Notes on Econometrics

Victor Li

Autumn Sememster, 2023

Contents

Part I

THE REGRESSION FRAMEWORK

1 Preparation and foundational knowledge

Reading this note book, you should understand all the basics of higher math.

1.1 Core of econometrics

A student studying econometrics should be able to differentiate that

- causal relationship
- correlative relationship

are two different things. The latter may be intuitive in live but is essentially misguiding in true meaning. [1]

The core pursuit of econometrics is to move beyond simple observation to rigorously estimate causal relationships.

We often observe that two things move together—a correlation—but the goal is to determine if a change in one causes a change in the other. For example, an online retailer might see that on days with high advertising spending, they also have high sales. This is a correlation. But did the advertising cause the increase in sales, or did both rise because of an external factor, like a holiday weekend? Econometrics provides the theoretical framework and practical tools to answer such questions with data.

Only with this kind of understanding, you can bear in mind that the core mission of econometrics is to use statistical methods and mathematical models to give empirical content to economic theory. In simpler terms, it's about **turning broad economic ideas into testable, quantifiable statements**. This is what sets econometrics apart from other causal studies.

1.2 Data structure

The data we use dictates the methods we can apply. Econometric data is typically organized in one of three ways:

- **Cross-Sectional Data**: A snapshot of many different entities at a single point in time. Example: Data on daily sales and advertising expenditure for 500 different online stores on December 1st, 2023. Each row is a different store.
- **Time Series Data**: Observations of a single entity over multiple time periods. Example: Data on the daily sales and advertising expenditure for one specific online store from January 1st to December 31st, 2023. Each row is a different day.

[1] Modern day science has brought us intuitive idea. The simple and commonly accepted idea of science is that agnosticism and determinism are required if you want to be rational and truth-seeking. Which, is mostly true.

- **Panel Data (Longitudinal Data)**: A combination of the two, observing multiple entities over multiple time periods. Example: Data on the daily sales and advertising expenditure for 500 different online stores, tracked each day for the entire year of 2023. This is incredibly powerful as it allows us to control for factors unique to each store that don't change over time.

1.3 Understanding the Shape of Your Data: Moments

Before modeling, we must understand the fundamental characteristics of our variables. The shape of a variable's distribution can be summarized by its statistical moments.

Skewness Skewness measures the asymmetry of a distribution. A perfectly symmetric distribution has zero skewness.

Definition: Skewness (Standardized 3rd Central Moment)

$$\tilde{\mu}_3 = \frac{E[(Y - \bar{Y})^3]}{\sigma_Y^3} \tag{1}$$

The skewness of a random variable Y is the average of its cubed standardized deviations. Cubing the deviations preserves their sign.

- $\approx 0 \iff$ Symmetric, almost like normal distribution.
- $> 0 \iff$ Right Skew, right side lower, meaning more outliers ar right side.
- $< 0 \iff$ Left Skew, left side lower, meaning more outliers ar left side.

Kurtosis Kurtosis measures the "tailedness" of a distribution. It tells us how much of the data's variance is driven by infrequent, extreme events (fat tails) versus frequent, modest deviations.

Definition: Kurtosis (Standardized 4th Central Moment)

$$Kurt = \frac{E[(Y - \bar{Y})^4]}{\sigma_Y^4} \tag{2}$$

The kurtosis of Y is the average of its standardized deviations raised to the fourth power. The fourth power makes extreme values dominate the calculation.

For a normal distribution, the kurtosis is 3. Based on this standard,

- $\approx 3 \iff$ (Mesokurtic): The distribution has tails similar to a normal distribution.
- $> 3 \iff$ (Leptokurtic): "Fat tails." The distribution has more mass in its tails than a normal distribution. In finance, this implies that extreme market movements (crashes or booms) are more likely than a normal model would predict.
- $< 3 \iff$ (Platykurtic): "Thin tails." Extreme events are less likely than in a normal distribution.

2 Linear Regression Model

The linear regression model is the workhorse of econometrics. It provides a simple yet powerful way to model how a dependent variable, Y changes in response to an independent (or explanatory) variable, X.

2.1 The Population and the Sample

It is crucial to distinguish between the unobservable reality we wish to understand and the limited data we have to work with. The reality is the population of data, the part of reality that we are able to direct observe is the sample of the population.

The Population Regression Function (PRF) This is the true, underlying relationship that governs how $Y$is determined. It is a theoretical ideal that we can never observe directly.

For a simplified version of function format (that is the simple form of linear function), the PRF can be stated as:

$$E(Y|X) = \beta_0 + \beta_1 X = E(Y|X) \text{ (regression equation)} \tag{3}$$

$$Y = \beta_0 + \beta_1 X + \underbrace{\mu}_{\text{disturbance}} = E(Y|X) + \mu \text{ (regression model)} \tag{4}$$

- β_0 (Intercept) and β_1(Slope) are the population parameters. They are fixed, unknown constants. β_1 is typically the object of our interest; it represents the true causal effect on Y of a one-unit change in X.
- μ is the unobservable disturbance or error term. It captures all other factors that affect Yapart from X, as well as any inherent randomness. In our retail example, if $Y$is sales and $X$is ad spend, μincludes competitor actions, news events, website glitches, and customer mood.

The Sample Regression Function (SRF) Since we cannot see the entire population, we use a random sample of data to estimate the PRF. The SRF is the estimated relationship for our specific sample.

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \text{ (regression equation)} \tag{5}$$

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \underbrace{e}_{\text{error}} = \hat{Y}_i + e_i \text{ (regression model)} \tag{6}$$

- $\hat{\beta}_0$ and $\hat{\beta}_1$ are estimators (or coefficients). They are our data-driven ”best guesses” for the true population parameters β_0 and β_1. The ”hat” notation ($\hat{\cdot}$) always denotes an estimate.
- e_i is the residual. It is the sample counterpart of the disturbance μ and represents the difference between the actual value Y_i and the predicted value from our model, $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$. Thus, $e_i = Y_i - \hat{Y}_i$.

2.2 OLS Estimator

The Ordinary Least Squares (OLS) Estimator How do we choose the best estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ to draw a line through our data points? The OLS method provides the answer: we choose the values that minimize the sum of the squared residuals.

The goal is to minimize the deviation of estimation from the real world

$$\min \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \tag{7}$$

$$= \min \sum_{i=1}^{n} e_i^2 \tag{8}$$

$$= \min \sum_{i=1}^{n} [Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)]^2 \tag{9}$$

The OLS Principle The goal is to find the line that is, on the whole, ”closest” to all the data points. We define ”closeness” as the vertical distance (e_i). By squaring each residual, we ensure that negative and positive deviations don’t cancel out and that larger errors are penalized more heavily.

$$\min_{\hat{\beta}_0, \hat{\beta}_1} \sum_{i=1}^{n} e_i^2 = \min_{\hat{\beta}_0, \hat{\beta}_1} \sum_{i=1}^{n} [Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)]^2 \tag{10}$$

So OLS is basically an optimization problem.

FOC: $\begin{cases} \frac{\partial \min}{\partial \hat{\beta}_0} = 0 \\ \frac{\partial \min}{\partial \hat{\beta}_1} = 0 \end{cases}$ $\Rightarrow$ yielding the optimal coefficients $\begin{cases} \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \\ \hat{\beta}_1 = \frac{S_{XY}}{S_X^2} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \end{cases}$

Using OLS, we would have fitted value $\hat{Y}_i$ and residual value $\hat{e}_i$

$$\begin{cases} \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i, i = 1, 2 \ldots n \\ \hat{e}_i = Y_i - \hat{Y}_i, i = 1, 2 \ldots n \end{cases} \tag{11}$$

2.3 The Gauss-Markov Theorem and BLUE

Why should we prefer the OLS method over any other way of fitting a line? The Gauss-Markov theorem provides the theoretical justification. It states that if a set of assumptions holds, then the OLS estimator is the Best Linear Unbiased Estimator (BLUE).

The Gauss-Markov Assumptions (Classical Linear Regression Model - CLRM) is required by:

1. Linearity in Parameters: The model is linear in β_0 and β_1.
2. Random Sampling: The data is a random sample from the population.

3. Variation in X: The sample outcomes for X are not all the same value.
4. Zero Conditional Mean ($E(\mu|X) = 0$): This is the most critical assumption. It states that the unobserved factors in μ are, on average, unrelated to the value of X. In our example, it means that a competitor's promotion (part of μ) is not systematically launched on days when we happen to increase our ad spend (X). A violation of this assumption leads to biased estimates.
5. Homoskedasticity ($var(\mu|X) = \sigma^2$): The variance of the unobserved factors is constant for all values of X. This means the "unpredictability" of sales is the same on high-spend ad days as it is on low-spend ad days.

> **Note 1** (What is BLUE?)**.** *If the five Gauss-Markov assumptions hold, the OLS estimator has the following desirable properties:*
>
> - **Best:** It has the minimum variance among all linear unbiased estimators. This means OLS is the most precise or efficient.
> - **Linear:** The estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are linear functions of the dependent variable Y.
> - **Unbiased:** On average, the estimator will equal the true population parameter. Formally, $E(\hat{\beta}) = \beta$. Your estimate from one sample may be high or low, but if you could repeat the sampling process infinitely, the average of your estimates would be the true value.
> - **Estimator:** It is a rule that tells us how to use data to compute an estimate of a population parameter.
>
> In essence, the theorem gives us confidence that, under ideal conditions, OLS is the optimal choice. Much of advanced econometrics is concerned with what to do when one or more of these assumptions are violated.

