Chapter 1: Outline of Causal Inference

Frontier Topics in Empirical Economics

Lecture notes based on slides by Zibin Huang
College of Business, Shanghai University of Finance and Economics

September 2025

Contents

[1]The author thanks Professor Chao Fu (University of Wisconsin-Madison) for introducing this example.

1 What Is This Course About?

Empirical economics is built on a simple ambition: to understand how the world works by using data. But data are silent about causation on their own. They record correlations—associations between variables—but the question of whether A causes B, and by how much, requires more than a regression coefficient. This course is about the gap between *statistical association* and *causal inference*, and the modern toolkit that allows researchers to bridge it responsibly.

A central theme is what we might call the "regression monkey" problem. A regression monkey runs regressions without understanding why. They use whatever variables are available, add controls when reviewers complain, and report results without any sense of what variation in the data is actually doing the identification work. They confuse the tools of statistics with economic reasoning. This is not an economist—it is a bad statistician borrowing economic vocabulary.

What separates serious empirical work from mechanical computation? Three things. First, a researcher must understand the *logic* behind every estimator they use: why it targets a causal parameter under some assumptions, and what those assumptions mean economically. Second, they must know the menu of tools available—instrumental variables, difference-in-differences, regression discontinuity, matching—well enough to match the method to the research question. Third, they must use economic theory as a guide: theory tells us which variables belong in a model, which direction the bias should run, and whether a result makes sense.

This chapter lays the conceptual foundation for the course. We begin with a motivating example that shows how a simple regression model is connected to a deeper structural model. We then introduce the two broad traditions in empirical economics—the structural (model-based) approach and the reduced-form (design-based) approach—and compare their relative strengths. The bulk of the chapter develops the potential outcome framework, the conditional expectation function and its relationship to regression, the problem of bad controls, and the logic of matching estimators.

2 Motivating Example: Female Labor Participation[2]

2.1 Two Levels of Empirical Analysis

To fix ideas, consider a classic question in labor economics: how does the wage rate affect the decision of women to enter the labor force? A researcher facing this question might immediately reach for a regression:

$$y_i = \beta_0 + \beta_1 w_i + u_i, \tag{1}$$

where $y_i \in \{0, 1\}$ is a labor-participation indicator and w_i is the offered wage. Estimating (1) by OLS gives what is called a *linear probability model* (LPM). This is an example of a *reduced-form* regression: it directly relates an observable outcome to an observable treatment variable, without modeling the mechanism connecting them.

But the name "reduced-form" already hints at something deeper. The regression is the visible tip of an iceberg. Underneath it lies a behavioral model of individual decision-making that the regression *reduces* to a single equation. Understanding this connection is important, because it clarifies what β_1 is estimating, when the regression is valid, and what it cannot tell us.

2.2 The Behavioral Model

Following the model due to Professor Chao Fu, suppose that female i chooses labor supply $l_i \in \{0, 1\}$ to maximize utility. Her problem is:

$$\max_{l_i} \; U_i(c_i,\; 1 - l_i) + \epsilon_{il}, \qquad \text{subject to} \quad c_i = w_i l_i, \tag{2}$$

where c_i is consumption, $1 - l_i$ is leisure, w_i is the wage, and ϵ_{il} is a taste shock that is unobservable to the researcher (though observed by the individual). The budget constraint $c_i = w_i l_i$ says that income equals wages times labor supply.

Because l_i is binary, we compare utilities under the two options. Individual i chooses to work ($l_i = 1$) if and only if the utility from working plus her taste shock for working

[2]The author thanks Professor Chao Fu (University of Wisconsin-Madison) for introducing this example.

exceeds the utility from not working:

$$U_i(w_i, 0) + \epsilon_{i1} \;\geq\; U_i(0, 1) + \epsilon_{i0}. \tag{3}$$

Rearranging, individual i works if and only if the difference in taste shocks falls below a threshold:

$$\epsilon_{i0} - \epsilon_{i1} \;<\; \epsilon^*(w_i), \qquad \text{where} \quad \epsilon^*(w_i) \equiv U_i(w_i, 0) - U_i(0, 1). \tag{4}$$

The threshold ϵ^* is the net utility gain from working (evaluated at the given wage) and it is increasing in w_i: higher wages make work more attractive, raising the threshold below which someone chooses to work.

2.3 The Working Probability and Two Research Approaches

Let $F_{\epsilon|w}$ denote the CDF of the shock difference $\epsilon_{i0} - \epsilon_{i1}$ conditional on w_i. The probability that individual i participates in the labor force, given her wage, is:

$$G(w) \;=\; \Pr(l = 1 \mid w) \;=\; \int_{-\infty}^{\epsilon^*(w)} dF_{\epsilon|w} \;=\; F_{\epsilon|w}(\epsilon^*(w))\,. \tag{5}$$

The function $G(w)$ is the object of interest—it tells us how labor supply responds to wages at the population level. Two very different empirical strategies can be used to learn about it.

The Reduced-Form / Design-Based Approach. The simplest approach is to assume that G is well approximated by a linear function:

$$G(w) \approx \beta_0 + \beta_1 w. \tag{6}$$

This gives the Linear Probability Model in (1). Regression then estimates $\beta_1 \approx \partial G/\partial w$—the effect of a one-unit increase in wage on the probability of participation. The key word is "reduced": this equation is derived (*reduced*) from the underlying structural model by imposing a linearity assumption and ignoring the internal structure. The researcher uses a research "design"—some source of exogenous variation in wages, such as a minimum wage change or a policy experiment—to estimate this parameter credibly. The method is called "design-based" because the identification

strategy is the center of the analysis.

The Structural / Model-Based Approach. An alternative is to directly estimate the utility function U and the shock distribution F from the data. If we parameterize these objects—say, with parameter vectors Θ^U for the utility function and Θ^F for the shock distribution—then each observation contributes to the likelihood:

$$L(\Theta^U, \Theta^F;\ \text{data}) \ = \ \prod_{i=1}^{N} F_\epsilon(\epsilon^*(w_i))^{l_i} \left[1 - F_\epsilon(\epsilon^*(w_i))\right]^{1-l_i}, \tag{7}$$

which is maximized by MLE to recover Θ^U and Θ^F simultaneously. This approach retrieves the primitive parameters directly.

