Notes on Applied Econometrics

From Dr. Aimee Chin's Lectures, Compiled by Brian Murphy*†

Contents

*These notes are from my time as a student in the University of Houston PhD Economics program.
†Typos may exist in these notes. If any are found, please contact me.

1 Introduction

These notes present key concepts from a Ph.D. course in Applied Econometrics. Topics covered include the selection problem, regression discontinuity design, controlling for confounding variables, fixed-effect regressions, difference-in-differences and instrumental variables.

2 OLS Review

2.1 Data Forms

Economic data can be collected and organized in several structural forms, each with distinct characteristics and analytical implications. Cross-sectional data consists of observations across different units (individuals, firms, countries) at a single point in time, enabling comparisons between different entities. Time series data tracks the same variable(s) for a single unit over multiple time periods, facilitating the analysis of temporal patterns and dynamics. Panel data (or longitudinal data) combines both dimensions, following multiple units over time, thereby capturing both cross-sectional heterogeneity and temporal evolution.

2.2 Data Collection Methods

The quality and reliability of econometric analysis fundamentally depend on the methods used to collect the underlying data. Experimental data arises from controlled experiments where researchers randomly assign subjects to treatment and control groups, providing the strongest basis for causal inference. Observational data, more common in economics, is collected without researcher intervention in the assignment process, requiring more sophisticated econometric techniques to establish causality.

Additionally, data may be classified as primary (collected directly by the researcher for the specific analysis) or secondary (obtained from external sources that collected the data for other purposes). While primary data collection offers greater control over the research design, secondary data often provides larger samples and broader coverage.

2.3 Simple Regression Model

The foundational linear regression model is expressed as:

$$Y = \beta_1 + \beta_2 X + u.$$

This equation represents the relationship between a dependent variable Y and an independent variable X, where β_1 is the intercept parameter, β_2 is the slope parameter indicating the effect of a one-unit change in X on Y, and u is the error term capturing all other factors affecting Y not explicitly included in the model.

The Ordinary Least Squares (OLS) estimators for the parameters are derived by minimizing the sum of squared residuals, yielding:

$$\beta_{2,\text{OLS}} = \frac{\text{Cov}(X,Y)}{\text{Var}(X)},$$

$$\beta_{1,\text{OLS}} = E[Y] - \beta_{2,\text{OLS}}\, E[X].$$

These formulas highlight that $\beta_{2,\text{OLS}}$ represents the covariation between X and Y scaled by the variance of X, while $\beta_{1,\text{OLS}}$ ensures that the regression line passes through the point of means.

2.4 Assumptions

The validity of OLS estimation rests on several critical assumptions:

1. Linear in Parameters: The dependent variable must be a linear function of the parameters (though not necessarily of the explanatory variables).

2. Identifiability: The matrix of explanatory variables X must have full rank, meaning no perfect multicollinearity exists among the regressors. This ensures the OLS estimator is uniquely defined.

3. Orthogonality (or exogeneity): The error term must be uncorrelated with the explanatory variables, formally expressed as $E[u \mid X] = 0$. This assumption is crucial for the consistency of OLS estimates and represents the absence of omitted variable bias, simultaneity, and measurement error.

4. Sphericality: The error terms must have constant variance (homoskedasticity) and be uncorrelated with each other (no serial correlation). Formally, $\text{Var}(u \mid X) = \sigma^2$ and $\text{Cov}(u_i, u_j \mid X) = 0$ for $i \neq j$.

5. Normality: The error term follows a normal distribution conditional on X, i.e., $u \sim N(0, \sigma^2)$. While not necessary for consistency, this assumption enables exact inference in finite samples.

2.5 Violations

Violations of the orthogonality assumption are particularly problematic as they render OLS estimates inconsistent, meaning they do not converge to the true parameter values even with infinite sample sizes. These violations generally arise from three sources:

Omitted Variables: When relevant explanatory variables are excluded from the model but are correlated with included variables, the estimated coefficients absorb the effects of these omitted factors, leading to biased and inconsistent estimates.

Simultaneity: When causality runs in both directions between the dependent and one or more independent variables, creating a feedback loop that violates the exogeneity assumption.

Measurement Error: When variables (particularly explanatory variables) are measured with error, introducing a correlation between the observed values and the error term.

Violations of the sphericality assumption, while not affecting consistency, impact the efficiency of OLS estimates and invalidate conventional inference procedures:

Heteroskedasticity: When the variance of the error term varies across observations, often related to the values of the explanatory variables.

Serial Correlation: When error terms are correlated across observations, particularly common in time series data.

Both violations necessitate adjustments to standard errors for valid inference.

2.6 Hypothesis Testing

Statistical hypothesis testing plays a central role in econometric inference, allowing researchers to evaluate the empirical validity of theoretical predictions. Two types of errors can occur in this process:

Type I Error: Rejecting a true null hypothesis, representing a false positive. The probability of committing this error is the significance level α of the test.

Type II Error: Failing to reject a false null hypothesis, representing a false negative. The probability of avoiding this error is the power of the test.

For testing hypotheses about individual parameters, such as $H_0 : \beta_i = c$ versus $H_a : \beta_i \neq c$, the t-test employs the following test statistic:

$$t_i = \frac{\beta_i - c}{\text{SE}(\beta_i)}.$$

Under the null hypothesis and assuming the model assumptions hold, this statistic follows a t-distribution with $(n-k)$ degrees of freedom, where n is the sample size and k is the number of parameters estimated.

3 The Selection Problem & Random Assignment

3.1 Effect of X on Y

A primary objective in applied econometrics is to identify the causal effect of a variable X (often a treatment or policy) on an outcome Y. This task is complicated by the fact that individuals or units with different values of X may also differ systematically in other characteristics that affect Y, making it difficult to isolate the pure effect of X.

3.2 SUTVA

The Stable Unit Treatment Value Assumption (SUTVA) forms a foundational premise for causal inference. It encompasses two critical conditions:

First, the potential outcome for any unit must be independent of the treatment status of other units, ruling out spillover or interference effects. For example, when analyzing the effect of a job training program, the employment prospects of a trained individual should not depend on whether other individuals received training.

Second, there should be no different versions of the treatment that would lead to different potential outcomes. All treated units receive the same treatment, and all control units receive the same (or no) treatment.

SUTVA dramatically simplifies causal analysis by reducing the number of potential outcomes that must be considered. Without it, each unit would have 2^N potential outcomes (considering all possible combinations of treatment assignments to the N units), making analysis intractable. With SUTVA, each unit has only two potential outcomes: one under treatment and one under control.

3.3 The Selection Problem

The fundamental challenge in causal inference is that for each unit, we observe only one of the potential outcomes—either the outcome under treatment or the outcome under control—not both. This creates the selection problem: units that receive a treatment may systematically differ from those that do not.

The observed difference in outcomes between treated and untreated groups can be expressed as:

$$\begin{aligned} E[Y_i \mid D_i = 1] - E[Y_i \mid D_i = 0] = &(E[Y_i(1) \mid D_i = 1] - E[Y_i(0) \mid D_i = 1]) \\ &+ (E[Y_i(0) \mid D_i = 1] - E[Y_i(0) \mid D_i = 0]) \end{aligned}$$

The first term represents the average treatment effect on the treated (ATT), which is the causal parameter of interest. The second term represents selection bias—the difference in potential untreated outcomes between the treated and untreated groups.

With random assignment of treatment, this selection bias term disappears because:

$$E[Y_i(0) \mid D_i = 1] = E[Y_i(0) \mid D_i = 0] = E[Y_i(0)]$$

Thus, under random assignment, the observed difference equals:

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

This is the average treatment effect (ATE), which equals the ATT under random assignment.

3.4 Analytical Steps

Analyzing data from randomized experiments typically involves two key steps:

First, researchers verify the balance of covariates between treatment and control groups. If treatment was truly randomly assigned, there should be no systematic differences in observable characteristics between the two groups. Statistical tests or standardized differences can assess this balance.

Second, researchers analyze the differences in outcomes between the treatment and control groups. The simplest approach is to compare mean outcomes, but regression analysis can also be employed to adjust for any remaining imbalances and improve precision.

3.5 Randomness Level

The level at which randomization occurs has important implications for analysis and interpretation. Treatment may be randomized at the individual level (each person independently assigned), the cluster level (e.g., all individuals in a city assigned together), or higher levels (e.g., states).

The randomization level affects the appropriate statistical analysis, particularly the calculation of standard errors. When treatment is assigned at a cluster level, standard errors must account for within-cluster correlation of outcomes. The level also influences the interpretation of treatment effects, as higher-level randomization may capture general equilibrium effects that individual-level randomization might miss.

3.6 Natural Experiments & Imperfect Compliance

Natural experiments occur when treatment assignment approximates randomness due to natural processes, policies, or administrative rules rather than researcher intervention. Lotteries, for instance, provide a clear mechanism for random assignment, such as in studies of military draft effects or school choice programs.

Imperfect compliance arises when not all units adhere to their assigned treatment status. Let D_i indicate the treatment offered (the assignment) and Z_i indicate the treatment actually received (the uptake). The compliance rate is defined as $\Pr(Z = 1 \mid D = 1)$, the proportion of those offered treatment who actually take it up. Non-random take-up can reintroduce selection bias.

With D randomly assigned, the intent-to-treat (ITT) effect—the effect of being offered treatment—can be estimated unbiasedly. The effect of actually receiving treatment (the treatment effect on the treated, TT) can then be calculated as:

$$\text{TT} = \frac{\text{ITT}}{\text{compliance rate}}.$$

This adjustment is valid under the assumption that treatment assignment affects outcomes only through actual treatment receipt (the exclusion restriction).

In analyzing experiments with imperfect compliance, researchers should use data from all study participants, regardless of their compliance status. Restricting analysis to compliers would reintroduce selection bias.

3.7 Missing Data

Missing data presents challenges in experimental studies through two main mechanisms:

Attrition occurs when participants drop out of the study over time, potentially creating differential loss between treatment and control groups. If attrition is related to treatment status or potential outcomes, it can introduce bias.

Non-response refers to missing answers on specific questions in surveys or other data collection instruments. Like attrition, differential non-response between groups can bias estimates.

Researchers employ various strategies to address missing data, including imputation methods, bounds analysis, and sensitivity tests to assess the potential impact of missing data on conclusions.

3.8 Random Assignment Conditional on Covariates

In some studies, treatment is not unconditionally random but is randomly assigned within groups defined by certain covariates. This design, known as stratified randomization or conditional random assignment, ensures that the treatment is unrelated to potential outcomes within each stratum:

$$E[Y_0 \mid X, D = 1] - E[Y_0 \mid X, D = 0] = 0.$$

Conditional random assignment is common when treatment is targeted to specific populations (e.g., based on income thresholds) or when stratification is used to improve precision. Proper analysis of such designs requires conditioning on the stratification variables.

3.9 Heterogeneity in Treatment Effects

Treatment effects often vary across subgroups of the population, a phenomenon known as treatment effect heterogeneity. Some groups may experience larger impacts than others due to differences in baseline characteristics, implementation quality, or complementarities with other factors.

Researchers can explore heterogeneity by examining treatment effects separately for different subgroups or by including interaction terms in regression models. For example, to compare treatment effects between older and younger subjects:

$$Y = \beta_0 + \beta_1 D + \beta_2 (D \cdot \text{old}) + \beta_3 \text{old} + u,$$

Here, β_1 represents the treatment effect for younger subjects, while $\beta_1 + \beta_2$ gives the effect for older subjects. Similarly, with continuous variables like age:

$$Y = \beta_0 + \beta_3 D + \beta_4 (D \cdot \text{age}) + \text{age} + u,$$

Where β_4 captures how the treatment effect changes with each year of age.

