What Is Econometrics?

From Economic Questions to Empirical Answers

Jake Anderson

March 26, 2026

Outline

1. Why Econometrics?
2. From Economic Models to Econometric Models
3. Correlation Is Not Causation
4. Data Types
5. The Econometrics Pipeline
6. What Lies Ahead

A Simple Economic Question

Suppose a household earns an extra $100 per week.

How much more will they spend on food?

You could guess:

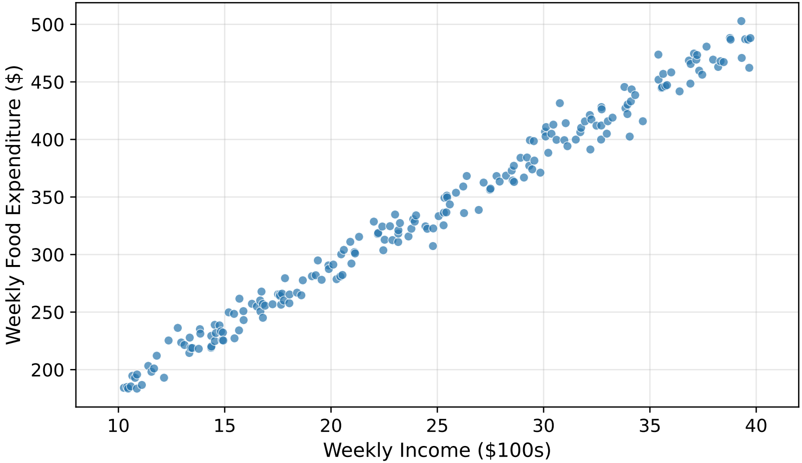
- Economic theory says food is a normal good, so spending goes up
- But by how much? $5? $15? $40?

$\Longrightarrow$ Theory tells us the *direction*, but not the *magnitude*. We need data and a method to pin down a number.

That method is **econometrics**.

Let's Look at Some Data

Here is food expenditure data from 200 households with weekly income between $1,000 and $4,000:



What Is Econometrics?

Econometrics uses economic theory, data, and statistical tools to:

1. **Estimate** economic relationships (how much does x affect y?)
2. **Predict** economic outcomes (what will GDP be next quarter?)
3. **Test hypotheses** (does a minimum wage increase reduce employment?)

You already know statistics from Econ 41: means, variances, hypothesis tests.
$\Longrightarrow$ Econometrics adds an **economic model** to guide which statistics to compute and how to interpret them.

"How Much" Questions Are Everywhere

Decision-makers across economics need magnitudes, not just directions:

- **Federal Reserve:** How much should interest rates change to lower inflation without stalling growth?
- **Business owner:** How much additional revenue does $1,000 in advertising generate?
- **University:** How many fewer students enroll if tuition rises by $500?
- **Transit authority:** How much does ridership fall when fares increase by 25 cents?

The values that answer these questions (elasticities, multipliers, marginal effects) are unknown **parameters**. Econometrics estimates them from data.

Outline

1. Why Econometrics?
2. From Economic Models to Econometric Models
3. Correlation Is Not Causation
4. Data Types
5. The Econometrics Pipeline
6. What Lies Ahead

Step 1: The Economic Model

Economics gives us a **functional relationship**:

$$\text{food expenditure} = f(\text{income})$$

Here $f(\cdot)$ is a general function saying "food expenditure depends on income." It doesn't tell us:

- The exact *form* of f (linear? curved?)
- The *size* of the effect
- Why two households with the *same income* spend different amounts

The economic model is a starting point. To get numbers from data, we need to be more specific.

A First Attempt: Averages by Income Group

We could split households into income bins and compute the average food expenditure in each:

This confirms the pattern: food spending rises with income

The Limitation of Bin Averages

Bin averages tell us the *average* food expenditure for households within each group. But consider:

- Can you predict food spending for a household earning $2,750/week? That falls between two bins.
- What about a household earning $5,000/week? That is outside the data entirely.

