subsidies (2)

Caliendo/Künn (2011): Start-Up Subsidies for the Unemployed: Long-Term Evidence and Effect Heterogeneity

- Research question: Long-term effects of start-up subsidies for unemployed?
- Results: Positive and significant effects on employment (ATT=23.5 months) and income 56 months after participation.
- Effect Heterogeneity: Effects are higher for low educated participants and participants above the age of 30.

⇒ Matching estimators allow you to identify effect heterogeneity!



Evaluation Methods

4.3 Difference-in-Differences

Difference-in-Differences (1)

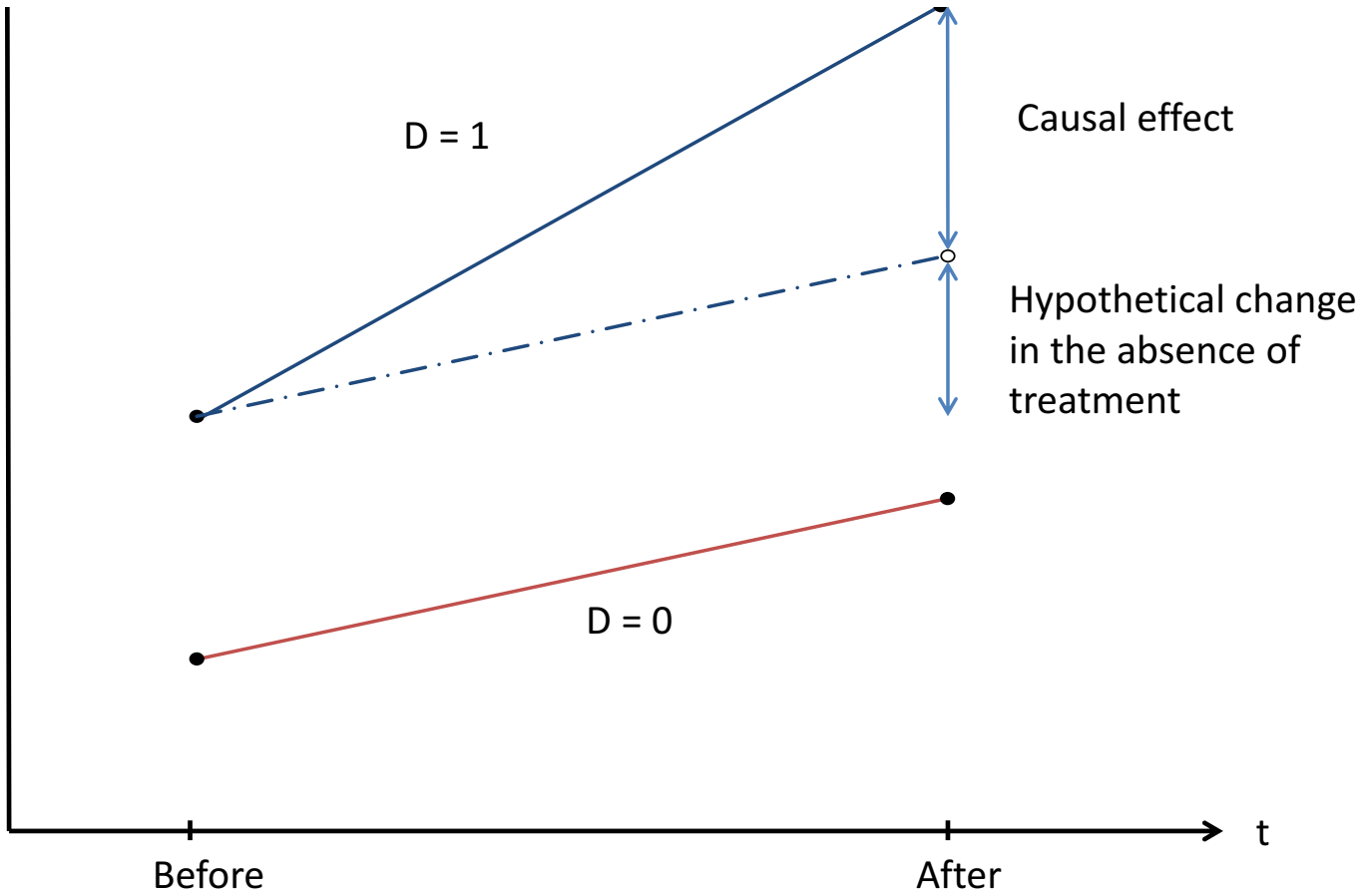
- Difference-in-Differences (DiD) set-ups often exploit some kind of "natural experiment" that occurs because of some policy change, where one group of units is affected by the treatment and one group is unaffected.
  - For example: One state raises the minimum wage, but the neighbouring state does not.
- Important: DiD assumes parallel time trends (PTT) for treatment and control group in absence of the treatment and allows for different pre-treatment levels ("baseline bias").
- Validity of the PTT:
  - Inspecting the similarity of pre-treatment trends provides some indication on the likelihood that the PTT assumption holds.
  - Significantly different pre-treatment trends cast serious doubt on the reliability of estimates.

