unicef | for every child

Implementing Causal Machine Learning in 🇰🇿

Online Workshop 2: Overview of research design & corresponding classical estimators

August / September / October 2024

Michael Lechner
Professor of Econometrics | University of St. Gallen | Switzerland

Lechner
Swiss Journal of Economics and Statistics (2023) 159:8
https://doi.org/10.1186/s41937-023-00113-y

Swiss Journal of
Economics and Statistics

CONFERENCE KEY NOTE Open Access

Causal Machine Learning and its use for public policy

Michael Lechner[1*]

The workshop series | 1

4 online workshops

- August 27, 13-15: Correlation & causation
- September 4, 13-15: Available identification strategies & classical estimation
- September 10, 13-15: Machine Learning
- September 11, 13-14: The data for the Astana workshop & useful descriptive statistics

Lechner
Swiss Journal of Economics and Statistics (2023) 159:8
https://doi.org/10.1186/s41937-023-00113-y

Swiss Journal of
Economics and Statistics

CONFERENCE KEY NOTE Open Access

Causal Machine Learning and its use for public policy

Michael Lechner[1*]

The workshop series | 2

The Astana Workshop September 30 to October 4, 2024 (10-13, 14:30-17:30)

- Monday
  - Morning: Identification with experiments & selection on observables
  - Afternoon: Discussion of potential programmes to be evaluated
- Tuesday: Causal Machine Learning (theory)
- Wednesday
  - Morning: 2 empirical examples
  - Afternoon: The mcf package – how to use it & how to interpret the results
- Thursday: Doing an empirical study in groups with the data introduced in Online Workshop 4
- Friday: Discussion of programmes to be evaluated continued (core team only)

Quick introduction

Participants

- Professional background?
- Knowledge in the estimation of causal effects, machine learning, Python?

Myself

- Professor of Econometrics at the University of St. Gallen
- Co-head of The Swiss Institute for Empirical Economic Research at the University of St. Gallen
- Empirical Economic Research | SEW-HSG | University of St.Gallen (unisg.ch)

  www.michael-lechner.eu
- Research interest in Causal Machine Learning, AI, programme evaluation, …

Plan for today’s workshop: **Research Designs**

Experiments

Unconfoundedness

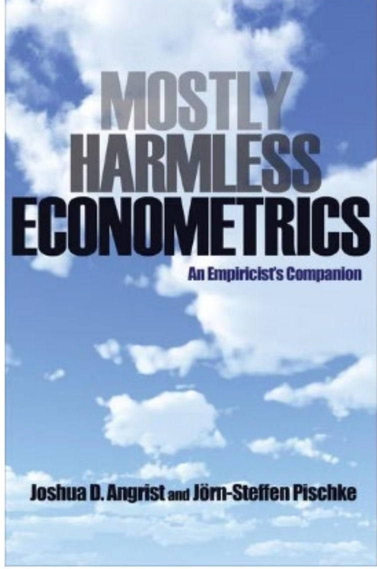
Instrumental variables

Difference-in-differences

Regression-discontinuity design

Synthetic control design

Recommended reading



MOSTLY HARMLESS ECONOMETRICS

An Empiricist's Companion

Joshua D. Angrist and Jörn-Steffen Pischke

ANNUAL REVIEWS

Annual Review of Statistics and Its Application

Causal Inference in the Social Sciences

Guido W. Imbens

Department of Economics and Graduate School of Business, Stanford University, Stanford, California, USA; email: imbens@stanford.edu

Nice, not very technical survey that starts from the basics of causal analysis. Focus mainly on section 1, 2, and the introductory parts of sections 3, 4, 5, 5.1, 5.2, 5.3). Just stop reading when it becomes too technical (if not, read all of it).