3.10 Quantile Regression

Quantile regression offers another approach to examining treatment effect heterogeneity by estimating effects across different points in the outcome distribution rather than focusing solely on the mean. It summarizes the conditional quantile function $Q_\tau(Y \mid X)$, where τ represents the quantile of interest (e.g., 0.25 for the first quartile, 0.5 for the median).

At the 50th percentile ($\tau = 0.5$), quantile regression yields the median outcome as a function of covariates, offering a robust alternative to mean regression when the outcome distribution is skewed or contains outliers. By examining effects at multiple quantiles, researchers can identify whether treatments have differential impacts across the outcome distribution, potentially revealing important patterns not apparent from mean effects alone.

3.11 Treatment Effects in Experiments

In experiments with perfect compliance, the treatment effect is directly estimated as the difference in outcomes between treatment and control groups. This represents the average treatment effect (ATE) when assignment is random.

With imperfect compliance, researchers estimate the intent-to-treat (ITT) effect first—the effect of being offered treatment regardless of uptake. The effect on those who actually receive

treatment (the treatment effect on the treated, TT) can then be calculated by scaling the ITT by the compliance rate:

$$\text{TT} = \frac{\text{ITT}}{\text{compliance rate}}.$$

This adjustment, formally derived using instrumental variables methods, provides a consistent estimate of the treatment effect for compliers under the assumption that treatment assignment affects outcomes only through treatment receipt.

3.12 Randomization Inference

Randomization inference offers a powerful approach to hypothesis testing in randomized experiments that does not rely on large-sample approximations or distributional assumptions. The method leverages the known randomization process to construct the exact distribution of test statistics under the null hypothesis.

For example, with 8 individuals and 4 treated, there are $\binom{8}{4} = 70$ possible treatment assignments. By computing the test statistic (e.g., the difference in means between treated and control groups) for each possible assignment and determining where the observed statistic falls in this distribution, exact p-values can be calculated.

This approach is particularly valuable for small samples where asymptotic approximations may be unreliable and when treatment is assigned at the cluster level with few clusters.

3.13 Cost Issues

Randomized controlled trials (RCTs), while methodologically rigorous, entail substantial costs that must be considered in research planning. These include:

1. Direct financial costs of implementation, including personnel, equipment, and participant compensation. 2. Administrative burdens of obtaining Institutional Review Board (IRB) approval and adhering to ethical guidelines. 3. Opportunity costs when more cost-effective methods might address the research question adequately.

The high costs of RCTs may constrain sample sizes, limiting statistical power, or restrict the complexity of interventions that can be tested.

3.14 Issues of Power

Statistical power—the probability of detecting a true effect when it exists—is a critical consideration in experimental design. Underpowered studies may fail to identify meaningful effects, leading to false negatives and publication bias when only statistically significant results are reported.

Power depends on several factors, including: 1. Sample size—larger samples provide greater power 2. Effect size—larger effects are easier to detect 3. Outcome variance—lower variance increases power 4. Significance level—less stringent thresholds increase power but also increase the risk of false positives

Power calculations should be conducted during study design to determine the appropriate sample size for detecting effects of interest with reasonable probability.

3.15 Internal Validity

Internal validity refers to the extent to which a study's results represent a causal relationship between treatment and outcome within the study population. Several threats to internal validity exist even in randomized experiments:

Spillovers occur when treatment of one unit affects the outcomes of other units, violating the SUTVA assumption. For example, an educational intervention for some students might indirectly benefit their untreated peers through knowledge sharing.

Non-response and attrition can create differential missing data patterns between treatment and control groups, potentially biasing results if not properly addressed.

Researchers should compare characteristics of dropouts versus stayers to assess whether attrition is related to treatment status or potential outcomes. Bounding techniques, such as Lee bounds described below, can quantify the potential impact of selective attrition on estimates.

3.16 Lee Bounds

Lee bounds provide a method for addressing selective attrition in experiments by calculating upper and lower bounds on treatment effects under worst-case scenarios about the missing data. The approach involves the following steps:

1. Determine which group (treatment or control) has higher attrition. 2. For the group with lower attrition, trim the sample by removing observations with the highest (for upper bound) or lowest (for lower bound) outcomes. 3. Trim until the proportion of observed outcomes is equal across groups. 4. Calculate treatment effects using these trimmed samples to obtain bounds.

This method makes minimal assumptions about the missing data process and provides a range within which the true treatment effect must lie, assuming monotonicity in the relationship between treatment and observation of outcomes.

3.17 External Validity

External validity concerns the generalizability of study findings beyond the specific context and population in which the research was conducted. Several factors may limit the external validity of RCTs:

Placebo effects occur when participants respond to the mere fact of being treated rather than to the treatment itself, potentially overstating treatment effects in experimental settings compared to real-world implementation.

John Henry effects arise when control group participants exert extra effort to compensate for not receiving the treatment, potentially understating treatment effects.

General equilibrium effects emerge when treatments are scaled up, creating feedback mechanisms and spillovers not captured in small-scale experiments.

Hawthorne effects occur when participants modify their behavior simply because they are being observed, potentially creating artificial responses not representative of natural settings.

Researchers can enhance external validity by conducting experiments in diverse settings, using representative samples, and complementing experiments with observational studies that examine similar questions in natural contexts.

4 Regression Discontinuity

4.1 Overview

Regression Discontinuity Design (RDD) represents a powerful quasi-experimental method for estimating causal effects when true experimental data is unavailable. The central question addressed by RDD remains consistent with broader causal inference goals: What is the effect of a treatment D on an outcome Y? RDD exploits situations where treatment assignment is determined by a continuous variable crossing a known threshold, creating a discontinuity in the probability of treatment that approximates local random assignment.

4.2 Basic Concept

The foundational premise of RDD is the existence of a running (or forcing) variable X_f that exhibits a smooth relationship with the outcome Y but a discontinuous relationship with the

treatment assignment D. This scenario frequently arises in policy settings where eligibility for a program or intervention is determined by a clear cutoff value c on some continuous measure.

The crucial insight of RDD is that individuals just above and below this cutoff are likely to be similar in all relevant characteristics except for their treatment status. As one approaches the cutoff from either direction, the assignment of treatment becomes increasingly similar to random assignment, creating what can be conceptualized as a local randomized experiment centered at the cutoff.

This local randomization occurs because, while individuals may have some control over their value of X_f, they typically cannot precisely manipulate it around the threshold. Consequently, whether an individual falls just above or just below the cutoff can be considered as good as random for those very close to the threshold.

4.3 Sharp vs. Fuzzy RDD

RDD methodologies are classified into two categories based on the nature of the discontinuity in treatment probability at the threshold:

In a Sharp RDD, the treatment assignment changes deterministically at the cutoff, meaning:

$$D_i = \mathbf{1}\{X_i \geq c\}$$

where $\mathbf{1}\{\cdot\}$ is an indicator function equal to 1 when the condition inside the brackets is true and 0 otherwise. Examples include age-based eligibility for Medicare (at age 65) or scholarship awards based on minimum test scores.

In a Fuzzy RDD, the treatment probability changes discontinuously at the cutoff but not from 0 to 1. Instead:

$$\lim_{x \downarrow c} \Pr(D = 1|X = x) \neq \lim_{x \uparrow c} \Pr(D = 1|X = x)$$

This occurs when the cutoff rule is imperfectly followed, leading to partial compliance. For instance, income-based program eligibility where some eligible individuals do not participate (non-takeup) or some ineligible individuals receive exceptions (crossovers).

4.4 Bandwidth Selection

A critical methodological decision in RDD analysis is determining the bandwidth—the range of data around the cutoff to include in the estimation. This selection involves a fundamental tradeoff:

Narrower bandwidths increase the plausibility of the local randomization assumption but reduce the sample size, decreasing precision.

Wider bandwidths increase the sample size and precision but may introduce bias if the relationship between the running variable and the outcome is not properly modeled.

Modern approaches to bandwidth selection employ data-driven methods that minimize the mean squared error of the RDD estimator, such as those developed by Imbens and Kalyanaraman (2012) and Calonico, Cattaneo, and Titiunik (2014).

4.5 Estimating Treatment Effects

4.5.1 Narrow Bandwidth Approach

The most transparent approach to RDD estimation uses a narrow bandwidth around the cutoff, effectively treating the data as a randomized experiment within this window.

For a Sharp RDD, the treatment effect is estimated as the difference in mean outcomes just above and below the cutoff:

$$\Delta_Y = \lim_{x \downarrow c} E[Y \mid X = x] - \lim_{x \uparrow c} E[Y \mid X = x]$$

For a Fuzzy RDD, the treatment effect is estimated by dividing the jump in outcomes by the jump in treatment probability:

$$\text{Effect of } D \text{ on } Y = \frac{\Delta_Y}{\Delta_D}$$

where

$$\Delta_D = \lim_{x \downarrow c} E[D \mid X = x] - \lim_{x \uparrow c} E[D \mid X = x]$$

This ratio represents a Wald estimator and can be implemented through two-stage least squares (2SLS) estimation.

4.5.2 Regression Approach: Narrow Bandwidth

With a sufficiently narrow bandwidth, RDD can be implemented through simple regression models.

The structural equation of interest is:

$$Y_i = \beta_0 + \beta_1 D_i + \epsilon_i$$

For a Fuzzy RDD, this is estimated through instrumental variables, where the instrument is the indicator for being above the cutoff.

The first stage captures how the cutoff affects treatment uptake:

$$D_i = \gamma_0 + \gamma_1 \, \mathbf{1}\{X_i \geq c\} + u_i$$

The reduced form shows how the cutoff directly affects outcomes:

$$Y_i = \delta_0 + \delta_1 \, \mathbf{1}\{X_i \geq c\} + v_i$$

In a Sharp RDD, $\gamma_1 = 1$ by definition, so $\beta_1 = \delta_1$. In a Fuzzy RDD, $\beta_1 = \delta_1 / \gamma_1$, representing the treatment effect for compliers.

4.5.3 Regression Approach: Wider Bandwidth

With a wider bandwidth, it becomes necessary to control for the smooth effect of the running variable on the outcome. This is achieved by including a function of the running variable in the regression:

$$Y_i = \beta_0 + \beta_1 D_i + f(X_i - c) + u_i$$

The first stage becomes:

$$D_i = \alpha + \zeta \, \mathbf{1}\{X_i \geq c\} + \tau(X_i - c) + e_i$$

And the reduced form:

$$Y_i = \delta_0 + \delta_1 \, \mathbf{1}\{X_i \geq c\} + \eta(X_i - c) + v_i$$

The function $f(\cdot)$ can take various forms, including linear, quadratic, or higher-order polynomials, as well as more flexible specifications like splines or local polynomials.

4.5.4 Local Linear Specification

A particularly important specification allows the slope of the running variable to differ on each side of the cutoff, capturing potential differences in the relationship between X and Y for treated and untreated units:

$$Y_i = \beta_0 + \beta_1 D_i + h(X_i - c) + f_r \, \mathbf{1}\{X_i \geq c\} \, (X_i - c) + u_i$$

The corresponding first stage is:

$$D_i = \alpha + \zeta \, \mathbf{1}\{X_i \geq c\} + \tau(X_i - c) + \theta \, \mathbf{1}\{X_i \geq c\} \, (X_i - c) + e_i$$

This specification provides robustness against misspecification of the functional form, particularly when the slopes differ substantially across the threshold.

4.6 Additional Considerations

Several specialized techniques address challenges that may arise in RDD applications:

Donut RDD involves dropping observations very close to the cutoff if there is evidence of manipulation or measurement error in this region. This creates a "donut hole" around the threshold, using observations that are near but not immediately adjacent to the cutoff.

Endogeneity Checks are crucial to verify the validity of the RDD. These include testing for: - Manipulation of the running variable (e.g., using McCrary's density test) - Balance of pre-treatment covariates around the cutoff - Discontinuities at placebo thresholds where no treatment change occurs

Violations of these checks may indicate that the fundamental assumption of RDD—that assignment near the cutoff approximates random assignment—does not hold.

