Introduction and Overview

EC 421, Set 1

Edward Rubin

Prologue

Why?

Motivation

Let's start with a few **basic, general questions:**

Why?

Motivation

Let's start with a few **basic, general questions:**

1. What is the goal of econometrics?

2. Why do economists (or other people) study or use econometrics?

Why?

Motivation

Let's start with a few **basic, general questions:**

1. What is the goal of econometrics?

2. Why do economists (or other people) study or use econometrics?

One simple answer: Learn about the world using data.

Why?

Motivation

Let's start with a few **basic, general questions:**

1. What is the goal of econometrics?

2. Why do economists (or other people) study or use econometrics?

One simple answer: Learn about the world using data.

- *Learn about the world* = Raise, answer, and challenge questions, theories, assumptions.

- *data* = Plural of datum.

Why?

Example

One might (reasonably) guess a company's sales are a function of its advertising spending, price, and intesity of competitors.

Why?

Example

One might (reasonably) guess a company's sales are a function of its advertising spending, price, and intesity of competitors.

So, one might hypothesize a model $\text{Sales} = f(\text{Ad}, \text{Price}, \text{Comp})$

where

- Ad represents dollars spent on advertising,
- Price is the product's price,
- Comp gives the product's competition.

Example

One might (reasonably) guess a company's sales are a function of its advertising spending, price, and intesity of competitors.

So, one might hypothesize a model $\text{Sales} = f(\text{Ad}, \text{Price}, \text{Comp})$

where

- Ad represents dollars spent on advertising,
- Price is the product's price,
- Comp gives the product's competition.

We expect that sales $\uparrow$ with advertising and $\downarrow$ with price and competition.

But who needs to *expect*?

We can *test* these hypotheses **using regression**.

But who needs to *expect*?

We can *test* these hypotheses **using regression**.

More importantly: Regression estimates the *size* of these effects

- *How much* does an additional dollar of advertising increase sales?
- *How much* does a one-dollar increase in price decrease sales?
- *How much* does an additional competitor reduce sales?

These (causal) questions are central to efficient decision-making and are the bread and butter of econometrics.

Why?

Example, cont.

Regression model:

$$\text{Sales}_i = \beta_0 + \beta_1 \text{Ad}_i + \beta_2 \text{Price}_i + \beta_3 \text{Comp}_i + \varepsilon_i$$

With this basic regression model, we can test/estimate/quantify the (linear) relationship between sales and advertising, price, and competition.

Why?

Example, cont.

Regression model:

$$\text{Sales}_i = \beta_0 + \beta_1 \text{Ad}_i + \beta_2 \text{Price}_i + \beta_3 \text{Comp}_i + \varepsilon_i$$

(Review) Questions

Why?

Example, cont.

Regression model:

$$\text{Sales}_i = \beta_0 + \beta_1 \text{Ad}_i + \beta_2 \text{Price}_i + \beta_3 \text{Comp}_i + \varepsilon_i$$

(Review) Questions

- **Q:** How do we interpret β_1?

Example, cont.

Regression model:

$$\text{Sales}_i = \beta_0 + \beta_1 \text{Ad}_i + \beta_2 \text{Price}_i + \beta_3 \text{Comp}_i + \varepsilon_i$$

(Review) Questions

- **Q:** How do we interpret β_1?
- **A:** An additional dollar of advertising corresponds with a β_1-unit change in sales (holding price and competition fixed).

Why?

Example, cont.

Regression model:

$$\text{Sales}_i = \beta_0 + \beta_1 \text{Ad}_i + \beta_2 \text{Price}_i + \beta_3 \text{Comp}_i + \varepsilon_i$$

(Review) Questions

- **Q:** Are the β_k terms population parameters or sample statistics?

Why?

Example, cont.