Bin averages are *discrete*. They give us a handful of points, not a continuous prediction.

$\Longrightarrow$ We need a **model** that summarizes the relationship with a formula we can evaluate at *any* income level.

Step 2: The Econometric Model

To get a “how much” answer, we specify a **functional form** and add a **random error**:

$$y = \beta_1 + \beta_2 x + e$$

- y = weekly food expenditure ($)
- x = weekly income ($100s)
- β_1 = food spending when income is zero (intercept)
- β_2 = extra food spending per additional $100 of income (slope)
- e = random error

Note: some textbooks label the intercept and slope β_0 and β_1. Our textbook numbers them β_1 and β_2, but the idea is identical.

β_1 and β_2 are unknown **parameters**. Econometrics estimates them from data.

Two Components of Every Econometric Model

$$y = \underbrace{\beta_1 + \beta_2 x}_{\text{systematic component}} + \underbrace{e}_{\text{random error}}$$

Systematic component: the predictable part, derived from economic theory.

- Tells us the *average* relationship between x and y

Random error (e): the unpredictable part. It captures:

- Omitted factors (household size, dietary preferences, location)
- Measurement error in the data
- Inherent randomness in human behavior

(We will define “omitted variables” formally in later chapters. For now: factors that affect y but are not included in the model.)

$\Longrightarrow$ The error explains why two households with the same income spend different amounts on food.

Estimation: Fitting the Line to Data

Using data from 200 households, econometric methods give us **estimates** $\hat{\beta}_1$ and $\hat{\beta}_2$. A hat ($\hat{}$) denotes a value computed from data, as opposed to the true unknown parameter.

Outline

1. Why Econometrics?
2. From Economic Models to Econometric Models
3. Correlation Is Not Causation
4. Data Types
5. The Econometrics Pipeline
6. What Lies Ahead

A Different Regression, A Different Problem

We just estimated how food spending relates to income. Now consider a different example to see where regression estimates can mislead us.

Suppose we estimate:

$$\text{GRADE} = \beta_1 + \beta_2\, \text{SKIP} + e$$

We find $\hat{\beta}_2 < 0$: students who skip more classes tend to get lower grades.

Question: Does skipping *cause* lower grades?

Why $\hat{\beta}_2$ May Not Be Causal

Students who skip may also:

- Work long hours at a job
- Be less motivated to study
- Have personal circumstances affecting both attendance and grades
- Already know the material well enough that they don't need help
- Prefer learning at their own pace from a book or other resources

These are **omitted variables**: factors left out of the model that are correlated with both SKIP and GRADE. Because they sit in the error term e, the estimate $\hat{\beta}_2$ absorbs their influence.

$\implies$ $\hat{\beta}_2$ captures the *association* between skipping and grades, not necessarily the causal effect of skipping alone.

Prediction vs. Causation

Not every "how much" question requires a causal answer:

- **Prediction:** "If a student has skipped 5 classes, what grade do we expect?" $\Longrightarrow$ $\hat{\beta}_2$ is fine for this.
- **Causation:** "If we *force* a student to attend 5 more classes, how much will their grade improve?" $\Longrightarrow$ $\hat{\beta}_2$ may be misleading.

Much of this course builds tools to move from association toward causation:

- Adding control variables (multiple regression)
- Hypothesis testing and model specification
- Indicator variables and interaction terms

$\Longrightarrow$ Distinguishing correlation from causation is a central challenge of econometrics.

Outline

1. Why Econometrics?
2. From Economic Models to Econometric Models
3. Correlation Is Not Causation
4. Data Types
5. The Econometrics Pipeline
6. What Lies Ahead

How Are Economic Data Generated?

