Lecture 2 - Simulations for regression analysis

Topics in Econometrics

Vincent Bagilet

2025-09-16

Housekeeping

- Tomorrow's class moved to next week
- Graded assignment 1 due next week
- Reading: I will post the paper online. Due next Wednesday
- Replication exercise: format TBA

Take-away points from last week

- Applied economics aim to produce **accurate causal estimates** (*eg* inform public policy)
- *Objective of the class*: discuss **practical issues** that may prevent us from doing so
- Can arise in any of the steps of research: **design**, **modeling** and **analysis**
- There are some fundamental **hurdles** to estimating causal effects
- **Simulations** can help spot and understand these hurdles

Order of concern

- There are different type of hurdles
- Each type only matters to the extent that the previous one are addressed
- We need to have, in that order of concern:

    1. A good **research question**, grounded in theory
    2. A good **identification strategy** to avoid some fundamental hurdles (reverse causality, confounders, etc)
    3. A specification that allow us to estimate the quantity we want to estimate
    4. Avoided econometric hurdles

Simulations for regression analysis

What, Why and How?

Lessons from last week's simulation

- What was the **idea behind** the implementation of simulations last week?
  - *Objective*: explore how several parameters affect the estimate of interest
  - *Approach*: Generate fake data (we thus know the whole DGP) and run an analysis
- Were they **useful?** If so, in what way?
  - Understand how various parameters affect the estimate of interest, without deriving the maths
  - Help to shape intuition and understanding

What is a simulation for regression analysis?

- A process in which we:
    1. Generate **artificial data**
    2. Run an analysis on this data
    3. Repeat the process many times
- We **know the data generating process**
- Can simulate data:
    - From scratch (**fake data simulation**) or
    - On top of an existing data set (**real data simulation**)

Overall principles

- Whole game in our metrics analyses: **approximate the DGP**
- With a simulation, we know the true DGP (at least to some extent)
- We can assess the performance of our analysis:
  - **Can we accurately estimate the true effect of interest?**
  - Are there hurdles to doing so and can we overcome them?

Why do simulations?

- To understand econometric concepts
- To design a study, before having the data
- To design a study, once having the data
- Tests and checks, after running the analysis
- As a rhetorical tool

Why do simulations?

To understand econometric concepts

- No maths required and allows to consider many general cases easily
- Useful to get intuition on how econometric aspects work
- Understand general concepts:
  - *eg* what happens in general when we omit a variable, or highlight issues with TWFE
  - Can explore this with naive fake-data simulations (*eg* $x \sim \mathcal{N}(0, 1)$)
- Understand conceptual hurdles specific to our context:
  - *eg* what happens if there is autocorrelation in *this* particular variable
  - Can explore this with calibrated fake-data or real-data simulations

The example of leverage and influence

- What did you learn from the exercise you had to do?
- What affects leverage? How does it affect the parameter of interest?
- Present the intuition behind leverage and influence
- How did you implement your simulation?
- Any cool graphs/outputs?

Why do simulations?

To design a study, before having the data

- Useful to get started on a concrete reflection about:
  - The setting
  - What we want to estimate, *exactly*
  - The data need and its granularity
  - The identification strategy
- As a proof of concept (to apply for grants, data access, etc)

Why do simulations?

To design a study, once having the data

- Useful to think about:
  - Threats to identification and important assumptions
  - The statistical power of our study (difficult to do without a simulation)
- Explore **where to best invest resources**:
  - Larger sample
  - Improved data precision (reduce measurement error)

Why do simulations?

Tests and checks, after running the analysis

- Does your analysis detects the effect you are interested in, in a **pristine setting**?
- If the analysis faces issues in simulations, it will probably also in an actual setting
- What happens to the product of our analysis if the setting is slightly more complex?
- What happens if some hypotheses do not hold?
- All this can be discussed **even after the analysis has been run**

Why do simulations?

As a rhetorical tool

- Simplify what we are working on to the bare minimum
- What is **the simplest way of pitching** the analysis and the identification strategy?
- Can help build simple visualizations
- Can be useful to **illustrate why a given approach does not work**
  - To argue why we chose a certain approach
  - In a referee report

General approach to simulations

- Start with a simple DGP:
  - Simple correlation structure
  - Our model represents the actual DGP
- Does our analysis recover the effect in a rather "pristine" setting?
- Then complexify the DGP

Steps of the simulation approach

1. Define a DGP and the distribution of variables
2. Set parameters values
3. Generate a data set
4. Estimate the effect in the generated data set
5. Repeat many times
6. Compute the measure of interest

Next steps

- **Change parameters values**
  - Understand how the measure of interest is affected by a given parameter
  - *eg* how does bias evolve with the correlation between x_1 and $x2$?
- **Complexify the DGP**
  - Would our method still performs well if the DGP was more and more complex?
- Repeat

Exercise

Simulating an RCT

Setting

- Impact of receiving extra lessons on students' grades
- Simulate an experiment (RCT):

  $$\forall i \in \{1, .., n\}, \quad Grade_i = \alpha_0 + \beta_0 Treat_i + u_i$$

- Which sample size and proportion of treated to have a high probability of detecting the effect?

How?

- Simulate many experiments
- Compute the proportion of effects detected

Switch to Quarto document for coding

Thank you!