Regression model:

$$\text{Sales}_i = \beta_0 + \beta_1 \text{Ad}_i + \beta_2 \text{Price}_i + \beta_3 \text{Comp}_i + \varepsilon_i$$

(Review) Questions

- **Q:** Are the β_k terms population parameters or sample statistics?
- **A:** Greek letters denote **population parameters**. Their estimates get hats, *e.g.*, $\hat{\beta}_k$. Population parameters represent the **average** behavior across the population.

Example, cont.

Regression model:

$$\text{Sales}_i = \beta_0 + \beta_1 \text{Ad}_i + \beta_2 \text{Price}_i + \beta_3 \text{Comp}_i + \varepsilon_i$$

(Review) Questions

- **Q:** Can we interpret the estimates for β_2 as causal?

Why?

Example, cont.

Regression model:

$$\text{Sales}_i = \beta_0 + \beta_1 \text{Ad}_i + \beta_2 \text{Price}_i + \beta_3 \text{Comp}_i + \varepsilon_i$$

(Review) Questions

- **Q:** Can we interpret the estimates for β_2 as causal?
- **A:** Not without making more assumptions and/or knowing more about the data-generating process.

Why?

Example, cont.

Regression model:

$$\text{Sales}_i = \beta_0 + \beta_1 \text{Ad}_i + \beta_2 \text{Price}_i + \beta_3 \text{Comp}_i + \varepsilon_i$$

(Review) Questions

- **Q:** What is ε_i?

Example, cont.

Regression model:

$$\text{Sales}_i = \beta_0 + \beta_1 \text{Ad}_i + \beta_2 \text{Price}_i + \beta_3 \text{Comp}_i + \varepsilon_i$$

(Review) Questions

- **Q:** What is ε_i?
- **A:** An individual's random deviation/disturbance from the population parameters.

Population parameters are averages; individuals are rarely average.

Why?

Example, cont.

Regression model:

$$\text{Sales}_i = \beta_0 + \beta_1 \text{Ad}_i + \beta_2 \text{Price}_i + \beta_3 \text{Comp}_i + \varepsilon_i$$

(Review) Questions

- **Q:** Which assumptions do we impose when estimating with OLS?

Why?

Example, cont.

Regression model:

$$\text{Sales}_i = \beta_0 + \beta_1 \text{Ad}_i + \beta_2 \text{Price}_i + \beta_3 \text{Comp}_i + \varepsilon_i$$

(Review) Questions

- **Q:** Which assumptions do we impose when estimating with OLS?
- **A:**
  - The relationship between the sales and the explanatory variables is linear in parameters, and ε enters additively.
  - The explanatory variables are **exogenous**, *i.e.*, $E[\varepsilon|X] = 0$.
  - You've also typically assumed something along the lines of: $E[\varepsilon_i] = 0$, $E[\varepsilon_i^2] = \sigma^2$, $E[\varepsilon_i \varepsilon_j] = 0$ for $i \neq j$.
  - And (maybe) ε_i is distributed normally.

Assumptions

How important can they be?

You've learned how **powerful and flexible** ordinary least squares (**OLS**) regression can be.

Assumptions

How important can they be?

You've learned how **powerful and flexible** ordinary least squares (**OLS**) regression can be.

However, the results you learned required assumptions.

Assumptions

How important can they be?

You've learned how **powerful and flexible** ordinary least squares (**OLS**) regression can be.

However, the results you learned required assumptions.

Real life often violates these assumptions.

EC421 asks "**What happens when we violate these assumptions?**"

- Can we find a fix? (Especially: How/when is β *causal*?)
- What happens if we don't (or can't) apply a fix?

OLS still does some amazing things—but you need to know when to be **cautious, confident, or dubious**.

Not everything is causal

But what *is*?

Suppose you estimate our sales model for your boss.

$$\text{Sales}_i = \hat{\beta}_0 + \hat{\beta}_1 \text{Ad}_i + \hat{\beta}_2 \text{Price}_i + \hat{\beta}_3 \text{Comp}_i + e_i$$

Can you trust that $\hat{\beta}_2$ gives you the actual effect of price on sales?