To see a concrete example, suppose utility is linear in the relevant arguments: $U_i(w_i l_i, 1 - l_i) = \alpha w_i l_i + \phi(1 - l_i)$, and the shock difference $\epsilon_{i0} - \epsilon_{i1}$ follows a Type-I Extreme Value (T1EV) distribution. Under these assumptions, the probability of working becomes the familiar logistic function, and the likelihood (7) reduces to the logit log-likelihood:

$$L(\Theta^U, \Theta^F;\ \text{data}) = \prod_{i=1}^{N} \left(\frac{e^{\alpha w_i}}{e^{\alpha w_i} + e^{\phi}} \right)^{l_i} \times \left(\frac{e^{\phi}}{e^{\alpha w_i} + e^{\phi}} \right)^{1-l_i}. \tag{8}$$

The structural approach thus *recovers the choice structure directly*, giving us estimates of the utility parameters α and ϕ rather than a reduced-form slope.

3 Two Approaches: Structural vs. Reduced-Form

3.1 Internal and External Validity

Before comparing the two approaches in detail, we need two foundational concepts that recur throughout applied work.

Definition 3.1 (Internal Validity)**.** A causal estimate is *internally valid* if it correctly identifies the causal effect within the specific population and setting under study. Internal validity is about getting the right answer for the particular context from which the data were drawn.

Definition 3.2 (External Validity)**.** A causal estimate is *externally valid* if the finding can be credibly extrapolated to populations, settings, or time periods other than those

from which the data were drawn. External validity is about whether the result travels.

These two concepts are often in tension. An estimate that is internally valid for a specific experimental population may say little about a broader population. Conversely, a structural model that makes strong assumptions to achieve external generalizability may have low internal validity because those assumptions are unverifiable.

Heckman and Vytlacil (2007) usefully distinguishes three layers of policy evaluation, which map onto this taxonomy. Consider China's One Child Policy (OCP) as an example:

1. **Evaluation of a historical intervention.** What was the causal impact of the OCP on China's fertility rate during the period it was in force? This is an internal question: we want a credible estimate for the specific historical context.
2. **Forecasting the effect of a similar intervention in a new environment.** What would the demographic impact be if China were to reinstate the one-child policy in 2023? This is an external question: we are extrapolating a previous policy from past conditions to a new time period and possibly a different economic environment.
3. **Evaluation of a novel counterfactual never observed in history.** What would happen if the government mandated that all women give birth to at least one child? This is a further external question: we are considering a counterfactual that has never been observed anywhere. Thus, answering it requires a model of behavior—simply extrapolating from observed data is insufficient.

The first question can, in principle, be answered by a good design-based study. The second and third require some structural content, because we must extrapolate to situations outside the historical record.

3.2 The Structural Approach: Advantages and Limitations

The structural approach aims to recover the primitive parameters of the economy—utility functions, production functions, market structures—and use them to simulate counterfactuals. Its primary advantages are:

Deeper economic content. Structural estimates recover the decision-making process itself. We learn not just *what* happens when a policy changes, but *why*: how agents

respond given their preferences, constraints, and information. This economic depth makes results easier to interpret and harder to misuse.

External validity and resistance to the Lucas critique. The Lucas critique (1976) warns that reduced-form behavioral equations estimated under one policy regime will shift when the policy changes, because agents adjust their expectations and behavior. Structural models, by targeting deep parameters (preferences, technology) that are invariant to policy, are in principle immune to this critique. They can therefore support credible counterfactual analysis far from the observed data.

Limitations. The main cost is the need for strong, often untestable assumptions. The researcher must specify the functional form of utility, the distribution of unobserved shocks, and the market structure etc. These choices are not determined by the data; they reflect modeling decisions. If the model is misspecified, the resulting estimates inherit that misspecification. Internal validity can therefore be low, in the sense that the structural estimates may not be robust to changes in modeling assumptions.

3.3 The Reduced-Form / Design-Based Approach: Advantages and Limitations

The design-based approach targets a specific marginal effect: what is the causal impact of treatment variable A on outcome B? It does not model the mechanism connecting them; the causal effect is, in a sense, a "black box." The approach builds its credibility from a research design—a source of exogenous variation in A that allows identification of its causal effect.

High internal validity. Because the approach relies on clean exogenous variation rather than modeling assumptions, its estimates are credible for the population from which the data are drawn, provided the identifying variation is genuinely exogenous.

Limitations. The effects identified are typically *local*—they apply to the individuals whose treatment status was affected by the source of variation used for identification. External validity requires additional assumptions. Moreover, because the mechanism is not recovered, the results are hard to use for out-of-sample counterfactuals. And the approach remains vulnerable to the Lucas critique when the researcher tries to use a local reduced-form estimate to predict the effect of a large-scale policy change.

3.4 The Target of This Book

The two approaches introduced above reflect two fundamental modes of scientific inquiry: deduction and induction. These represent two complementary ways in which humans seek to understand the world. The structural approach is grounded in deduction: it incorporates economic theory and subjective prior knowledge into a model and derives implications from general principles to specific cases. In contrast, the reduced-form approach is based on induction: it relies primarily on observed data and aims to uncover empirical relationships with minimal reliance on theoretical structure. This philosophy is often summarized by the phrase "letting the data speak" in statistics.

Neither approach is inherently superior; each has its own strengths and limitations. In practice, they often complement one another. A well-trained economist should be able to work with both approaches and understand which framework is more appropriate for a given research question. A recent trend in empirical research is to combine these two approaches in order to leverage the strengths of both. In particular, researchers may first construct a structural model and then estimate its key parameters using design-based methods such as randomized controlled trials (RCTs) or difference-in-differences (DID). This strategy allows the model to be grounded in economic theory while ensuring that parameter estimates are credibly identified from the data.

This course focuses primarily on the reduced-form, or design-based, approach. Our goal is to provide students with a starting point for engaging in empirical economic research. In the later weeks of the course, we will also give a very brief introduction of the structural approach, with a particular focus on discrete choice models.

4 Potential Outcomes and the Randomized Controlled Trial

4.1 Correlation Is Not Causality

Consider the following data on health status and hospitalization drawn from a hypothetical survey: The data show that people who went to hospital have worse average health status than those who did not. Should we conclude that going to hospital makes you sicker? Of course not. People go to hospital *because* they are already sick. The

Group	Sample Size	Mean Health Status
Hospital	7,774	3.21
No hospital	90,049	3.93

Table 1: Health status by hospitalization status.

comparison is contaminated by the fact that hospitalization and health status are both driven by an underlying illness— the correlation between hospitalization and health status reflects selection into the hospital, not the causal effect of the hospital on health. This is a textbook example of *selection bias*: the treated and untreated groups are systematically different even before treatment.

This example crystallizes a fundamental challenge of empirical work. Observational data record outcomes that are realized under actual treatment decisions. But to identify a causal effect, we would need to observe the same individual both treated and untreated—a logical impossibility. The potential outcome framework, which we now develop, provides a formal language for thinking about this problem.