4.7 External Validity

A significant limitation of RDD is that it estimates a local average treatment effect (LATE) specifically for individuals near the cutoff. This effect may not generalize to individuals far from the threshold, especially if treatment effects are heterogeneous.

For example, a scholarship based on a minimum test score may have different effects on students who barely qualified compared to those who scored well above the threshold. Researchers should carefully consider this limitation when interpreting and applying RDD results to broader populations or policy contexts.

5 Controlling for Confounding Variables

5.1 Introduction

The fundamental question in causal analysis remains consistent: how does a treatment D affect an outcome Y? In the absence of experimental data, where treatment is randomly assigned, researchers must employ various strategies to establish causal relationships using observational data. The success of these approaches depends critically on the nature of the selection process into treatment and the available data.

The econometric toolkit for addressing selection includes several major approaches: controlling for confounding variables, fixed effects, difference-in-differences (DiD), instrumental variables (IV), and regression discontinuity (RD). Each method addresses specific types of selection processes and carries its own assumptions and limitations. This section focuses specifically on controlling for confounding variables as a strategy for causal inference.

5.2 Controlling for Confounding Variables

The challenge of causal inference in observational studies stems from the presence of confounding variables—factors that affect both the treatment assignment and the outcome. These confounders can be observable (measured in the data) or unobservable (unmeasured), and their influence must be addressed to isolate the causal effect of the treatment.

A common temptation is to include a large number of control variables in a regression — sometimes called a "kitchen sink regression"—in an attempt to account for all possible confounders. However, this approach has several drawbacks: it can lead to overfitting, reduce statistical power, make results difficult to interpret, and may introduce post-treatment bias if some controls are themselves affected by the treatment.

Instead, a more principled approach begins by carefully investigating the treatment assignment process. The nature of this process determines which identification strategy is most appropriate:

If treatment was randomly assigned, experimental methods can be applied directly. If treatment was not randomly assigned, the key question becomes whether all variables affecting both treatment and outcome are observable and available in the data.

When all confounders are observable, we have "selection on observables" (also called unconfoundedness or conditional independence). When some confounders remain unobservable, we have "selection on unobservables," requiring more sophisticated techniques like fixed effects, instrumental variables, or difference-in-differences.

5.3 Selection on Observables

Under selection on observables, once we condition on a set of covariates X, treatment assignment becomes as good as random with respect to potential outcomes. Formally, this assumption is expressed as:

$$Y_i(0), Y_i(1) \perp D_i \mid X_i$$

This means that, within groups defined by the same values of X, variation in treatment is unrelated to potential outcomes. If this assumption holds and we correctly specify the functional relationship between X and the outcome, we can estimate the causal effect of treatment using a regression model:

$$Y_i = \alpha + \beta D_i + \Pi X_i + u_i$$

Here, β represents the causal effect of the treatment. However, the success of this approach hinges critically on two factors:

1. Having measured all relevant confounding variables 2. Correctly specifying the functional form of their relationship with the outcome

In practice, researchers often augment the basic linear specification with additional terms to capture more complex relationships, such as:

$$Y_i = \alpha + \beta D_i + \Pi X_i + (\text{quadratic terms, cubic terms, interaction terms}) + u_i$$

While regression provides a parametric approach to controlling for confounders, matching methods offer a non-parametric alternative that may be more robust to functional form misspecification.

5.4 The Idea of Matching Estimators

Matching estimators directly compare outcomes between treated and untreated units that have similar or identical values of confounding variables. This approach mimics the logic of a randomized experiment by constructing appropriate comparison groups post hoc.

Consider a simple example with two binary confounders: sex (male/female) and education level (high school dropout, high school graduate, college graduate). These variables create six

distinct groups or "cells." Within each cell, the difference in outcomes between treated and untreated units provides an estimate of the treatment effect for that specific subgroup:

$$E[Y \mid D = 1, X = x] - E[Y \mid D = 0, X = x]$$

Under the selection on observables assumption, this difference represents the conditional average treatment effect for individuals with characteristics $X = x$. Importantly, within each cell, the average treatment effect (ATE) equals the treatment effect on the treated (TT) because conditioning on X removes selection bias.

To obtain overall treatment effects, these cell-specific effects are weighted and averaged:

For the average treatment effect (ATE) across the entire population:

$$E[Y_1 - Y_0] = E_X\Big[E[Y_1 - Y_0 \mid X = x]\Big] = \sum_x E[Y_1 - Y_0 \mid X = x] \cdot P(X = x)$$

For the treatment effect on the treated (TT):

$$E[Y_1 - Y_0 \mid D = 1] = E_X\Big[E[Y_1 - Y_0 \mid X = x, D = 1]\Big] = \sum_x E[Y_1 - Y_0 \mid X = x] \cdot P(X = x \mid D = 1)$$

The ATE and TT generally differ because the distribution of confounders may vary between the treated population and the overall population. Additionally, practical challenges arise when some cells contain only treated or only untreated units, creating a "common support" problem.

Fixed effects regression is sometimes used to implement matching estimators, particularly with panel data, but it's important to note that this does not directly estimate either ATE or TT without additional assumptions or weighting.

5.5 Propensity Score Matching

As the number of confounding variables increases or when continuous confounders are present, exact matching becomes increasingly difficult due to the "curse of dimensionality"—the rapid proliferation of cells requiring matches. Propensity score matching offers a solution by reducing the matching problem to a single dimension.

Rosenbaum and Rubin (1983) demonstrated a remarkable result: if matching on the full set of covariates X is valid for removing selection bias, then matching on the probability of treatment given X—the propensity score—is equally valid. Formally, the propensity score is defined as:

$$p(x) = \Pr(D = 1 \mid X = x)$$

This approach rests on two key assumptions:

1. Selection on observables (also called unconfoundedness, ignorability, or conditional independence):

$$Y_i(0), Y_i(1) \perp D_i \mid X_i$$

2. Overlap (or common support):

$$0 < p(x) < 1$$

This ensures that for each value of X, there is a positive probability of both treatment and non-treatment, making comparisons possible.

Two critical properties make propensity score matching effective:

1. Balancing property: Conditional on the propensity score, the distribution of covariates X is the same in the treated and untreated groups. Mathematically, if $X \perp D \mid p(X)$, then in a regression of D on X and $p(X)$, the coefficients on X should be statistically insignificant.

2. Unconfoundedness property: If treatment assignment is unconfounded given X, then it is also unconfounded given the propensity score:

$$Y_i(0), Y_i(1) \perp D_i \mid p(X_i)$$

Together, these properties allow researchers to estimate treatment effects by conditioning on the propensity score rather than the potentially high-dimensional set of covariates.

5.6 Overview of the Propensity Score Matching Procedure

Implementing propensity score matching typically involves the following steps:

Step 1: Assess Feasibility

Determine whether the selection on observables assumption is plausible in the specific context. This requires detailed knowledge of the treatment assignment process and the availability of data on all relevant confounders. Additionally, verify that there is sufficient overlap in the propensity score distribution between treated and untreated units.

Step 2: Estimate the Propensity Score

Select covariates X that affect both treatment assignment and outcomes, potentially including functions of these variables (interactions, polynomials) to capture complex relationships. Estimate the propensity score using logit or probit models:

$$\Pr(D_i = 1 \mid X_i) = F(X_i\beta)$$

where $F(\cdot)$ is the logistic or normal CDF. The predicted values from this model, $\hat{p}(X_i)$, serve as estimates of the propensity score.

Step 3: Stratify the Sample and Check Balance

Divide the sample into strata (typically quintiles) based on the estimated propensity score. Within each stratum, verify that the covariates are balanced between treated and untreated units—there should be no statistically significant differences in the distributions of covariates conditional on the propensity score. If imbalances persist, refine the propensity score model by adding interactions or higher-order terms.

Step 4: Estimate Treatment Effects

Various methods can be used to estimate treatment effects based on the propensity score:

- Stratification: Compute the treatment effect within each propensity score stratum and take a weighted average. - Nearest Neighbor Matching: Match each treated unit to the untreated unit(s) with the closest propensity score. - Radius Matching: Match each treated unit to all untreated units within a specified distance (caliper). - Kernel Matching: Match each treated unit to all untreated units, with weights determined by the distance between their propensity scores. - Inverse Probability Weighting: Weight observations by the inverse of their propensity score (for untreated units) or the inverse of one minus their propensity score (for treated units).

The choice of method involves tradeoffs between bias and efficiency, with nearest neighbor methods minimizing bias at the cost of higher variance, while kernel methods improve efficiency but may introduce some bias.

6 Panel Data & Fixed Effects

6.1 Overview

Panel data analysis represents a powerful approach to causal inference that leverages repeated observations of the same units over time. This longitudinal structure enables researchers to control for time-invariant unobserved heterogeneity that might otherwise confound the relationship between variables of interest. Panel data is characterized by observations indexed by both unit (i) and time (t) dimensions:

- Y_{it}, X_{it}: Outcome and explanatory variables for unit i at time t.
- N: Number of cross-sectional units (individuals, firms, countries, etc.).
- T: Number of time periods.

The resulting dataset can be organized in different formats:

- Long format: Each row represents a unique unit-time combination, resulting in $N \times T$ rows.
- Wide format: Each row represents a unique unit, with separate columns for each time period's observations, resulting in N rows.

In microeconometric applications, the asymptotic properties of estimators typically rely on N approaching infinity while T remains fixed or grows at a slower rate. This differs from time series econometrics, where asymptotics are driven by T approaching infinity.

6.2 Variation in Panel Data

A fundamental advantage of panel data is its ability to capture different sources of variation:

Within variation (or "time variation") reflects changes in a variable over time for the same unit. For a variable X_{it}, the within variation is measured around each unit's mean $\bar{X}_i$, capturing temporal dynamics while holding unit identity constant.

Between variation (or "cross-sectional variation") reflects differences across units, typically measured using unit means $\bar{X}_i$. This captures stable differences between units while abstracting from temporal fluctuations.

Fixed effects estimation, the focus of this section, primarily exploits within variation to identify causal effects, effectively controlling for all time-invariant differences between units.

6.3 Fixed Effects Estimation

The core insight of fixed effects estimation is that time-invariant unobserved heterogeneity can be eliminated through within-unit comparisons. Consider a model where the outcome depends on observed variables X_{it} and unobserved time-invariant characteristics C_i:

$$Y_{it} = X_{it}\beta + C_i + u_{it} \tag{1}$$

If C_i is correlated with X_{it}, pooled OLS estimates of β will be biased due to omitted variable bias. Fixed effects estimation addresses this by exploiting the time dimension of the data.

In the simplest case with $T = 2$, we can difference the equation across time periods:

$$(Y_{i2} - Y_{i1}) = (X_{i2} - X_{i1})\beta + (C_i - C_i) + (u_{i2} - u_{i1}) \tag{2}$$

Since C_i is time-invariant, it cancels out in the differencing, allowing consistent estimation of β even when C_i is correlated with X_{it}. More generally, the fixed effects estimator can be implemented by demeaning each variable within units:

$$(Y_{it} - \bar{Y}_i) = (X_{it} - \bar{X}_i)\beta + (u_{it} - \bar{u}_i) \tag{3}$$

This "within transformation" removes all time-invariant factors, both observed and unobserved, allowing identification of β solely from within-unit variation. The transformation is equivalent to including a separate dummy variable for each unit in the regression:

$$Y_{it} = X_{it}\beta + d_i + u_{it} \tag{4}$$

where d_i represents a set of unit-specific dummy variables. This approach directly controls for all time-invariant characteristics of units without requiring explicit specification of these factors.