The way data are collected affects what conclusions we can draw:

1. **Experimental data:** the researcher controls the treatment
   - Tennessee's Project Star: students randomly assigned to small vs. large classes
2. **Quasi-experimental data:** a policy or natural event creates treatment and control groups, though the researcher did not assign them
   - Card and Krueger: NJ raised the minimum wage, neighboring PA did not
3. **Observational (nonexperimental) data:** we observe the world as it is
   - Current Population Survey: monthly survey of $\sim$50,000 U.S. households

Most economic data are observational. $\implies$ We cannot assume that differences in x are "as good as random," which makes causal inference harder.

Some experiments are also **unethical**: you cannot randomly assign people to get cancer, or randomly sentence people to incarceration, just to measure the effect.

Cross-Section Data

Observations on **many units at one point in time**.

Individual	Race	Education	Sex	Wage ($/hr)
1	Black	12	Female	15.50
2	White	16	Male	27.00
3	Hispanic	10	Female	11.25
⋮	⋮	⋮	⋮	⋮

Example: CPS March 2013: wages, education, demographics for thousands of workers, all measured in the same month.

$\implies$ We see variation *across* individuals, but we have no "before and after."

Time-Series Data

The **same variable** recorded at regular intervals over time.

Year	**Real GDP (Billions, 2009 $)**
2010	14,784
2011	15,021
2012	15,355
2013	15,612
2014	15,982

Example: quarterly GDP, monthly unemployment, daily stock prices.

$\implies$ We see how things change *over time*, but observations are typically not independent (today's GDP depends on yesterday's).

Panel (Longitudinal) Data

The **same units** observed over **multiple time periods**.

Farm	Year	Output	Area	Labor
1	1995	510	13	28
1	1996	530	13	30
2	1995	720	18	42
2	1996	690	18	38

Example: the same rice farms tracked year after year.

Panel data combine the strengths of cross-section and time-series:

- Variation *across* units (like cross-section)
- Variation *within* units over time (like time-series)

$\Longrightarrow$ This combination is powerful for controlling unobserved differences between units.

Outline

1. Why Econometrics?
2. From Economic Models to Econometric Models
3. Correlation Is Not Causation
4. Data Types
5. The Econometrics Pipeline
6. What Lies Ahead

The Research Process

Econometric analysis follows a structured pipeline:

1. **Economic theory** identifies variables and hypothesized relationships
2. **Econometric model** specifies a functional form and adds an error term
3. **Data** are collected (cross-section, time-series, or panel)
4. **Estimation** produces numerical estimates of the parameters
5. **Diagnostics** check whether the model's assumptions hold
6. **Inference** draws conclusions, makes predictions, tests hypotheses

This course walks through each step. By the end, you will be able to run this pipeline yourself.

The Pipeline Applied: Food Expenditure

1. **Theory:** food is a normal good; spending increases with income
2. **Model:** $y = \beta_1 + \beta_2 x + e$, where y = food expenditure, x = income ($100s)
3. **Data:** survey of 200 households (cross-section)
4. **Estimation:** $\hat{\beta}_2 \approx 10.4$ (an extra $100 of income $\implies$ about $10.40 more spent on food)
5. **Diagnostics:** Is linearity reasonable? Is the spread of errors constant?
6. **Inference:** Is β_2 statistically different from zero? Can we construct a confidence interval?

We will return to this example throughout the course as we develop each step.

Outline

1. Why Econometrics?
2. From Economic Models to Econometric Models
3. Correlation Is Not Causation
4. Data Types
5. The Econometrics Pipeline
6. What Lies Ahead

Where This Course Is Going

Today you saw the big picture. The rest of the course fills in the details:

- **Chapters 2–3:** The simple linear regression model and its properties
- **Chapters 4–6:** Inference, prediction, and the multiple regression model
- **Chapters 7–9:** Indicator variables, heteroskedasticity, regression specification
- **Later chapters:** Time series, simultaneous equations, panel data

We estimated $\hat{\beta}_2 \approx 10.4$ today. But how confident should we be in that number? Could it be 5, or 15, and we just got unlucky with our sample? That is where we start next time.

Thank you!

jakeanderson@g.ucla.edu