What is Econometrics?

Put simply, it is the application of statistics to economic data. The econometrician uses the tools of probability & statistics to
define and test hypotheses about economic phenomena from the data

But, unlike the statistician, our
decision-making is driven by economic theory and the limitations inherent to the field.

One defining limitation is the
impossibility of running economy wide experiments. We

must deal with observational (in contrast with experimental) data, which presents challenges when trying to establish causal relationships.

To a large degree, this lack of control in our experiments, still holds true. But there are some exceptions:

1. Experimental economics in applied micro (i.e., behavioral, neuro, etc.)

2. Computer simulations of the economic system

Nevertheless, we won't run any

experiments in this course. Our goal is to master fundamental concepts of probability & statistics to be educated consumers of observational data, specially economic data.

I have to warn you that the more statistics, econometrics, or data wrangling techniques in general that you learn the more you will see limitations of data-driven statements in the real-world. You already know that an average is sensitive to outliers and that any statistic is pretty uninformative without details of the underlying distribution. This is a blessing and a curse. But better to be aware than to live a life of ignorance.

Basic Methodology

There are often two general purposes of an econometric model

1- Verify qualitative economic theory using data

2- Discover and develop new economics from empirical observ.

As a general guideline, the econometrician might follow these 8 steps:

1 - State economic theory to be tested

$$educ \rightarrow wages$$

2 - Specify math model of the theory

$$wage_i = \beta_0 + \beta_1 \cdot educ_i$$

3 - Specify econometric model

$$wage_i = \beta_0 + \beta_1 \cdot educ_i + \varepsilon$$

4 - Collect data for each variable in your econometric model

wage := Avg. hourly salaries,

educ := Years of schooling

5 - Estimate the parameters of

the econometric model

MLE, OLS, NLOLS, VAR, GD, ...

6 - Test hypotheses suggested by economic theory and related to the economic parameters

educ $\uparrow$ $\rightarrow$ wage $\uparrow$

7- Forecasts and predicts

$$wage = \hat{\beta}_0 + \hat{\beta}_1 (5)$$

8- Apply econometric model for control and/or policy purposes

Grace McCormack (Harvard) outlines it more succintly

1. Population Data Generating Process (DGP)

$$y_i = g(\underbrace{x_i}_{\text{Data}} \mid \underbrace{\theta}_{\text{Params.}})$$

Sampling

2. Observe data from a sample of N observations of $i : 1 \ldots N$

$$\{y_i, x_{1i}, x_{2i}, x_{Ki}\}$$

Estimating

3. Characterize parameters of the model using some econometric method

OLS, NLLS, VAR, ...

1. The population DGP is like the recipe that Nature used to create the data we observed

- We will never know this accurately. Our measurements are imperfect, models are incomplete, and our understanding limited.

The DGP is defined by a set of population parameters θ. The $g(\cdot)$ function specifies the math/metrics model.

2. We then sample as many observations from thes DGP as possible. For ex, if the DGP is the causal effect of education on income, we could have a sample of N people with data on their education and income

3. The last step is to use the sample

of data to understand the true population parameter. Two ways of "characterizing" θ, depending on the context, are:

1. Hypothesis test.
2. Estimator.

Important Concepts

- Statistical vs Deterministic:

  Statistical = Stochastic = Random

  Deterministic = Non-random

  Our variables of interest are always stochastic. We deal with Random Variables that have a probability distribution. This means that we can never know for sure what the outcome of a variable of interest will be.

Statistics gives us tools to make inferences about the possible random outcomes. Specifically, it gives us tools to determine likely (or expected) outcomes. Hence the use of averages, confidence intervals, and so on.

- Correlation vs Causation:

  Causality $\Rightarrow$ Correlation

  Correlation $\not\Rightarrow$ Causation

  Correlation is a mathematical construct

that measures the degree of linear dependence between two variables.

Causality is often more qualitative.

thus requires solid grounding in

theory. We can compute causal estimates, but in practice these mean nothing if they are not supported by a sound and validated rationale.

There are a bunch of definitions of causality, and an equal number of causality tests. It is all a philosophical debate, which I invite you to explore

on your own. Some of you may even subscribe to David Hume's ideas that there doesn't exist such a thing as causation. But the whole scientific endeavor relies on our ability to

disentangle cause-and-effect, so practice a healthy level of skepticism.

A modern view, from complexity sciences,
is that rather than direct cause-and-effect we have feedback loops and co-evolution. It's a new way of doing science.

- Regression vs Causation

Regression $\not\Rightarrow$ Causation

Regression is a tool to explore the dependence of one variable on others, but it doesn't imply that it is caused by them. We will study regression analysis in detail this semester, and you will all the ways in which it may fail to provide unbiased results.

- Regression vs Correlation:

Correlation → linear relationship between two variables

Regression → Expected value of a random variable conditioned on other exogenous factors

Correlations coefficients are symmetric

Regression coefficients enforce (or assume) a directed relationship. More specifically, setting up a regression equotion implies an assumption that the independent variables have a direct effect on the dependent variable.

For example, $wage_i = \beta_0 + \beta_1 educ_i + \varepsilon$ states that education affect wage and not the other way around.

As we will learn, almost everything in economics goes the other way around as well. Regression is a powerful tool, but you must use it carefully and mindfully.

Types of data

Time series x_t
Cross-section x_i
Panel (longitudinal) x_{it}