6.4 First-Difference Estimation

An alternative to the within transformation is first-differencing, which also eliminates time-invariant factors:

$$(Y_{it} - Y_{i,t-1}) = (X_{it} - X_{i,t-1})\beta + (C_i - C_i) + (u_{it} - u_{i,t-1}) \tag{5}$$

Both fixed effects and first-difference estimators are consistent when $N \to \infty$ with fixed T, but they differ in their efficiency properties:

- If u_{it} is serially uncorrelated, the fixed effects estimator is more efficient. - If u_{it} follows a random walk (perfect serial correlation), the first-difference estimator is more efficient. - For intermediate cases of serial correlation, their relative efficiency depends on the specific correlation structure.

The choice between these approaches may also be influenced by other considerations, such as the pattern of missing data or the presence of predetermined (rather than strictly exogenous) regressors.

6.5 Alternative Panel Data Estimators

Several alternative estimators exist for panel data, each making different assumptions about the relationship between unobserved heterogeneity and the regressors:

Pooled OLS treats the panel as a large cross-section, ignoring the panel structure:

$$Y_{it} = X_{it}\beta + C_i + u_{it} \tag{6}$$

Consistency requires:

$$E[X_{it}u_{it}] = 0, \quad E[X_{it}C_i] = 0 \tag{7}$$

The second condition is the key distinction from fixed effects—pooled OLS assumes no correlation between the regressors and the unobserved heterogeneity. Even when consistent, pooled OLS requires clustered standard errors to account for within-unit correlation in the composite error term $C_i + u_{it}$.

Between Estimation uses only cross-sectional variation, averaging the data over time for each unit:

$$\bar{Y}_i = \bar{X}_i\beta + C_i + \bar{u}_i \tag{8}$$

Consistency requires:

$$E[\bar{X}_i\bar{u}_i] = 0, \quad E[\bar{X}_iC_i] = 0 \tag{9}$$

The between estimator is efficient if these conditions hold but inconsistent if unobserved heterogeneity correlates with the regressors. It cannot identify the effects of time-invariant variables, as these are absorbed into the unit effects.

Random Effects (RE) treats C_i as a random variable uncorrelated with X_{it} but accounts for the resulting error structure. The model can be written as:

$$Y_{it} = X_{it}\beta + v_{it}, \quad v_{it} = C_i + u_{it} \tag{10}$$

The composite error term v_{it} has a specific variance-covariance structure:

$$\Omega = \sigma_c^2 I_T + \sigma_u^2 j_T j_T' \tag{11}$$

where I_T is the identity matrix and j_T is a vector of ones. This structure captures the fact that observations from the same unit share the common component C_i and are therefore correlated.

RE estimation uses generalized least squares (GLS) to account for this correlation structure, resulting in a weighted average of the within and between estimators. When the assumption $E[X_{it}C_i] = 0$ holds, RE is more efficient than FE; when this assumption fails, RE is inconsistent.

6.6 Hausman Specification Test

The Hausman test provides a formal method for choosing between fixed effects and random effects estimators. The test compares the coefficients from both estimators under the null hypothesis that the RE assumptions are valid:

- H_0: $E[X_{it}C_i] = 0$, implying both FE and RE are consistent but RE is more efficient.
- H_A: $E[X_{it}C_i] \neq 0$, implying only FE is consistent.

The test statistic follows a chi-squared distribution under the null hypothesis:

$$H = (\hat{\beta}_{FE} - \hat{\beta}_{RE})'[\text{Var}(\hat{\beta}_{FE}) - \text{Var}(\hat{\beta}_{RE})]^{-1}(\hat{\beta}_{FE} - \hat{\beta}_{RE}) \tag{12}$$

Rejecting the null hypothesis suggests that the correlation between regressors and unobserved heterogeneity is significant, favoring the fixed effects estimator. However, it's worth noting that fixed effects estimation is always consistent (though potentially inefficient) under both hypotheses, making it a conservative choice when in doubt.

6.7 Standard Errors in Panel Data Models

Proper inference in panel data models requires accounting for potential correlation structures in the error terms:

Fixed Effects: Standard errors should be robust to heteroskedasticity and, if T is moderately large, to serial correlation. This can be implemented using cluster-robust standard errors, clustering at the unit level.

Pooled OLS: Clustering at the unit level is essential to account for within-unit correlation induced by the unobserved heterogeneity component.

Between Estimation: Robust standard errors should be used to address potential heteroskedasticity across units.

Random Effects: Both heteroskedasticity and within-unit correlation should be addressed, typically through a combination of robust and clustered standard errors.

In all cases, the appropriate standard error specification depends on the assumed error structure and the dimensions of the panel. When N is large relative to T (as in most microeconometric applications), clustering at the unit level addresses the most important sources of correlation.

7 Difference-in-Differences

7.1 Panel Data and Causal Inference

Panel data enables researchers to address evaluation problems through difference-in-differences (DiD or DD) estimation. This approach does not necessarily require true longitudinal data on individuals; repeated cross-sections suffice. The DiD approach is particularly useful when studying a policy, intervention, program, or event beginning at a specific time t^* that affects treatment status. With data collected both before and after t^* for both treatment-affected and unaffected units, difference-in-differences provides a strategy for evaluating policy impacts. The resulting DiD estimate quantifies the direct impact of the policy on the outcome variable of interest, while also providing information about the treatment effect, though not necessarily at the precise scale of a one-unit increase in the treatment variable.

7.2 Potential Outcomes Framework for Difference-in-Differences

The difference-in-differences approach can be understood through the potential outcomes framework. Consider a setting where no units initially experience a policy intervention, after which a treatment group becomes exposed beginning at time t^*. For every group j and time period t, we can define potential outcomes $Y_{j,t}(0)$ and $Y_{j,t}(1)$ representing outcomes without and with the policy, respectively. This notation extends the standard potential outcomes framework to explicitly represent group- and time-specific potential outcomes. Within this framework, we can define an average treatment effect for each group-time combination: $ATE_{j,t} = E[Y_{j,t}(1) - Y_{j,t}(0)]$.

In the standard DiD setup, we have two groups $j \in \{Treatment\ (\mathrm{T}), Comparison\ (\mathrm{C})\}$ and two time periods $t \in \{Before, After\}$. For example, $Y_{T,After}(0)$ represents the outcome for the treatment group in the post-intervention period had they not been treated. The average treatment effect on the treated in the post-period is $ATT_{T,After} = E[Y_{T,After}(1) - Y_{T,After}(0)]$. Since we cannot directly observe $Y_{T,After}(0)$, estimating this treatment effect requires additional assumptions and estimation strategies.

7.3 Simple Difference Estimators and Their Limitations

A simple pre/post difference estimator compares outcomes of the treatment group before and after the policy change:

$$\begin{aligned}
\beta_{d1} &= E[y_{it}|j = T, t \geq t^*] - E[y_{it}|j = T, t < t^*] && (13)\\
&= E[y_{After} - y_{Before}|j = T] && (14)
\end{aligned}$$

Examining the expected value of this estimator reveals that it combines the true causal effect with a potential bias term:

$$\begin{aligned}
E(\beta_{d1}) &= E[y_i|j = T, t = After] - E[y_i|j = T, t = Before] && (15)\\
&= E[Y_{T,After}(1)] - E[Y_{T,Before}(0)] && (16)\\
&= E[Y_{T,After}(1) - Y_{T,After}(0)] + E[Y_{T,After}(0) - Y_{T,Before}(0)] && (17)\\
&= \text{True causal effect} + \text{Bias} && (18)
\end{aligned}$$

This estimator can be implemented using OLS regression with data from the treatment group before and after intervention:

$$y_{it} = \alpha + \beta_{d1} After_t + \varepsilon_{it} \tag{19}$$

where $After_t$ indicates observations in the post-intervention period. The key identifying assumption is that, absent treatment, outcomes would remain unchanged over time ($\beta_{d1} = 0$). This assumption rarely holds in practice since outcomes typically change over time for numerous reasons unrelated to the policy under study.

An alternative approach uses a simple treatment/control difference estimator that compares outcomes between treatment and comparison groups after the policy change:

$$\begin{aligned}
\beta_{d2} &= E[y_{it}|j = T, t \geq t^*] - E[y_{it}|j = C, t \geq t^*] && (20)\\
&= E[y_T - y_C|t \geq t^*] && (21)
\end{aligned}$$

The expected value of this estimator also combines the true causal effect with a bias term:

$$\begin{aligned}
E(\beta_{d2}) &= E[y_i|j = T, t = After] - E[y_i|j = C, t = After] && (22)\\
&= E[Y_{T,After}(1)] - E[Y_{C,After}(0)] && (23)\\
&= E[Y_{T,After}(1) - Y_{T,After}(0)] + E[Y_{T,After}(0) - Y_{C,After}(0)] && (24)\\
&= \text{True causal effect} + \text{Bias} && (25)
\end{aligned}$$

This estimator can be implemented using OLS with post-intervention data from both groups:

$$y_{ij} = \alpha + \beta_{d2} Treatment_j + \varepsilon_{ij} \tag{26}$$

where $Treatment_j$ indicates membership in the treatment group. The key identifying assumption is that, absent treatment, outcomes would be identical between groups ($\beta_{d2} = 0$). This assumption is problematic because treatment and comparison groups often differ systematically in ways that correlate with outcomes.

7.4 The Difference-in-Differences Estimator

The difference-in-differences estimator addresses limitations of both simple difference approaches by comparing changes in outcomes over time between treatment and comparison groups:

$$\begin{aligned} \beta_{DID} = &\text{ (change in outcome over time for Treatment group)} \\ &- \text{(change in outcome over time for Comparison group)} \end{aligned} \tag{27}$$

$$\begin{aligned} \beta_{DID} = &\underbrace{E[y_i|j = T, t = After] - E[y_i|j = T, t = Before]}_{\text{Treated group change}} \\ &- \underbrace{E[y_i|j = C, t = After] - E[y_i|j = C, t = Before]}_{\text{Control group change}} \end{aligned} \tag{28}$$

$$\tag{29}$$

Taking the expectation reveals:

$$\begin{aligned} E(\beta_{DID}) &= (E[Y_{T,After}(1)] - E[Y_{T,Before}(0)]) \\ &\quad - (E[Y_{C,After}(0)] - E[Y_{C,Before}(0)]) \qquad (30) \\ &= \underbrace{E[Y_{T,After}(1) - Y_{T,After}(0)]}_{\text{ATT}} \\ &\quad + (E[Y_{T,After}(0) - Y_{T,Before}(0)] - E[Y_{C,After}(0) - Y_{C,Before}(0)]) \qquad (31) \end{aligned}$$

The crucial parallel trends assumption states that in the absence of treatment, both groups would experience identical changes in outcomes over time:

$$E[Y_{T,After}(0) - Y_{T,Before}(0)] = E[Y_{C,After}(0) - Y_{C,Before}(0)] \tag{32}$$

Under this assumption, the bias term becomes zero, making the DiD estimator an unbiased estimator of the average treatment effect on the treated:

$$E(\beta_{DID}) = E[Y_{T,After}(1) - Y_{T,After}(0)] = \text{True causal effect} \tag{33}$$

7.5 Implementation and Calculation

The difference-in-differences estimator requires calculating four group means: (1) treatment group after intervention, (2) treatment group before intervention, (3) comparison group after intervention, and (4) comparison group before intervention. The DiD estimate is then computed as the difference between the over-time change for the treatment group and the over-time change for the comparison group.

Alternatively, researchers can implement DiD estimation using regression analysis with all observations in the sample:

$$y_{ijt} = \alpha + \beta_{DID}(After_t \times Treatment_j) + \delta After_t + \gamma Treatment_j + \varepsilon_{ijt} \tag{34}$$

In this specification, $After_t$ indicates post-intervention periods, $Treatment_j$ indicates membership in the treatment group, and the interaction term $After_t \times Treatment_j$ captures the treatment effect. The coefficient β_{DID} represents the difference-in-differences estimate. The regression approach offers advantages including straightforward computation of standard errors and the ability to incorporate additional control variables.