Difference-in-Differences (2)

- Intuition of the DiD Estimator: Combine before-after estimates for the treatment and the control group.
  - By comparing changes within groups, we implicitly control for time-constant unobserved factors.
  - By comparing these changes across groups, we also control for time-trends in outcomes.
- Estimator:

$$\text{DiD} = \underbrace{E[Y^{after} - Y^{before} \mid D = 1]}_{\text{BAE for the affected}} - \underbrace{E[Y^{after} - Y^{before} \mid D = 0]}_{\text{BAE for the unaffected}}$$

Illustration



DiD: Pros, Cons, Pitfalls and Requirements

- Pros and Cons:
  - (+) Intuitive method using a "natural experiment"
  - (+) Similarity of pre-treatment trends can easily be compared
  - (+) Allows for time-constant unobserved factors
  - (-) Results may be sensitive to which time-frame is used around the policy shift
- Pitfalls:
  - Pre-treatment trends may be very different between two groups.
  - Treatment may contaminate the group definitions (e.g. the minimum wage hike may result in restaurants setting up shop across the state border).
- Requirements:
  - We need data over several time-periods.
  - More data on pre-treatment years helps with inspecting the parallel trends assumption.

Example DiD - Minimum Wages (1)

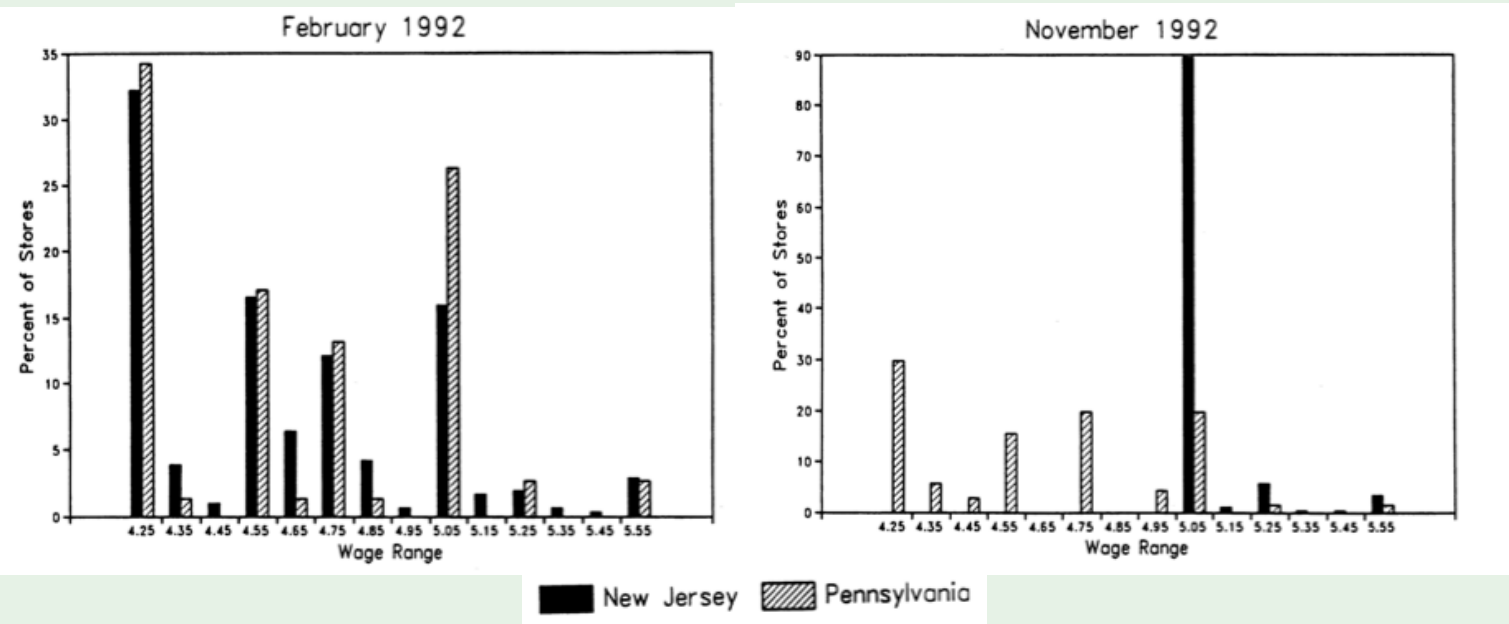
Card/Krüger (1994): Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania

- Research Question: Impact of minimum wage increase on low-wage employment?
- Treatment: Rise of minimum wage from \$4.25 to \$5.05 per hour in New Jersey in April 1992.
- Data: Survey data on wages and employment for $N = 410$ fast food restaurants in New Jersey and Pennsylvania.
- Source of selection bias: Unaffected restaurants in New Jersey may serve to different customers and offer more pricey meals.
- Method: Compare the evolution of full-time employment in fast-food restaurants in NJ and neighboring state PA.

Example DiD - Minimum Wages (2)

Card/Krüger (1994): Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania

- Descriptive comparison of pre- and post-treatment wages.