2.4 Measures of Fit

Once we have estimated a regression model using OLS, a natural question arises: how well does our model actually fit the data? We need metrics to quantify the model's explanatory power.

Decomposing Variance The foundation of the most common goodness-of-fit measure is the decomposition of the total variation in the dependent variable, Y. The total variation is the sum of the squared deviations of each Y_i from its mean $\bar{Y}$. This is called the **Total Sum of Squares (TSS)**.

This total variation can be broken into two parts: the portion that is explained by our model, called the **Explained Sum of Squares (ESS)**, and the portion that is left unexplained, which is captured by the residuals and is called the **Sum of Squared Residuals (SSR)**.

$$\underbrace{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}_{\text{TSS}} = \underbrace{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}_{\text{ESS}} + \underbrace{\sum_{i=1}^{n} e_i^2}_{\text{SSR}} \tag{12}$$

Degree of Freedom In statistics, degrees of freedom (df) refers to the number of values in a final calculation that are free to vary. A good way to think about it is as the number of independent pieces of information that you can use to estimate a parameter.

> **Note 2** (Degree of freedom)**.**
>
> For the decomposition, degree of freedom is actually $\begin{cases} TSS : n-1 \\ ESS : k \\ SSR : n-k-1 \end{cases}$.
>
> Denominators below are actually degree of freedom. Only in large samples approximated by n.

For Y,

$$\frac{TSS}{n} = var(Y) \tag{13}$$

$$TSS = n \cdot var(Y) = n \cdot \frac{\sum_i^n (Y_i - \bar{Y})^2}{n} = \sum_i^n (Y_i - \bar{Y})^2 \tag{14}$$

$$SE(Y) = \sqrt{var(Y)} = \sqrt{\frac{TSS}{n}} = \sqrt{\frac{\sum_i^n (Y_i - \bar{Y})^2}{n}} \tag{15}$$

For $\hat{Y}$,

$$\frac{ESS}{n} = var(\hat{Y}) \tag{16}$$

$$ESS = \sum_i^n (\hat{Y}_i - \bar{Y})^2 \tag{17}$$

$$SE(\hat{Y}) = \sqrt{var(\hat{Y})} = \sqrt{\frac{ESS}{n}} = \sqrt{\frac{\sum_i^n (\hat{Y}_i - \bar{Y})^2}{n}} \tag{18}$$

For e,

$$\frac{SSR}{n} = var(e) \tag{19}$$

$$SSR = \sum_i^n (e_i - \bar{e})^2 \tag{20}$$

$$SER = SE(e) = \sqrt{var(e)} = \sqrt{\frac{SSR}{n}} = \sqrt{\frac{\sum_i^n (e_i - \bar{e})^2}{n}} \text{ (also the SE of the regression)} \tag{21}$$

R-squared A common and easy way to measure goodness of fit is by using R^2. It is considered an indicator to judge a model.

The R^2, or the coefficient of determination, formalizes this decomposition into a single, intuitive metric. It measures the fraction of the total variance in Y that is explained by the explanatory variable(s) in the model.

$$R^2 = \frac{ESS}{TSS} = \frac{\sum_i^n (\hat{Y}_i - \bar{Y})^2}{\sum_i^n (Y_i - \bar{Y})^2} \tag{22}$$

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS} = \frac{\text{explained by the estimated model}}{\text{total sample data}} \tag{23}$$

Standard Error of the Regression (SER) While R^2 is a relative measure of fit, the SER is an absolute measure. It estimates the standard deviation of the regression disturbance μ. In practical terms, it tells us the typical size of the regression error, or how far our predictions typically are from the actual outcomes.

Definition: Standard Error of the Regression (SER)

$$SER = s_e = \sqrt{\frac{SSR}{n-2}} = \sqrt{\frac{\sum_i^n e_i^2}{n-2}} \tag{24}$$

The SER is measured in the same units as the dependent variable, Y.

A lower SER implies a more accurate model in terms of prediction. In our retail example, if sales (Y) are measured in dollars, an SER of $500 means our model's predictions of daily sales are typically off by about $500.

3 Hypothesis and Test

Remember the metrics before are used to test how good a model is. When a number is calculated, is it 100 percent convincing? No, they are not. Because they are calculated under assumptions and simplization of reality.

Would a different sample produce a different estimate? The fundamental question of statistical inference is: how confident are we that our estimated relationship is real and not just a fluke of our particular sample? For instance, is the true effect of advertising on sales, β_1, actually zero?

We care about significance in statistics because it provides a way to quantify the likelihood that an observed result in a study is not due to random chance.

Hypothesis testing provides a formal framework to answer this.

3.1 The t-test

The most common method for testing a hypothesis about a single regression coefficient is the t-test.[2] It follows a structured process to determine whether to accept or reject a claim about the true population parameter.

Step 1: State the Hypotheses

We begin by stating a **null hypothesis (H_0)**, which represents the "status quo" or a benchmark of no effect, and an **alternative hypothesis (H_1)**, which is what we are trying to establish. The most common test is for statistical significance:

$$\begin{cases} H_0 : \beta_1 = 45812 \\ H_1 : \beta_1 \neq 45812 \end{cases}$$

This is a two-sided test, as we are interested in deviations from zero in either direction (positive or negative).

Step 2: Calculate the t-statistic

The t-statistic (or t-value) measures how many standard errors our estimated coefficient, $\hat{\beta}_1$, is away from the value hypothesized under the null. A larger t-statistic implies that our estimate is less likely to have occurred by random chance if the null hypothesis were true.

Definition: t-statistic

$$t = \frac{\hat{\beta}_1 - \beta_{1,H_0}}{se(\hat{\beta}_1)} = \frac{\text{Estimation} - \text{Hypothesized Value}}{\text{Standard Error of Estimation}} \tag{25}$$

where $se(\hat{\beta}_1)$ is the standard error of our coefficient estimate, a measure of its sampling variability. When testing for significance, β_{1,H_0} is 0.

Step 3: Make a Decision

We have two common, and equivalent, ways to decide whether our t-statistic is "large enough" to reject the null hypothesis.

The p-value Approach: The p-value is the probability of observing a t-statistic as extreme as, or more extreme than, the one we calculated, assuming the null hypothesis is true.

$$p = 2\Phi(-|t|) \tag{26}$$

Here, Φ is the cumulative distribution function of the standard normal distribution (a good approximation for the t-distribution in large samples). We compare the p-value to a pre-determined **significance level (α)**, usually 0.05 (5%), 0.01 (1%), or 0.10 (10%).

- If $p < \alpha$, we **reject the null hypothesis**. The result is "statistically significant at the α level." We have strong evidence that β_1 is not zero.
- If $p \geq \alpha$, we **fail to reject the null hypothesis**. The result is "not statistically significant." We do not have sufficient evidence to claim that β_1 is different from zero.

[2] Take a good look at the word "single" and remember what it stands.

$$p = 2\Phi(-|t|) \begin{cases} < \alpha \iff \text{at reject area} \iff \text{reject null hypothesis } H_0 \iff X \text{ is significant} \\ > \alpha \iff \text{at accept area} \iff \text{accept null hypothesis } H_0 \iff X \text{ is not significant} \end{cases}$$

The Critical Value Approach: Alternatively, we can find a critical value, t_c, from a t-distribution table (or software) that corresponds to our chosen significance level α and degrees of freedom ($df = n - 2$).

- If $|t| > t_c$, our calculated statistic falls in the "rejection region." We **reject the null hypothesis**.
- If $|t| \leq t_c$, our statistic falls in the "acceptance region." We **fail to reject the null hypothesis**.

$$|t| \begin{cases} > t_{\frac{\alpha}{2}} \iff \text{at reject area} \iff \text{reject null hypothesis } H_0 \iff X \text{ is significant} \\ < t_{\frac{\alpha}{2}} \iff \text{at accept area} \iff \text{accept null hypothesis } H_0 \iff X \text{ is not significant} \end{cases}$$

3.2 Confidence Interval

While a t-test gives a yes/no answer about a single hypothesized value, a confidence interval provides a more informative range of plausible values for the true population parameter, β_1.

Constructing a Confidence Interval A 95% confidence interval is constructed by taking our point estimate and adding and subtracting a margin of error, which is determined by the critical t-value and the standard error of the estimate.

Formula: (1-α)% Confidence Interval

$$CI = [\hat{\beta_1} - t_c \cdot se(\hat{\beta_1}), \quad \hat{\beta_1} + t_c \cdot se(\hat{\beta_1})] \tag{27}$$

For a 95% confidence interval, $\alpha = 0.05$, and t_c is the critical value leaving $\alpha/2 = 2.5\%$ in each tail of the t-distribution. For large samples, $t_c \approx 1.96$.

A confidence interval can also be used for hypothesis testing. To test the null hypothesis $H_0 : \beta_1 = 0$, we simply check if 0 lies within the interval.

- If the interval **does not** contain 0, we can reject H_0 at the corresponding significance level.
- If the interval **does** contain 0, we fail to reject H_0.