For data in wide format (one observation per unit rather than unit-time observations), researchers can implement DiD by estimating a model with the change in the outcome as the dependent variable:

$$\Delta y_{ij} = \alpha + \beta_{DID} Treatment_j + \Delta \varepsilon_{ijt} \tag{35}$$

where $\Delta y_{ij} = y_{ij,after} - y_{ij,before}$ represents the change in the outcome for each unit.

7.6 Variations on the Basic Framework

An important variation of the standard DiD approach accommodates scenarios where a policy is initially present and then eliminated. In such cases, the DiD estimator can be formulated as:

$$y_{ijt} = \alpha + \beta(\text{Before}_t \cdot \text{Treatment}_j) + \gamma \text{Before}_t + \delta \text{Treatment}_j + \epsilon_{ijt} \tag{36}$$

Here, Before_t indicates the pre-elimination period, and β measures the differential impact of the policy on the treatment group before its elimination. Alternatively, one could estimate the standard DiD equation and interpret the coefficient of $\text{After}_t \cdot \text{Treatment}_j$ as the effect of eliminating the policy.

7.7 Cohort-Based Difference-in-Differences

When longitudinal data are unavailable but cohort variation exists, researchers can implement a cohort-based DiD strategy. This approach redefines the temporal dimension in terms of cohorts (c) rather than time periods, such as birth cohorts or school cohorts. The estimating equation becomes:

$$y_{ijc} = \alpha + \beta(\text{Affected Cohort}_c \cdot \text{Treatment}_j) + \gamma \text{Affected Cohort}_c + \delta \text{Treatment}_j + \epsilon_{ijc} \tag{37}$$

Where Affected Cohort_c equals 1 for cohorts exposed to the policy change and 0 otherwise. This strategy is particularly valuable as it can be implemented with a single cross-section of data. However, its applicability is limited to outcomes involving age-specific investments or experiences that would be differentially affected across cohorts.

7.8 Enhanced Control Specifications

7.8.1 Additional Covariates

While the DiD framework inherently controls for time-invariant group characteristics and group-invariant time effects, researchers often incorporate additional covariates that vary across both groups and time to strengthen identification. This enhanced specification can be written as:

$$y_{ijt} = \alpha + \beta(\text{After}_t \cdot \text{Treatment}_j) + \gamma \text{After}_t + \delta \text{Treatment}_j + \omega w_{ijt} + \epsilon_{ijt} \tag{38}$$

Where w_{ijt} represents an additional explanatory variable. Including such covariates may help satisfy the parallel trends assumption conditional on these variables, increasing the credibility of the causal interpretation of β.

7.8.2 Fully Controlling for Main Effects

When dealing with multiple groups and time periods, researchers can implement a more flexible DiD specification that fully controls for group and time main effects through fixed effects:

$$y_{ijt} = \alpha + \beta(\text{After}_t \cdot \text{Treatment}_j) + \delta_t + \gamma_j + \epsilon_{ijt} \tag{39}$$

In this equation, δ_t represents time fixed effects and γ_j represents group fixed effects. With J groups, one can include up to $J-1$ group dummies, allowing the mean outcome to differ across groups in a more flexible manner than would be possible with a single treatment dummy. Similarly, with T time periods, one can include up to $T-1$ time dummies, providing an unrestricted way to account for temporal variations in outcomes that affect all groups similarly.

7.9 Heterogeneous Treatment Effects

7.9.1 Event Study Specifications

To examine treatment effect dynamics and assess the validity of the parallel trends assumption, researchers frequently employ event study DiD specifications. These specifications allow the treatment effect to vary by time:

$$y_{ijt} = \alpha + \sum_k \beta_k[I(t=k) \cdot \text{Treatment}_j] + \delta_t + \gamma_j + \epsilon_{ijt} \tag{40}$$

In this formulation, k indexes time periods, with one period (typically either the earliest period or the period immediately preceding the intervention) designated as the reference period and omitted from the summation. The coefficients β_k capture the year-specific DiD effects relative to this reference period. Each β_k represents $[\bar{y}_{\text{treatment, time } t} - \bar{y}_{\text{treatment, reference time}}] - [\bar{y}_{\text{comp, time } t} - \bar{y}_{\text{comp, reference time}}]$.

The event study approach serves multiple purposes. First, it allows researchers to visualize treatment effect dynamics, potentially revealing how effects evolve over time. Second, it provides a test of the parallel trends assumption: coefficients for pre-intervention periods should not significantly differ from zero if the assumption holds. Third, it may reveal anticipatory effects or implementation lags that a standard DiD specification would miss.

7.9.2 Heterogeneity by Observable Characteristics

Researchers may also be interested in exploring how treatment effects vary across observable characteristics. This can be accomplished by introducing triple interactions:

$$\begin{aligned} y_{ijt} = \alpha &+ \beta(\text{After}_t \cdot \text{Treatment}_j) + \phi(X_{ijt} \cdot \text{After}_t \cdot \text{Treatment}_j) + \gamma\text{After}_t \\ &+ \delta\text{Treatment}_j + \theta X_{ijt} + \rho(X_{ijt} \cdot \text{After}_t) + \lambda(X_{ijt} \cdot \text{Treatment}_j) + \epsilon_{ijt} \end{aligned} \tag{41}$$

Where X_{ijt} represents an observable characteristic of interest (e.g., race, gender, or educational attainment). In this specification, β captures the treatment effect for observations with $X_{ijt} = 0$, while ϕ measures the differential treatment effect for each unit increase in X_{ijt}. For binary characteristics, ϕ represents the difference in treatment effects between the two groups defined by X_{ijt}.

This approach facilitates formal hypothesis testing regarding effect heterogeneity through t-tests on the coefficient ϕ. Notably, when all coefficients are allowed to vary by X_{ijt}, this specification is equivalent to estimating separate DiD models for each value of X_{ijt}.

7.10 The Parallel Trends Assumption and Violations

The key identifying assumption in difference-in-differences estimation, known as the parallel trends assumption, states that in the absence of treatment, the average change in outcomes would be identical between treatment and comparison groups. This does not require that the levels of outcomes be identical across groups, merely that the trends over time would be parallel without intervention.

Violations of the parallel trends assumption can lead to biased DiD estimates. One form of violation occurs when treatment group outcomes would naturally grow faster than comparison group outcomes absent intervention, perhaps due to mean reversion or catch-up effects. In this scenario, the DiD estimator would attribute this differential trend to the treatment effect, resulting in an upwardly biased estimate.

Conversely, if treatment group outcomes would naturally grow more slowly than comparison group outcomes absent intervention, possibly due to increasing inequality between groups, the DiD estimator would produce a downwardly biased estimate of the treatment effect.

7.11 Multiple Policy Changes

7.11.1 Policies with Identical Timing

When multiple policy changes occur simultaneously across different treatment units, researchers can either analyze each policy change separately or pool the data to estimate an average treatment effect. For the latter approach, one can define a treatment group encompassing all units exposed to any of the policy changes and implement the standard DiD specification.

7.11.2 Two-Way Fixed Effects Models for Staggered Adoption

For settings with staggered policy adoption, where treatment units adopt policies at different times, the two-way fixed effects (TWFE) specification offers a convenient estimation framework:

$$y_{ijt} = \alpha + \beta \text{HasPolicy}_{jt} + \delta_t + \gamma_j + \epsilon_{ijt} \tag{42}$$

Where HasPolicy_{jt} equals 1 when unit j has implemented the policy by time t, and 0 otherwise. This specification efficiently leverages all available variation in treatment timing, with units that have not yet adopted the policy serving as comparisons for early adopters.

However, it is important to recognize that this TWFE specification imposes homogeneous treatment effect assumptions. The coefficient β represents a weighted average of all possible two-group/two-period DiD estimators, with weights that can be negative under certain conditions. Recent econometric literature has highlighted potential biases in TWFE estimators when treatment effects vary over time or across groups, leading to the development of alternative estimators that accommodate such heterogeneity.

When employing a TWFE specification, the key identifying assumption remains the parallel trends assumption: in the absence of treatment, treated and untreated units would have experienced parallel trends in outcomes. This assumption becomes more complex in staggered adoption settings, as it must hold across multiple treatment and control groups defined by adoption timing.

7.12 Staggered Treatment Adoption

7.12.1 Dynamic Treatment Effects and Event Studies

When policy effects potentially evolve over time, researchers often employ event study specifications that allow for dynamic treatment effects relative to policy adoption. Rather than focusing on calendar time, these models center on event time—periods relative to treatment initiation.

Let t_j^* denote the adoption time for group j, then $l \equiv t - t_j^*$ represents time relative to adoption, with $l = 0$ indicating the implementation period, $l = -1$ one period before implementation, and so forth.

The dynamic TWFE specification typically takes the form:

$$y_{ijt} = \alpha + \sum_{l=-K, l\neq -1}^{L} \beta_l I(t - t_j^* = l) + \delta_t + \gamma_j + \varepsilon_{ijt} \tag{43}$$

where $I(t - t_j^* = l)$ indicates that unit j at time t is l periods away from its treatment. By convention, the period immediately preceding treatment ($l = -1$) serves as the reference category. The coefficients β_l capture the treatment effect l periods relative to implementation, allowing researchers to trace the temporal evolution of policy impacts and assess pre-treatment parallel trends.

7.13 Recent Methodological Advances

7.13.1 Challenges with Conventional TWFE Estimators

Recent econometric literature has demonstrated important limitations of conventional TWFE estimators in settings with staggered treatment adoption and heterogeneous treatment effects. Even under valid parallel trends assumptions, standard TWFE estimators may yield biased estimates of average treatment effects because they implicitly use already-treated units as control groups for later-treated units. When treatment effects vary over time—for instance, if effects grow or diminish with exposure duration—these comparisons can introduce substantial bias, potentially yielding estimates with incorrect magnitudes or even signs.

Furthermore, pre-treatment coefficients in event studies may be contaminated by treatment effect dynamics, complicating their interpretation as tests of parallel trends. These issues stem from the "forbidden comparisons" problem, where previously treated units with evolving treatment effects serve as inappropriate counterfactuals for newly treated units.

7.13.2 Heterogeneity-Robust Estimators

To address these methodological challenges, econometricians have developed several heterogeneity-robust DiD estimators. These approaches avoid problematic comparisons by carefully selecting valid control groups and employing appropriate weighting schemes. The central principle involves using only never-treated or not-yet-treated units as comparison groups when estimating treatment effects for a given cohort at a specific time.

Within the potential outcomes framework, we can define the average treatment effect for adoption cohort g at time t as:

$$ATT(g,t) = E[Y_{i,t} - Y_{i,g-1}|G_i = g] - E[Y_{i,t} - Y_{i,g-1}|G_i = g'], \text{ for any } g' > t \tag{44}$$

where G_i denotes the cohort to which unit i belongs. This formulation compares outcome changes for units treated at time g to those for units not yet treated at time t. The approach extends to using any pool of not-yet-treated units as comparisons:

$$ATT(g,t) = E[Y_{i,t} - Y_{i,g-1}|G_i = g] - E[Y_{i,t} - Y_{i,g-1}|G_i \in \mathcal{G}_{comp}] \tag{45}$$

where $\mathcal{G}_{comp}$ represents a set of comparison cohorts all satisfying $g' > t$.

Empirical implementation involves estimating cohort-time-specific effects and then aggregating these estimates into policy-relevant parameters, such as event-study coefficients measuring average effects l periods after adoption:

$$ATT_l^w = \sum_g w_g ATT(g, g + l) \tag{46}$$

The weights w_g could balance cohorts equally or reflect their relative population frequencies.