Example DiD - Minimum Wages (3)

Card/Krüger (1994): Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania

– Calculating the sample averages yields (s.e. in parentheses):

Variable	Stores by state: PA (i)	NJ (ii)	Difference, NJ − PA (iii)
1. FTE employment before, all available observations	23.33 (1.35)	20.44 (0.51)	−2.89 (1.44)
2. FTE employment after, all available observations	21.17 (0.94)	21.03 (0.52)	−0.14 (1.07)
3. Change in mean FTE employment	−2.16 (1.25)	0.59 (0.54)	2.76 (1.36)

Source: Card/Krueger (1994), p. 780

Evaluation Methods

4.3 Synthetic Control Method

Synthetic Control Method

- What if we are faced with a policy that only affects one unit, e.g., a region, state or country?
- Idea of the Synthetic Control Method (SCM): Re-weight unaffected units to obtain a synthetic control unit.
- How to find the weights? Data-driven algorithm, that assigns weights to control units such that ...
  - the synthetic control unit looks like the treated unit before the policy was in place ...
  - ... both in terms of trends in pre-treatment outcomes and characteristics.
- Estimator: Difference between outcome of treated unit and synthetic control unit.

SCM: Pros, Cons, Pitfalls and Requirements

- Pros and Cons:
  - (+) Can be applied for treatments at aggregate level
  - (+) Very transparent through data-driven algorithm
  - (+) Quality of weights are easy to assess graphically
  - (−) Unobserved factors may cause bias
  - (−) It may be hard to find suitable control units that were not affected by the (same or similar) policy shift
- Pitfalls:
  - The algorithm may fail to produce acceptably similar pre-treatment trends if the treated unit and the control units are too different.
- Requirements:
  - Data required can usually be obtained through (inter)national statistical offices.
  - Sufficient data on pre-treatment trends needs to be available in order to get a credible match.

Example SCM - Industrial Policy (1)

Castillo/Figal Garone/Maffioli/Salazar (2017): The causal effects of regional industrial policies on employment

- Research Question: Can state-level tourism policy raise regional employment?
- Treatment: In 2003, the Argentinian state of Salta implemented tax-credits for the hospitality sector and invested in infrastructure, restoration of historical sights and marketing for tourism abroad.
- Data: Monthly, aggregate data on all Argentinian states published by the Ministry of Labour. Years 1996-2013.
- Source of selection bias: Salta was a state with relatively poor population with low employment rates before the introduction!
- Method: Weight control states to construct a synthetic control unit that has similar pre-treatment characteristics and outcome trends.

Example SCM: Industrial Policy (2)

Castillo/Figal Garone/Maffioli/Salazar (2017): The causal effects of regional industrial policies on employment

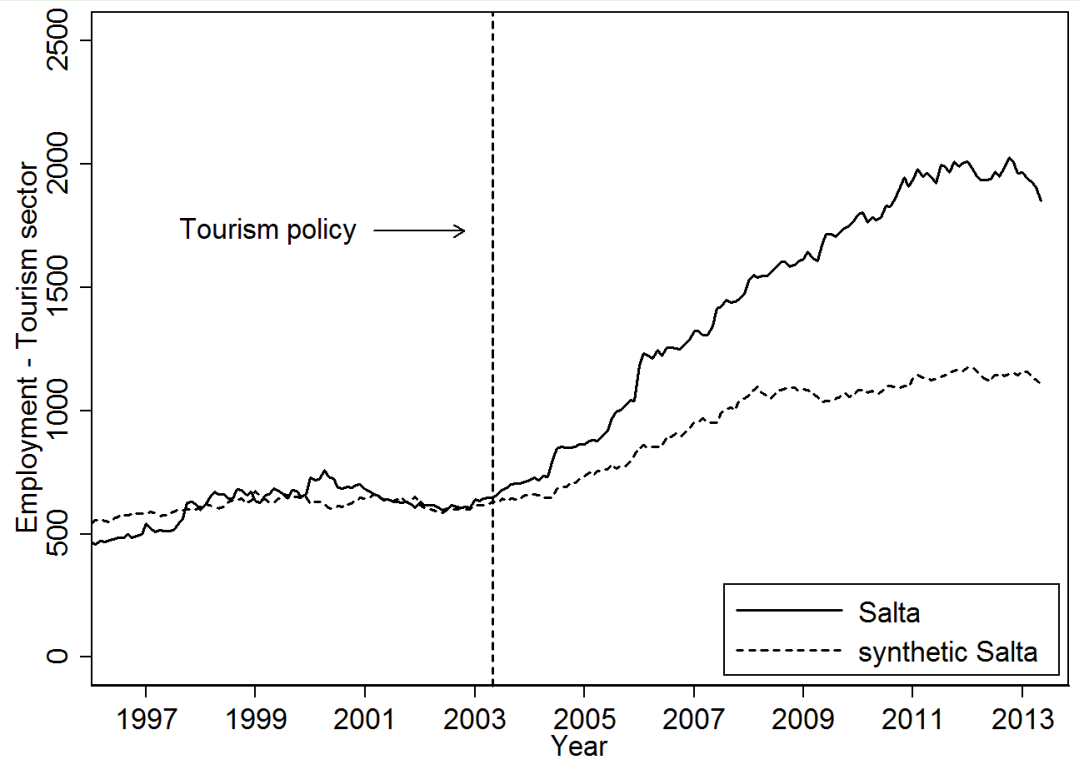
Table 1. Province weights in the synthetic Salta

Province	Weights
Buenos Aires	-
Autonomous City of Buenos Aires	-
Catamarca	0
Córdoba	-
Corrientes	0
Chaco	0
Chubut	0
Entre Ríos	0
Formosa	0.114
Jujuy	0.393
La Pampa	0
La Rioja	0
Mendoza	0
Misiones	0
Neuquén	0.064
Río Negro	-
San Juan	0
San Luis	0
Santa Cruz	0
Santa Fé	0.222
Santiago del Estero	0
Tucumán	0.207
Tierra del Fuego	0

Example SCM - Industrial Policy (3)

Castillo/Figal Garone/Maffioli/Salazar (2017): The causal effects of regional industrial policies on employment

– Results: The tourism policy led to a significant increase in employment, not just in the hospitality sector (as shown below) but also in other sectors.