Not everything is causal

But what *is*?

Suppose you estimate our sales model for your boss.

$$\text{Sales}_i = \hat{\beta}_0 + \hat{\beta}_1 \text{Ad}_i + \hat{\beta}_2 \text{Price}_i + \hat{\beta}_3 \text{Comp}_i + e_i$$

Can you trust that $\hat{\beta}_2$ gives you the actual effect of price on sales?

You should be asking several questions...

Not everything is causal

But what *is*?

Suppose you estimate our sales model for your boss.

$$\text{Sales}_i = \hat{\beta}_0 + \hat{\beta}_1 \textbf{Ad}_i + \hat{\beta}_2 \textbf{Price}_i + \hat{\beta}_3 \textbf{Comp}_i + e_i$$

Can you trust that $\hat{\beta}_2$ gives you the actual effect of price on sales?

You should be asking several questions...

1. *Where* does the *variation in price* come from?
   - Is it *random* (*exogenous*)?
   - *Why* are some products (or times) *more* expensive than others?
2. *Whom* do the data represent? Are they *relevant* to your setting?
3. How *confident* are you in your answer?

Econometrics

Applied econometrics, data science, analytics require:

1. Intuition for the **theory** behind statistics/econometrics
   (assumptions, results, strengths, weaknesses).

2. Practical knowledge of how to **apply theoretical methods** to data.

3. Efficient methods for **working with data**
   (cleaning, aggregating, joining, visualizing).

Econometrics

Applied econometrics, data science, analytics require:

1. Intuition for the **theory** behind statistics/econometrics (assumptions, results, strengths, weaknesses).
2. Practical knowledge of how to **apply theoretical methods** to data.
3. Efficient methods for **working with data** (cleaning, aggregating, joining, visualizing).

This course aims to deepen your knowledge in each of these three areas.

Econometrics

Applied econometrics, data science, analytics require:

1. Intuition for the **theory** behind statistics/econometrics (assumptions, results, strengths, weaknesses).
2. Practical knowledge of how to **apply theoretical methods** to data.
3. Efficient methods for **working with data** (cleaning, aggregating, joining, visualizing).

This course aims to deepen your knowledge in each of these three areas.

- 1: As before.
- 2–3: **R**

Econometrics

My "**big-picture takeaways**" (the *intuition* that I hope you form)

Econometrics

My "**big-picture takeaways**" (the *intuition* that I hope you form)

- *most* interesting questions are causal;
- selection into treatment dominates correlation (esp. cross-sectional);
- measurement error can too;
- causality comes from design—not from models/assumptions;
- ask about the counterfactual;
- non-stationary time series will lead you to bad conclusions;

Econometrics

My "**big-picture takeaways**" (the *intuition* that I hope you form)

- *most* interesting questions are causal;
- selection into treatment dominates correlation (esp. cross-sectional);
- measurement error can too;
- causality comes from design—not from models/assumptions;
- ask about the counterfactual;
- non-stationary time series will lead you to bad conclusions;
- quantifying uncertainty is just as important as the effect estimate;
- consider *which* population your data represent;
- the mean is only one of many ways to summarize a population;
- don't mistake mean reversion for treatment effects/heterogeneity;

Econometrics

My "**big-picture takeaways**" (the *intuition* that I hope you form)

- *most* interesting questions are causal;
- selection into treatment dominates correlation (esp. cross-sectional);
- measurement error can too;
- causality comes from design—not from models/assumptions;
- ask about the counterfactual;
- non-stationary time series will lead you to bad conclusions;
- quantifying uncertainty is just as important as the effect estimate;
- consider *which* population your data represent;
- the mean is only one of many ways to summarize a population;
- don't mistake mean reversion for treatment effects/heterogeneity;
- many maps are just *population*;
- graphs should clearly communicate a *message*… beautifully.

Next: R basics + (More) Metrics review(s)