This provides a measure of both statistical significance and the practical range of uncertainty around our estimate. A very wide interval, even if it excludes zero, signals that our estimate is imprecise.

Note 3 (one-tale or two-tale?)**.** *depending on the hypothesis*

4 Multi-variate linear regression

4.1 Multiple Variables

The simple linear regression model is a powerful starting point, but reality is rarely so simple. The outcomes we seek to explain, like sales, wages, or economic growth, are influenced by more than just one factor. The multivariate linear regression (MLR) model is a significant step forward, allowing us to estimate the effect of one variable while simultaneously *controlling for* the effects of others.

MLR The population model with k independent variables is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + u \tag{28}$$

Each coefficient, β_j, now represents the partial effect of X_j on Y, holding all other variables constant. This ”ceteris paribus” interpretation is the core strength of MLR.

4.2 Omitted Variable Bias (OVB)

The primary motivation for moving from simple to multiple regression is to avoid **Omitted Variable Bias**. OVB occurs when we leave out a variable from our model that is both:

1. A determinant of the dependent variable Y (i.e., it belongs in the true model).
2. Correlated with one or more of the included independent variables X.

The Consequence of OVB When both conditions hold, the omitted variable becomes part of the error term, u. Since it is also correlated with an included X, this violates the crucial Zero Conditional Mean assumption ($E(u|X) \neq 0$). The result is that the OLS estimator for the included variable's coefficient becomes biased and inconsistent—it does not converge to the true population value, even with an infinitely large dataset.

The bias in the simple regression coefficient $\hat{\beta}_1$ can be expressed as:

$$E(\hat{\beta}_1) = \beta_1 + \beta_2 \cdot \delta_1 \tag{29}$$

where β_2 is the true effect of the omitted variable on Y, and δ_1 is the slope coefficient from a regression of the omitted variable on the included variable X_1. This formula clearly shows that the bias is zero only if the omitted variable is irrelevant ($\beta_2 = 0$) or if it is uncorrelated with our variable of interest ($\delta_1 = 0$).

$$\hat{\beta} \xrightarrow{P} \beta + \underbrace{\frac{\sigma_u}{\sigma_X}\rho_{uX}}_{\text{effect of OVB}} \tag{30}$$

> **Example of OVB:** Imagine a simple regression of daily ice cream sales (Y) on the number of beach visitors (X_1). We would likely find a strong positive relationship. However, we have omitted temperature (X_2). Temperature affects sales ($\beta_2 > 0$) and is also highly correlated with the number of beach visitors ($\delta_1 > 0$). Therefore, our estimate for the effect of beach visitors will be biased upwards, incorrectly attributing the effect of the warm weather to the mere presence of people on the beach.

Overcoming OVB The most direct solution to OVB is to include the omitted variable in the regression, turning it into an MLR model. By adding control variables, we can isolate the effect of our primary variable of interest. Other advanced methods for tackling OVB when data on the omitted variable is unavailable include Instrumental Variables (IV) and Panel Data Fixed Effect models, which are discussed in later sections.

4.3 Fitting

Information criterion how many variables should i add into the model?

AIC

BIC

OLS in Matrix Form With multiple regressors, it is convenient to express the OLS problem using matrix algebra. Let Y be an $n \times 1$ vector of outcomes, X be an $n \times (k+1)$ matrix of regressors (including a column of ones for the intercept), and β be a $(k+1) \times 1$ vector of parameters. The minimization problem $\min(Y - X\beta)'(Y - X\beta)$ yields the compact solution:

$$\hat{\beta} = (X'X)^{-1}X'Y \tag{31}$$

The OLS is to

$$\min_{\{\beta_0,\ldots,\beta_k\}} \sum_{i}^{n} (Y_i - \hat{Y}_i)^2 \tag{32}$$

This could end in the result of

$$\hat{\beta} = (X^T X)^{-1} X^T Y \tag{33}$$

4.4 Measures of Fit in MLR

Adjusted R-Squared In a multiple regression setting, the standard R^2 has a critical flaw: it mechanically increases every time we add a new variable to the model, regardless of whether that variable is truly relevant. This makes it a poor tool for comparing models with different numbers of variables.

The Adjusted R-Squared ($\bar{R}^2$) corrects this problem by penalizing the inclusion of irrelevant variables. It adjusts both the SSR and TSS by their respective degrees of freedom.

Definition: Adjusted R-Squared ($\bar{R}^2$)

$$\bar{R}^2 = 1 - \frac{SSR/(n-k-1)}{TSS/(n-1)} = 1 - \frac{n-1}{n-k-1} \cdot \frac{SSR}{TSS} \tag{34}$$

where n is the sample size and k is the number of independent variables.

Key properties of $\bar{R}^2$:

- $\bar{R}^2$ is always less than R^2.

- Adding a new variable will only increase $\bar{R}^2$ if that variable's contribution to explaining Y is large enough to offset the penalty for losing a degree of freedom.
- $\bar{R}^2$ can be negative, which is a strong signal of a very poorly fitting model.

For model selection, a higher $\bar{R}^2$ is generally preferred. Other common model selection criteria that impose different penalties for complexity include the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC).

4.5 Inference in MLR

Hypothesis testing in MLR extends the concepts from simple regression, but with the added ability to test more complex hypotheses involving multiple coefficients.

Joint Hypothesis Testing: The F-test Often, we want to test a hypothesis that involves multiple coefficients simultaneously. For example, we might want to test if a group of variables *as a whole* has no effect on Y. This is called a joint hypothesis. The null and alternative hypotheses take the form:

- $H_0 : \beta_1 = 0, \beta_2 = 0, \ldots, \beta_q = 0$ (q restrictions)
- H_1 : At least one of the β_j in H_0 is non-zero.

We cannot test this by simply looking at individual t-tests, as they don't account for the covariance between the coefficient estimates. The correct tool is the F-test. The F-statistic compares the fit of the "unrestricted" model (with all variables) to the "restricted" model (where the null hypothesis is forced to be true by excluding the variables).

Definition: The F-statistic

$$F = \frac{(SSR_{\text{restricted}} - SSR_{\text{unrestricted}})/q}{SSR_{\text{unrestricted}}/(n - k_{\text{unrestricted}} - 1)} = \frac{(R^2_{\text{unrestricted}} - R^2_{\text{restricted}})/q}{(1 - R^2_{\text{unrestricted}})/(n - k_{\text{unrestricted}} - 1)} \tag{35}$$

where q is the number of restrictions (coefficients set to zero) in the null hypothesis. A large F-statistic provides evidence against the null, suggesting the tested variables are jointly significant.

5 Functional Form and Diagnostic Issues in Regression

The linear regression model is more flexible than its name suggests. The "linear" part refers to the model being linear in its parameters (β_j), not necessarily in its variables. By transforming the variables (Y and X), we can model a wide variety of non-linear relationships. This section also addresses common practical issues that can invalidate our standard OLS inference.

5.1 Modeling Non-linear Relationships

Real-world economic relationships are often non-linear. The effect of an additional year of experience on wages, for instance, is likely much larger for a new graduate than for a late-career professional.

Logarithmic Transformations The most common method for modeling non-linearities is the natural logarithm ('log' or 'ln'). It allows us to interpret coefficients in terms of percentage changes, which is often more intuitive than unit changes.

- **Log-Lin Model**: $\log(Y) = \beta_0 + \beta_1 X + u$.
  - *Interpretation*: A one-unit increase in X is associated with a $(100 \cdot \beta_1)\%$ change in Y. This is a standard model for wage equations, where Y is wages and X is years of education.
- **Lin-Log Model**: $Y = \beta_0 + \beta_1 \log(X) + u$.
  - *Interpretation*: A one-percent increase in X is associated with a $(\beta_1/100)$-unit change in Y. For example, modeling the effect of a percent change in advertising spend (X) on the number of units sold (Y).
- **Log-Log Model**: $\log(Y) = \beta_0 + \beta_1 \log(X) + u$.
  - *Interpretation*: A one-percent increase in X is associated with a $\beta_1\%$ change in Y. In this specification, β_1 is an **elasticity**. This is the standard functional form for estimating demand curves.

Polynomial Regression To model relationships that may increase and then decrease (or vice-versa), we can include polynomial terms. A quadratic model is the most common:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + u \tag{36}$$

The partial effect of X on Y is now $\frac{\partial Y}{\partial X} = \beta_1 + 2\beta_2 X$, which depends on the level of X. If $\beta_1 > 0$ and $\beta_2 < 0$, the relationship is an inverted U-shape. This is often used to model the effect of age or experience on income.

5.2 Using Qualitative Information: Dummy Variables

Often, important explanatory factors are categorical (e.g., gender, region, industry). We can incorporate this information using dummy variables(or indicator variables).

Single Dummy Variable A dummy variable, D, is a binary variable that takes the value 1 if a certain condition is met and 0 otherwise. Consider the model:

$$Y_i = \beta_0 + \delta_0 D_i + u_i \tag{37}$$

Here, the average value of Y for the group where $D = 0$ is $E(Y_i|D_i = 0) = \beta_0$. For the group where $D = 1$, the average value is $E(Y_i|D_i = 1) = \beta_0 + \delta_0$. Thus, δ_0 represents the difference in the average value of Y between the two groups.

