Introduction to Structural Models

Gabriel Petrini

August 25th, 2020

Contents

1 Introduction

Pre-class reading:

- **lewbel2019zoo**

1.1 Reading Quiz

1. Explain what identification means, according to Arthur Lewbel
2. Correctly order the following econometric actions in their logical sequence:
    - estimation
    - hypothesis testing
    - identification
    - inference
3. What are the two characteristics of reduced-form causal methods, according to Lewbel? How is this different from structural methods?

1. What does Lewbel refer to as "crude structural modeling"?
2. What does Lewbel suggest is a way to overcome the external validity problem?

1.2 Causality: the goal of econometrics

In any econometric endeavor, the goal is to uncover causal relationships. Correlations, otherwise, is contaminated by omitted variable bias.

1.2.1 Causality requires a counterfactual

Causality is defined in terms of a **counterfactual**

- this is the notion of ceteris paribus in principles of economics

"Causal effect": difference between reality and the most plausible counterfactual

1.3 Structural empirical work

Structure: A structure is a data generating process, i.e. a set of functional or probabilistic relationships between observable and latent variables which implies a joint distribution of the observables

- The goal of structural estimation, then, is to estimate the parameters of the DGP
- This allows us to make counterfactual comparisons, i.e. perform causal inference
- Note that "structural" here refers to basically all of modern econometrics

Estimate parameters of a data generating process (DGP) which are assumed to be invariant to policy changes or other counterfactuals

- Once we know the DGP, we can make causal inferences
- *e.g.* estimate a DGP that specifies how cognitive ability and family background relate to the decision to enroll in college and to post-graduation earnings

1.3.1 Brief history of the term "reduced form"

- The term reduced form refers to solving a structural model
  - The structural model may have endogenous variables on both sides of the equation
  - But the reduced form puts all endogenous variables on the left hand side
  - All exogenous variables and error terms are on the right hand side
- Reduced form tends to refer to linear models estimated by RCT, IV, DID, RDD, etc.
  - Methods that try to exploit randomization (or quasi-randomization)
  - Synonymous with the phrase "identification strategy"
- Structural tends to refer to **non-linear** models that are more difficult to estimate
  - Methods that make explicit the (typically large) set of maintained assumptions
  - Methods that focus on settings where RCTs would be infeasible
  - Very little randomization is involved

2 Randomize Control-Trial (RCT) as structural estimation

- All causal inference is structural in nature (as correctly defined)
- An RCT is a structural model that can be evaluated descriptively
  - No fancy econometrics needed: just compute $\overline{y}_{\text{Treatment}} - \overline{y}_{\text{Control}}$
- This is because great effort was expended at the randomization step
- The experimenters had a (structural) model in mind when defining treatment

2.1 Not every DGP can be evaluated by RCT

- Without randomization, we have to rely on observational data
- This requires more complex econometric methods to estimate the DGP
  - Need to explicitly specify how unobservables relate to other parts of model
  - *e.g.* can't randomize a merger of two large firms, or a person's height

$$\log \text{wages} = \beta_0 + \beta_1 \text{educ} + \underbrace{\qquad\qquad}_{\varepsilon}$$
$$\text{familly background, genetics}$$

3 Key parameters of any economic model

- As mentioned above, we assume that DGP parameters are policy-invariant
- These parameters tend to be related to **economic fundamentals**:
  - commodities
  - demographics
  - preferences
  - production technology
  - information and expectations
  - space (includes networks & social interactions)

4 Reading-to-Children Example

4.1 Reduced-form example

- A reduced-form (as misused today) approach would look like the following:

    1. recruit a group of families to participate in a reading study
    2. randomize into "no-read" and "read" groups
    3. after some period of time, give their children a cognitive test
    4. compare the average scores of children across each of the groups

4.2 Structural approach

- A structural (as misused today) approach would look like the following:

    1. write a model of child skill formation (**cunha_al2010**)
    2. gather data on parental and child time use and child test scores
    3. estimate the parameters of the child skill formation model
    4. use model to simulate counterfactual policies (e.g. where reading is set to 0)
    5. compare average scores in counterfactual and *status quo*