4.2 The Potential Outcome Framework

The potential outcome framework—also known as the Rubin Causal Model—is the dominant formalism for causal inference in economics (Rubin, 1974). Its central objects are the *potential outcomes* associated with each possible treatment status.

Definition 4.1 (Potential Outcomes)**.** Let $D_i \in \{0, 1\}$ be a binary treatment indicator for individual i, and let Y_i be an observable outcome. Define:

- Y_{0i}: the **potential outcome** that individual i would experience if *not* treated ($D_i = 0$), regardless of their actual treatment status.
- Y_{1i}: the **potential outcome** that individual i would experience if *treated* ($D_i = 1$), regardless of their actual treatment status.

The key feature of these objects is that they are defined independently of the actual treatment received. Whether we observe Y_{0i} or Y_{1i} depends on D_i, but both potential outcomes are, in principle, well-defined properties of individual i. The connection be-

tween potential outcomes and the observed outcome is given by the *switching equation*:

$$Y_i = \begin{cases} Y_{1i} & \text{if } D_i = 1 \\ Y_{0i} & \text{if } D_i = 0 \end{cases} = Y_{0i} + (Y_{1i} - Y_{0i})\, D_i. \tag{9}$$

The individual treatment effect for person i is the difference $Y_{1i} - Y_{0i}$: how much better or worse off would i be if treated versus untreated? This quantity is **fundamentally unobservable**. For any given individual, we observe the world either under treatment or under no treatment—never both simultaneously. This is the *fundamental problem of causal inference*: individual-level causal effects cannot be identified from a single observation. We can only work with averages.

4.3 Average Treatment Effects

Two population averages of the individual treatment effect are of primary interest.

Definition 4.2 (Average Treatment Effect and ATT)**.** The *Average Treatment Effect* (ATE) is:

$$\text{ATE} = E[Y_{1i} - Y_{0i}],$$

the expected treatment effect over a randomly drawn individual from the population. The *Average Treatment Effect on the Treated* (ATT) is:

$$\text{ATT} = E[Y_{1i} - Y_{0i} \mid D_i = 1],$$

the expected treatment effect conditional on actually receiving treatment.

The ATE and ATT differ when treated and untreated individuals have different potential outcomes on average—which is typical in observational settings where treatment is not random.

4.4 The Selection Bias Decomposition

In an observational study, the researcher typically observes the difference in mean outcomes between treated and untreated groups:

$$E[Y_i \mid D_i = 1] - E[Y_i \mid D_i = 0].$$

Using the switching equation (9), this naive comparison can be decomposed as follows. First, expand using the definitions:

$$E[Y_i \mid D_i = 1] - E[Y_i \mid D_i = 0] = E[Y_{1i} \mid D_i = 1] - E[Y_{0i} \mid D_i = 0].$$

Add and subtract $E[Y_{0i} \mid D_i = 1]$:

$$= \underbrace{E[Y_{1i} \mid D_i = 1] - E[Y_{0i} \mid D_i = 1]}_{\text{ATT}} + \underbrace{E[Y_{0i} \mid D_i = 1] - E[Y_{0i} \mid D_i = 0]}_{\text{Selection bias}}. \tag{10}$$

This decomposition is fundamental. The first term is the ATT—the causal effect of treatment on the treated individuals. The second term is *selection bias*: the difference in untreated potential outcomes between the group that chose to be treated and the group that did not. In the hospitalization example, $E[Y_{0i} \mid D_i = 1]$ is the health status the hospitalized individuals would have had if they had not gone to hospital. This counterfactual health status is lower than the health status of those who actually chose not to go to hospital, because people self-select into hospitalization based on being sick. Selection bias is therefore negative in that example—it causes the naive comparison to understate any beneficial effect of the hospital, and indeed to flip its sign entirely.

4.5 Randomization Eliminates Selection Bias

The elegance of a Randomized Controlled Trial (RCT) is that it eliminates selection bias by design. When treatment assignment is random—determined by a coin flip or lottery that is independent of any individual characteristic—the treatment indicator D_i is statistically independent of all individual attributes, including the potential outcomes:

$$D_i \perp\!\!\!\perp Y_{0i},\ Y_{1i}. \tag{11}$$

This is the *independence condition* for an ideal RCT. It implies that the two groups—treated and untreated—are on average identical in all respects, including their potential outcomes:

$$E[Y_{0i} \mid D_i = 1] = E[Y_{0i} \mid D_i = 0] = E[Y_{0i}].$$

The selection bias term in (10) therefore vanishes, and the simple difference-in-means identifies the average treatment effect:

$$\begin{aligned} E[Y_i \mid D_i = 1] - E[Y_i \mid D_i = 0] &= E[Y_{1i} \mid D_i = 1] - E[Y_{0i} \mid D_i = 1] \\ &= E[Y_{1i} - Y_{0i} \mid D_i = 1] \\ &= E[Y_{1i} - Y_{0i}] = \text{ATE}. \end{aligned} \tag{12}$$

The last equality holds because, under randomization, D_i is independent of (Y_{0i}, Y_{1i}), so the distribution of potential outcomes is the same in the treated and untreated groups. ATT equals ATE under randomization.

This is why the RCT is the "gold standard" of causal inference. It does not require any assumptions about how treatment affects outcomes; it does not require any model of individual behavior. The only requirement is that treatment is genuinely randomly assigned.

5 Regression, the CEF, and Causal Inference

The potential outcome framework clarifies the *target* of causal inference. But in practice, most empirical work uses regression rather than a simple difference-in-means, because we rarely have clean randomization and we need to control for confounding variables. This section develops the relationship between regression, the Conditional Expectation Function (CEF), and causal inference, and asks: when does a regression coefficient carry a causal interpretation?

5.1 The Conditional Expectation Function

Definition 5.1 (Conditional Expectation Function)**.** Given a random outcome Y_i and a vector of conditioning variables X_i, the Conditional Expectation Function (CEF) is:

$$E[Y_i \mid X_i = x] = \int t \; f_Y(t \mid X_i = x) \; dt, \tag{13}$$

where $f_Y(\cdot \mid X_i = x)$ is the conditional density of Y_i given $X_i = x$.

The CEF is a population-level object. It answers the question: "in the population, what is the average value of Y among all individuals who share the covariate value

$X_i = x$?" It is a description of the data-generating process, not an estimator.

Two important properties deserve emphasis.