1 | Introduction

2 | Experiments

3 | Unconfoundedness

4 | Instrumental variables

5 | Differences-in-differences

6 | Regression-discontinuity design (RDD)

7 | Synthetic control design

8 | Conclusions

Example (run in one version at University of SG)

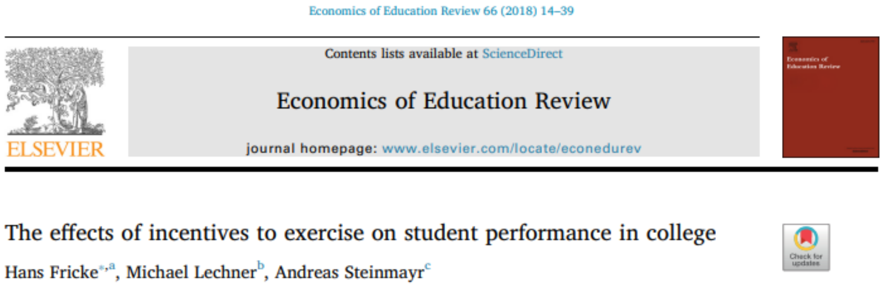
What is the effect of using university sports facilities on grades of students?

Treatment (*D*): Participating in University Sports (US)

Outcome (*Y*): Grade at the end of the year

Possible confounding

- Better students tend to be more physically active (selection bias)
  - Documented in the literature



Economics of Education Review 66 (2018) 14–39

Contents lists available at ScienceDirect

ELSEVIER

Economics of Education Review

journal homepage: www.elsevier.com/locate/econedurev

The effects of incentives to exercise on student performance in college

Hans Fricke[∗,a], Michael Lechner[b], Andreas Steinmayr[c]

Different research designs can be motivated by different selection processes

Experiments | Example | 1

Suppose that some students are randomly forced to participate in US, while all the others are prevented from using US

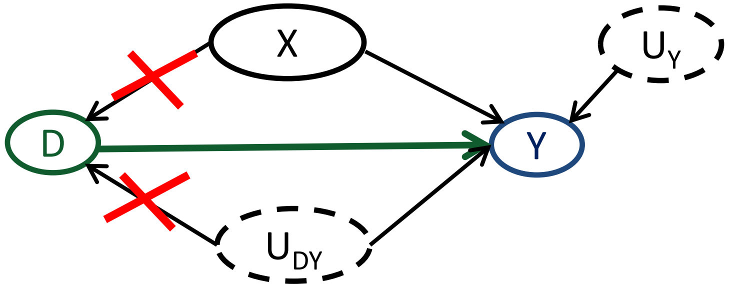
- Of course, this is not what happened in our study

Experiments | Identification

Assumption

- Treatment is randomly assigned

Unconfoundedness | The corresponding causal graphs



Unconfoundedness | Implication of identifying assumption

Any potentially confounding factors have the same distribution for participants & non-participants

Thus, unadjusted comparisons of the outcomes of participants with the outcomes of non-participants reveal causal effects

Experiments | Estimation

Mean outcome of participants – mean outcome of non-participants

- $ATE = ATET = E(Y|D=1) - E(Y|D=0)$

More fancy estimators (e.g. linear regression) may be used only to …

- potentially gain precision
- estimate effect heterogeneity

Potential role of Causal Machine Learning

- Might be more effective in gaining efficiency & estimating heterogeneity due to a relaxation of functional form assumptions of classical regression estimators

Experiments | Advantages & disadvantages in practise

Advantages

- If experiments were implemented cleanly, causal effects have very high credibility
- Usual causal parameters of interest are identified

Disadvantages

- Not every variation in *D* can be generated by a reasonable experiment
- Experiments may be expensive
- … may take long (waiting time from starting treatment until measuring outcomes)
- … may be messed up by administrators or subjects not obeying the rules
- Results may be not externally valid
  - Internal vs external validity

1 | Introduction

2 | Experiments

3 | Unconfoundedness

4 | Instrumental variables

5 | Differences-in-differences

6 | Regression-discontinuity design (RDD)

7 | Synthetic control design

8 | Conclusions

Unconfoundedness | Example

Use rich observational (i.e. non-experimental) data for empirical analysis

Get access to additional data (*X*) about students

- Past grades, entry examination results, past sports activity, measure about motivation & ability
- These data are not influenced by current US participation

unicef for every child

Unconfoundedness | Identification | Assumption 1

Potential outcomes are conditionally independent of treatment for any given values of the confounding variables

- Need data that captures all factors jointly influencing potential outcomes & selection into *D*
  - Identifying such data requires substantial knowledge about assignment process
- Quasi-experimental interpretation: Stratified experiment
  - *D* is as good as randomly assigned for units with the same features (*X*)