Evaluation Methods

4.5 Instrumental Variables

Instrumental Variables (1)

Hypothetical Example: RCT with non-compliance

- Again, imagine you want to evaluate the effects of a training program on the individuals' subsequent employment probabilities.
- You randomly assign whether applicants receive a voucher for the program or not.
- After you run the experiment and analyze your data, you find that ...
  - ... 10 % of the people assigned the voucher ($Z = 1$) never took part in the program and ...
  - ... 10% of the people assigned to control ($Z = 0$) got access to the program anyway.

Instrumental Variables (2)

Hypothetical Example: RCT with non-compliance

- Result: Actual participants and non-participants of the training program are again selected groups!
- Therefore, simple comparisons between those two groups will suffer from selection bias. But:
  - Mean comparisons between those assigned to receive the voucher and those without vouchers give a credible estimate of the effect of voucher receipt on employment outcomes $\Rightarrow$ Intention-to-treat (ITT) analysis.
  - The true effect of taking part in the program will be larger, because the ITT analysis ignores, that some individuals in the voucher group ($Z = 1$) did not receive the benefits of the program, while some of the other group ($Z = 0$) group did.
- How do we get an estimate of the local average treatment effect (LATE) of the program for those that actually receive treatment, but only if assigned the voucher ("compliers")?

Instrumental Variables (3)

Hypothetical Example: RCT with non-compliance

	Group assigned to treatment	Group not assigned to treatment	Impact
	Percent enrolled = 90% Average Y for those assigned to treatment = 110	Percent enrolled = 10% Average Y for those not assigned to treatment = 70	Δ% enrolled = 80% ΔY = ITT = 40 LATE = 40/80% = 50
Never enroll			—
Only enroll if assigned to treatment			
Always enroll			—

Instrumental Variables (4)

- In the hypothetical example, the random assignment indicator Z for the voucher serves as an instrumental variable (IV).
- Definition of an IV: An instrumental variable is one that has a causal impact on selection into treatment.
- Crucial assumptions:
  - The IV is unrelated to unobserved factors!
  - It must not have a direct impact on the outcome of interest!
- Local Average Treatment Effect:
  - Under these assumptions, an IV estimate gives the local average treatment effect for units affected by the treatment (compliers).
  - For units that always or never receive treatment, whatever value the instrument takes on, IV methods provide no information.

Instrumental Variables (5)

- In the case described – with a randomly assigned binary instrument Z – the IV Estimator can be written as

$$\hat{\Delta}_{IV}^{LATE} = \frac{E[Y \mid Z = 1] - E[Y \mid Z = 0]}{E[D \mid Z = 1] - E[D \mid Z = 0]},$$

  where ...
  - ... the nominator gives the ITT effect of the instrument and ...
  - ... the denominator represents the fraction of compliers.
- Intuitively, the IV estimator scales up the ITT estimate to account for the fact that not everyone in the sample is affected by the instrument.

IV: Pros, Cons, Pitfalls and Requirements

- Pros and Cons:
  - (+) With a valid instrument, the method provides very credible estimates.
  - (–) Without a randomly assigned instrument, it may still be related to unobserved factors!
  - (–) Compliers may not be your population of interest.
  - (–) Method hard to communicate.
- Pitfalls:
  - Some instruments have only a small impact on the treatment status despite plausible theoretical effects.
  - Other instruments may have a direct impact on the outcome of interest.
- Requirements:
  - Typically, IV methods need very large samples in order to give precise estimates!

