Statistical Modeling and Causal Inference with R

Week 1: Introduction

Manuel Bosancianu

Max Schaub

September 7, 2020

Hertie School of Governance

Today's focus

- ✓ Welcome and introductions
- ✓ Course logistics
- ✓ Why study causal inference?

Welcome

Thank you for joining us in the course!

The team (I)

Manuel

- Research fellow at the WZB Berlin Social Science Center
- Political Economy of Development
- Interests: political inequality, economic inequality, Leftist parties, data visualization
- Methods: multilevel modeling, Bayesian analysis, experimental (recent), causal inference

Max

- Research fellow at the WZB Berlin Social Science Center
- Migration and Diversity
- Interests: migration, violence, poverty, political participation
- Methods: causal inference, survey design, field and lab-in-the-field experiments

What’s your profile?

How many...graduated a quantitative-focused program in economics, political science, etc?

...are familiar with Stata (wrote the analysis for a paper with it)?

...are familiar with R (had a class with it)?

...can name the assumptions needed to make OLS a BLUE?

The course: goals

- ✓ To train your ability to look at research designs from a causal perspective
- ✓ To increase your discernment regarding the causal foundations of empirical analyses
- ✓ To provide you with a standard toolbox of approaches to causal inference
- ✓ To accustom you with how these are implemented in *R*

The course: topics

1. Potential outcomes framework & Causal graphs
2. IV estimation
3. Matching
4. Regression discontinuity designs
5. Difference-in-Difference designs & Synthetic controls
6. Panel data & Fixed Effects
7. Moderation & Mediation
8. Field experiments

A “fox”, rather than a “hedgehog” (Sir Isaiah Berlin).

The course: readings

Three tiers:

1. **required**: copies of *Mastering 'Metrics* have been bought and put on reserve by the library
2. **optional**: if you're particularly interested in the topic
3. **applied**: practical example for method (focus only on underlined one in syllabus)

All required readings, and one of the applied ones, are to be done before Monday's sessions.

Logistics

The class setting

You will have lectures in video format available on Friday the week before our Monday meeting.

In class, we hope you will be the stronger force in pushing the discussion:

- ✓ If readings or video lecture weren't clear, please ask questions!
- ✓ If tempo is too rapid (or slow), please let us know!

Many opportunities to re-visit material: Monday sessions & initial part of labs.

Attendance

You've been allocated to 3 groups, and switching is generally **not possible**. Let us know in advance if you cannot join your group, and want to switch.

12 sessions in total, meaning **you cannot skip more than 2**.

If you cannot make it at all (e.g. health reasons, personal emergencies), please let the Examination Office know beforehand—they will inform us.

We have to take attendance, partly for public health reasons.

Other points of contact

Labs (drop-in) sessions are not mandatory, but highly encouraged—lots of *R* practice.

Labs also allow one more opportunity to ask questions about lecture material.

Finally, we offer office hours, via Zoom or Teams video call:

- ✓ Max: Thursdays, 14–16
- ✓ Manuel: Thursdays, 10–12

Please send us an email in advance with what you want to discuss (in case we need to prepare a bit beforehand).

Bi-weekly Assignments

Problem sets that combine conceptual and applied tasks.

Collaboration is encouraged while learning, but bi-weekly assignments are individual work.

Assignment “live”	**Deadline** (always 11:59 PM CET)
14.09	23.09
28.09	07.10
12.10	28.10
02.11	11.11
16.11	25.11

Bi-weekly Assignments

Submit your answers via Moodle: (1) `.Rmd` file with answers, code, and output; (2) Knitted HTML file.

Final grade for this component is weighted average of 5 grades, with equal weights.

Each day of delay in submission results in a 10% drop in the grade.

Additional assessments

Type	Deadline	% of final grade
Bi-weekly assignment	See above	40%
Final exam	TBC	25%
Replication task	22.12 11:59 PM CET	35%

For all of these, submission is via the Moodle system.

Final exam is an at-home 120-minute test with open-ended questions, multiple choice questions, model output interpretation, graphs, and very simple calculations.

Additional assessments

Replication task based on set of recent published articles.

The challenge is to replicate the analyses in the article, and go beyond them, by exploring additional questions.

Final submission in paper (6–8 pages) and *Rmd* file. Short, but time consuming!

Labs (drop-in sessions)

Run entirely by **Adelaida Barrera** and **Sebastian Ramirez Ruiz**.

Voluntary attendance, but encouraged:

- ✓ Covering lecture content, if need be;
- ✓ Going over tasks similar to what assignments will cover
- ✓ Additional R practice

Health considerations

In-person teaching

Participants are asked to wear masks when moving to and from desks. Masks can be taken off at the desk.

Instructors will teach without mask, from fixed position.

Minimum: 1.5 meters (without mask), 1 meter (with mask).

If you have any concerns about this policy, please let us know early, so we can address them.

Attendance sheets will be used by Berlin health authorities to track movement, in case of infection.

Room use

Airing to take place at least once every 30 minutes.

Disinfectant spray and paper towels are available for use.

In case of any Covid-19-like symptoms, please don't come to class, and immediately notify Student Life office and Berlin health authorities (see Hertie **Hygiene Policy, v2**).

Instructors are subject to same rules in case of symptoms.

Chance to ask questions before we move to causality…

Why study causal inference?

Why study causal inference?