➔ comparisons of outcomes of treated & controls with the same value of *X* are causally valid

Unconfoundedness | Identification | Assumption 2

For any given value of the confounding variables, a unit could be observed with *D=1* or *D=0* (***common support***)

- After controlling for all confounders, there must still be some randomness in participation
  - Otherwise, it is impossible to compare treated & non-treated with same *X*
  - This implies that participation rule must not be deterministic
- If some subpopulations have deterministic participation rules → must be removed from analysis
  - Data is not informative about causal effects for these subpopulations

Unconfoundedness | Identification | Assumption 3

The confounding variables are not influenced by the treatment in a way that is related to the outcome variables (***exogeneity*** of confounders)

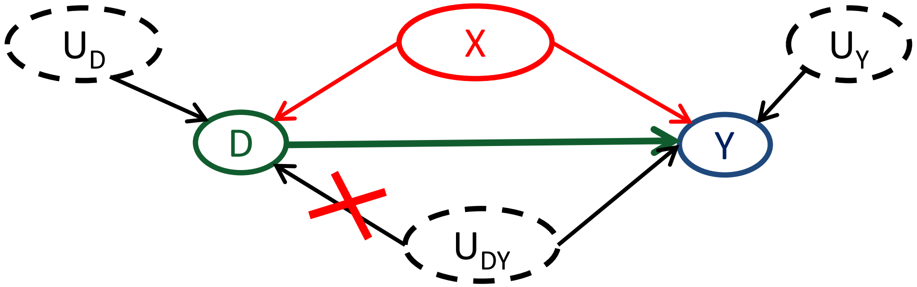
Unconfoundedness | Identification | Assumption 4

The observed outcomes in one treatment state correspond to the potential outcomes of that treatment state for the participants in that state (stable unit treatment value assumption, ***SUTVA***)

- $Y = D\,Y(1) + (1-D)\,Y(0)$
- Participation of others must not affect own potential outcomes
- If SUTVA is violated, *Y(1), Y(0)* is no longer a valid notation to capture causal effects

Unconfoundedness | The corresponding causal graphs

Compared to experiment: *X* is a *confounder,* i.e., *X* influences *Y* **&** *D*



Unconfoundedness | Implication of identifying assumption

Potential confounders may be distributed differently for participants & non-participants

Thus, unadjusted comparisons of outcomes of participants & outcomes of non-participants do not reveal causal effects

Experiment-like comparisons for units with the same values of *X* are causally valid

Unconfoundedness | Estimation | 1

All estimators must do the following (implicitly or explicitly)

- Estimate causal effects for all different observed values of *X* → aggregate them to obtain ATE
  - Matching estimators do this explicitly
- Estimate weights that would make the distribution of the confounders among treated & non-treated identical → use these weights for weighted mean comparison of the outcomes of treated & non-treated
  - Special case 1: Methods that remove the effects of other variables (*X*) (e.g. linear regressions) $Y = D\alpha + X\beta$
  - Special case 2: (1) Weight outcomes of treated by estimated *P(D=1|X)* [=: propensity score]

    (2) Weight outcomes of non-treated by *1-P(D=1|X)*

    (3) Mean of (1) minus mean of (2)

Unconfoundedness | Estimation | 2

The value of Causal Machine Learning

- Avoid additional assumptions in the estimation steps (e.g., for regressions, propensity scores)
  - As would be required by classical regression-type & weighting type estimators
- More powerful in estimating heterogeneities

Unconfoundedness | Advantages & disadvantages in practise

Advantages

- Credibility could be high
- Usual causal parameters of interest are identified

Disadvantages

- Substantial knowledge about assignment process is needed to identify relevant confounders
- Data hungry strategy (many features - *X*)
- More fancy estimators needed to perform confounder adjustments
  - Loss of precision compared to experiments → more observations needed (*N* larger)

1 | Introduction

2 | Experiments

3 | Unconfoundedness

4 | Instrumental variables

5 | Differences-in-differences

6 | Regression-discontinuity design (RDD)

