Mostly Harmless Econometrics

Reading Group Discussion Guide

Chapters 1-2

Chapter 1: Questions about Questions

The authors emphasize three key points when formulating research questions:

- Questions should focus on causal effects rather than associations
- Research design matters more than sophisticated statistical methods
- Simple comparisons of means often provide the most compelling evidence

Four Questions About Your Research Questions

1. What is the causal relationship of interest?
2. What is the ideal experiment to capture the causal relationship of interest?
3. What is your identification strategy?
4. What is your mode of statistical inference?

The Fundamental Problem of Causal Inference

Causal inference requires comparing outcomes under different conditions, but we cannot observe counterfactuals. For instance, when studying the effect of college on earnings, we cannot observe what college graduates would have earned without attending college.

Discussion Questions

1. What distinguishes causal questions from purely descriptive ones?
2. What is a *counterfactual*?
3. What does *identification* mean?
4. Can all causal questions be answered by a hypothetical ideal experiment, as the authors asset?

Chapter 2: The Experimental Ideal

Randomized experiments represent the gold standard for causal inference because:

- Random assignment creates comparable treatment/control groups
- Large samples ensure statistical precision
- Clear treatment conditions enable interpretation
- Control groups provide valid counterfactuals

Ingredients of an Experiment

Experiments typically have two group, treated and control. The treated group gets a treatment while the control group does not. Treatment is the manipulation of interest administered by the researcher. For example:

- Medical trials: A new drug vs placebo
- Education: A teaching intervention vs standard curriculum
- Labor: Job training program vs no training
- Development: Cash transfer vs no transfer

The key is that the treatment must be well-defined, consistently administered, and under the researcher's control.

Selection Bias

Selection bias occurs when treatment and control groups differ systematically. In the college example, graduates likely differ from non-graduates in ability, motivation, and family background. These differences make it difficult to isolate education's effect on earnings.

Question 1: What is an example of selection bias?

Question 2: Sec 2.3 says, "selection bias amounts to correlation between the regression error term, n_i, and the regressor, D_i" – what is another word economists often use for this?

Why Randomization Solves Selection Bias

Randomization ensures that treatment assignment is independent of all observed and unobserved characteristics. When we randomly assign units to treatment and control groups:

- All systematic differences between groups are eliminated in expectation
- Any remaining differences are due to chance and diminish with sample size
- Both observed and unobserved confounders are balanced across groups

Discussion Questions

1. What makes a natural experiment "good enough" to approximate randomization?
2. Why are 'balance tests' used?
3. When is random assignment feasible or ethical?
4. What are the key threats to validity in your research context?

Next Week: Chapter 3, Making Regression Make Sense

Read 3.1 (Regression Fundamentals) and 3.2 (Regression and Causality) for next week. We will discuss 3.3 and 3.4 the week after.

Getting Ready for Regression

A linear equation / model of Y in terms of X:

$$Y = \alpha + \beta X + u. \tag{1}$$

Terminology:

- Y: Dependent Variable, Left Hand Side (LHS) variable
- X: Independent Variables, RHS variables, controls or control variables, regressors
- (α, β): Parameters, coefficients; α - intercept, β - slope, marginal effect
- u: unobservable, residual, error term... do not think of this as random noise!

Rant on Unobservables: The unobservable of the regression is everything you *do not* control for. This is my preferred term when I am thinking about the theoretical object of interest; I saw residual when I am literally talking about the prediction/fit error of an estimated model. It is a term for our model's error; that is, an *error term*. Economists always remember that there are economic agents making rational decisions behind the data. The 'error term' is not noise or randomness, but the result of actions we do not fully understand (or need to worry about in a given application). Assumptions we make about the error term (e.g., homoskedasticity, finite second moments, conditional exogeneity) are assumptions about the preferences and decisions of the agents of the underlying data.

Big Picture Idea: All data is from an underlying process of equilibrium forces, policies, constrains, preferences, and technology. Most things we want to estimate are not 'fixed' but functions of the context. When you go to estimate something, remember that the estimate is also an economic outcome. Sometimes that is exactly what we want (e.g., what is the effect of EITC payments on labor supply?) and sometimes we want to go deeper (e.g., how do we use the EITC policy change to estimate what would happen if we had a Negative Income Tax?).