4.3 Hybrid approach

- A hybrid approach would do the following:

    1. estimate the skill formation parameters
    2. leverage randomization to better estimate/validate the model
        - e.g. by allowing for identification of a parameter previously not identifiable
        - e.g. recover randomization-implied ATE using structural parameter estimates
    3. use the validated structural model to explore other counterfactuals

5 What is identification? (lewbel2019zoo)

- **Identification:** model parameters being uniquely determined from the observable population that generates the data
    - identification is never a question about a sample of data
        - it is a question about the population from which the sample is drawn

5.1 More formal definition

Let θ denote a set of unknown parameters that we would like to learn about, and ideally, estimate

- e.g. regressor coefficients, average treatment effects, or error distributions
- identification asks what could be learned about parameters θ from observable data
- if we knew the population that data are drawn from, would θ be known?
- if not, what could be learned about θ?

5.2 Why is identification important?

- The study of identification logically precedes estimation, inference, and testing
- For θ to be identified, alternative values of θ must imply different distributions of the observable data
- If θ is not identified, then we cannot hope to find a consistent estimator for θ
- More generally, identification failures complicate statistical analyses of models, so recognizing lack of identification, and searching for restrictions that suffice to attain identification, are fundamentally important problems in econometric modeling
  - If the DGP is not known, it is not possible to do hypothesis test (inference)

5.3 Reduced-form vs. Structural Identification

- In reduced-form econometrics (a.k.a. causal modeling):
  - Typically talk of an "identification strategy" (i.e. randomization setup)
  - Focus is on estimation of treatment effects, not "deep parameters"
  - Relies on randomization from some kind of randomized or natural experiment
- In structural econometrics:
  - Typically talk of "establishing identification" (i.e. sufficient variation in data)
  - In complex models, can be difficult to do without imposing more assumptions

5.4 The Credibility Revolution

- What makes an identification strategy credible?
  - Identification means separating selection from treatment
  - This is best done when treatment is randomized
  - The closer a reduced-form model is to an RCT, the better

5.4.1 Examples of Identification Strategies

- Randomized experiments, field experiments, lab experiments
- Instrumental variables, regression discontinuity
- Difference in differences, synthetic control methods
- Matching methods (nearest neighbor, propensity score, ...)
- OLS that does not suffer from omitted variable bias
- These are almost exclusively estimated using linear econometric models
- Credibility is proportional to the **"cleanliness" of randomization**

5.4.2 Credible Structural Models

What makes a structural model credible?

- At the very least, the model should "fit the data" (i.e. reproduce key patterns)
- But that is usually a low bar to clear, so additional criteria are required
- Results should also "make sense" (i.e. conform to economic theory)
  - e.g. An upward-sloping demand curve would violate this criterion
- Typically requires modeling heterogeneity in preferences or productivity
  - Another difficulty: separating preferences from constraints

5.4.3 Structural Methods

Unlike reduced-form methods, there is not a set "toolkit" of techniques

- Rather, structural modeling is a bit ad hoc or a bit "Wild West"
- Whereas RF methods almost exclusively focus on linear econometric models, Structural methods overwhelmingly require use of **non-linear** econometric models
  - Structural models are typically estimated by GMM or Maximum Likelihood
  - Computational know-how helps speed up the process of estimating these models

6 Internal and External Validity

- **Internal validity** refers to "how causal" an estimated parameter is
  - "This approach is internally valid" $\Rightarrow$ no selection bias
- **External validity** refers to generalizability of estimates to new contexts
- Typically, RF approaches are very good at internal validity but not at external validity
  - On the other hand, if economic agents behave similarly across contexts, structural models can be externally valid
  - RF and structural methods used together can improve both internal and external validity

6.1 Example: Internal vs. External Validity

Suppose we want to measure earth's gravitational force, g

- We can measure g by timing how long it takes various objects to fall some distance
- We can do this with objects of varying mass and of varying fall distances
  - Using this data, we can estimate earth's g
- But what about the g on Mars? Or some other planet?
- For this we need a model of what exactly determines g
  - (A planet's mass and proximity to other large objects)
- This model will tell us what g is on planets we haven't yet visited