Several heterogeneity-robust estimators have been developed in recent years, including the Callaway and Sant'Anna (2021) estimator, the Sun and Abraham (2021) interaction-weighted estimator, the de Chaisemartin and D'Haultfoeuille (2020) difference-in-differences with multiple groups at multiple times approach, and the imputation-based estimator of Borusyak, Jaravel, and Spiess. Each implements the principle of avoiding forbidden comparisons while differing in their precise implementation details and efficiency properties.

7.14 Treatment Intensity Variation

7.14.1 Continuous Treatment Variables

The DiD framework extends naturally to continuous treatment variables, allowing researchers to leverage variation in treatment intensity rather than binary treatment status. With continuous treatments, the estimated coefficient captures the marginal effect of increased treatment "dosage" rather than the average effect of a binary intervention.

Treatment intensity may vary along multiple dimensions. On the extensive margin, units differ in whether they receive any treatment. On the intensive margin, treated units experience different treatment magnitudes. For example, some regions may experience larger policy shocks than others due to pre-existing conditions or implementation differences. Similarly, treatment exposure may vary temporally, with some units experiencing longer duration of treatment than others.

To incorporate treatment intensity, researchers typically construct a policy exposure measure that equals zero for untreated units and increases with treatment dosage for treated units:

$$y_{ijt} = \alpha + \beta \text{PolicyExposure}_{jt} + \delta_t + \gamma_j + \varepsilon_{ijt} \tag{47}$$

This approach offers distinct advantages. First, it does not require a completely untreated comparison group, as identification leverages comparisons between units with different exposure levels. Second, it enables estimation of dose-response relationships, potentially revealing nonlinear effects or threshold phenomena. However, this specification identifies the marginal effect of increased exposure rather than the overall policy effect; if all units receive some treatment, the baseline effect remains unidentified.

The identifying assumption shifts accordingly: without treatment, outcomes for higher-intensity groups would have evolved in parallel with those for lower-intensity groups. This continuous difference-in-differences approach requires careful consideration of what generates variation in treatment intensity and whether this variation is plausibly exogenous conditional on controls.

Treatment intensity can be operationalized in various ways depending on the context. Common measures include the magnitude of policy changes (e.g., tax rate differentials), pre-treatment characteristics that moderate exposure (e.g., baseline infection rates before a health intervention), or treatment duration (e.g., years of exposure to an educational reform). In some applications, researchers combine multiple dimensions of intensity, such as both cross-sectional variation in potential treatment magnitude and temporal variation in exposure duration.

7.15 Assessing the Validity of the Parallel Trends Assumption

When working with a minimal data structure consisting of two groups measured at only two time points, researchers face significant limitations in evaluating the validity of the parallel trends assumption. The most viable approach in such cases involves thoroughly investigating the contextual factors surrounding the policy implementation. This includes articulating a compelling case that policy adoption can be considered conditionally random after accounting

for time-invariant group characteristics (controlled for through group fixed effects) and relevant covariates at both the time-specific and group-time-specific levels.

Particular attention should be paid to potential confounding factors at the group-time level, such as contemporaneous policy changes that might also influence the outcome of interest. While policies affecting both treatment and comparison groups uniformly do not typically threaten identification, policies with differential effects across groups can compromise the ability to isolate the impact of the focal intervention. In such cases, controlling for these additional policies becomes essential to support the causal interpretation of the DiD estimates.

The assessment of parallel trends becomes substantially more feasible when richer data are available. With observations spanning multiple pre-intervention time periods or cohorts, researchers can implement empirical tests of pre-treatment trend similarities. Similarly, data encompassing multiple groups enables additional validation strategies that leverage variation across unaffected populations.

7.16 Methodological Approaches with Extended Pre-Treatment Data

When researchers have access to data covering multiple pre-intervention periods, they can implement various methods to assess the credibility of the parallel trends assumption. The fundamental approach involves conducting "placebo" or "control" experiments that utilize only pre-treatment observations. A straightforward visual inspection entails graphing the mean outcomes for treatment and comparison groups across pre-intervention periods. If these trends appear parallel prior to the intervention, this provides intuitive evidence supporting the identifying assumption.

Some researchers enhance this graphical analysis by first regressing the outcome variable on relevant covariates and then plotting the residuals, arguing that conditional parallel trends may exist even when unconditional trends diverge. Beyond visual inspection, more formal statistical tests can be implemented to evaluate pre-treatment trend similarities.

7.17 Leveraging Unaffected Groups for Validation

Another powerful validation strategy involves utilizing additional groups that remain unaffected by the policy throughout the study period. For instance, if examining the impact of maternity benefit mandates on wages of women of childbearing age, researchers might consider women beyond childbearing age or single men as placebo groups. These groups presumably experience the same broader economic trends as women of childbearing age but remain unaffected by the maternity policy.

By estimating the standard DiD model on these unaffected populations:

$$y_{ijt} = \alpha + \beta \times \text{After}_t \times \text{Treatment}_j + \gamma \times \text{After}_t + \delta \times \text{Treatment}_j + \epsilon_{ijt} \tag{48}$$

researchers can assess whether differential trends exist between treatment and comparison regions even in populations unaffected by the policy. A coefficient β statistically indistinguishable from zero would support the parallel trends assumption, as it indicates that group-specific time trends unrelated to the policy were similar across treatment and comparison regions.

7.18 Addressing Violations of the Parallel Trends Assumption

If pre-intervention data reveal differential trends between treatment and comparison groups, several methodological approaches can be employed to address this violation of the standard DiD identifying assumption. First, researchers might include additional time-varying covariates to condition on factors driving the differential trends. While time-invariant group characteristics are already controlled through group fixed effects, and aggregate time effects through time

fixed effects, variables that vary at the group-time level may absorb the differential trends if appropriately specified.

Second, the DiD estimates can be adjusted to account for pre-existing differential trends. One approach involves implementing a triple differences (DDD) estimator, which incorporates an additional dimension of comparison to difference out the differential trends. Alternatively, and more commonly, researchers may augment the standard DiD specification with group-specific time trends:

$$y_{ist} = \alpha + \beta \times \text{After}_t \times \text{Treatment}_s + \gamma_t + \delta_s + \sum_{k=1}^{S} \theta_k I(s = k) \times t + \epsilon_{ist} \tag{49}$$

where θ_k captures the linear time trend specific to state k. This approach allows each group to follow its own trend over time, with the treatment effect measured as deviations from these group-specific trajectories. In data-limited settings, more aggregated group-specific trends (e.g., at the regional rather than state level) may be implemented as a compromise solution.

7.19 The Triple Differences (DDD) Methodology

The triple differences or difference-in-difference-in-differences (DDD) estimator provides a structured approach to addressing differential trends in DiD designs. This method introduces a third dimension of comparison—typically a demographic group unaffected by the policy but subject to similar background conditions as the affected group.

In a standard implementation with two groups (treatment and comparison), two time periods (before and after), and two demographic categories (affected and unaffected), the DDD approach involves estimating separate DiD models for each demographic group. For the affected demographic group, the standard DiD equation is:

$$\begin{aligned} y_{ijt} = {} & \alpha^{AFFECTED} + \beta^{AFFECTED} \times \text{After}_t \times \text{Treatment}_j \\ & + \gamma^{AFFECTED} \times \text{After}_t + \delta^{AFFECTED} \times \text{Treatment}_j + \epsilon_{ijt}^{AFFECTED} \end{aligned} \tag{50}$$

For the unaffected demographic group, a parallel placebo DiD equation is:

$$\begin{aligned} y_{ijt} = {} & \alpha^{UNAFFECTED} + \beta^{UNAFFECTED} \times \text{After}_t \times \text{Treatment}_j \\ & + \gamma^{UNAFFECTED} \times \text{After}_t + \delta^{UNAFFECTED} \times \text{Treatment}_j + \epsilon_{ijt}^{UNAFFECTED} \end{aligned} \tag{51}$$

The DDD estimate is then calculated as $\beta^{AFFECTED} - \beta^{UNAFFECTED}$, effectively differencing out any differential trends between treatment and comparison regions that are common across demographic groups. This estimate can be obtained directly through a single regression specification:

$$\begin{aligned} y_{ijt} = {} & \alpha + \beta_{ddd} \times \text{AffectedGroup}_{ijt} \times \text{After}_t \times \text{Treatment}_j \\ & + \delta_t + \gamma_j + \pi_1 \times \text{AffectedGroup}_{ijt} + \pi_2 \times \text{AffectedGroup}_{ijt} \times \text{After}_t \\ & + \pi_3 \times \text{AffectedGroup}_{ijt} \times \text{Treatment}_j + \pi_4 \times \text{After}_t \times \text{Treatment}_j + \epsilon_{ijt} \end{aligned} \tag{52}$$

The triple interaction coefficient β_{ddd} represents the policy effect purged of differential trends, provided that all relevant main effects and two-way interactions are included as controls. This approach assumes that any differential trends between treatment and comparison regions would have been similar across affected and unaffected demographic groups in the absence of the policy.

7.20 Statistical Inference in DiD Applications

While DiD research often emphasizes consistent estimation of policy effects, appropriate inference requires attention to standard error estimation. In DiD papers, clustered standard errors have become standard practice, with clustering at the level of cross-sectional variation in the policy measure. For instance, with state-time level policy variables, standard errors should be clustered at the state level. This approach accounts for both heteroskedasticity and correlation between error terms within clusters. Conventional standard errors would inappropriately ignore both heteroskedasticity and serial correlation, while heteroskedasticity-robust standard errors would address only the former.

The asymptotic justification for clustered standard errors relies on the number of clusters approaching infinity. Consequently, small numbers of clusters create inferential challenges, typically resulting in over-rejection of null hypotheses. The mathematical framework for clustered variance estimation illustrates this limitation:

$Var(b) = (X'X)^{-1} \sum_{g=1}^{G} X_g' \hat{\Psi}_g X_g (X'X)^{-1}$

where $\hat{\Psi}_g$ represents the estimated cluster-specific variance-covariance matrix of residuals, G is the number of clusters, and M_g denotes observations within cluster g.

Several methodological approaches address inference with few clusters. Cameron and Miller (2015) summarize solutions including finite sample bias correction, t-distribution critical values, and cluster bootstrap methods with asymptotic refinement. Abadie et al. (2023) provide guidance on when standard error clustering is appropriate. Roth et al. (2023) synthesize recent literature on inference with few clusters and appropriate clustering levels. With particularly small numbers of clusters, wild cluster bootstrap procedures have become increasingly common, as developed by Cameron, Gelbach, and Miller (2008). Even with apparently large numbers of clusters, few treated clusters can produce similar inferential problems, addressed by approaches such as MacKinnon and Webb's (2020) randomization inference for DiD with few treated clusters.

7.21 Bootstrap Methods for Inference in DiD Settings

Bootstrap methods estimate the distribution of an estimator or test statistic through data resampling, as articulated by Horowitz (1999). These approaches employ the empirical distribution function (EDF) of an estimator based on observed data to approximate its true distribution through resampling procedures. The resulting EDF can be used to calculate the estimator's variance, yielding standard errors for inference. Alternatively, recognizing that empirical distributions may be asymmetric or otherwise non-normal, researchers may employ percentile methods to derive critical values for hypothesis testing. This approach orders the bootstrap estimates and identifies appropriate percentiles (e.g., 2.5th and 97.5th percentiles for a two-tailed test at 5% significance).

For independent observations, paired bootstrap involves resampling both dependent and independent variables. With N independent observations of data (y, x), the procedure entails drawing B bootstrap samples of size N by sampling with replacement from the original sample. For each bootstrap sample (y^*, x^*), the statistic of interest is calculated, ultimately yielding B estimates of this statistic. Inference then proceeds based on the empirical distribution of these estimates.

