Models, Estimation, and the Language of Empirical Economics

Hyoungchul Kim[1]

The Wharton School, University of Pennsylvania

January 20, 2025

[1] Most of the resources in this slide are either from slides by Philip Haile or Peter Hull. Kudos to them all!

Disclaimers

- Lot of the information in this slides are directly from slides by Philip Haile or Peter Hull.
- Do not cite any of this as my work.
- Think of this just as a summary I made from studying their slides and lecture notes.

Why this slide?

- TL;DR: I read some lecture slides by Philip Haile (Yale) and Peter Hull (Brown) in my free time.
- I think they provide a very nice framework for precisely understanding the language of empirical economics.
- So I made my review slide.

Caution

- I am a Ph.D. student still in the process of “learning by doing.”
- Do not think all of the information in this slide is flawless.
- Be critical and please give feedback!
- Contact info: hchulkim[at]wharton.upenn.edu

Motivation

The language of empirical economics can be quite confusing.

- e.g. What is “reduced-form?”
- e.g. What is so-called “Structural estimation?”
- e.g. What is “descriptive evidence?”
- e.g. Is the term “identification strategy” differ from “identification” in metrics?

Goal of this slide

This slide...

1. aims to provide a more logically consistent framework for empirical economics (at least in my understanding).
2. points out some abuse of notation in empiric economics that you should remember.

Descriptive vs. Structural

All empiric work in economics is either “Descriptive” or “Structural.”

1. **Descriptive work ⇒ tends to characterize or describe (descriptive) relations between observed variables.**
    - e.g. college grads earn 98% more per hour than others; income inequality higher now than 30 years ago
    - There is no economic theory or model involved (at least in terms of implementation): We are just plotting correlation between some variables of possible interest.
    - But this is a nice starting place for academic research (more on next slide).
2. **Structural work ⇒ quantifies certain features of an assumed data generating process (i.e., a “structure”) that allow one to answer counterfactual questions.**
    - Any empiric research that involves some notion of “causal” can be understood as structural.

Why care about descriptive work?

While descriptive work might not be our end, it can still be a very useful starting point.

- Sometimes, correlation “strongly hints” causality.
- Striking descriptive result is a nice motivation to start with.
- If the descriptive result is not that prominent, it’s likely that the following structural result will also not be very promising.
- Sometimes, very striking descriptive result is enough to appeal to people.

What about “Descriptive evidence”

Recently there has been some trend where researchers would employ so-called “reduced-form” work in the start of their paper and call it “descriptive evidence.”

- These usually mean they do some DID, IV, event study-ish work in the front.
- In essence, this is an abuse of notation as any process involving identification of causal parameter of interest should be considered “structural.”
- Usually this seems to mean the authors are employing some sort of “loose” identification strategy where the assumption is weakly held or when they don’t want to impose the full specific model from the start.
- It might be better to say that you are going to first impose a simpler framework for the DGP.

Data-generating-process (DGP)

All structural work starts with some assumed Data-generating-structure (DGP).

- DGP refers to some researcher’s assumption about how the observed data is generated.
- To be more formal, it is a probabilistic or functional (causal) distribution of observed and latent variables.
- Think of like a full formula used by god to make some variable of our interest.
- This is where “structure” notation comes from: We are imposing some assumed restrictions on the data being generated.
- Usually this DGP is based on some general theoretical foundation in one’s field.
    - This is where economic theory comes into play! Using economic theory, we as economists have some assumed structure of this DGP.

Examples of DGP

There are many examples of DGP we use in our daily empirical analysis

- e.g. ATE is one of them!
- Using potential outcome framework, we are assuming average treatment effect of certain variable will be constructed as $E(Y(1)|D=1) - E(Y(0)|D=1)$.
- Under some assumptions (model) like parallel trend and anticipation, we are able to uniquely uncover this parameter of interest using DID. (We will talk more about this in identification).

Model

What about model then?

- Model can be thought of as abstraction of the DGP.
- To be more formal, it is a class of possible structures that is assumed to contain the true DGP of our interest.
- Basically, model is some maintained hypothesis about the DGP s.t. it will still contain the true DGP but allow flexibility to uncover the parameter of our interest.

Parameters

Somehow, I have started using the term "parameters."

- Parameter is quantity of interest derived from the model.
- This is just a coefficient in the structural model we are interested in.
- Be aware that this is different from estimand.
- Estimand is just a function of population distribution. This happens to be identical to the parameter when certain assumptions are met. (More about this later).

Structure and model: Example

Suppose we are interested in learning about gains from education.