The Dummy Variable Trap When a categorical variable has k mutually exclusive categories (e.g., a company's industry: Manufacturing, Retail, Tech), we must include only $k-1$ dummy variables in the regression. If we include a dummy for every single category, the sum of these dummies will be a constant (equal to 1 for every observation), which is perfectly collinear with the intercept term. This is the **dummy variable trap**, and it makes OLS estimation impossible due to perfect multicollinearity.

The category for which we omit the dummy becomes the base category, and the coefficients on the included dummies are interpreted as the difference relative to this base.

Interaction Terms The effect of one variable may depend on the value of another. We can model such dependencies using interaction terms.

- **Dummy-Continuous Interaction**: Does the effect of advertising (X_{cont}) depend on whether it's a holiday ($D_{holiday}$)?

$$Y = \beta_0 + \beta_1 X_{cont} + \delta_1 D_{holiday} + \delta_2 (X_{cont} \cdot D_{holiday}) + u \tag{38}$$

  The effect of an extra dollar of advertising is β_1 on a non-holiday, but it is $\beta_1 + \delta_2$ on a holiday. A t-test on δ_2 can determine if this difference is statistically significant.
- **Continuous-Continuous Interaction**: To see if the effect of X_1 depends on the level of X_2, we can include their product: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 \cdot X_2) + u$. The effect of X_1 on Y is now $\beta_1 + \beta_3 X_2$.

5.3 Violations of the OLS Assumptions

Remember a OLS can be and should be used in certain conditions? In reality they are not so ideal and perfect.

Multicollinearity Multicollinearity occurs when independent variables are highly correlated with each other.

- **Perfect Multicollinearity**: One regressor is a perfect linear combination of another (e.g., falling into the dummy variable trap). The model cannot be estimated.
- **Imperfect Multicollinearity**: Regressors are strongly but not perfectly correlated. OLS estimates remain unbiased, but their standard errors become inflated. This makes it hard to distinguish the individual impacts of the collinear variables, leading to statistically insignificant t-statistics even when the variables are jointly significant.

Heteroskedasticity Heteroskedasticity is a violation of the Gauss-Markov assumption of constant variance.

> **Definition: Heteroskedasticity** The variance of the error term, conditional on the independent variables, is not constant.
>
> $$Var(u|X_1, X_2, \ldots, X_k) \neq \sigma^2 \tag{39}$$
>
> For example, the variance in food consumption may be much larger for high-income households than for low-income households.

- **Consequences**: The OLS estimators ($\hat{\beta}_j$) are still unbiased and consistent. However, the formulas for their standard errors are incorrect. Consequently, t-statistics, F-statistics, and confidence intervals are unreliable. You might conclude an effect is significant when it is not, or vice-versa.
- **Solution**: The modern, standard solution is to use Heteroskedasticity-Robust Standard Errors (often called White, Huber-White, or simply robust standard errors). These standard errors are calculated in a way that is valid even in the presence of heteroskedasticity of an unknown form. Most statistical software packages can compute them easily.

Note 4 (Modern Practice). *In applied econometrics, it is now standard practice to report robust standard errors by default, as heteroskedasticity is a common feature of economic data, particularly in cross-sectional studies. Assuming homoskedasticity without evidence is often considered unrealistic.*

Part II

EXTENSIONS OF THE REGRESSION FRAMEWORK

6 Models for Limited Dependent Variables

In many economic scenarios, the outcome we want to explain is not a continuous variable. It might be a binary choice (to buy a product or not), an ordinal ranking (a credit rating), or a count (the number of patents filed). When the dependent variable is ”limited” in this way, OLS is no longer the appropriate tool. This section introduces models designed specifically for the most common case: a binary dependent variable.

6.1 The Linear Probability Model (LPM)

The most direct approach to modeling a binary outcome ($Y \in \{0, 1\}$) is to simply use OLS. This is called the Linear Probability Model (LPM).

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + u \tag{40}$$

Because the expected value of a binary variable is the probability that it equals 1, we have $E[Y|X] = Pr(Y = 1|X)$. The LPM therefore models the probability directly as a linear function of the regressors.

Interpretation and Flaws The coefficient β_j in an LPM is interpreted as the change in the probability of success ($Y = 1$) for a one-unit change in X_j. While simple to estimate and interpret, the LPM has several serious drawbacks:

1. **Predicted Probabilities Out of Bounds:** The linear functional form can produce predicted probabilities that are less than 0 or greater than 1, which is nonsensical.
2. **Constant Marginal Effects:** The model assumes that the effect of X_j on the probability is constant, which is often unrealistic. The impact of an extra $1000 in income on the probability of buying a car is likely much larger for a low-income individual than for a millionaire.
3. **Inherent Heteroskedasticity:** The variance of the error term in an LPM depends on the values of X, violating the homoskedasticity assumption by construction. While this can be fixed with robust standard errors, the other issues remain.

The LPM is useful as a simple benchmark, but its flaws motivate the need for more sophisticated models.

6.2 Probit and Logit Models

Instead of modeling the probability as a linear function, Probit and Logit models use a non-linear S-shaped curve that is bounded between 0 and 1. This is achieved by linking the probability to the Cumulative Distribution Function (CDF) of a probability distribution.

$$Pr(Y = 1|X) = G(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k) \tag{41}$$

where $G(\cdot)$ is a chosen CDF. The term inside, $Z = \beta_0 + \beta_1 X_1 + \ldots$, is often called the index.

The Probit Model The Probit model assumes that $G(\cdot)$ is the CDF of the standard normal distribution, denoted by $\Phi(\cdot)$.

$$Pr(Y = 1|X) = \Phi(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k) \tag{42}$$

As the index Z goes from $-\infty$ to $+\infty$, the resulting probability smoothly moves from 0 to 1.

The Logit Model The Logit model uses the CDF of the standard logistic distribution, denoted by $\Lambda(\cdot)$.

$$Pr(Y = 1|X) = \Lambda(\beta_0 + \cdots + \beta_k X_k) = \frac{e^{\beta_0 + \cdots + \beta_k X_k}}{1 + e^{\beta_0 + \cdots + \beta_k X_k}} \tag{43}$$

In practice, the Probit and Logit models yield very similar results. The logistic distribution has slightly fatter tails than the normal distribution, but the choice between them is often a matter of convention in a particular field.

6.3 Estimation and Interpretation

Maximum Likelihood Estimation (MLE) Because these models are non-linear, they cannot be estimated by OLS. Instead, they are estimated using the principle of **Maximum Likelihood Estimation (MLE)**. The intuition of MLE is to find the parameter values ($\hat{\beta}$) that make the observed data (our sample of Ys and Xs) most likely to have occurred. It is a general and powerful estimation method used throughout econometrics.

Interpreting Coefficients: Marginal Effects A critical point to understand is that the coefficients (β_j) in a Probit or Logit model are **not** the marginal effects. They show the effect of a one-unit change in X_j on the latent index Z, not on the probability itself.

To find the effect on the probability, we must compute the derivative of $Pr(Y = 1|X)$ with respect to X_j, which is:

$$\frac{\partial Pr(Y = 1|X)}{\partial X_j} = g(\beta_0 + \beta_1 X_1 + \ldots) \cdot \beta_j \tag{44}$$

where $g(\cdot)$ is the Probability Density Function (PDF) corresponding to the CDF $G(\cdot)$. This shows that the marginal effect is not constant; it depends on the values of all the X variables.

In practice, we don't report β_j. Instead, we report one of two summary measures, which are easily calculated by statistical software:

- **Marginal Effect at the Mean (MEM):** The marginal effect calculated at the mean values of all X variables.
- **Average Marginal Effect (AME):** The marginal effect is calculated for each individual observation in the sample, and then the average of these effects is taken. The AME is generally preferred as it is more representative of the entire sample.

Note 5 (Random Utility and the Logit Model)**.**

A powerful theoretical justification for the Logit model comes from discrete choice theory, specifically Daniel McFadden's work on Random Utility Models. Imagine an individual must choose between two options, 0 and 1. The utility they get from each choice is:

$$U_{i0} = V_{i0} + \epsilon_{i0}$$
$$U_{i1} = V_{i1} + \epsilon_{i1}$$

Here, V_{ij} is the deterministic part of the utility (which can be modeled as a function of observed characteristics, e.g., $V_{i1} = X_i\beta$), and ϵ_{ij} is a random, unobserved component.

The individual will choose option 1 if $U_{i1} > U_{i0}$. McFadden showed that if the random error terms, ϵ_{i0} and ϵ_{i1}, are assumed to be independently and identically drawn from a Type-I extreme value distribution, then the probability of choosing option 1 takes the exact form of the Logit model:

$$Pr(Y_i = 1|X_i) = Pr(V_{i1} - V_{i0} > \epsilon_{i0} - \epsilon_{i1}) = \frac{e^{V_{i1} - V_{i0}}}{1 + e^{V_{i1} - V_{i0}}} \tag{45}$$

This framework generalizes to choices among multiple alternatives (the multinomial logit), where the probability of choosing option j out of J total options becomes the **softmax function**:

$$Pr(Y_i = j) = \frac{e^{V_{ij}}}{\sum_{k=1}^{J} e^{V_{ik}}} \tag{46}$$

Part III

THE CAUSAL INFERENCE TOOLKIT

Traditional models are excellent at finding correlation—showing how variables move together—but they struggle to prove that one variable causes a change in another.