When data exhibit clustering, as in many DiD applications, modified bootstrap approaches become necessary. Paired bootstrap with clustered data resamples at the cluster level rather than the individual observation level. With N observations distributed across G clusters, resampling occurs by drawing G clusters with replacement from the original sample. While each bootstrap sample contains the same number of clusters, the number of observations may vary if clusters differ in size. The statistic is calculated for each bootstrap sample, and inference follows from the resulting empirical distribution.

For DiD applications with few clusters, wild cluster bootstrap provides an alternative

approach. Rather than resampling both dependent and independent variables, this method fixes the independent variables and resamples only the dependent variable. The procedure begins by estimating the model under the null hypothesis, generating parameter estimates $\hat{\beta}_{H_0}$ and residuals $\hat{e}_{ig} = y_i - x_i\hat{\beta}_{H_0}$. For each of B iterations, weights w_g are randomly assigned to each cluster (with all observations within a cluster receiving identical weights), typically following a Rademacher distribution ($w_g = -1$ or 1, each with probability 0.5). These weights generate pseudo-residuals $e^*_{ig} = w_g \cdot \hat{e}_{ig}$ and resampled dependent variables $y^*_{ig} = x_i\hat{\beta}_{H_0} + e^*_{ig}$. The statistic is calculated for each bootstrap sample, and inference proceeds based on the empirical distribution. For extremely small numbers of clusters, Webb's six-point distribution provides an alternative to the Rademacher distribution, employing weights $\{-\sqrt{\frac{3}{2}}, -1, -\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 1, \sqrt{\frac{3}{2}}\}$, each with probability $\frac{1}{6}$.

8 Synthetic Control Method

8.1 Introduction and Motivation

The Synthetic Control Method addresses a fundamental challenge in causal inference: identifying the impact of policy interventions or events that affect aggregate entities such as countries, regions, or firms. Traditional difference-in-differences approaches often rely on subjectively selected comparison groups based on researcher intuition about similarities between treated and control units, such as geographic proximity or cultural affinity. In contrast, the Synthetic Control Method, introduced by Abadie and Gardeazabal (2003) in their American Economic Review paper on the economic costs of conflict in the Basque Country, employs a data-driven approach to construct comparison groups. This method creates a "synthetic control" for the treatment unit as a weighted average of untreated units, with weights selected to maximize pre-intervention similarity between the treated unit and its synthetic counterpart. This approach enhances the credibility of counterfactual scenarios and reduces researcher discretion in the selection of comparison units.

8.2 Methodological Framework

The formal framework of the Synthetic Control Method, as detailed in Abadie's 2021 Journal of Economic Literature article, begins with a setting of $J + 1$ units where only unit 1 receives an intervention after time period T_0, while the remaining J units serve as potential controls. For any unit j at time t, we denote the outcome as Y_{jt}. The causal effect of the intervention on unit 1 at time $t > T_0$ is defined as $\tau_{1t} = Y^I_{1t} - Y^N_{1t}$, where Y^I_{1t} represents the observed outcome with intervention and Y^N_{1t} represents the counterfactual outcome without intervention. Since Y^N_{1t} is unobservable, the synthetic control approach constructs an estimate using a weighted average of outcomes from the untreated units. This provides an estimate of the treatment effect: $\hat{\tau}_{1t} = Y_{1t} - \sum_{j=2}^{J+1} w^*_j Y_{jt}$.

8.3 Weight Determination Process

The determination of optimal weights in the Synthetic Control Method requires rich pre-intervention data and proceeds through a two-stage optimization process. Let $X_1 = (Z_1, Y_1)$ represent the vector of predictor variables for the treated unit, and $X_0 = (Z_j, Y_j)$ for the non-treated units. The predictor variables typically include outcome lags and other relevant covariates. The first stage selects weights w^* to minimize the distance between characteristics of the treated unit and the synthetic control: $\|X_1 - X_0W\| = \left(\sum_{h=1}^{k} v_h(X_{h1} - w_2X_{h2} - \cdots - w_{J+1}X_{h,J+1})^2\right)^{1/2}$, where weights w_j are restricted to be non-negative and sum to one. The parameters v_h represent the relative importance of each predictor variable.

In the second stage, the method selects the vector V of variable weights that minimizes the mean squared prediction error (MSPE) in the pre-intervention period. For each potential V, we compute the corresponding optimal $W(V)$ through the minimization described above. The final selection among these combinations aims to minimize the pre-intervention prediction error: $\sum_{t \in T_0} (Y_{1t} - w_2(V)Y_{2t} - \cdots - w_{J+1}(V)Y_{J+1,t})^2$ for some set $T_0 \subseteq \{1, 2, \ldots, T_0\}$ of pre-intervention periods. This two-stage process ensures that the synthetic control closely tracks the trajectory of the treated unit before intervention, strengthening the plausibility of the counterfactual.

8.4 Inference and Validation

Statistical inference in the Synthetic Control Method primarily relies on placebo tests rather than traditional asymptotic inference, given the limited number of units typically involved. One prominent approach involves iteratively applying the method to each control unit as if it had received the treatment, generating a distribution of "placebo" treatment effects. The magnitude of the actual treatment effect can then be evaluated relative to this placebo distribution. Additionally, researchers may conduct "in-time" placebo tests by artificially backdating the intervention to a pre-treatment period; significant effects in these tests would suggest pre-existing differences rather than causal impacts.

8.5 Empirical Implementation

The Synthetic Control Method has been implemented in various statistical packages. In Stata, the procedure is executed using the `synth` command. A typical implementation, as in the replication of the California tobacco control program analysis, would involve commands such as:

```
use synth_smoking
tsset state year
synth cigsale beer lnincome retprice age15to24 cigsale(1988)
cigsale(1980) cigsale(1975) , trunit(3) trperiod(1989)
xperiod(1980(1)1988) nested fig
```

In this specification, the intervention unit (California) is identified as unit 3, with 1989 as the intervention year. The pre-intervention period spans 1980-1988, and the model incorporates both contemporaneous predictors and lagged outcome variables.

8.6 Recent Methodological Advances

Recent methodological developments have extended the Synthetic Control framework to address various practical challenges. A notable advancement is the Synthetic Difference-in-Differences (SDID) approach proposed by Arkhangelsky et al. (2021) in the American Economic Review. This method integrates elements of both difference-in-differences and synthetic control approaches. While standard synthetic control methods focus on matching levels of outcomes, SDID aims to construct a weighted combination of control units that satisfies the parallel trends assumption. This innovation addresses situations where no convex combination of control units can adequately match the treated unit's level, while standard difference-in-differences methods cannot identify suitable comparison groups for the parallel trends assumption.

The SDID estimator can be implemented in Stata using the `sdid` package, available through the command `ssc install sdid`. Researchers interested in applying this method can find further guidance and code repositories at resources such as the GitHub repository maintained by Daniel Pailanir.

8.7 Conclusion

The Synthetic Control Method represents a significant advancement in the causal analysis of aggregate-level interventions. By offering a data-driven approach to counterfactual construction, it reduces researcher discretion in comparison group selection while maintaining transparency in weight determination. The method has found applications across various domains in economics and political science, and continues to evolve through methodological refinements. Comprehensive resources for researchers interested in this approach include Abadie's 2021 JEL article, as well as review pieces by Abadie and Cattaneo (2018) and Athey and Imbens (2017), which situate the method within the broader landscape of program evaluation techniques.

9 IV Estimation

9.1 Instrumental Variables Estimation

The central question in instrumental variables (IV) estimation is determining the causal effect of a variable x on an outcome y, formalized in the structural equation:

$$y_i = \alpha + \beta x_i + u_i. \tag{53}$$

A core concern is the potential endogeneity of x, which arises when x is correlated with the error term u_i, thereby rendering ordinary least squares (OLS) estimates biased and inconsistent. Suppose a variable z exists that satisfies two critical conditions: it must be correlated with the endogenous regressor x (relevance), and it must be uncorrelated with the error term u_i (exogeneity). Formally, these conditions are:

$$\text{(IV.A1)} \quad \text{Cov}(z_i, x_i) \neq 0 \tag{54}$$

$$\text{(IV.A2)} \quad \text{Cov}(z_i, u_i) = 0. \tag{55}$$

If both conditions are satisfied, z serves as a valid instrument for x, and IV estimation provides a consistent estimate of β.

In the simple bivariate case, the OLS estimator is given by

$$b_{OLS} = \frac{\text{Cov}(x, y)}{\text{Var}(x)}. \tag{56}$$

Expanding the numerator yields

$$\text{Cov}(x, y) = \beta \text{Var}(x) + \text{Cov}(x, u), \tag{57}$$

which implies that

$$b_{OLS} = \beta + \frac{\text{Cov}(x, u)}{\text{Var}(x)}. \tag{58}$$

When $\text{Cov}(x, u) \neq 0$, OLS estimates are biased. In contrast, the IV estimator in the bivariate case takes the form

$$b_{IV} = \frac{\text{Cov}(z, y)}{\text{Cov}(z, x)}. \tag{59}$$

Substituting in the structural equation, and under the assumption of valid instruments, this simplifies to $b_{IV} = \beta$, ensuring consistency.

9.2 The Wald Estimator

When the instrument z is binary, the IV estimator simplifies to the Wald estimator:

$$b_{IV} = \frac{E[y_i|z_i = 1] - E[y_i|z_i = 0]}{E[x_i|z_i = 1] - E[x_i|z_i = 0]}. \tag{60}$$

This expression provides an intuitive ratio of differences in means across treatment groups and is particularly useful in evaluating randomized controlled trials and natural experiments.

9.3 Terminology in IV Estimation

The structural equation represents the economic relationship of interest, for instance:

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i, \tag{61}$$

where x_1 is endogenous and x_2 is exogenous. The first stage equation regresses the endogenous regressor on all exogenous variables and instruments:

$$x_{1i} = \pi_1^{first} + \pi_2^{first} x_{2i} + \pi_3^{first} z_i + u_i^{first}. \tag{62}$$

Identification relies on exclusion restrictions, which assume that instruments z influence y only through their effect on x_1. The explanatory power of instruments should be reported, typically via the F-statistic on the instruments' coefficients in the first stage regression.

The reduced form equation relates the outcome y directly to all exogenous variables:

$$y_i = \pi_1^{RF} + \pi_2^{RF} x_{2i} + \pi_3^{RF} z_i + u_i^{RF}. \tag{63}$$

This equation, along with the first stage, can be estimated using OLS. Whether or not they are of direct interest, their coefficients are integral to understanding IV estimation.

The system is said to be just identified if the number of instruments equals the number of endogenous regressors, overidentified if there are more instruments than endogenous regressors (enabling overidentification tests), and not identified if there are too few instruments.

9.4 Indirect Least Squares

In cases with one endogenous regressor and one instrument, the IV estimator can be obtained as a ratio of reduced form to first stage coefficients:

$$\beta_1 = \frac{\pi_3^{RF}}{\pi_3^{first}}. \tag{64}$$

This approach is known as the indirect least squares (ILS) estimator. With multiple instruments, alternative ILS estimators can be computed by substituting in different instruments, as in the case of using z_i and w_i separately.

9.5 Two-Stage Least Squares Estimation

Among linear IV estimators, two-stage least squares (2SLS) is typically the most efficient. In the first stage, the endogenous variable is regressed on all instruments and exogenous variables, and fitted values are computed. In the second stage, the structural equation is estimated using these fitted values in place of the endogenous regressor:

1. First stage: $x_{1i} = \hat{\pi}_1^{first} + \hat{\pi}_2^{first} x_{2i} + \hat{\pi}_3^{first} z_i$.
2. Second stage: $y_i = \alpha + \beta_1 \hat{x}_{1i} + \beta_2 x_{2i} + u_i$.