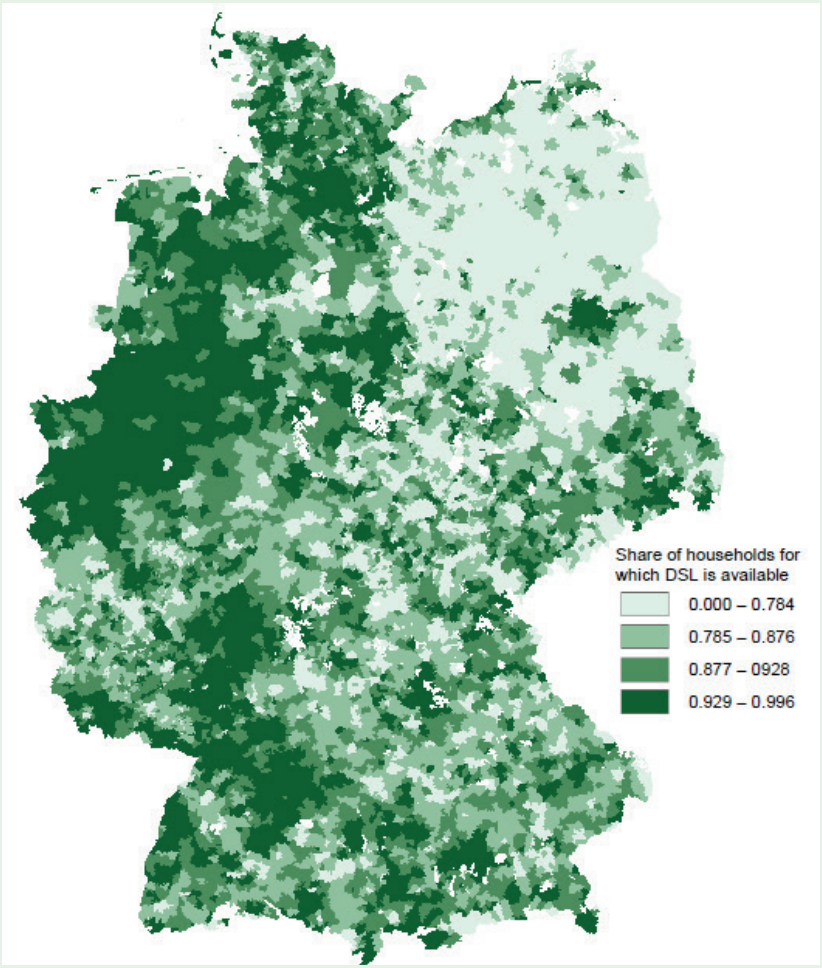
Example IV - Internet and Employment Growth (1)

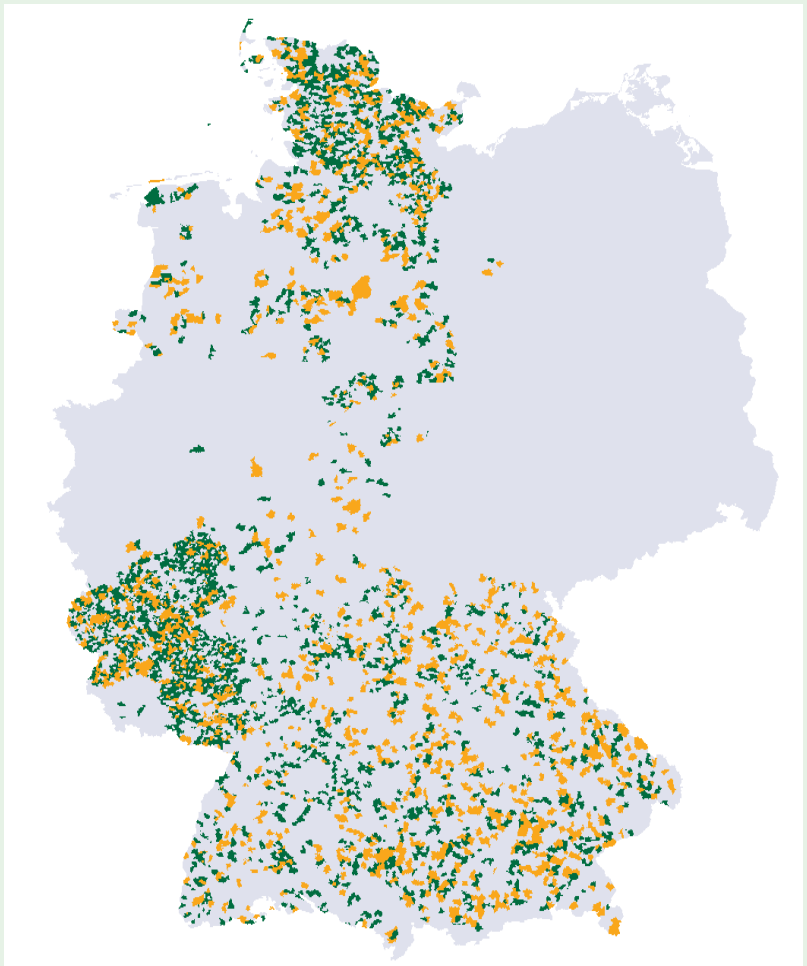
Stockinger (2017): The effect of broadband internet on establishments' employment growth: evidence from Germany

- Research Question: Does broadband internet access affect employment growth of German establishments?
- Treatment: Roll out of broadband internet access across Germany in the 2000s.
- Source of selection bias: Firms that get internet access more quickly might be more productive.
- Data: Combination of the IAB Establishment Survey with administrative data on telephone networks. $N = 25,000$ establishments, years 2005-2009.

Example IV - Internet and Employment Growth (2)

Stockinger (2017): The effect of broadband internet on establishments' employment growth: evidence from Germany





Legend for right graph: green $<=$ 4.2 km, yellow $>$ 4.2 km

Example IV - Internet and Employment Growth (3)

Stockinger (2017): The effect of broadband internet on establishments' employment growth: evidence from Germany

- Method:
  - Compare outcomes of establishments that are below 4.2 km distance to their next main telephone distribution frame (installed: 1960s) with other establishments.
  - For technological reasons, establishments below the 4.2 km threshold are more likely to have broadband internet access.
- Results: Broadband internet access increase employment growth in the service sector and decreased employment growth in the manufacturing sector in western Germany.

Evaluation Methods

4.6 Regression Discontinuity Designs

Regression Discontinuity Designs (1)

– Many programs operate with some eligibility cut-off with respect to some index.