Some assumption are really unrealistic. This means endogeneity being inevitable. That's why we have to forfeit some of the rules and be practical.

Stuff in this part, they are explicitly designed to identify and isolate a causal effect. Or in simple terms, they are designed to overcome endogeneity problems.

7 The Challenge of Causal Inference

This section is the philosophical heart of modern applied econometrics.

The previous sections focused on fitting models and describing relationships within a dataset. We now shift our focus to the central ambition of modern econometrics: estimating the causal effect of a variable or intervention. While a traditional regression can show us that advertising and sales are correlated, a causal approach seeks to answer the question: "By how much *would sales increase* if we were to increase our advertising budget by $100?" Answering this "what if" question requires moving beyond mere association to a more structured way of thinking about cause and effect.

IF ALL CRITERIONS ARE MET IN THE OLS, they are purely causal.

7.1 The Potential Outcomes Framework

The most influential framework for thinking about causality is the Rubin Causal Model (RCM), also known as the Potential Outcomes framework. It provides a precise language for defining a causal effect, even if that effect is fundamentally unobservable.

Let's consider a binary treatment, D_i. For each individual i, there are two potential outcomes:

- $Y_i(1)$: The potential outcome for individual i *if they receive the treatment* ($D_i = 1$).
- $Y_i(0)$: The potential outcome for individual i *if they do not receive the treatment* ($D_i = 0$).

The individual-level causal effect is the difference between these two potential outcomes:

$$\delta_i = Y_i(1) - Y_i(0) \tag{47}$$

The Fundamental Problem of Causal Inference The core dilemma is that for any given individual i, we can only ever observe one of their potential outcomes. The one we do not observe is called the counterfactual. We observe the realized outcome, Y_i, which is:

$$Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0) \tag{48}$$

Since we can never observe δ_i for any individual, our goal must be to estimate the **average** causal effect across a population.

Average Causal Effects There are several population-level parameters of interest:

- **Average Treatment Effect (ATE)**: The average effect for an individual chosen at random from the population.
$$\tau_{ATE} = E[\delta_i] = E[Y_i(1) - Y_i(0)] \tag{49}$$
- **Average Treatment Effect on the Treated (ATT)**: The average effect specifically for those who actually received the treatment.
$$\tau_{ATT} = E[\delta_i|D_i = 1] = E[Y_i(1)|D_i = 1] - E[Y_i(0)|D_i = 1] \tag{50}$$

7.2 Selection Bias: Why Simple Comparisons Fail

A naive approach to estimating the causal effect is to simply compare the average outcomes of the treated group and the untreated group. This simple difference in means is:

$$\Delta = E[Y_i|D_i = 1] - E[Y_i|D_i = 0] \tag{51}$$

Using the potential outcomes framework, we can decompose this simple difference:

$$\begin{aligned}\Delta &= E[Y_i(1)|D_i = 1] - E[Y_i(0)|D_i = 0] && (52)\\ &= E[Y_i(1)|D_i = 1] - E[Y_i(0)|D_i = 1] + E[Y_i(0)|D_i = 1] - E[Y_i(0)|D_i = 0] && (53)\\ &= \underbrace{\tau_{ATT}}_{\text{True Causal Effect}} + \underbrace{E[Y_i(0)|D_i = 1] - E[Y_i(0)|D_i = 0]}_{\text{Selection Bias}} && (54)\end{aligned}$$

The simple difference in means is the sum of the true causal effect we want and a selection bias term. This bias term represents the difference in the *no-treatment potential outcome* between those who chose treatment and those who did not.

> **Example of Selection Bias:** Consider estimating the effect of a job training program on wages. If more motivated individuals are more likely to sign up for the program, they would likely have earned higher wages than the non-participants *even if they had not taken the training*. This means $E[Y_i(0)|D_i = 1] > E[Y_i(0)|D_i = 0]$, leading to a positive selection bias. A simple comparison would overstate the program's true effect.

The central challenge of causal econometrics is finding a research design that eliminates or accounts for selection bias. A randomized controlled trial (RCT) does this by design, as random assignment ensures that, on average, the treatment and control groups are identical before treatment ($E[Y_i(0)|D_i = 1] = E[Y_i(0)|D_i = 0]$). The methods we will discuss in the following sections (Panel Data, IV, DID, RDD) are all strategies for approximating an RCT and eliminating selection bias when we only have observational, non-experimental data.

7.3 Causal Identification Strategies

An ”identification strategy” is the set of assumptions we rely on to argue that our chosen method has isolated a true causal effect, free from selection bias. It is the story that connects our statistical model to the causal parameter we wish to estimate.

From Potential Outcomes to Regression The problem of selection bias can be reframed in the regression context. In the model $Y = \beta_0 + \beta_1 D + u$, the coefficient β_1 only has a causal interpretation if the treatment assignment D is uncorrelated with the error term u. However, if individuals who are more likely to have a high Y for other reasons (e.g., motivation, ability, which are part of u) are also more likely to select into treatment (have $D = 1$), then $Cov(D, u) \neq 0$. This is the problem of endogeneity, and selection bias is a primary cause. The goal of causal identification strategies is to find a way to make $Cov(D, u) = 0$.

Note 6 (Directed Acyclic Graphs, DAGs)**.**

Another increasingly popular framework for reasoning about causality is the Directed Acyclic Graph (DAG). DAGs provide a visual way to represent the assumed causal relationships between variables.

- **Nodes and Arrows:** Each variable is a node, and a causal relationship is represented by a directed arrow ($A \to B$ means A causes B).
- **The Goal:** To estimate the causal effect of a treatment T on an outcome Y, we need to block all ”backdoor paths” between T and Y. A backdoor path is a non-causal path of association that can create bias.
- **Controlling for Confounders:** A common backdoor path is created by a ”confounder” C, a variable that causes both T and Y ($T \leftarrow C \to Y$). This path creates a spurious correlation between T and Y. By controlling for C in a regression, we ”block” this backdoor path.

DAGs are a powerful tool for making our causal assumptions explicit. An OVB arises when we fail to control for a confounder that opens a backdoor path. The potential outcomes framework and DAGs are two complementary languages for describing the same fundamental challenges of causal inference.

8 Matching Methods: Propensity Score Matching (PSM)

After formalizing the problem of selection bias in the potential outcomes framework, our first line of attack is to ask: can we eliminate this bias by simply ”controlling for” all relevant observable differences between the treated and untreated groups? Matching methods, particularly Propensity Score Matching (PSM), are designed to do exactly this.

8.1 The ”Selection on Observables” Assumption

Matching methods are built upon a key identifying assumption known as selection on observables, or conditional independence.

The Conditional Independence Assumption (CIA) Conditional on a set of observable pre-treatment characteristics, X, the assignment to treatment is independent of the potential outcomes. Formally:

$$(Y_i(1), Y_i(0)) \perp D_i | X_i \tag{55}$$

The Intuition The CIA states that once we have controlled for a rich enough set of observable variables (X), there are no remaining unobserved differences (like motivation, ability, or risk tolerance) that are correlated with both treatment status and the outcome. In essence, it assumes that the selection bias we identified earlier is entirely due to observable factors. If CIA holds, then within any group of individuals with the same characteristics X, the assignment to treatment is ”as good as random.”

This is a very strong assumption, and it is the primary weakness of matching methods. Its credibility hinges entirely on the researcher’s ability to argue that their dataset contains *all* the key variables that jointly determine treatment and outcomes.

8.2 From High Dimensions to a Single Score

In principle, if CIA holds, we could estimate the treatment effect by matching each treated individual to an untreated individual with the exact same values for all variables in X. However, if X contains many variables, especially continuous ones, finding exact matches becomes impossible. This is known as the **curse of dimensionality**.

The **propensity score**, introduced by Rosenbaum and Rubin, is an ingenious solution to this problem.

> **Definition: The Propensity Score** The propensity score, $p(X_i)$, is the conditional probability of an individual receiving the treatment, given their vector of pre-treatment characteristics X_i.
>
> $$p(X_i) = Pr(D_i = 1|X_i) \tag{56}$$

In practice, the true propensity score is unknown and must be estimated, typically using a Logit or Probit model where the dependent variable is the treatment dummy D_i and the independent variables are the characteristics in X_i.

Rosenbaum and Rubin proved that if CIA holds conditional on X, then it also holds conditional on the one-dimensional propensity score $p(X)$. This means that instead of having to match on a potentially huge vector of variables, we only need to match individuals based on their estimated probability of being treated.

8.3 Implementing PSM: A Practical Guide

The process of estimating a treatment effect using PSM typically involves four steps.

Step 1: Estimate the Propensity Score Run a Logit or Probit regression of the treatment indicator D_i on the set of pre-treatment covariates X_i to get the estimated propensity score, $\hat{p}(X_i)$, for every observation in both the treatment and control groups.

Step 2: Check for Common Support A crucial diagnostic step is to check for **common support** (or overlap). The CIA is only useful in regions where we have both treated and untreated individuals. If, for example, the control group only has individuals with propensity scores between 0.1 and 0.5, while the treatment group has scores from 0.6 to 0.9, the groups are too different to be compared, and we have no common support. In practice, observations that fall outside the common support region are often discarded.