7 | Synthetic control design

8 | Conclusions

Instrumental Variables | Example

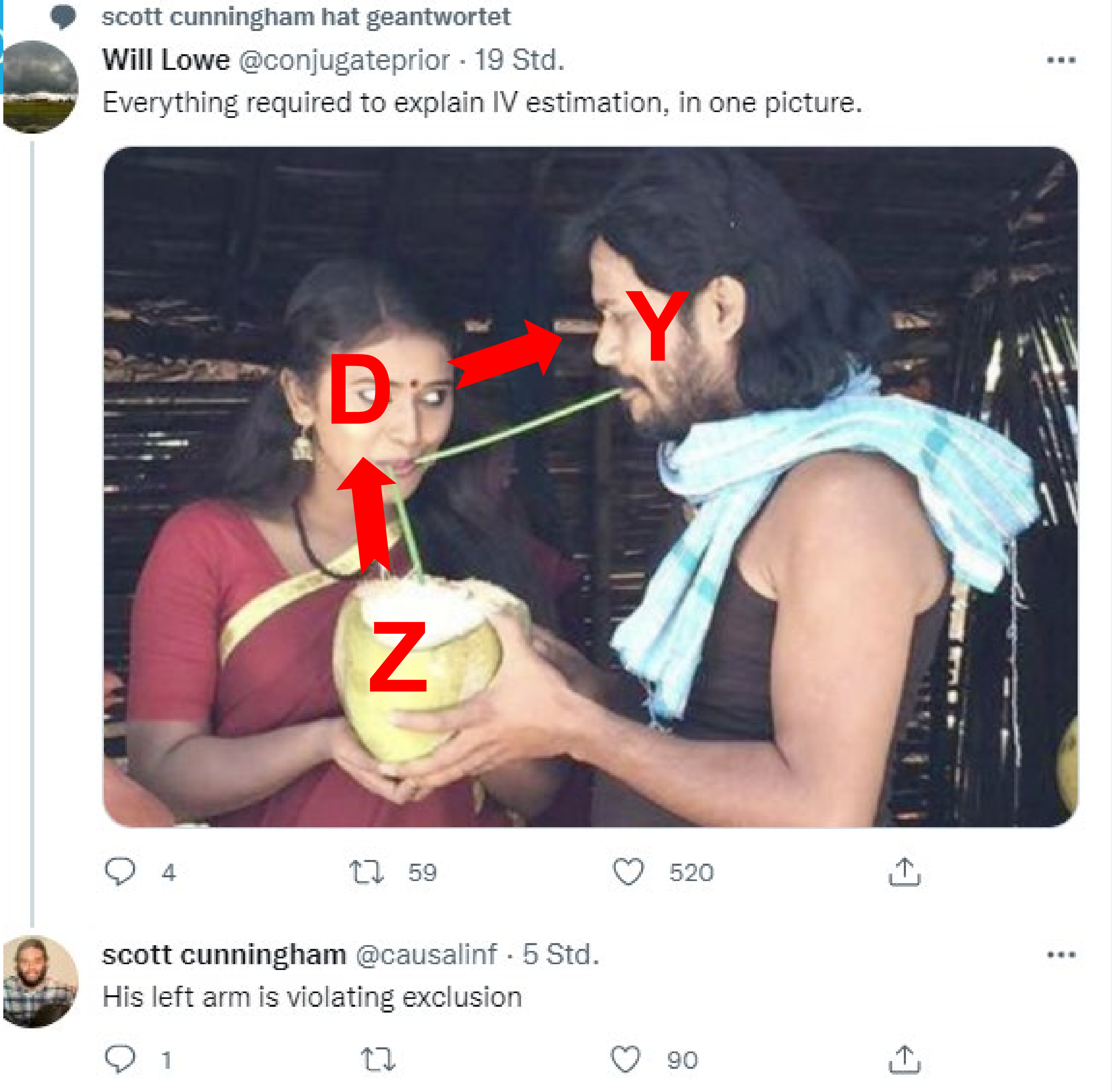
Instead of randomizing access to US, an incentive to use US is randomized

- The outcome of this randomization process is called an instrument (*Z*)
  - In our paper, we randomized a financial incentive to participate in US (*D*)
- Note
  - *Z* does only influence the outcome *Y*, because it influences *D*
  - If students do not care about this financial incentive, their observations will not be useful in any empirical analysis

Many famous examples in the econometric literature

- E.g. Angrist & Evans (AER, 1998): Same sex siblings unicef | for every c

Such pictures appear, when econometricians having fun on Twitter 😂😂



scott cunningham hat geantwortet

Will Lowe @conjugateprior · 19 Std.