1. Overcoming flaws in traditional statistical methods
2. A new way of thinking about the world
3. Answer questions we care about – in a rigorous way

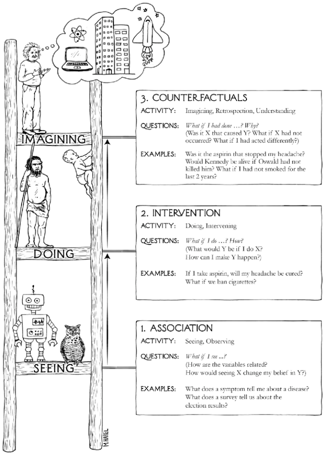
Questions we care about...

- ✓ Does micro-finance **affect** individuals' poverty status?
- ✓ Can personal experiences of extreme weather **shift** people's perception of climate change?
- ✓ Can we **demonstrate** that policing is racially biased?
- ✓ Can technological innovations **improve** political accountability and infant health?

"affect", "shift", "demonstrate", "improve" all imply we are interested in **causal relationships**

Overcoming flaws in traditional statistical methods

The ladder of abstraction (Pearl & Mackenzie, 2018)



Overcoming flaws in traditional statistical methods

- ✓ Tailored towards 'seeing' instead of 'doing' (Pearl & Mackenzie, 2018)
- ✓ Non-directionality: $y = a + bx \leftrightarrow x = (y - a)/b$
- ✓ Threat of confounding by third variables – and no guidance how to avoid it
- ✓ Lack of 'shoe-leather'/ dominance of out-of-the-box solutions (Freedman, 1991)
- ✓ "Correlation is not causation"
  ...but what, then, is causation?

A new way of thinking about the world

Counterfactual thinking and the potential outcomes framework

A new way of thinking about the world

Counterfactual thinking and the **potential outcomes framework**

Causal effect: $\Delta_i = Y_{i1} - Y_{i0}$

Answer questions we care about – in a rigorous way

Example: Effect of micro-finance on poverty status

- Imagine you were asked to assess the effect of receiving a micro-finance loan on income 5 years later?
- What would you do to find out?

Answer questions we care about – in a rigorous way

Conduct experiments/ randomized control trial (RCTs)

1. Intervention (e.g. access to micro-finance)
2. Randomly assign individuals to treatment and control
3. Compare outcomes (taking into consideration treatment compliance issues)

Answer questions we care about – in a rigorous way

Example (cont'd): Effect of micro-finance on poverty status

- ✓ Imagine you are given a dataset with data on individuals' incomes who received/did not receive loan
- ✓ Data shows that individuals who received a loan were doing better
- ✓ Can we conclude that micro-finance had a positive causal effect?

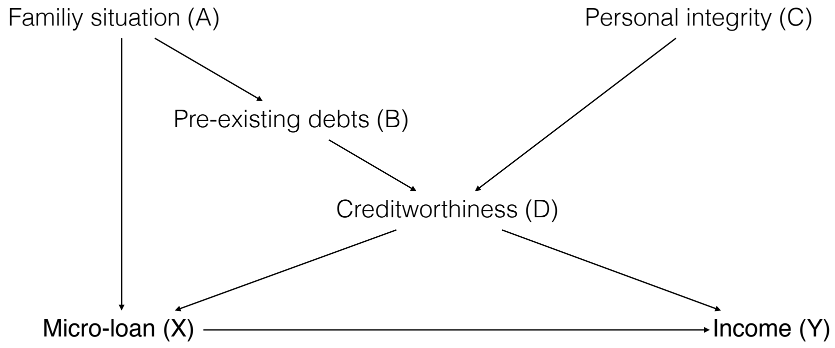
Answer questions we care about – in a rigorous way

Analyze observational data in terms of hypothetical experiments (Rubin, 2008)

1. What would the ideal experiment have looked like?
2. What was the process of treatment assignment, and how did it deviate from random?
3. Account for these deviations (if at all possible), and compare outcomes

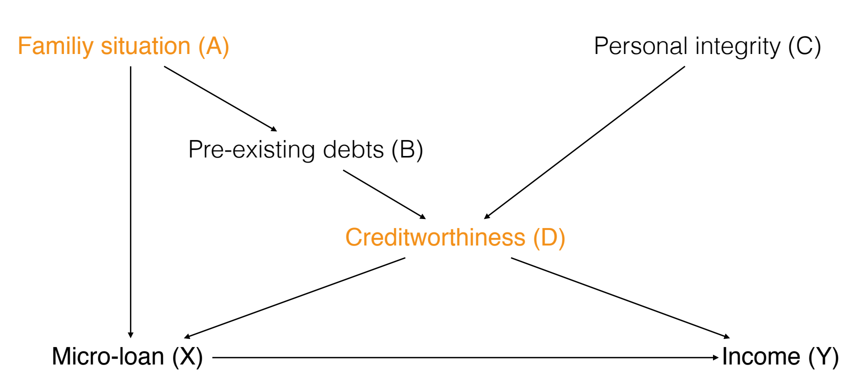
Answer questions we care about – in a rigorous way

Use causal diagrams/directed acyclic graphs (DAGs) (Pearl & Mackenzie, 2018)



Answer questions we care about – in a rigorous way

Use causal diagrams/directed acyclic graphs (DAGs) (Pearl & Mackenzie, 2018) – and try to shut backdoors to potential confounders



Thank you for the kind attention!

References

Freedman, D. A. (1991). Statistical Models and Shoe Leather. *Sociological Methodology*, *21*, 291–313.

Pearl, J., & Mackenzie, D. (2018). *The Book of Why: The New Science of Cause and Effect*.

Rubin, D. B. (2008). For Objective Causal Inference, Design Trumps Analysis. *The Annals of Applied Statistics*, *2*(3), 808–840.

Answer questions we care about – in a rigorous way

Conduct experiments/ randomized control trial (RCTs): *do* micro-loan