Step 3: Choose a Matching Algorithm Once scores are estimated, each treated unit must be matched to one or more control units. Common algorithms include:

- **Nearest Neighbor Matching:** Each treated unit is matched to the control unit with the closest propensity score. ”Matching with replacement” is often preferred, as it allows a good control unit to be used as a match multiple times, which can improve match quality at the cost of some precision.
- **Caliper Matching:** A variant of nearest neighbor that imposes a tolerance level: a match is only made if the propensity score difference is smaller than a pre-specified amount (the caliper).
- **Kernel Matching:** Each treated unit is matched to a weighted average of all control units, where the weights are inversely proportional to the distance in propensity scores. This uses more data but can be sensitive to the choice of kernel function.

Step 4: Estimate the Treatment Effect and Check Balance After matching, the Average Treatment Effect on the Treated (ATT) is estimated by taking the mean difference in the outcome variable between the treated units and their matched control units.

Finally, one must perform a **balance check**. The whole point of matching was to make the treatment and control groups comparable on their observable characteristics X. We must verify that this was successful by comparing the means of the X variables between the two groups *after* matching. If significant differences remain, the matching procedure has failed.

> **Note 7** (PSM vs. Regression)**.**
>
> You might wonder how PSM differs from just running an OLS regression of Y on D and X. OLS imposes a linear functional form on the relationship between X and Y. PSM is non-parametric in this regard; it does not make assumptions about how X affects Y. By focusing on the *treatment assignment process* (the propensity score), it can be more robust if the OLS model is misspecified. However, both methods are equally reliant on the strong, untestable Conditional Independence Assumption.

9 Panel Data Methods

Panel data (or longitudinal data), which tracks the same entities (e.g., individuals, firms, countries) over time, is one of the most powerful tools for causal inference. The ability to observe the same unit under different circumstances allows us to control for unobservable factors that are constant over time, thereby mitigating a critical source of omitted variable bias.

9.1 The Power of Fixed Effects

Consider a simple regression to estimate the effect of X on Y for an entity i at time t:

$$Y_{it} = \beta_0 + \beta_1 X_{it} + u_{it} \tag{57}$$

The key challenge is that the error term u_{it} contains unobserved factors. Some of these factors may be constant for a given individual over time (e.g., innate ability, a firm's foundational culture, a country's geography), while others may vary over time. We can decompose the error term:

$$u_{it} = \alpha_i + \nu_{it} \tag{58}$$

where α_i is the unobserved, time-invariant individual-specific effect, and ν_{it} is the idiosyncratic error that changes over time.

The endogeneity problem arises if the variable of interest, X_{it}, is correlated with the fixed effect α_i. For example, when estimating the effect of education (X_{it}) on wages (Y_{it}), an individual's innate ability (α_i) is likely correlated with both their educational attainment and their wage potential. This leads to biased OLS estimates.

The Fixed Effects (FE) Estimator The Fixed Effects (or "within") estimator provides a brilliant solution to this problem. It eliminates the time-invariant effect α_i by using only the variation *within* each individual over time. This is typically done through one of two equivalent procedures:

1. **De-meaning:** For each individual i, calculate the average of Y, X, and u over time ($\bar{Y}_i, \bar{X}_i, \bar{u}_i$). Subtracting these averages from the original equation yields:

   $$(Y_{it} - \bar{Y}_i) = \beta_1(X_{it} - \bar{X}_i) + (\nu_{it} - \bar{\nu}_i) \tag{59}$$

   The individual fixed effect α_i is perfectly constant over time, so it drops out of the equation ($\alpha_i - \bar{\alpha}_i = \alpha_i - \alpha_i = 0$). We can then run OLS on this "de-meaned" data to get an unbiased estimate of β_1.
2. **Least Squares Dummy Variable (LSDV):** An equivalent approach is to include a dummy variable for every single individual i in the regression (except one, to avoid the dummy variable trap). This is computationally intensive for large datasets but yields the exact same estimate for β_1.

Real-World Application and Interpretation The FE estimator is a workhorse in applied economics for estimating the effect of policies or changing circumstances. A key feature is that it cannot estimate the effect of any time-invariant variable (e.g., gender, race, a firm's industry). Because these variables do not change within an individual, their effects are absorbed into the fixed effect and cannot be separately identified. The interpretation of β_1 is therefore: "When individual i changes its value of X by one unit, its value of Y is expected to change by β_1 units, holding all other time-varying factors constant and controlling for all stable, unobserved characteristics of the individual."

9.2 The Two-Way Fixed Effects (TWFE) Model

The standard FE model controls for unobserved factors that are constant within an individual. However, there may also be unobserved factors that are constant across all individuals at a specific point in time, such as a nationwide macroeconomic shock, a new regulation, or a change in public sentiment. These are called time fixed effects.

The Two-Way Fixed Effects (TWFE) Model To control for both types of unobserved confounders simultaneously, we augment the FE model with time dummies. This is the standard model used in a vast amount of modern empirical research.

$$Y_{it} = \beta_1 X_{it} + \alpha_i + \gamma_t + \nu_{it} \tag{60}$$

- β_1: The coefficient of interest.
- α_i: A full set of individual fixed effects (entity dummies). Controls for all time-invariant confounders.
- γ_t: A full set of time fixed effects (time period dummies). Controls for all common shocks that affect all entities in a given period.

Practical Implementation In practice, nobody calculates de-meaned data by hand or includes thousands of individual dummies manually. All major statistical software packages have specialized commands that estimate FE and TWFE models efficiently. The standard practice is to specify the model and declare which variables represent the ”individual” and ”time” dimensions. **A crucial and often overlooked best practice is to cluster standard errors at the individual level.** This accounts for the fact that the errors for a given individual are likely to be correlated across time (serial correlation).

Example: Minimum Wage and Employment A classic application of TWFE is estimating the effect of minimum wage increases on employment. Researchers use a panel of all 50 U.S. states over many years.

- Y_{it}: Employment in the restaurant industry in state i in year t.
- X_{it}: The minimum wage in state i in year t.
- α_i (State Fixed Effects): Control for time-invariant differences between states (e.g., industrial structure, political culture, geography).
- γ_t (Year Fixed Effects): Control for national trends affecting all states in a given year (e.g., recessions, changes in federal policy, technological progress).

By using TWFE, the estimate for β_1 is isolated from both stable differences across states and common national trends, bringing us closer to a causal estimate.

9.3 Random Effects and the Hausman Test

An alternative to the FE model is the Random Effects (RE) model. The RE model assumes that the unobserved individual effects α_i are random and, crucially, uncorrelated with the explanatory variables.

A Conceptual Distinction: Fixed vs. Random Effects

Beyond the statistical mechanics of the Hausman test, there is a conceptual difference in the research question being asked by FE and RE models.

> The Fixed Effects approach is fundamentally agnostic about the distribution of the individual effects, α_i. It treats them as fixed, unknown constants that we need to control for and eliminate in order to get a clean estimate of β_1. The inference is conditional on the specific individuals in our sample. We are primarily interested in the causal effect *within* this sample, making it the preferred model for causal inference where selection bias is a concern. It answers the question: ”For the firms in our dataset, what was the effect of this policy?”
>
> The Random Effects approach, by contrast, assumes that the individuals in our sample were randomly drawn from a larger population and treats the individual effects, α_i, as random variables that are part of the composite error term. The goal is to make an inference about the entire population from which the sample was drawn. This makes it more suitable for predictive modeling or for situations where the entities are genuinely interchangeable and random sampling from a large population can be credibly assumed (a scenario less common in economics than in, for example, biology or psychology). By treating the effects as random, it ”uses” both the within-individual and between-individual variation, making it more efficient—but this comes at the cost of the strong and often untenable assumption of zero correlation.

$$Cov(X_{it}, \alpha_i) = 0 \tag{61}$$

If this assumption holds, the RE model is more efficient (has lower variance) than FE. However, this assumption is often considered unrealistic in economics—it's the econometric equivalent of saying there is no selection bias.

Choosing Between FE and RE: The Hausman Test How do we decide which model to use? The Hausman test provides a formal procedure. It compares the coefficient estimates from the FE model and the RE model.

- **Null Hypothesis (**H_0**)**: The RE model is appropriate ($Cov(X_{it}, \alpha_i) = 0$). Both FE and RE estimators are consistent, but RE is more efficient.
- **Alternative Hypothesis (**H_1**)**: The RE model is inappropriate ($Cov(X_{it}, \alpha_i) \neq 0$). Only the FE estimator is consistent.

If the p-value of the Hausman test is small (e.g., < 0.05), we reject the null hypothesis and conclude that the unobserved effects are correlated with our regressors. In this case, the RE estimates are biased, and we must use the FE model. In most applied microeconomic settings, the Hausman test rejects the null, and **Fixed Effects is the preferred, more conservative, and more credible approach.**

10 Instrumental Variables and Two-Stage Least Squares

We often face situations where our primary explanatory variable, X, is endogenous—correlated with the error term u—due to omitted variables, measurement error, or simultaneous causality. In these cases, OLS is biased and inconsistent, and a fixed effects strategy may not be possible (e.g., in a cross-sectional study). The Instrumental Variables (IV) method is a powerful, general-purpose solution to this problem. It is a technique that, under the right conditions, allows us to isolate a sliver of ”clean” variation in our endogenous variable and use it to estimate a causal effect.