Examples

– Anti-poverty program: Only households below some poverty index are eligible for transfers.

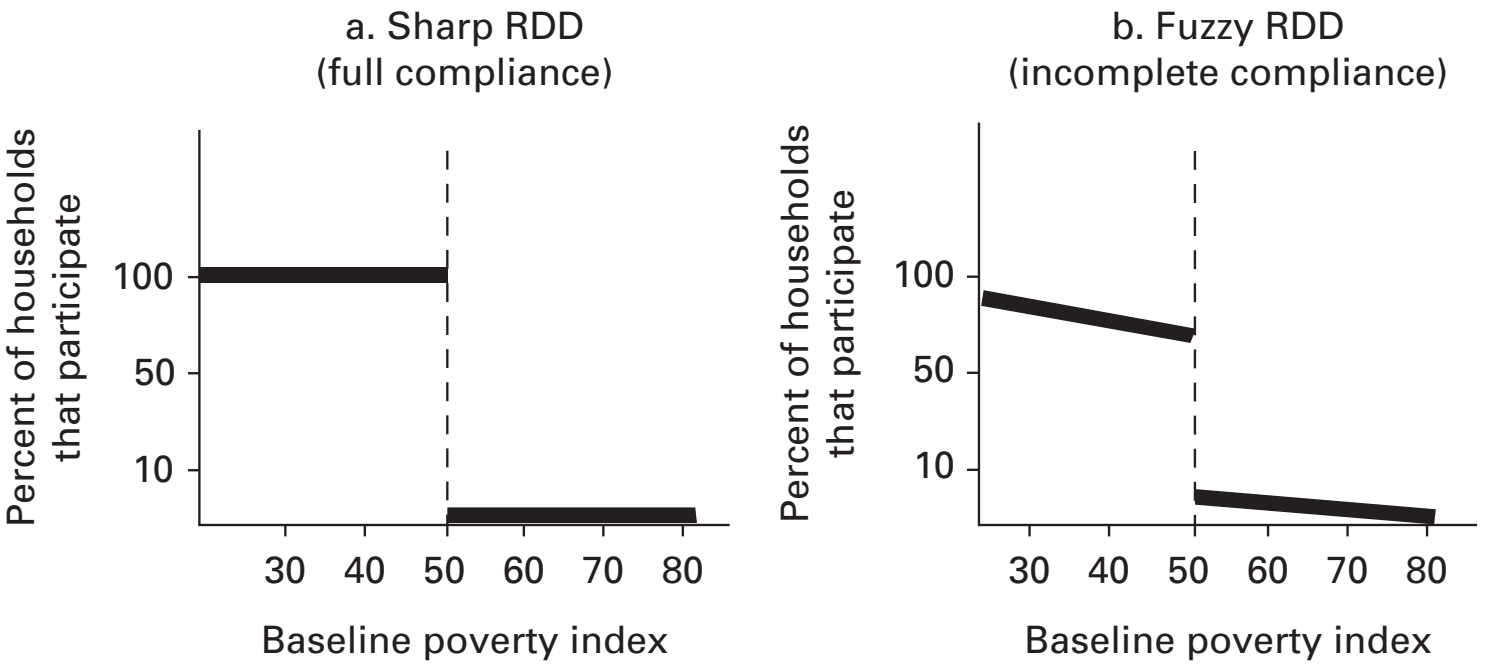
– Unemployment benefits: Workers above a certain age receive unemployment benefits for a longer duration.

– University education: A certain university only admits applicants if they score above a certain threshold on their standardized math test.

– Structural funds: A region/country gets support only if the GDP is below a certain threshold.

Regression Discontinuity Designs (2)

- For an anti-poverty program, two types of set-ups can be thought of:
  - Sharp Regression Discontinuity Design: Households below the threshold automatically receive tax deductions.
  - Fuzzy Regression Discontinuity Design: Households below the threshold are eligible for tax deductions but have to apply for it.



Source: Impact Evaluation in Practice

Regression Discontinuity Designs (3)

- Both sharp and fuzzy RDD make use of the discontinuity in the eligibility/assignment rule.
- Sharp RDD:
  - Compares average outcomes of units just below and just above the threshold.
  - The difference gives an estimate of the local average treatment effects of the program for people at the cut-off.
- Fuzzy RDD:
  - Uses the eligibility rule as an IV for treatment receipt.
  - Resulting estimates are a LATE for compliers at the cut-off!

RDD: Pros, Cons, Pitfalls and Requirements

- Pros and Cons:
  - (+) Intuitive method.
  - (+) Often applicable.
  - (+) Credible estimates.
  - (−) Provides only local effect estimates (for compliers) at the cut-off.
- Pitfalls:
  - RDD estimates fail if the same eligibility cut-off is used for different programs.
  - Sometimes, there is manipulation around the cut-off if individuals have control over the relevant index used for assignment.
- Requirements:
  - The program must have a specific cut-off based on an index of observed characteristic(s).
  - The evaluators' measure of the index must be precise.