Property 1: The CEF decomposition. We can always write:

$$Y_i = E[Y_i \mid X_i] + \varepsilon_i, \tag{14}$$

where the residual ε_i satisfies $E[\varepsilon_i \mid X_i] = 0$ by construction. This is not an assumption—it is an algebraic identity. It holds for any joint distribution of (Y_i, X_i).

Property 2: The CEF is not necessarily causal. The CEF is the best statistical predictor of Y_i given X_i, but prediction is not causation. $E[Y_i \mid X_i]$ tells us what value of Y to expect for a person with covariates X_i; it does not tell us what would happen to that person's Y if we changed their X_i.

Theorem 5.2 (CEF Minimizes MSE)**.** *Let $m(X_i)$ be any measurable function of X_i. Then:*

$$E[Y_i \mid X_i] = \underset{m(X_i)}{\arg\min}\ E\left[(Y_i - m(X_i))^2\right]. \tag{15}$$

That is, the CEF is the Minimum Mean Squared Error (MMSE) predictor of Y_i given X_i, within the class of all measurable functions.

Proof. Write $Y_i = E[Y_i \mid X_i] + \varepsilon_i$. For any function m:

$$\begin{aligned} E[(Y_i - m(X_i))^2] &= E[(E[Y_i \mid X_i] - m(X_i) + \varepsilon_i)^2] \\ &= E[(E[Y_i \mid X_i] - m(X_i))^2] + 2\,E[(E[Y_i \mid X_i] - m(X_i))\varepsilon_i] + E[\varepsilon_i^2]. \end{aligned}$$

The cross-product term is zero because $E[\varepsilon_i \mid X_i] = 0$ implies $E[(E[Y_i|X_i] - m(X_i))\varepsilon_i] = 0$. Therefore:

$$E[(Y_i - m(X_i))^2] = \underbrace{E[(E[Y_i \mid X_i] - m(X_i))^2]}_{\geq 0} + E[\varepsilon_i^2].$$

This is minimized by setting $m(X_i) = E[Y_i \mid X_i]$, which makes the first term zero. □

5.2 Linear Regression as Best Linear Predictor

The CEF minimizes MSE over *all* functions of X_i. Linear regression restricts attention to *linear* functions and finds the best one. Formally, the population regression coefficient β is defined as:

$$\beta = \arg\min_{b \in \mathbb{R}^k} \; E\left[(Y_i - X_i'b)^2\right]. \tag{16}$$

The first-order condition (the *moment condition* of regression) is obtained by differentiating with respect to b and setting to zero:

$$E[X_i(Y_i - X_i'\beta)] = 0. \tag{17}$$

Let $\varepsilon_i = Y_i - X_i'\beta$. Then (17) says $E[X_i\varepsilon_i] = 0$: the residual is uncorrelated with every regressor. Solving directly:

$$\beta = E[X_iX_i']^{-1}E[X_iY_i], \tag{18}$$

provided $E[X_iX_i']$ is invertible (the rank condition).

An important notational distinction must be kept clear. The *population* coefficient β in (18) is an identification concept: it exists in the population and characterizes the best linear predictor. The *OLS estimator* $\hat{\beta}_{\text{OLS}}$ is an estimation concept:

$$\hat{\beta}_{\text{OLS}} = (X'X)^{-1}X'Y, \tag{19}$$

where X is the $n \times k$ matrix of observed data. The vector X_i is a random variable; X is its matrix of realizations. As sample size grows, $\hat{\beta}_{\text{OLS}} \to \beta$ under standard regularity conditions.

We usually use the OLS estimator to estimate the regression coefficient. Though, in principle, there can be other estimators. The sample OLS formula $\hat{\beta}_{\text{OLS}} = (X'X)^{-1}X'Y$ is also called the *sample analogue* of the population formula $\beta = E[X_iX_i']^{-1}E[X_iY_i]$: it replaces population expectations with their sample counterparts.

5.3 The Hierarchy of Assumptions

The difference between regression, the CEF, and a causal model lies in the strength of the assumption placed on the residual e in the model $Y = f(D) + e$:

- **Linear regression** requires only that the residual is *uncorrelated* with the regressors: $E[De] = 0$. This is the weakest condition.
- **The CEF** requires *mean independence*: $E[e \mid D] = 0$. This is stronger than uncorrelatedness. It says the expected residual is zero not just on average but for every value of D.
- **A causal model under the CIA** requires *full independence*: $e \perp\!\!\!\perp D$, equivalently $D_i \perp\!\!\!\perp Y_{0i}, Y_{1i}$. This is the strongest condition.

Hierarchy of Assumptions

$$\underbrace{e \perp\!\!\!\perp D}_{\text{Causal model (CIA)}} \succ \underbrace{E[e \mid D] = 0}_{\text{CEF}} \succ \underbrace{E[De] = 0}_{\text{Linear regression}}$$

Each condition implies all weaker conditions below it, but not conversely.

What is the relationship between the CEF and linear regression? When the CEF is itself a linear function of X_i, then the two coincide exactly: linear regression recovers the CEF. When the CEF is nonlinear, linear regression still has a useful interpretation: it is the best linear approximation to the CEF, in the sense that it minimizes the expected squared error $E[(E[Y_i \mid X_i] - X_i'\beta)^2]$ over all linear functions. This means regression is always meaningful as a prediction tool, even if the CEF is nonlinear.

It is also worth noting a useful special case: when D is a binary variable (a dummy), linear regression with only D as a regressor is identical to the CEF. This is because the CEF of Y on a binary D must be a linear function of D—there are only two values of D, so any function of D is linear in D.

5.4 When Does a Regression Coefficient Identify a Causal Effect?

A researcher can always compute a regression coefficient, provided the data satisfy the rank condition. But the fundamental question is: what is β *identifying*? Is it a causal parameter? The answer depends on whether the assumptions of the relevant causal

model are satisfied. We consider two important cases.

Case 1: Unconditional Randomization. Suppose we have data from an experiment in which treatment D_i was randomly assigned, so that $D_i \perp\!\!\!\perp Y_{0i}, Y_{1i}$. Consider the regression:

$$Y_i = \alpha + \rho D_i + \varepsilon_i. \tag{20}$$

Under randomization (and with D_i binary), the CEF is linear in D_i, so the regression coefficient ρ equals the CEF slope. Computing directly:

$$\begin{aligned}
\rho &= E[Y_i \mid D_i = 1] - E[Y_i \mid D_i = 0] \\
&= E[Y_{1i} \mid D_i = 1] - E[Y_{0i} \mid D_i = 0] \\
&= E[Y_{1i}] - E[Y_{0i}] \quad \text{(by independence)} \\
&= E[Y_{1i} - Y_{0i}] = \text{ATE}.
\end{aligned} \tag{21}$$

Thus, under unconditional randomization, the regression coefficient on treatment directly estimates the ATE. Note that we also have $\rho = \text{ATT}$ under randomization, since the distribution of (Y_{0i}, Y_{1i}) is independent of D_i.