Everything required to explain IV estimation, in one picture.

4 59 520

scott cunningham @causalinf · 5 Std.

His left arm is violating exclusion

1 90

d, Dec, 14, 2021

Why do instruments work? | 1

The starting point of IV estimation: The *Wald* estimator

Assume simplest case: *Z* (instrument), *D* (treatment) are binary

The *Wald* estimator solves the following population problem (for binary *Z*):

$$LATE = \frac{E(Y \mid Z = 1) - E(Y \mid Z = 0)}{E(D \mid Z = 1) - E(D \mid Z = 0)}$$

This parameter can be estimated consistently by substituting averages for expectations.

IV is an *indirect* estimation approach.

$$\beta = \frac{E(Y \mid Z=1) - E(Y \mid Z=0)}{E(D \mid Z=1) - E(D \mid Z=0)}$$

Why do instruments work? | 2

Why does this expression make sense?

- Suppose that numerator (denominator) can be interpreted as the causal effects of *Z* on *Y* (*Z* on *D)*
- Suppose there is *only* a causal effect of *Z* on *Y because Z changes D*
- This resulting causal effect on *D* changes *Y*
  - There is no other 'channel' through which *Z* can affect *D*
  - Indirect estimate
- Example: *Z increases D by 10%-points on average*
  - *But we are looking for the effect of a 100%-points change of D (0 → 1)*
  - If there is no direct effect of *Z* on *Y*, then the effect of *Z* on *Y* should be multiplied by 10 to account for a 100%-points change in *D*! (=divide by 0.1)

Instrumental Variables | Assumptions

Exclusion restriction

- *Z* influences *Y* only because it influences *D* (& *D* influences *Y*)
  - This also requires that the relation between *Z* & *Y* as well *D* & *Y* is unconfounded

Relevance of instrument

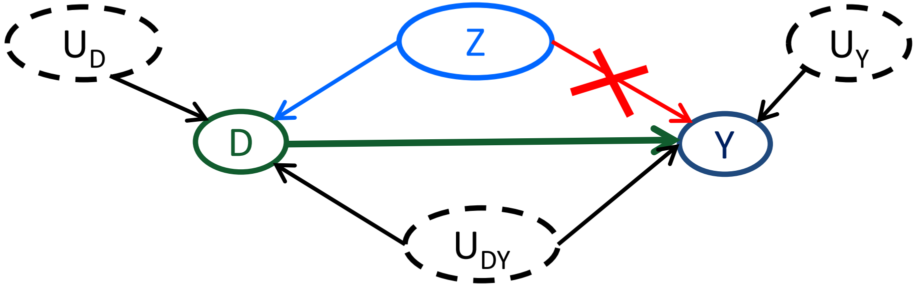
- *Z* does influence *D* at least for some units

Monotonicity

- If *Z* influences *D*, this influence goes in the same direction for all units that are influenced
  - It is ok when a change in *Z* increases *D* for some units, & does not change *D* for others
  - *It is not ok* when a change in *if Z increases D for some, & decreases D for others*

Instrumental variables | The corresponding causal graphs

Compared to unconfoundedness: Unobserved confounder (U_{DY}) allowed



Instrumental Variables | Advantages & disadvantages in practise

Advantages

- A way to deal with confounders that are **un**observable

Disadvantages

- Effect is only identified for subpopulation that reacts to a change in *Z* with a change in *D* (compliers): Local average treatment effect (LATE) ➔ ATE etc. is usually not identified
- Numerical issues (weak instrument problem) when effect of *Z* on *D* is too small
  - Because of division by this effect

1 | Introduction

2 | Experiments

3 | Unconfoundedness

4 | Instrumental variables

5 | Differences-in-differences

6 | Regression-discontinuity design (RDD)

7 | Synthetic control design

8 | Conclusions

Introduction

Controlling for observable covariates may not be sufficient to control for confounding

Data has time dimension

Time dimension of data can be used for identification if …

- … impact of confounder is constant over time
- … some functional form assumptions hold