Example RDD - EU Structural Funds and Growth (1)

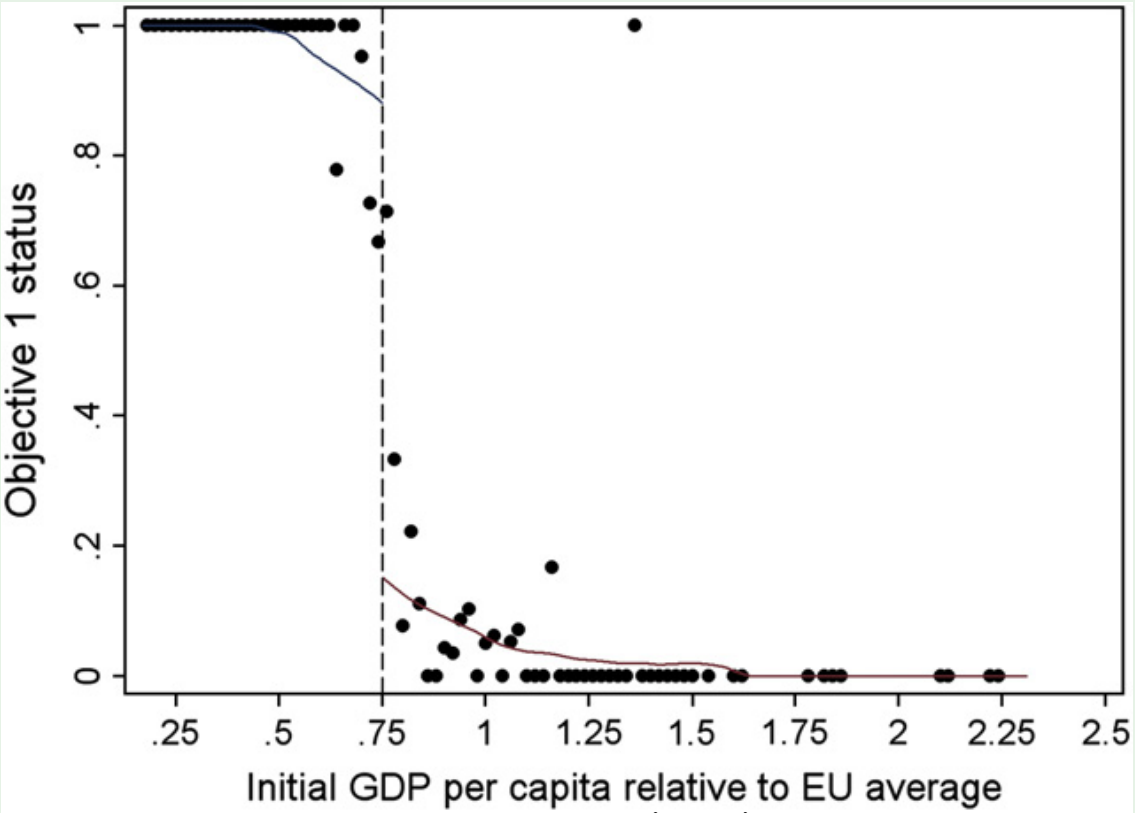
Becker/Egger/Ehrlich (2010): Going NUTS: The effect of EU Structural Funds on regional performance

- Research Question: What are the effects of receiving structural funds transfers on GDP and employment growth for disadvantaged regions?
- Treatment: Receipt of Objective 1 transfers to enhance GDP per capita growth in poorer regions.
- Source of selection bias: Poorer regions may be less populated and have less educated workers.
- Data: Aggregate data (NUTS-2 and NUTS-3 level) from Cambridge Econometrics' Regional Database and the European Commission, years 1989-2006.
- Method:
  - Regions are eligible to receive Objective 1 transfers if their GDP per capita in Purchasing Power Parities is less than 75% of the EU average.
  - Use this eligibility rule as an IV for transfer receipt (fuzzy RDD).

Example RDD - EU Structural Funds and Growth (2)

Becker/Egger/Ehrlich (2010): Going NUTS: The effect of EU Structural Funds on regional performance

- Jump in treatment probability at the cut-off:

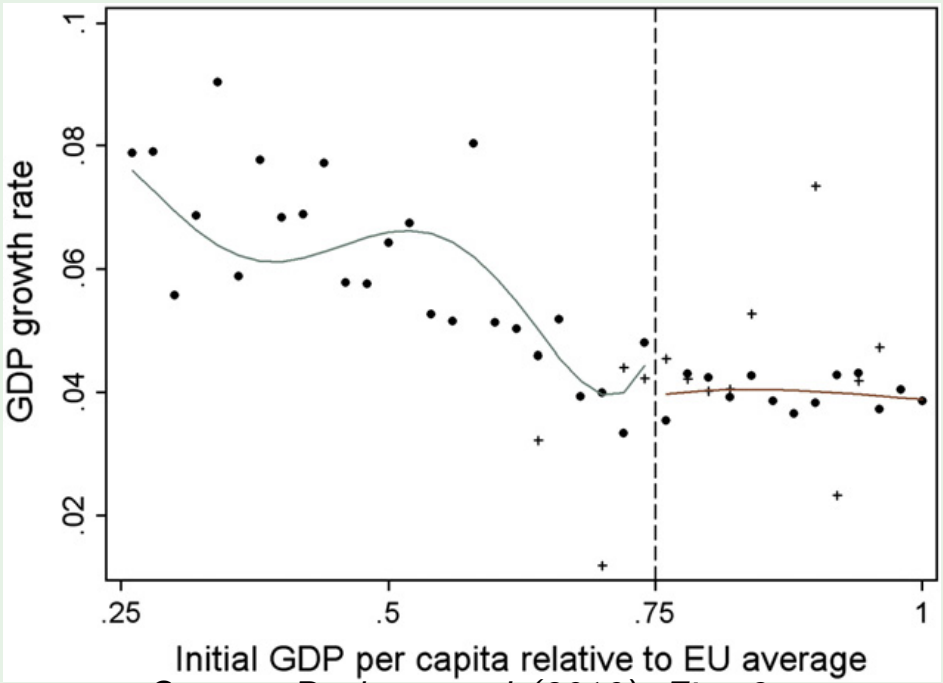


Source: Becker et. al (2010), Fig. 2

Example RDD - EU Structural Funds and Growth (3)

Becker/Egger/Ehrlich (2010): Going NUTS: The effect of EU Structural Funds on regional performance

- Jump in GDP growth at the cut-off:



Source: Becker et. al (2010), Fig. 3

- Results: Transfer receipt significantly increased GDP growth for compliers at the cut off.

Let us summarize . . .

Summary of Counterfactual Methods (1)

- Randomised Controlled Trials ...
  - ... solve the selection problem by randomly assigning willing individuals into treatment and control group.
- (Propensity Score) Matching ...
  - ... mimics RCTs by balancing observed characteristics through picking statistical twins as comparison individuals.
- Difference-in-Differences ...
  - ... differences out time-constant selection bias due to unobserved characteristics.
- The Synthetic Control Method ...
  - ... constructs a synthetic control unit by reweighing control units so that they look like the treated unit before the treatment took place.

Summary of Counterfactual Methods (2)

- Instrumental Variables ...
  - ... use exogenous variation in the selection process and compare outcomes of individuals whose treatment decision depends on the value of the instrument.
- Sharp RDD
  - ... exploits assignment mechanisms based on a cut-off rule for some observed characteristic and compares outcomes of individuals just below and just above the cut off.
- Fuzzy RDD ...
  - ... exploits eligibility cut-offs as IV for participation.

Evidence-Based Policy Making

- What empirical evidence should be used when deciding if and how to implement a certain policy?
- Evidence Hierarchy (Leigh, 2009):
  1. Systematic Review of Multiple RCTs
  2. High-Quality RCT
  3. Systematic Review of Multiple Non-experimental studies
  4. Non/Quasi-Experimental studies (Matching, DiD, SCM, IV, RDD)
  5. ...
  6. Before-After Comparison
- Systematic reviews of multiple RCTs/Non-experimental studies can increase external validity.

Conclusions

- "You can do anything. But you can't do everything and you certainly can't do everything at once."
- Quasi- and non-experimental methods to infer the missing counterfactual are well-established but data hungry.
  - In an ideal world, the evaluator is already involved at the design stage for a certain policy and process evaluation can guide the following impact evaluation.
  - There is no magic bullet! Each estimator relies on some identifying assumptions, none of which will always hold (even RCTs can fail).
  - Looking at effect heterogneity/mechanisms helps with improving future programs.
- Important: Better data helps a lot! The combination of different data sources (e.g. administrative and survey data) can be helpful in many situations (but may also take time)!

Additional References:

- Leigh, A. (2009): What evidence should social policymakers use? *Australian Treasury Economic Roundup*, Vol. 1, pp. 27 – 43.
- List, J. A. (2011): Why economists should conduct field experiments and 14 tips for pulling one off. *Journal of Economic perspectives*, Vol. 25, No. 3, pp. 3 –16.

Examples:

- Bauernschuster, S. (2013). Dismissal protection and small firms' hirings: Evidence from a policy reform. *Small Business Economics*, 40(2), 293-307.
- Becker, S. O., Egger, P. H., and Von Ehrlich, M. (2010): Going NUTS: The effect of EU Structural Funds on regional performance, *Journal of Public Economics*, Vol. 94, No. 9 – 10, pp. 578-590.
- Caliendo, M., Künn, S., and Weißenberger, M. (2016): Personality traits and the evaluation of start-up subsidies. *European Economic Review*, Vol. 86, pp. 87 – 108.
- Caliendo, M., and Künn, S. (2011). Start-up subsidies for the unemployed: Long-term evidence and effect heterogeneity. *Journal of Public Economics*,Vol. 95, No. 3-4, pp. 311 – 331.
- Card, D., and Krueger, A. B. (1993): Minimum wages and employment: A case study of the fast food industry in New Jersey and Pennsylvania, *American Economic Review*, Vol. 84, No. 4, pp. 772 –793.
- Castillo, V., Garone, L. F., Maffioli, A., and Salazar, L. (2017): The causal effects of regional industrial policies on employment: A synthetic control approach, *Regional Science and Urban Economics*, Vol. 67, pp. 25-41.
- Schultz, T. P. (2004): School subsidies for the poor: evaluating the Mexican Progresa poverty program. *Journal of development Economics*, Vol. 74, No. 1, pp. 199 – 250.
- Stockinger, Bastian (2017): The effect of broadband internet on establishments' employment growth: evidence from Germany. (IAB-Discussion Paper, 19/2017), Nürnberg, 54 S.