Although the process appears two-step, software such as Stata typically implements 2SLS in a single step to ensure correct standard errors and minimize user error.

9.6 Sources of Valid Instruments

The credibility of IV estimation rests on finding valid instruments that satisfy both relevance and exogeneity. Instruments must be strongly correlated with the endogenous regressor and must not independently affect the outcome variable. Theoretical reasoning and structural models can inform the selection of instruments by addressing the specific sources of endogeneity. Systems of simultaneous equations often provide natural sources of instruments through exclusion restrictions.

Randomized experiments and natural experiments are valuable sources of exogenous variation. These include policy changes, institutional rules, or exogenous shocks like weather deviations. However, even if an instrument provides variation in x, researchers must ensure it satisfies the exclusion restriction. If z affects y through multiple channels, it cannot be considered valid.

9.7 Randomized Experiments and IV

Randomized controlled trials can be interpreted through an IV framework, especially when there is non-compliance. Suppose z is a binary indicator for treatment assignment, and x denotes actual treatment received. The reduced form regression:

$$y_i = \pi_1^{RF} + \pi_3^{RF} z_i + u_i^{RF} \tag{65}$$

provides the intention-to-treat (ITT) effect. If compliance is imperfect, IV methods can recover the treatment effect for the treated. The Wald estimator:

$$\frac{E[y_i|z_i = 1] - E[y_i|z_i = 0]}{E[x_i|z_i = 1] - E[x_i|z_i = 0]} \tag{66}$$

reflects the local average treatment effect (LATE), which is the effect of treatment for compliers.

9.8 The Local Average Treatment Effect (LATE)

The IV estimator identifies the LATE, defined as the average treatment effect for individuals who comply with the instrument (e.g., take treatment when assigned to treatment and not otherwise):

$$\text{LATE} = E[y_1 - y_0|D_1 > D_0]. \tag{67}$$

Let z be the indicator for treatment assignment, D the actual treatment received, and y_1 and y_0 denote potential outcomes with and without treatment. Individuals fall into one of four categories: never-takers, always-takers, compliers, and defiers. Under the assumption of monotonicity (no defiers), LATE is identified.

When treatment effects are heterogeneous or monotonicity is violated, IV no longer estimates the average treatment effect (ATE) or the treatment effect on the treated (TT). However, LATE remains policy-relevant and informative, particularly in the presence of imperfect compliance and heterogeneous responses.

9.9 Where Do Valid Instruments Come From?

Policy changes can serve as a valuable source of exogenous variation in the explanatory variable of interest, typically denoted as x. These changes, particularly when implemented at different times or in different locations, offer a framework that resembles difference-in-differences (DiD) designs. In such cases, the policy can be used to construct an interaction term (e.g., After $\times$ Treatment), capturing the differential exposure to the policy. A first-stage regression might take the form $x = \pi_1 + \pi_2(\text{After} \times \text{Treatment}) + \pi_3\text{After} + \pi_4\text{Treatment} + u$, where π_2 represents the impact of the policy on x. This variation can then be used in an instrumental variables (IV)

strategy to estimate the causal effect of x on an outcome y. However, for the instrument to be valid, one must ensure that the policy-induced variation in x is not correlated with other determinants of y, necessitating appropriate control for main effects and potential differential trends. Even if a policy shifts x, it may still fail to meet the exclusion restriction required for a valid instrument. Therefore, careful background research on the policy is essential. If the policy plausibly affects only x and not y directly or through other channels, then it may serve as a valid instrument.

Another fruitful source of instruments is administrative or institutional rules that create sharp thresholds or discontinuities in treatment assignment. These setups often align with regression discontinuity (RD) designs. A typical case involves a running variable x_f that determines eligibility for a treatment x based on a cutoff. A dummy variable indicating whether the running variable exceeds the cutoff can be used as an instrument for x to estimate its effect on y, provided that the running variable is adequately controlled for in the analysis. This design helps isolate local exogenous variation near the threshold.

This approach is well illustrated by Angrist and Krueger (1991), who used quarter of birth as an instrument for educational attainment to estimate its effect on earnings. Another example is Bleakley and Chin (2004), who examined the impact of English-language proficiency on earnings by instrumenting language skills with a variable capturing whether an immigrant arrived in the United States at a young age from a non-English-speaking country, controlling for both components individually.

Hoekstra (2009) provides another RD-based example by examining the earnings effect of attending a flagship state university. The instrument in this case is a dummy variable indicating whether an applicant's test score was above the admission cutoff. The credibility of the RD-IV approach hinges on the assumption that no other factors jump discontinuously at the cutoff, aside from the treatment of interest.

Angrist and Lavy (1999) offer an additional illustration, using Maimonides' rule for determining class size as an instrument to estimate the effect of class size on student achievement. In this context, the rule creates predictable variation in class size based on enrollment numbers, generating exogenous variation suitable for IV estimation.

Across all these examples, the core principle remains consistent: for an instrument to be valid, it must satisfy relevance (correlation with x) and exogeneity (no correlation with the error term in the outcome equation). The choice of instrument must be guided by theory, institutional knowledge, and empirical validation.

Valid instruments are central to credible instrumental variable (IV) estimation. These instruments must satisfy two core requirements: they must be correlated with the endogenous regressor of interest (relevance) and uncorrelated with the structural error term (exogeneity). A variety of empirical strategies have emerged to identify valid instruments, including those based on geography, weather, historical factors, and shift-share designs.

Geographic instruments leverage spatial variation in features that are plausibly exogenous. While geographic variation can offer powerful identification, caution is warranted, as many geographic factors may directly influence the outcome variable. Instruments based on geography are more credible when justified by clear, exogenous mechanisms such as historical engineering decisions or nonlinearities tied to natural constraints or policy changes.

Weather-based instruments rely on the stochastic nature of weather variation. Crucially, it is often the deviations from usual weather patterns, rather than weather levels themselves, that serve as valid instruments. This distinction is essential for ensuring that the instrument captures exogenous variation rather than being correlated with latent factors affecting the outcome.

Historical instruments draw on the persistence of institutions and practices over time. Institutional structures that originated for reasons unrelated to the present context may continue to exert influence on current variables of interest, providing quasi-random variation. A prominent example of this is the class of shift-share instruments. In such designs, the instrument is

constructed as the product of historical shares (e.g., employment composition by sector in a base year) and aggregate shocks (e.g., national sector growth), predicting local exposure to exogenous shifts.

Several recent studies offer detailed guidance on implementing and evaluating shift-share instruments. Notable references include Borusyak, Hull, and Jaravel (2025), who provide a practical framework for applying shift-share methods, and earlier contributions such as Goldsmith-Pinkham, Sorkin, and Swift (2019); Adão, Kolesár, and Morales (2019); and Borusyak et al. (2022). Key identification challenges in shift-share settings include assessing whether the shifts are truly exogenous and whether the historical shares can be treated as predetermined.

While IV estimation is mechanically straightforward, identifying a truly valid instrument remains challenging. Researchers are encouraged to review existing literature for instruments used in similar contexts, even if the outcome differs. Additionally, policy changes that affect the endogenous variable are promising sources of exogenous variation. Reduced-form effects stemming from such changes can inform the plausibility of an instrument and may, in some cases, be leveraged directly for IV estimation.

9.10 Interpretation of IV Estimates

The IV estimator identifies the local average treatment effect (LATE), which represents the causal effect of the treatment on compliers—individuals whose treatment status is influenced by the instrument. This concept was formalized by Imbens and Angrist (1994) and is discussed extensively in MHE Chapter 4.

It is important to distinguish LATE from the average treatment effect (ATE) and the average treatment effect on the treated (ATT). In general, IV estimation does not recover ATE or ATT unless additional assumptions hold. For instance, under constant treatment effects, LATE coincides with both ATE and ATT. Similarly, if all untreated individuals are never-takers or if treatment can only be obtained in one way, then LATE may approximate ATT.

Nonetheless, the external validity of IV estimates can be limited. Although LATE may be informative, it applies only to a specific subpopulation. In some settings, however, this subpopulation (i.e., the compliers) may be of particular policy interest.

Recent work by Blandhol, Bonney, Mogstad, and Torgovitsky (NBER Working Paper No. 29709) explores the conditions under which the two-stage least squares (2SLS) estimate corresponds to LATE, highlighting potential limitations of conventional interpretations.

9.11 Alternative IV Estimators: LIML

In overidentified settings, several IV estimators are available. While the 2SLS estimator is the most common, it can suffer from finite sample bias. An alternative is the limited information maximum likelihood (LIML) estimator, which is consistent and often has reduced finite-sample bias relative to 2SLS. In Stata, the LIML estimator can be implemented via the `ivregress 2sls, liml` command. When the model is just-identified, all IV estimators—2SLS, LIML, and others—coincide.

9.12 Pitfalls in Instrumental Variable Estimation

One of the most frequent issues in IV estimation is the use of weak or invalid instruments. The validity of an instrument hinges on the satisfaction of two conditions: (1) the instrument must be correlated with the endogenous regressor (relevance) and (2) the instrument must not affect the outcome variable except through its effect on the endogenous regressor (exogeneity). Angrist and Krueger (2001) emphasize that a good instrument is one whose relationship to the endogenous regressor is both strong and theoretically justified, and whose exclusion from the outcome equation is credible.

Weak instruments lead to biased IV estimates and unreliable inference. The relevance of instruments can be assessed via the first-stage regression. A common rule of thumb is that if the F-statistic for the excluded instruments is below 10, the instruments are weak. Seminal contributions on this issue include Staiger and Stock (1997) and Bound, Jaeger, and Baker (1995).

To mitigate the issues arising from weak instruments, researchers are advised to report first-stage results and consider inference methods robust to weak identification. For example, the Anderson-Rubin test offers inference that remains valid even in the presence of weak instruments. Keane and Neale (2023), Mikusheva (2013), and Andrews, Stock, and Sun (2019) offer detailed treatments of inference under weak instruments.

A separate concern arises when the exclusion restriction is violated. If the instrument affects the outcome through channels other than the endogenous regressor, it cannot be considered valid. Overidentification tests, such as the NR^2 test, offer a way to evaluate the joint validity of instruments when the model is overidentified ($K > G$). However, these tests have limitations: they cannot confirm validity, and rejections may arise from treatment effect heterogeneity rather than instrument invalidity.

Random assignment, while useful, does not automatically produce a valid instrument. An instrument generated through experimental or quasi-experimental variation must still satisfy the exclusion restriction. If the instrument has multiple causal pathways to the outcome, it cannot serve as a valid instrument for isolating the effect of a single regressor.

The external validity of IV estimates remains an important issue. Since IV identifies the effect for compliers, researchers must consider the representativeness of this group. Comparing compliers to the general population, treated individuals, or policy-relevant subpopulations helps assess the generalizability of the findings. If the study offers the only credible causal estimate available, or if the compliers are substantively interesting, the results may still carry significant value.

9.13 Elements of a Convincing IV Study

A compelling IV paper begins with a well-motivated research question, typically framed as the causal effect of a treatment or variable x on an outcome y. The author should provide a clear explanation of why OLS estimation may be biased, often due to endogeneity arising from omitted variables, measurement error, or reverse causality.

The core of the argument rests on the validity of the instrument. This includes a conceptual narrative linking the instrument to the endogenous regressor, empirical evidence of a strong first stage, and a plausible justification for the exclusion restriction. Visualizations of first-stage and reduced-form relationships can enhance credibility.

Empirical results are typically presented in a series of tables: first-stage regression estimates, reduced-form regression estimates, and the final 2SLS estimates of the structural equation. When appropriate, OLS estimates should be presented for comparison with the 2SLS results.

Finally, the interpretation of results should reflect on the likely characteristics of compliers, the external validity of the estimates, and the potential policy implications. A discussion of robustness checks and specification tests further strengthens the credibility of the findings.