Difference-in-differences: Introduction

Remove confounding by differencing outcome over time

- Time constant confounders with additive time constant effects drop out

Panel data or *repeated cross-sections* are needed

Popular applications analyse *changes in some law* if there exists a group not affected by these changes

Key identifying assumption: Potential outcomes of groups with different levels of *D* are subject to the same time trends (*common trend assumption*)

DiD: Example-2, Card & Krueger (AER, 1994)

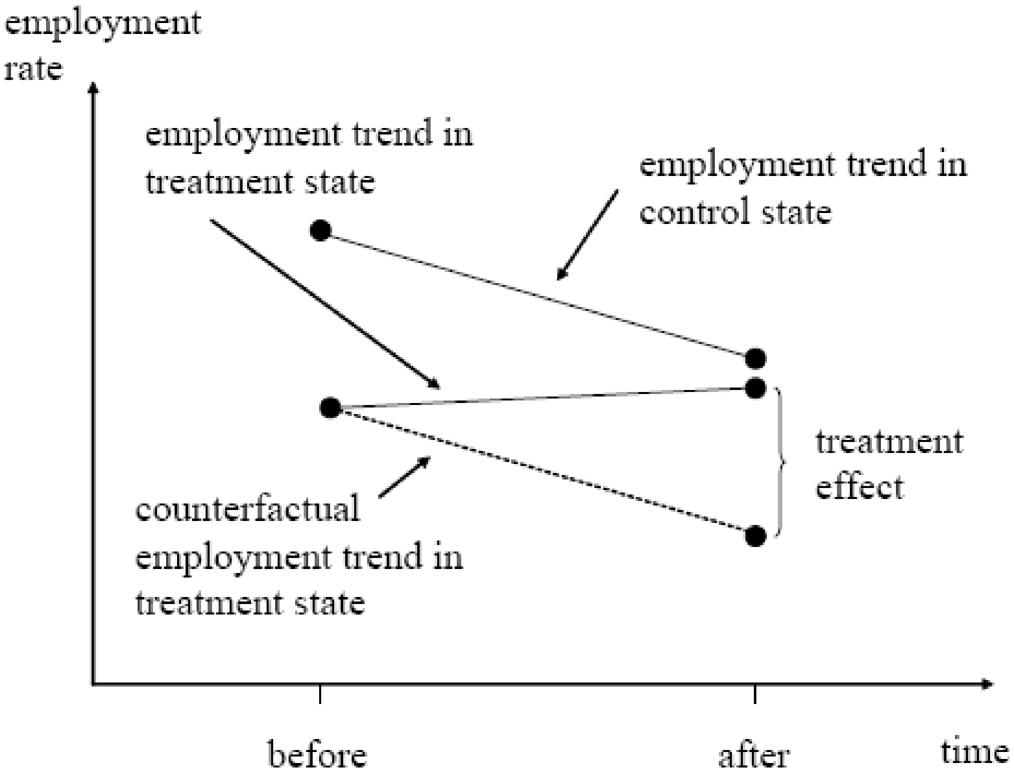
Goal: Understand employment effects of a change in the minimum wages

Basic ideas of the CK study

- Exploit that different US states have different minimum wages that change over time
- Comparing employment levels before & after a minimum wage change may be confounded by the business cycle
- Comparing states with low & high minimum wages may be confounded by other characteristics of the local economy (like sectoral skill distribution)
- Find a state that raised its minimum wage (New Jersey) & compare development of employment to a state that did not raise the minimum wage (Pennsylvania)
- But: The effect on the different sectors of the economy should be very different, because minimum wage regulations affect only low skilled workers
- Concentrate on one sector which has a high share of minimum wages workers
  - fast food restaurants

DiD: Basic ideas | 2

Causal effects in the difference-in-differences model


employment rate
employment trend in treatment state
employment trend in control state
treatment effect
counterfactual employment trend in treatment state
before
after
time

Source: Card & Krueger (AER, 1994)

Differences-In-Differences | Identification

Key assumptions

- Common trend
  - The change of $Y_t(0)$ over time is the same for treated & non-treated
    - The same as assuming that the relation of $(Y_t(0)- Y_{t-1}(0))$ & D is unconfounded
- No anticipation
  - There is a period without treatment for everybody (& future treatment is not anticipated)