- In our mind, there will be some fully specified DGP that is unobservable.
- We know from economic theory that links the positive effect of gain in human capital (education) on outcome like future earning.
- We use this economic hypothesis to specify the model such as: $Y = \beta D + \varepsilon$.
- Y is earning and D is years in education.
- Note that this is different from the statistical model of OLS.
- While D and ε is correlated in our structural model, if we run OLS, the statistical model will automatically make explanatory variable uncorrelated with error term.
- This is where identification comes into play: identification links the economic parameter to the estimand.

Definition of identification

So, what is identification then?

- **Important**: Identification and estimation are different!
- Identification is set of (theoretic) assumptions that allows us to link structural parameter to the estimand derived from the statistics.
- Think of the example earlier: If we just run OLS, we cannot get the structural parameter β.
- But suppose now we added enough confounding variables to make D uncorrelated with the error term in the structural model.
- Now if we run OLS, we can retrieve the parameter value because parameter value $=$ estimand.

Identification vs. Identification strategy

To be honest, the term "Identification" I used until now should be distinguished from the term "identification" in econometrics.

- In metrics, identification means certain parameter of interest is always uniquely defined from the distribution of observables.
- e.g. If invertible, the OLS coefficient will be unique.
- "Identification strategy" seems to have stemmed from this idea of "uniqueness" but seems to be slightly different from the original terminology.
- It basically is a set of assumptions that allows us to retrieve the economic parameters of interest in the structural model.
- But both terms are quite similar if we think of identification strategy as specific case of identification: It is basically saying we can always retrieve the parameter value from set of maintained hypotheses which contain the true DGP.

Estimation: Statistical inference

Now we move from economic theory to statistic inference.

- This is no longer about our economic parameter of interest.
- It is just a process of approximating values estimated from sample data to the values in the population distribution.
- If you see sth about population and sample, this means you are now in the statistics zone.

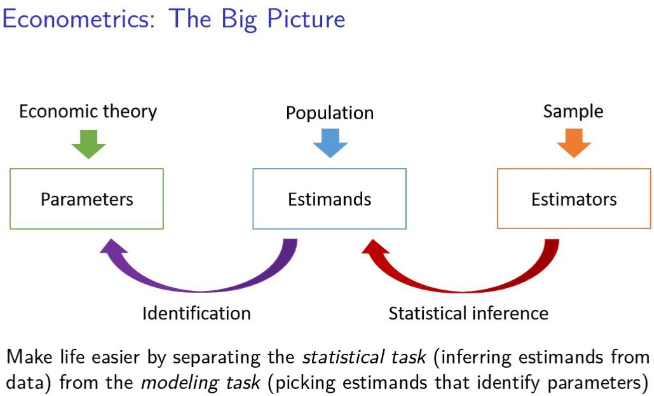
Estimand, estimate, estimator

We need to differentiate three key concepts: Estimand, esimate, estimator.

1. Estimand: Function of some population distribution. This is unobservable.
2. Estimate: Function of some sample data used to approximate the value of estimand (observable).
3. Estimator: Some formula to approximate the value of the estimand.

Statistics jargon: we use estimator to estimate the estimand.

One picture to rule them all



Econometrics: The Big Picture

Economic theory

Population

Sample

Parameters

Estimands

Estimators

Identification

Statistical inference

Make life easier by separating the *statistical task* (inferring estimands from data) from the *modeling task* (picking estimands that identify parameters)

Figure: Citation (Peter Hull notes)

Reduced-form

“Reduced-form” is one of the most abused terminology.

- In applied economics, this seems to mean we are not fully imposing all the underlying mechanisms or implied functional forms for some DGP.
- Thus, this is understood as setting some simplified relationships (model).
- But this is bit confusing as Phil says, “every model suppress underlying mechanisms!”
- Also problematic because this makes people think there is no “structure” imposed in the analysis. In fact, this is still a structural work.
- But everyone uses it, so just keep in mind that this is what they mean.

Structural model, or structural estimation

“Structural model” or “structural estimation” is also one of the most abused terminology.

- “Structural model” or “structural estimation” is also used as some antithesis of “Reduced-form.”
- Thus, this is understood as setting a full underlying mechanisms or functional forms.
- This is also confusing because no model is full: we always suppress some underlying mechanisms.
- Also confusing because we are always “structurally estimating” as long as we do some empiric work other than descriptive.
- IRL, this just kind of means one is using a complex model, or a model with many parametric/functional form assumptions.

Conclusion

- All empiric work is either descriptive or structural.
- Model is class of maintained hypotheses of the DGP.
- Identification links structural parameter to the estimand.
- Statistical inference links estimand to estimate.
- Do not use the terms “reduced-form” or “structural estimation.”