10.1 The Intuition of Instrumental Variables

Imagine the variation in our problematic endogenous variable, X, is composed of two parts: a "good" part that is exogenous and a "bad" part that is correlated with the error term u. OLS uses all the variation and is therefore biased. The goal of IV is to find a special third variable, Z, called an **instrument**, that can isolate only the "good" variation in X.

To be a valid instrument, Z must satisfy two core conditions:

1. **Instrument Relevance:** The instrument must be correlated with the endogenous explanatory variable.

$$Cov(Z, X) \neq 0 \tag{62}$$

The instrument needs to have some power to explain or predict the variable we are trying to fix. If it's unrelated to X, it's useless.

2. **Instrument Exogeneity (The Exclusion Restriction):** The instrument must be uncorrelated with the error term u. It can only affect the outcome variable Y through its effect on X.

$$Cov(Z, u) = 0 \tag{63}$$

This is the crucial, untestable assumption that gives IV its causal power. It demands that the instrument itself is not plagued by the same endogeneity problems as the original variable. It must not have any direct effect on Y, nor be correlated with any of the unobserved factors that affect Y.

Classic Example: The Effect of Education on Earnings A classic IV application estimates the return to education. Education (X) is likely endogenous because of unobserved "ability" (u), which affects both educational attainment and earnings.

- **Instrument (Z):** Proximity to a college or quarter of birth. Angrist and Krueger (1991) famously used an individual's quarter of birth as an instrument. Due to compulsory schooling laws, those born earlier in the year could legally drop out of school with slightly less education than those born later.
- **Relevance:** Quarter of birth is (weakly) correlated with years of schooling.
- **Exogeneity:** Quarter of birth is plausibly random and should not be correlated with an individual's innate ability or have a direct effect on their future earnings, other than through its effect on their schooling.

By using this instrument, they could isolate the variation in schooling that was driven only by the arbitrary schooling laws, not by ability, thus estimating a causal effect.

10.2 Estimation: Two-Stage Least Squares (2SLS)

The most common method for implementing IV is Two-Stage Least Squares (2SLS). It's an intuitive, two-step procedure that mechanically isolates the "good" variation.

Consider the system:

$$\begin{aligned} Y &= \beta_0 + \beta_1 X + u \quad \text{(Structural Equation)} \\ X &= \pi_0 + \pi_1 Z + \nu \quad \text{(First-Stage Equation)} \end{aligned}$$

The Two Stages

- **First Stage:** We regress the endogenous variable X on the instrument Z (and any other exogenous control variables).

  $$\text{Regress } X \text{ on } Z \Rightarrow \text{ Obtain predicted values } \hat{X} \tag{64}$$

  The predicted values, $\hat{X} = \hat{\pi}_0 + \hat{\pi}_1 Z$, represent the part of the variation in X that is explained by our instrument. Because our instrument Z is assumed to be exogenous, this predicted variation, $\hat{X}$, is now also exogenous—it has been "cleaned" of its correlation with u.
- **Second Stage:** We regress our original outcome variable Y on the *predicted values* $\hat{X}$ from the first stage.

  $$\text{Regress } Y \text{ on } \hat{X} \Rightarrow \text{ Obtain the IV estimate } \hat{\beta}_1^{IV} \tag{65}$$

  The resulting coefficient, $\hat{\beta}_1^{IV}$, is our consistent estimate of the causal effect. It's crucial to note that while this two-step procedure provides the correct coefficient, the standard errors calculated manually from the second stage are incorrect. Statistical software must be used to perform 2SLS, as it makes the necessary adjustments to compute the correct standard errors.

10.3 IV in Practice: Finding Instruments in the Real World

The credibility of an IV analysis rests entirely on finding a plausible instrument. This search is often a creative process that combines economic theory, institutional knowledge, and an understanding of how natural or historical processes can generate "as-if random" variation. Below are a few famous examples that illustrate the diversity of instrumental variables.

Quasi-Random Natural Events: Rainfall

- **Research Question:** What is the effect of a country's income on its likelihood of experiencing civil war?

- **Endogeneity Problem:** Income and conflict are simultaneously determined. Poor economic conditions might cause conflict, but conflict also destroys the economy. OLS cannot disentangle this.

- **Instrumental Variable (Z):** Variation in annual rainfall. In countries heavily reliant on agriculture, rainfall is a strong predictor of agricultural output and thus GDP growth (Instrument Relevance). However, rainfall itself is unlikely to have a direct effect on civil war, other than through its effect on the economy (Instrument Exogeneity). Miguel, Satyanath, and Sergenti (2004) used this strategy to find a strong causal link from negative economic shocks to increased conflict in Africa.

Institutional Rules and Administrative Boundaries

- **Research Question:** What is the causal effect of having more police on the crime rate?
- **Endogeneity Problem:** Police presence is not random. More police are typically assigned to high-crime areas, and crime rates might drive police deployment, creating simultaneity bias. OLS would likely find a positive correlation between police and crime, which is not the true causal effect.
- **Instrumental Variable (Z):** The timing of mayoral or gubernatorial elections. Levitt (1997) argued that politicians hire more police in election years to appear "tough on crime." This creates variation in the size of the police force that is driven by the political cycle, not by the current crime rate (Relevance). The timing of an election itself should have no direct effect on crime (Exogeneity).

Geographic "Lotteries": Land Suitability or Distance

- **Research Question:** Do a country's institutions (e.g., property rights, rule of law) have a causal effect on its long-run economic development?
- **Endogeneity Problem:** It's difficult to separate the effect of institutions from other factors like geography or culture. Rich countries may be able to "afford" better institutions, creating reverse causality.
- **Instrumental Variable (Z):** Mortality rates of early European colonial settlers. Acemoglu, Johnson, and Robinson (2001) argued that in places where settlers faced high mortality rates (e.g., due to malaria in Africa), they established "extractive" institutions designed to exploit resources quickly. Where they could settle safely, they set up institutions that protected property rights, mimicking their home countries. These early institutional choices persisted over centuries (Relevance). The historical mortality rates are unlikely to have a direct effect on a country's GDP *today*, except through their persistent effect on its institutions (Exogeneity).

These examples highlight the art of IV estimation: it requires identifying a source of variation—be it from nature, politics, or history—that affects the endogenous variable of interest but is plausibly independent of the unobserved factors that confound the causal relationship.

10.4 Practical Challenges and Diagnostics

Finding a valid instrument is one of the most difficult creative acts in econometrics. The credibility of any IV study rests entirely on the quality of the instrument.

The Weak Instrument Problem A "weak instrument" is one that satisfies the relevance condition ($Cov(Z, X) \neq 0$) but is only weakly correlated with the endogenous variable.

- **Consequences:** If the instrument is weak, even a tiny violation of the exogeneity assumption ($Cov(Z, u) \approx 0$) can lead to large biases in the 2SLS estimate. Furthermore, weak instruments lead to imprecise (large standard error) estimates.
- **Diagnosis:** The standard diagnostic is the **F-statistic from the first-stage regression**. A common rule of thumb is that an **F-statistic less than 10 signals a weak instrument problem**. If the F-statistic is low, the 2SLS results should not be trusted.

Testing the Exogeneity Assumption The core exogeneity assumption ($Cov(Z, u) = 0$) cannot be formally tested when we have exactly one instrument for one endogenous variable (a case called ”exact identification”). This is why the justification for the instrument’s validity must be based on deep institutional knowledge and economic theory.

However, if we have more instruments than endogenous variables (”over-identification”), we can perform a **test of over-identifying restrictions** (e.g., the Sargan-Hansen J-test). This test checks if the ”extra” instruments are correlated with the residuals from the 2SLS regression.

- **Null Hypothesis (H_0):** All instruments are exogenous.
- If we reject the null, it’s a strong sign that at least one of our instruments is invalid. If we fail to reject, it provides some (but not definitive) confidence in our instruments’ validity.

> **Note 8** (What Does IV Actually Estimate?)**.**
>
> When the causal effect is heterogeneous (different for different people), the IV estimate does not recover the Average Treatment Effect (ATE). Instead, under certain assumptions, it estimates the **Local Average Treatment Effect (LATE)**. The LATE is the average causal effect for the specific sub-population of individuals whose treatment status was changed by the instrument. These individuals are called ”compliers.” In the Angrist and Krueger example, the LATE is the return to education specifically for those people who were induced to get more schooling because of the compulsory schooling laws, which may not be the same as the return for the population at large.

11 Quasi-Experiments I: Difference-in-Differences

Many of the most credible causal studies come from **quasi- or natural experiments**, situations where a real-world event, policy change, or administrative decision creates a setup that *mimics* a randomized controlled trial. The Difference-in-Differences (DID) design is one of the most common and powerful frameworks for analyzing such events.

> **Note 9** (Quasi Experiments)**.** *A research design that studies causal relationships without random assignment to treatment groups, often because it’s impractical or unethical to do so.*

11.1 The Intuition of DID

The DID method is used to estimate the causal effect of a specific intervention by comparing the change in outcomes over time between a group that receives the treatment (the **treatment group**) and a group that does not (the **control group**).

Imagine a new training program is introduced in one region of a country but not another. We have data on worker wages in both regions, both before and after the program was launched. A simple comparison would suffer from severe biases:

- A simple ”before-after” comparison in the treated region is biased if other things were also changing over time (e.g., a national economic boom).