Differences-In-Differences | Estimation

Without covariates & 2 periods, a linear regression with *Y* as dependent variables & the following independent variables

- Indicator variable for period 2
- Indicator variable for belonging to the treatment group
- Interaction of the period & treatment group indicator

The coefficient of the interaction term corresponds to the ATET.

unicef for every child

Differences-In-Differences | Advantages & disadvantages

Advantages

- Easy to estimate
- Allows unobserved confounding as long as it is time constant

Disadvantage

- ATE is not identified (only ATET)
- Common trend may not hold
- Subtracting quantities makes ‘common trend’ assumption measurement dependent
  - E.g. the logarithm of a difference does equal the difference of logarithm
    - A common trend in level does not the imply a common trend in logs of the same variables

1 | Introduction

2 | Experiments

3 | Unconfoundedness

4 | Instrumental variables

5 | Differences-in-differences

6 | Regression-discontinuity design (RDD)

7 | Synthetic control design

8 | Conclusions

Regression Discontinuity Design | Example

Sports example again …

Suppose, due to capacity constraints, access to US is restricted to the fittest

- Fitness is determined by a prior fitness test (with fitness levels 0-100)
- University decides to give access only to students that have fitness levels of 70+

Regression Discontinuity Design | Identification in example

Comparing grades of participants with grades of non-participants is confounded by fitness levels

- Overestimation of the effects (in this example)

Adjusting for this confounder is impossible, because there is no common support

Idea of RDD

- Students with fitness level 69 (not allowed to participate) are very similar to those with fitness level 70 (allowed to participate)
- Being allocated just above, or just below the cut-off is **as good as random**
- Thus, use comparison of outcomes of observations local to the cut-off
  - Mean of *Y* just above cut-off *minus* mean of *Y* just below cut-off

Regression **D**iscontinuity **D**esign | Advantages & disadvantages ...

Advantage

- Credible research design derived directly from participation rule
- Allows for unobserved confounding if stable around the cut-off

Disadvantage

- Valid only around cut-off
- In some settings, it is possible for units to influence whether own position is below or above cut-off
  - Additional confounding that invalidates RDD design
- Only causal effect of population located around the cut-off is identified ➔ ATE is not identified

1 | Introduction

2 | Experiments

3 | Unconfoundedness

4 | Instrumental variables

5 | Differences-in-differences

6 | Regression-discontinuity design (RDD)

7 | Synthetic control design

8 | Conclusions

Synthetic control design | Example

Suppose that …

- everybody at the University of St. Gallen participates in US
  - No group of non-participants available
- there are 5 other Swiss university that have no US facilities at all

Data available on past grades including the time when Uni St. Gallen had no US

Synthetic control design | Example

Possible approach

- Compare average of grades of other university (non-participants) to grades of University of St. Gallen
  - This comparison will be confounded by different student quality at different university

Synthetic control

- Use data from period before Uni St. Gallen introduced US
- Compute weights such that weighted mean of grades in other university are similar to those in St. Gallen BEFORE Uni St. Gallen had US facilities
  - Creates a synthetic control observation as weighted mean of the other 5 universities
- Apply these weight to current grades to estimate what would have happened to St. Gallen students had they no access to US

Synthetic control design | Advantages & disadvantages …

Advantage

- Allows for unobserved confounding (to some extent)
- Allows for cases with very, very few treated (only 1 ok)

Disadvantage

- Sometimes difficult to argue how & why past similarity of weighted control group translates into future similarity (time stability)

1 | Introduction

2 | Experiments

3 | Unconfoundedness

4 | Instrumental variables

5 | Differences-in-differences

6 | Regression-discontinuity design (RDD)

7 | Synthetic control

8 | Conclusions

Conclusion

Each research design depends on the validity of untestable assumptions

- Their validity must be argued based on knowledge about the assignment process

Only experiments & unconfoundedness identify causal effects for full population

- Most interest in CML focusses on these two methods
- If identification is achieved under these 2 methods → heterogeneity results can be used for algorithmic decision making

Causal machine learning

- does not help with identification
- helps with estimation of the effects (& improving assignments)

Next week: Some basics of Machine Learning

Michael Lechner

Swiss Institute for Empirical Economic Research (SEW)
University of St. Gallen | Switzerland