Case 2: Conditional Randomization (CIA / Selection on Observables). In observational studies, treatment is rarely unconditionally random. The *Conditional Independence Assumption* (CIA)—also known as the *selection-on-observables* assumption—is the key identifying assumption for regression-based causal inference:

$$D_i \;\perp\!\!\!\perp\; Y_{0i},\; Y_{1i} \;\Big|\; X_i. \tag{22}$$

This says: conditional on the observed covariates X_i, treatment assignment is as good as random. The potential outcomes are independent of treatment once we account for X_i. Equivalently, after controlling for X_i, there is no remaining selection bias.

Under the CIA, the appropriate regression model includes controls:

$$Y_i = \alpha + \rho_r D_i + X_i'\gamma + \nu_i. \tag{23}$$

Homogeneous treatment effects. If the treatment effect is the same for all individ-

uals (constant $\rho_r = Y_{1i} - Y_{0i}$), then under the CIA:

$$\begin{aligned}
\rho_r &= E[Y_i \mid X_i, D_i = 1] - E[Y_i \mid X_i, D_i = 0] \\
&= E[Y_{1i} \mid X_i, D_i = 1] - E[Y_{0i} \mid X_i, D_i = 0] \\
&= E[Y_{1i} \mid X_i] - E[Y_{0i} \mid X_i] \quad \text{(by CIA)} \\
&= E[Y_{1i} - Y_{0i} \mid X_i] = E[Y_{1i} - Y_{0i}] = \text{ATE}.
\end{aligned} \tag{24}$$

The regression coefficient identifies the treatment effect.

Heterogeneous treatment effects. When the treatment effect $\delta_x \equiv E[Y_{1i} - Y_{0i} \mid X_i = x]$ varies across individuals with different covariates, the analysis is more subtle. Define the within-cell treatment effect as:

$$\delta_x = E[Y_i \mid X_i = x, D_i = 1] - E[Y_i \mid X_i = x, D_i = 0],$$

and define the within-cell treatment variance:

$$\sigma_D^2(X_i) \equiv E\Big[(D_i - E[D_i \mid X_i])^2 \;\Big|\; X_i\Big].$$

Then, as shown in Angrist and Pischke (2009, Chapter 3.3.1), the regression coefficient from (23) is:

$$\rho_r = \frac{E[\sigma_D^2(X_i)\,\delta_x]}{E[\sigma_D^2(X_i)]}. \tag{25}$$

This is a **treatment-variance weighted average** of the group-level ATEs δ_x. The weight assigned to each cell x is proportional to the within-cell variance of treatment. Cells where all units are treated (or all are untreated) contribute zero weight because they contain no variation in treatment to exploit. Cells with roughly equal numbers of treated and untreated units—where D_i has the highest variance—contribute the most.

To understand the intuition behind this result: OLS is projecting Y_i onto D_i after partialing out X_i. The variation in D_i that regression uses is the within-cell deviation $D_i - E[D_i \mid X_i]$. Cells with more of this variation naturally receive more weight. When treatment effects are homogeneous, this weighting is harmless. When they are heterogeneous, the OLS coefficient is a weighted average of local treatment effects, with weights determined by the distribution of treatment propensity—not by the distribution of the population.

Summary: Regression and Causal Inference

- The CEF is the best predictor of Y given X over all functions.
- Linear regression is the best *linear* predictor; it is the best linear approximation to the CEF even if the CEF is nonlinear.
- A regression coefficient is a causal effect only when the CIA holds.
- Under CIA with homogeneous treatment effects, the regression coefficient equals the ATE.
- Under CIA with heterogeneous treatment effects, the regression coefficient is a treatment-variance weighted average of group-level ATEs—a meaningful but not necessarily population-representative causal parameter.

6 Simpson's Paradox, Omitted Variable Bias, and Bad Controls

The CIA is the most important assumption we need in causal inference. A natural question then arises: is it always better to control for more variables? The answer is no. What variables should we control for, and what variables should we not? In this section, we briefly discuss this issue.

6.1 Simpson's Paradox

A recurring source of confusion in empirical work is the phenomenon known as Simpson's Paradox: an association between treatment and outcome that reverses direction when a third variable is controlled for. Consider the following data on two COVID-19 treatments:

Treatment	Mild Condition	Severe Condition	Overall
A	15% (210/1400)	30% (30/100)	16% (240/1500)
B	10% (5/50)	20% (100/500)	19% (105/550)

Table 2: Death rates by treatment and patient condition (COVID-19 example).

Looking at overall death rates, treatment A appears better: 16% versus 19%. Yet within each condition group, treatment A is worse: 15% vs. 10% for mild patients, and

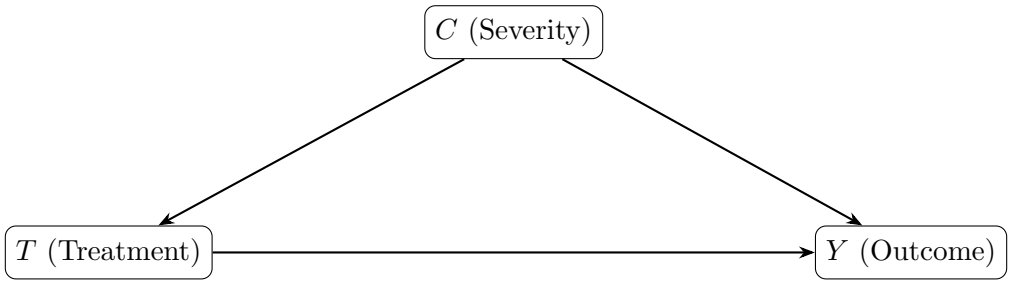
30% vs. 20% for severe patients. How can A be simultaneously worse in every subgroup but better overall?

The resolution lies in the composition of the groups. Treatment A is more associated with mild patients (1400 out of 1500), who have low baseline death rates. Treatment B is more associated with severe patients (500 out of 550), who have high baseline death rates.

Which comparison is correct? This is a question equivalent to whether we should control for mild/severe condition when regressing death rate on treatment. The answer depends on the causal structure among these variables.

