

# Introduction to Econometrics

William Brasic

The University of Arizona

# Introduction to Me



- Originally from Chicago (Bear Down)
- Did my bachelors and masters at UNLV in economics and data science
- Currently a 3rd year economics doctoral student studying industrial organization, machine learning, applied econometrics
- Big EDM guy (Illenium, Slander, Excision, Gryffin, Porter Robinson)

# Introduction to You Guys

Make an introduction post on the Slack course group chat by  
30 August at 11:59 P.M. for two points of extra credit!

- Click here to join the chat
  - ▶ Please use your first and last name as well as registering with your university email address

# Introduction to You Guys

Once in the chat, navigate to the introductions channel and answer the following questions:

1. What is your first and last name?
2. Where are you from?
3. What is a hobby of yours?
4. What is your major?
5. What programming experience do you have?
6. What do you know about econometrics?
7. What do you know about machine learning?

# What is Econometrics?

## Question 1: What is Econometrics?

What is Econometrics?

# What is Econometrics?

## Question 1: What is Econometrics?

What is **Econometrics**?

## Answer to Question 1

**Econometrics** is a set of statistical tools by which economists and others in the social sciences analyze data.

- We use **econometrics** to:
  - ▶ Estimate economic relationships
  - ▶ Test economic theories
  - ▶ Evaluate government and business policy

# Causality

## Definition 1: Causality

The causal effect of a variable  $x$  on a variable  $y$  is the impact  $x$  has on  $y$  holding all other factors constant.

# Correlation Versus Causation

## Property 1: Correlation Versus Causation

Correlation does not imply causation!

### Spurious Correlations

# Experimental Data

## Definition 2: Experimental Data

Experimental data is generated through controlled experiments where the researcher manipulates one or more variables to observe the effect on other variables.

- So long as the experiment is done in the right way (e.g., using random assignment), the causal effect can be obtained.

# Observational Data

## Definition 3: Observational Data

**Observational data** is collected without manipulating the environment or variables. We have no control over how the data was generated.

- When using **observational data**, it is much harder to obtain the causal effect due to potential confounding variables and lack of randomization.

# Econometrics Analysis Example

## Example 1: Econometrics Analysis Example

Examining the effect of school spending on student performance.

- Causality is difficult to determine
  - ▶ Reverse causality: If students perform better this may cause schools to want to spend more.
  - ▶ Endogeneity: There could be factors excluded from our model influencing both school spending and student performance

# Economic Theory

## Definition 4: Economic Theory

Economic theory is a way to identify *deterministic* relationships between variables.

- Theory usually tells us the direction of the change in one variable when another changes.
- Theory does not usually tell us the magnitude of that change.

# Econometrics

## Definition 5: Econometrics

Econometrics begins where theory concludes by trying to explain how much one variable changes when another changes. Econometrics requires:

1. Theory
2. Data
3. Statistical Methods

- Since relationships are stochastic in the real world, econometrics uses probability and statistics to make conclusions about them.

# Econometric Model

## Definition 6: Econometric Model

An **econometric model** starts with a theoretical proposition and tests this proposition through applying statistical methods to the data.

# Econometric Model

## Example 2: Econometric Model

Our theoretical proposition is that a person's education level should influence their income. Theory indicates as one's education level rises, their income should rise as well. So, we formulate the **econometric model**

$$\underbrace{\text{Income}}_{\text{Dependent Variable}} = \underbrace{\beta}_{\text{Parameter}} * \underbrace{\text{Education}}_{\text{Independent Variable}} + \underbrace{u}_{\text{Stochastic}}.$$

Then, we collect data on people's income and education to test this hypothesis.

- The goal of the **econometric model** is to obtain a good estimate of  $\beta$ .
- $u$  accounts for the stochasticity involved in the relationship between income and education.

# Econometric Model

## Question 2: Econometric Model

If  $y$  is a person's income and  $x$  is a person's education level, do you think the **econometric model**

$$y = x\beta + u$$

is good?

# Econometric Model

## Question 2: Econometric Model

If  $y$  is a person's income and  $x$  is a person's education level, do you think the **econometric model**

$$y = x\beta + u$$

is good?

## Answer to Question 2

No! There are likely confounding variables (contained in  $u$ ) correlated with  $y$  and  $x$  so we won't get a reliable estimate of the effect of education on income,  $\beta$ . This is called an **endogeneity** issue.

- We think of  $u$  as everything excluded from our model that impacts  $y$ .

# Hypothesis Testing

## Example 3: Hypothesis Testing

Suppose we collect data on an outcome  $y$  and covariates  $x_1$  and  $x_2$ . Then, we formulate the econometric model

$$y = \beta_1 x_1 + \beta_2 x_2 + u.$$

We can conduct hypothesis tests of  $H_0 : \beta_1 = 0$  versus  $H_A : \beta_1 \neq 0$ .

- The **null hypothesis** corresponds to  $x_1$  not having a statistically significant impact on  $y$ .
- The **alternative hypothesis** corresponds to  $x_1$  having a statistically significant impact on  $y$ .

# Types of Datasets

## Definition 7: Cross Sectional

A dataset is said to be **cross-sectional** if we have data on *multiple* agents at a *single* point in time.

- Gathering data on *multiple* individual's education and income levels at one point in their life.

# Types of Datasets

## Definition 8: Time Series

A dataset is said to be a **time series** if we have data on a *single* agent at *multiple* points in time.

- Gathering data on inflation over *multiple* years in history.

# Types of Datasets

## Definition 9: Panel Data

A dataset is said to be a **panel** if we have data on a *multiple* agents at *multiple* points in time.

- Gathering data on *multiple* countries' GDP over *multiple* years in history.

# Empirical Work

## Definition 10: Empirical Work

Empirical work falls into two categories

1. Causal Inference (e.g., econometrics)
  - ▶ Drawing causal conclusions among an independent variable and its covariates.
2. Predictive modeling (e.g., machine learning)
  - ▶ Only caring about predicting the outcome with high accuracy.

# Machine Learning

## Definition 11: Machine Learning

Machine Learning (ML) is a field of artificial intelligence that focuses on developing algorithms and statistical models that enable computers to learn and make predictions or decisions without being explicitly programmed.

- ML involves training models on data to identify patterns, make inferences, and improve performance over time.
- ML only cares about prediction and doesn't concern itself with any of the issues econometrics worries over (e.g., causality).

# Machine Learning

## Example 4: Machine Learning

Examples of machine learning include:

- Using linear regression to predict a person's income based on their education level.
- Using random forests to classify a patient has having cancer based on their individual characteristics.
- Using neural networks to predict the optimal listing price of a home based on its characteristics.
- Using reinforcement learning to construct a large language model (LLM) like ChatGPT.

# Thank You!