

# Mostly Harmless Econometrics (Ch.1–3): Quick & Intuitive Study Page

*Core frame: causal inference = faking an experiment*

## The one mental picture

Every empirical project is trying to fake an experiment. The whole game is:

1. **Define the causal question** (what “if we changed  $X$ ?” means).
2. **Imagine the ideal experiment** you wish you could run.
3. **Find an identification strategy** that mimics that experiment using real-world variation.
4. **Do inference correctly** (standard errors, clustering, robustness).

This is the “Four FAQs” frame from Chapter 1: everything else is a tool for making (3) credible.

## Pass 1 (30–45 min): Learn the core causality engine

### A. Potential outcomes (two parallel universes)

For each unit  $i$ , there are two potential outcomes:

- $y_{1i}$ : outcome if treated,
- $y_{0i}$ : outcome if not treated.

You observe only one, depending on treatment status  $d_i \in \{0, 1\}$ . A key identity is:

$$y_i = y_{0i} + (y_{1i} - y_{0i}) d_i.$$

This cleanly separates what is observable from what is fundamentally missing.

### B. Selection bias decomposition (why naive comparisons lie)

The difference in observed means

$$E[y \mid d = 1] - E[y \mid d = 0]$$

splits into:

- **treatment effect on the treated**, plus
- **selection bias** (treated and untreated differ even without treatment).

Intuition (hospital example): hospitalized people look sicker because they were sicker *before*, even if hospitals help.

### C. Why random assignment is the gold standard

Random assignment makes  $d$  independent of  $(y_0, y_1)$ , so selection bias disappears and differences-in-means become causal.

If you only learn one thing: Causal inference = removing selection bias.

## Pass 2 (45–75 min): Regression as “organized comparisons,” not magic

### A. Regression summarizes conditional averages

The key object is the **Conditional Expectation Function (CEF)**:

$$E[y \mid X].$$

Regression is the **best linear approximation** to the CEF (even if the CEF is nonlinear).

**Don’t worship regression. Always ask:**

- What CEF is it approximating?
- Is that CEF causal?

### B. “Regression anatomy” (partialling out)

Each coefficient reflects the relationship between  $y$  and the part of  $x_k$  left after removing the variation explained by other controls. This is why adding controls changes coefficients.

## Pass 3 (30–45 min): Three causal landmines

### 1) CIA / selection-on-observables (when regression can be causal)

Causal regression needs a story like:

“Conditional on  $X$ , treatment is as-good-as-random.”

Formally: potential outcomes are independent of treatment given  $X$ .

### 2) Omitted Variable Bias (OVB) = why coefficients shift

Memorize the sentence version:

**Short = Long + (effect of omitted)  $\times$  (correlation of omitted with included).**

Use it as a sign test: if the omitted factor raises  $y$  and is positively correlated with  $x$ , the estimate is biased upward.

### 3) Bad controls (don’t control for consequences)

Controlling for a mediator breaks causality. Example:

schooling  $\rightarrow$  occupation  $\rightarrow$  wages.

If you control for occupation, you condition on an outcome of schooling and reintroduce selection problems.

**Rule of thumb:**

- **Good controls:** fixed before treatment (pre-treatment covariates).
- **Bad controls:** affected by treatment (post-treatment variables).

Timing matters.

## The fastest way to “feel” the book: one reusable checklist

For any chapter/example, force it into this template:

1. **Unit:** person? firm? school? country?
2. **Treatment  $d$ :** what is being “changed”?
3. **Outcome  $y$ :** what do we care about?
4. **Threat:** why treated and untreated differ (source of selection bias)?
5. **Fix:** randomization / natural experiment / CIA with controls / RD / IV / etc.
6. **Don’t-do:** what are bad controls here?
7. **Inference:** what SEs are needed (robust/clustered)?

Chapter 1 is essentially telling you to live inside this checklist.

## Super-short study plan (2 sessions)

### Session 1 (90 min)

- Ch1: Four FAQs (write them as headings in your notes).
- Ch2: Hospital example + selection bias decomposition + why random assignment works.

**Output:** one page titled “*Causality = fixing selection bias.*”

### Session 2 (90 min)

- Ch3: CEF + regression–CEF approximation (regression is not magic).
- CIA, OVB, Bad controls (these explain 80% of mistakes).

**Output:** one page titled “*When regression is causal (and when it lies).*”