6.2 Confounders: When to Control

Case 1: The covariate C is a cause of treatment T. Suppose condition severity C affects both the treatment a patient receives T and the health outcome Y: the arrows run $C \to T$ and $C \to Y$, with $T \to Y$ as well. In this case, C is a *confounder*: it is a pre-existing characteristic that affects both treatment receipt and the outcome. If we compare treated and untreated patients without controlling for C, we are comparing groups that differed in C even before treatment was assigned. The difference in outcomes picks up not just the effect of T but also the effect of C—this is *Omitted Variable Bias*.

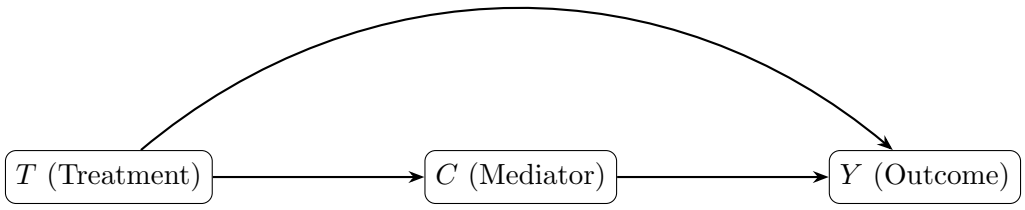


Case 1: C is a confounder. Must control for C.

The solution is to control for C: compare treated and untreated patients within each condition group. This is what the within-group comparisons in the table above show. Controlling for severity, treatment B is consistently better, and this comparison correctly isolates the causal effect of treatment.

Case 2: Treatment T is a cause of the covariate C. Now suppose instead that the treatment T itself causes the condition status C: for example, T may cause an

immune response that leads to a "severe" classification on some post-treatment health index. The causal graph now has arrows $T \to C$ and $T \to Y$ (direct effect) and $C \to Y$ (mediated effect through C). In this case, C is a *mediator*—it lies on the causal pathway from T to Y.



*Case 2: C is a mediator. Do **not** control for C.*

If we control for C in this setting, we are holding constant something that treatment itself caused. We are effectively asking: "what is the effect of T on Y, holding C fixed?"—but C cannot be held fixed independently of T, since T determines C. Controlling for C blocks the indirect causal channel from T to Y running through C, and the regression now captures only the direct effect of T on Y (not through C). If the goal is to estimate the total causal effect of T on Y, this is *wrong*. This is the **Bad Control Problem**.

Rule of Thumb: Pre-determined vs. Post-determined Variables

Pre-determined variables—those that are determined before the treatment is assigned and are not themselves caused by treatment—are generally safe to control for. Controlling for pre-determined confounders helps satisfy the CIA.

Post-determined variables—those that are caused by the treatment, or that happen after treatment assignment—should generally *not* be controlled for. Controlling for a mediator blocks the causal channel and introduces bias. Never control for a variable that lies on the causal pathway from treatment to outcome.

Let us apply this rule to three examples:

(a) Y = wage, D = education, X = natural ability. Natural ability is determined at birth, before any educational decision. It affects both education (more able individuals get more schooling) and wages (more able workers are more productive). X is a pre-determined confounder: we *should* control for it. Failing to do so overstates the return to education.

(b) Y = wage, D = education, X = labor participation decision. Labor market participation is a post-education choice: it happens *after* and is partly *caused by* D (education affects the decision to participate). X is post-determined: we should *not* control for it. Conditioning on labor participation selects a non-random subset of workers and introduces sample selection bias.

(c) $Y = \text{GDP}_{t+1}$, D = R&D expenditure at time t, X = trade volume at time t+1. Trade volume at $t+1$ is post-determined: it is realized after the R&D investment at t and may itself be a consequence of innovation. We should *not* control for it.

The DAG framework, introduced formally in Chapter 4, provides a systematic and rigorous tool for resolving such questions in general settings.

7 Matching

7.1 The Idea of Matching

Regression is only one way to use the CIA to estimate causal effects. A conceptually simpler and nonparametric alternative is *matching*. The basic idea is direct: find, for each treated unit, a control unit with *identical* observable characteristics X_i, and compare their outcomes. Because the two units have the same X_i, any difference in outcomes can be attributed to treatment alone (under the CIA), not to differences in the covariates.

Matching is attractive because it is nonparametric—it makes no functional form assumption about how X_i enters the outcome equation. It also has a simple and transparent interpretation: we are computing a within-cell average of the treatment effect, then averaging those within-cell effects across cells.

7.2 Formal Derivation of the Matching Estimator

Under the CIA $(Y_{0i}, Y_{1i} \perp\!\!\!\perp D_i \mid X_i)$, we can derive the matching estimators for ATT and ATE. For the ATT:

$$\begin{aligned}
\delta_{\text{ATT}} &= E[Y_{1i} - Y_{0i} \mid D_i = 1] \\
&= E\Big[E[Y_{1i} - Y_{0i} \mid X_i, D_i = 1] \,\Big|\, D_i = 1\Big] \quad \text{(law of iterated expectations)} \\
&= E\Big[E[Y_{1i} \mid X_i, D_i = 1] - E[Y_{0i} \mid X_i, D_i = 1] \,\Big|\, D_i = 1\Big] \\
&= E\Big[E[Y_i \mid X_i, D_i = 1] - E[Y_i \mid X_i, D_i = 0] \,\Big|\, D_i = 1\Big]. \qquad (26)
\end{aligned}$$

In the last step, we used the CIA: conditional on X_i, the treatment group and the control group have the same untreated potential outcome Y_{0i}, so $E[Y_{0i} \mid X_i, D_i = 1] = E[Y_{0i} \mid X_i, D_i = 0] = E[Y_i \mid X_i, D_i = 0]$. Defining the within-cell ATT as:

$$\hat{\delta}_x = E[Y_i \mid X_i = x, D_i = 1] - E[Y_i \mid X_i = x, D_i = 0],$$

the ATT becomes $\delta_{\text{ATT}} = E[\delta_x \mid D_i = 1]$, estimated by the sample analog:

$$\hat{\delta}_{\text{ATT}} = \sum_x \hat{\delta}_x \, \hat{P}(X_i = x \mid D_i = 1). \qquad (27)$$

The ATT matching estimator weights each cell by the fraction of treated units that fall in that cell. Similarly, the ATE matching estimator weights by the unconditional share of each cell in the population:

$$\hat{\delta}_{\text{ATE}} = \sum_x \hat{\delta}_x \, \hat{P}(X_i = x). \qquad (28)$$

7.3 Matching, Regression, and Weighting

A striking insight from Angrist and Pischke (2009) is that regression is itself a particular weighted matching estimator. All three estimators compute within- cell treatment effects $\hat{\delta}_x$ and then aggregate them, but they differ in the weights used:

The ATT matching estimator asks: "on average, what is the treatment effect for a randomly drawn treated individual?"—so it weights by the distribution of covariates among the treated. The ATE matching estimator asks: "on average, what is the

Estimator	Formula	Weighting Principle
Matching (ATT)	$\sum_x \hat{\delta}_x \, \hat{P}(X_i = x \mid D_i = 1)$	Treated units' covariate distribution
Matching (ATE)	$\sum_x \hat{\delta}_x \, \hat{P}(X_i = x)$	Overall covariate distribution
Regression	$\frac{\sum_x \hat{\sigma}^2_D(X_i)\, \hat{\delta}_x}{\sum_x \hat{\sigma}^2_D(X_i)}$	Within-cell treatment variance

Table 3: The three estimators as weighted averages of cell-level treatment effects $\hat{\delta}_x$.

treatment effect for a randomly drawn individual from the whole population?"—so it weights by the population distribution. The regression estimator up-weights cells where treatment variance is high (where roughly half the units are treated) and down-weights cells where almost everyone is treated or untreated.

When treatment effects are homogeneous (constant $\delta_x = \delta$ for all x), all three estimators identify the same parameter δ. When treatment effects are heterogeneous, they identify different weighted averages, and the choice of estimator should reflect the policy question of interest.

8 Propensity Score Matching

8.1 The Curse of Dimensionality

Matching on covariates X_i is straightforward when X_i is a scalar or a small discrete vector: we simply find pairs of treated and untreated units with the same value of X_i and compare them. The method runs into serious difficulties, however, when X_i is high-dimensional.

Consider estimating the college wage premium while satisfying the CIA. The covariates needed to make D_i (college attendance) conditionally independent of potential wages might include gender, race, nationality, birth weight, IQ, parental income, parental education, and many more. With ten such covariates, each taking many values, the number of possible combinations of X_i is enormous. In a dataset of even 10,000 observations, most covariate cells will contain zero or one observation. We cannot compute $\hat{\delta}_x$ because we cannot find treated and untreated units with the same X_i. This is the *curse of dimensionality*: the data become too sparse to support matching as the

dimension of X_i grows.

Regression side-steps this problem by imposing linearity: instead of matching within each cell, it extrapolates using the linear functional form. But matching is more flexible and transparent. Is there a way to do nonparametric matching even with high-dimensional covariates? The answer is yes, and it is provided by the Propensity Score Theorem.

8.2 The Propensity Score Theorem

Definition 8.1 (Propensity Score)**.** The *propensity score* is the conditional probability of receiving treatment given covariates:

$$P(X_i) = \Pr(D_i = 1 \mid X_i). \tag{29}$$

The propensity score is a single number that summarizes the relevant information in X_i for the purpose of matching. Its key property is given by the following theorem.

Theorem 8.2 (Propensity Score Theorem, Rosenbaum and Rubin 1983)**.** *Assume:*

(i) ***CIA****:* $Y_{1i}, Y_{0i} \perp\!\!\!\perp D_i \mid X_i$.
(ii) ***Overlap****:* $0 < P(D_i = 1 \mid X_i) < 1$ *for all* X_i.

Then $Y_{1i}, Y_{0i} \perp\!\!\!\perp D_i \mid P(X_i)$.

Proof sketch. It suffices to show $\Pr(D_i = 1 \mid Y_{1i}, Y_{0i}, P(X_i)) = P(X_i)$. By iterated expectations and CIA:

$$\begin{aligned}
E[D_i \mid Y_{1i}, Y_{0i}, P(X_i)] &= E\Big[E[D_i \mid Y_{1i}, Y_{0i}, X_i] \,\Big|\, Y_{1i}, Y_{0i}, P(X_i)\Big] \\
&= E\Big[E[D_i \mid X_i] \,\Big|\, Y_{1i}, Y_{0i}, P(X_i)\Big] \quad \text{(by CIA)} \\
&= E\Big[P(X_i) \,\Big|\, Y_{1i}, Y_{0i}, P(X_i)\Big] \\
&= P(X_i).
\end{aligned}$$

Since D_i is binary, $E[D_i \mid Y_{1i}, Y_{0i}, P(X_i)] = P(X_i)$ implies $D_i \perp\!\!\!\perp (Y_{1i}, Y_{0i}) \mid P(X_i)$. □

The theorem says: if we condition on the scalar propensity score $P(X_i)$ rather than the full vector X_i, the CIA is preserved. This is powerful: instead of matching on all

the dimensions of X_i (infeasible when X_i is high-dimensional), it suffices to match on the single scalar $P(X_i)$. The curse of dimensionality is resolved by collapsing X_i into a one-dimensional sufficient statistic for selection into treatment.

8.3 Implementing PSM

In practice, the propensity score $P(X_i)$ is unknown and must be estimated. A typical PSM implementation proceeds in two steps:

1. **Estimate the propensity score.** Regress D_i on X_i using a binary model—logistic regression (logit) is the most common choice, though probit or even the linear probability model are also used. Obtain predicted probabilities $\hat{P}(X_i)$.
2. **Match and compare.** For each treated unit, find one or more control units with similar values of $\hat{P}(X_i)$—either via nearest-neighbor matching, kernel matching, or stratification on $\hat{P}(X_i)$. Average the within-match differences in Y_i to estimate ATT or ATE.

8.4 PSM Compared with Regression

Regression and PSM are both valid estimators under the CIA, but they have different properties:

- Regression does not suffer from the curse of dimensionality because the linearity assumption acts as a regularizer: it extrapolates smoothly across covariate cells using the fitted linear surface.
- PSM avoids functional form assumptions about the outcome equation, at the cost of requiring a model for the treatment probability.
- A natural compromise is to run a regression of Y_i on D_i and $\hat{P}(X_i)$ (controlling for the propensity score rather than all covariates individually). This "double robustness" strategy provides some protection against misspecification.
- In practice, Angrist and Pischke (2009) argue that regression generally dominates PSM because the implementation of PSM involves discretionary choices (which estimator for $P(X_i)$? how many neighbors? what bandwidth?) that are not standardized and can affect results.

Critical Warning: PSM Cannot Solve Endogeneity

Both regression and PSM assume CIA. They reduce the *dimensionality* of the matching problem, but they do not relax the identifying assumption. If there are unobserved variables that affect both treatment assignment and the outcome—that is, if selection is not fully explained by the observed X_i—then neither OLS nor PSM identifies a causal effect.

PSM CANNOT solve the endogeneity problem. It is a method for implementing the CIA efficiently, not a solution to violations of the CIA. If you cannot defend the CIA on substantive grounds, neither regression nor PSM will save you.

A Homework Problems

1. **Weighted average under RCT.** Consider the expression for the regression coefficient under heterogeneous treatment effects:

$$\rho_r = \frac{E[\sigma_D^2(X_i)\,\delta_x]}{E[\sigma_D^2(X_i)]}.$$

   Show what this expression reduces to when unconditional independence holds, i.e., when $D_i \perp\!\!\!\perp Y_{1i}, Y_{0i}$ (as in an RCT without covariates). Interpret the result.

2. **Selection bias in observational data.** The data below show test scores for students who attended a tutoring program ($D_i = 1$) and those who did not ($D_i = 0$):

Group	n	Mean score
Tutored	200	75
Not tutored	800	70

   (a) What is the naive estimate of the effect of tutoring?

   (b) Students who sign up for tutoring are likely to be more motivated. Write down the selection bias formally using the potential outcome notation. What is the direction of the bias?

   (c) Describe an ideal RCT design that would identify the causal effect of tutoring.

Practice Example: Estimating the ATE in a Randomized Controlled Trial

Economic Setup

A government runs a job training program and randomly assigns 1,000 unemployed workers to either participate in the program ($D_i = 1$) or remain in the control group ($D_i = 0$). Monthly wages are measured six months after the program ends. Three pre-treatment characteristics are also recorded: age (years), education (years of schooling), and gender.

The goal is to estimate the *average treatment effect* (ATE) of the program on monthly wages. This example illustrates the main message of Chapter 1: under random assignment, a simple OLS regression of the outcome on the treatment indicator identifies the ATE.

Data-Generating Process

The true model used to simulate the data is:

$$\begin{aligned} \text{wage}_i &= 1500 + 500 \cdot D_i + 80 \cdot \text{educ}_i + 30 \cdot (\text{age}_i - 35) - 200 \cdot \text{female}_i + \varepsilon_i, \qquad (30) \\ \varepsilon_i &\sim \mathcal{N}(0,\ 300^2), \end{aligned}$$

with covariates drawn independently: $\text{age}_i \sim \text{Uniform}[18, 60]$, $\text{educ}_i \sim \text{Uniform}[8, 20]$, $\text{female}_i \sim \text{Bernoulli}(0.5)$.

The **true ATE is \$500 per month**. Treatment D_i is drawn independently of all covariates and the error term, so:

$$D_i \ \perp\!\!\!\perp \ (Y_{0i},\ Y_{1i},\ \text{age}_i,\ \text{educ}_i,\ \text{female}_i).$$

Regression Specifications

We estimate the ATE with two OLS specifications:

$$\text{(1)} \quad \text{wage}_i = \alpha + \rho\, D_i + e_i \qquad \text{(Simple OLS, no controls)} \tag{31}$$

$$\text{(2)} \quad \text{wage}_i = \alpha + \rho\, D_i + \beta_1\, \text{age}_i + \beta_2\, \text{educ}_i + \beta_3\, \text{female}_i + v_i \qquad \text{(OLS with controls)} \tag{32}$$

Under randomization, the coefficient ρ in specification (1) equals the ATE by the selection-bias decomposition: since $D_i \perp\!\!\!\perp Y_{0i}$, the selection bias term $E[Y_{0i} \mid D_i = 1] - E[Y_{0i} \mid D_i = 0] = 0$, and the naive difference-in-means recovers the ATT, which equals the ATE.

Adding controls in specification (2) does not change the point estimate of ρ (because D_i is already independent of the controls), but it absorbs the variance in wages attributable to age, education, and gender, reducing the residual variance $\text{Var}(v_i)$ and therefore shrinking the standard error of $\hat{\rho}$.

Expected Results

Table 4 presents the expected regression output (representative sample results; exact values will vary across simulation runs with different seeds).

	(1) Simple OLS	(2) With Controls
D_i (job training)	≈ 500	≈ 500
	(SE ≈ 19)	(SE ≈ 12)
age_i	—	≈ 30
educ_i	—	≈ 80
female_i	—	≈ -200
N	1,000	1,000
R^2	low	higher

Table 4: Expected regression results. True ATE = \$500/month. SE reported in parentheses. Coefficients on controls recover their true values from the DGP (30).

Key Learning Points

Running this example teaches three lessons that mirror the theoretical discussion in Sections 4 and 5.

1. **Identification under randomization.** The coefficient on D_i in specification (1) is approximately \$500, confirming that simple OLS identifies the ATE when treatment is randomly assigned. There is no need for controls to get an unbiased estimate.
2. **Invariance of the estimate to controls.** The coefficient on D_i does not change materially between specifications (1) and (2). This is the expected behavior under a genuine RCT: because D_i is independent of the covariates, adding them does not affect what variation in D_i is doing the work. A large change in the coefficient when controls are added would be a warning sign that the treatment was not truly random.
3. **Efficiency gain from pre-treatment controls.** The standard error of $\hat{\rho}$ falls substantially in specification (2). Age, education, and gender together explain a large share of the variance in wages. Removing this variance from the residual tightens the confidence interval on the treatment effect without introducing any bias.

Balance Check

Before running regressions, it is good practice to verify that randomization produced balanced groups. A balance table should report means of pre-treatment covariates by treatment status, together with p-values from t-tests for equality of means. Under a genuine RCT with $N = 1{,}000$ and a 50/50 split, we expect all p-values to exceed 0.1, indicating no statistically significant differences between treated and control groups before treatment.

Files

- **Replication code**: `src/chapter 1/ch1_rct.do`
  Generates the dataset, runs the balance check, estimates both regressions, and prints a comparison table. Set the `global root` path at the top of the file before running.

- **Simulated data**: `data/chapter 1/ch1_rct.dta`
  Created automatically by running `ch1_rct.do` with `set seed 20250912`. Contains 1,000 observations and five variables: `wage`, `D`, `age`, `educ`, `female`.

References

Angrist, J. D. and J.-S. Pischke (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Heckman, J. J. and E. Vytlacil (2007). Econometric evaluation of social programs, part I: Causal models, structural models, and econometric policy evaluation. In J. Heckman and E. Leamer (eds.), *Handbook of Econometrics*, Vol. 6B, pp. 4779–4874. Elsevier.

Lucas, R. E. Jr. (1976). Econometric policy evaluation: A critique. *Carnegie-Rochester Conference Series on Public Policy*, 1, 19–46.

Rosenbaum, P. R. and D. B. Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1), 41–55.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5), 688–701.