- A simple ”treated-control” comparison after the policy is biased if the two regions were different to begin with (e.g., the treated region was already richer).

The DID estimator brilliantly solves both problems by netting out these biases. The logic is as follows:

1. Calculate the change in wages in the treated region (After - Before). This difference captures the treatment effect *plus* any time trends.
2. Calculate the change in wages in the control region (After - Before). This difference captures *only* the time trends.
3. The causal effect is the difference between these two differences . By subtracting the change in the control group from the change in the treated group, we isolate the true effect of the treatment, assuming the time trends would have been the same for both groups.

The DID Estimator Let $\bar{Y}_{g,t}$ be the average outcome for group $g \in \{\text{Treatment, Control}\}$ in time period $t \in \{\text{Before, After}\}$.

$$\hat{\delta}_{DID} = (\bar{Y}_{\text{Treat, After}} - \bar{Y}_{\text{Treat, Before}}) - (\bar{Y}_{\text{Control, After}} - \bar{Y}_{\text{Control, Before}}) \tag{66}$$

11.2 The Crucial Assumption: Parallel Trends

The entire credibility of the DID design rests on one single, critical assumption: the **parallel trends assumption**.

The Parallel Trends Assumption In the absence of the treatment, the average outcome for the treatment group would have followed the same time trend as the average outcome for the control group.

This is the counterfactual we can never observe. We assume that the change experienced by the control group is a valid proxy for the change the treatment group *would have* experienced if they had not been treated. While this assumption is fundamentally untestable, we can build confidence in it by using data from multiple pre-treatment periods. If we can show that the two groups were trending in parallel *before* the treatment occurred, it lends strong support to the assumption that they would have continued to do so. A graph showing these pre-trends is a mandatory component of any serious DID study.

11.3 DID Estimation Using Regression

While the double-difference calculation is intuitive, the DID model is most often estimated using a regression framework, as it easily allows for the inclusion of control variables and is equivalent to the two-way fixed effects model in the simple 2x2 case.

Consider a panel dataset where we observe individuals i over time periods t. We define two dummy variables:

- $Treat_i$: A dummy equal to 1 if individual i is in the treatment group, 0 otherwise.
- $Post_t$: A dummy equal to 1 if the time period is after the treatment, 0 otherwise.

We then estimate the following regression model:

$$Y_{it} = \beta_0 + \beta_1 Treat_i + \beta_2 Post_t + \delta(Treat_i \cdot Post_t) + u_{it} \tag{67}$$

The coefficients have the following interpretation:

- β_0: The average outcome for the control group in the pre-period.
- β_1: The average difference between the treatment and control groups in the pre-period.
- β_2: The average change in the outcome for the control group from the pre- to the post-period (the time trend).
- δ: The coefficient on the interaction term. This is the DID estimator . It captures the additional change in the outcome for the treatment group in the post-period, over and above the general time trend. It is our estimate of the causal effect.

Real-World Practice and Extensions In modern empirical work, the simple 2x2 DID model is often extended, especially when researchers have many time periods and when treatment is rolled out to different units at different times.

- **Adding Control Variables**: The regression framework allows for the inclusion of other time-varying covariates, X_{it}, to absorb some of the residual variance and improve precision.
- **Staggered Adoption**: When different groups receive the treatment at different points in time, the model is often specified as a Two-Way Fixed Effects model. For years, this was the standard approach, though recent econometric research has highlighted potential biases in the TWFE estimator in this context and proposed alternative estimators (e.g., Callaway and Sant'Anna, 2021). This is an active area of research.
- **Standard Errors**: As with panel data, it is crucial to cluster standard errors. In DID, the standard practice is to cluster at the level of the group that was treated (e.g., at the state level if the policy was enacted by a state), to account for correlation within those groups.

Note 10 (A Classic DID Study: The Mariel Boatlift)**.**

One of the most famous DID studies is David Card's (1990) analysis of the Mariel Boatlift's effect on the Miami labor market.

- **Treatment:** A sudden, unexpected influx of low-skilled immigrants (the ”Marielitos”) to Miami in 1980.
- **Treatment Group:** The city of Miami.
- **Control Group:** Four similar cities (Atlanta, Houston, etc.) that did not experience the immigration shock.
- **Outcome:** Wages and unemployment rates for low-skilled workers.
- **Finding:** By comparing the change in labor market outcomes in Miami before and after 1980 to the change in the control cities, Card found surprisingly little effect on the wages or employment

> of existing low-skilled workers. The DID design was crucial for netting out the effects of the national recession that was happening at the same time.

12 Quasi-Experiments II: Regression Discontinuity Design

The Regression Discontinuity Design (RDD) is a powerful quasi-experimental method that can be used when a treatment is assigned based on whether an observation's value for a specific numeric variable—the **running variable**—is above or below a known **cutoff point**. The core idea is that individuals who are just barely on either side of the cutoff are likely very similar in all other respects, making any sharp difference in their outcomes a credible estimate of the treatment effect.

12.1 The Intuition of RDD

Imagine a university offers a merit scholarship to all students with an entrance exam score of 80 or higher. The exam score is the running variable, and 80 is the cutoff.

- A student with a score of 80.1 receives the scholarship (treatment).
- A student with a score of 79.9 does not.

It is highly plausible that these two students are virtually identical in terms of motivation, background, and ability. The only substantive difference between them is the scholarship, which was assigned based on a tiny, almost random difference in their scores. By comparing the average outcomes (e.g., graduation rates) of students in a very narrow band just above and just below the 80-point cutoff, we can estimate the causal effect of receiving the scholarship.

This design essentially uses the arbitrary nature of the cutoff to create a localized randomized experiment.

Sharp vs. Fuzzy RDD There are two main variants of the RDD.

- **Sharp RDD:** Treatment assignment is a deterministic function of the running variable. All units above the cutoff are treated, and all units below are not. The scholarship example above is a sharp RDD.
- **Fuzzy RDD:** Crossing the cutoff does not guarantee treatment but rather changes the **probability** of being treated. For example, a rule might make individuals over age 65 **eligible** for a program, but not all of them will sign up. The cutoff still creates a discontinuity in the likelihood of treatment, which can be leveraged using an IV approach.

12.2 The RDD Assumptions and Graphical Analysis

The credibility of an RDD study rests on two key conditions.

1. **No Manipulation of the Running Variable:** Individuals should not be able to precisely manipulate their running variable to place themselves on one side of the cutoff. If students who scored 79 could

somehow get their scores bumped up to 80, the sample around the cutoff would no longer be comparable. A key diagnostic test is to check for a ”bunching” or discontinuous jump in the number of observations right at the cutoff.

2. **Continuity of Potential Outcomes:** In the absence of the treatment, the relationship between the running variable and the outcome must be continuous at the cutoff. There should be no other reason for the outcome to jump suddenly at that exact point.

Graphical Evidence: The Heart of RDD The most compelling evidence for an RDD estimate is almost always graphical. The standard RDD graph plots the average outcome against the running variable. A clear, discontinuous ”jump” in the outcome at the cutoff is the visual signature of a treatment effect. This graphical transparency is one of the main strengths of the RDD method. It is standard practice to show this plot, often with a smooth line (e.g., a local polynomial) fitted to the data on both sides of the cutoff.

12.3 Estimation of the RDD Effect

While the effect can be seen graphically, it is formally estimated using regression models that control for the underlying relationship between the running variable and the outcome.

Local Linear Regression The modern state-of-the-art method for RDD estimation is to use **local linear regression**. This involves:

1. Choosing a narrow **bandwidth** , h, around the cutoff c.
2. Running a linear regression only on the observations within this bandwidth ($c - h < X_i < c + h$).

The model estimated is:

$$Y_i = \beta_0 + \tau D_i + \beta_1(X_i - c) + \beta_2 D_i(X_i - c) + u_i \tag{68}$$

- D_i is a dummy variable equal to 1 if $X_i \geq c$.
- $(X_i - c)$ is the centered running variable. This allows the regression line to have a slope.
- $D_i(X_i - c)$ is an interaction term that allows the slope to be different on either side of the cutoff.
- τ: This is the RDD estimate of the treatment effect. It captures the size of the jump in the regression line precisely at the cutoff.

The choice of bandwidth (h) is crucial. A narrow bandwidth reduces bias by only using observations very close to the cutoff but increases variance because it uses less data. In practice, researchers use data-driven methods to select an optimal bandwidth.

Fuzzy RDD as an IV A fuzzy RDD is estimated using a 2SLS approach, where the discontinuity is used as an instrument.

- **Endogenous Variable:** The actual treatment status, T_i (e.g., whether someone actually enrolled in the program).
- **Instrumental Variable:** The dummy variable for being above the cutoff, D_i. This instrument is relevant (crossing the cutoff changes the probability of treatment) and exogenous (the cutoff itself is arbitrary).

- The resulting 2SLS estimate is the LATE—the causal effect of the treatment for the ”compliers,” i.e., those who were induced to take up the treatment by crossing the cutoff.

Example: The Effect of Class Size on Student Performance One of the most famous RDD studies is Angrist and Lavy’s (1999) analysis of Maimonides’ Rule in Israel.

- **Rule:** A new class must be created whenever a school’s grade enrollment exceeds a multiple of 40 students.
- **Running Variable:** Grade enrollment.
- **Cutoff:** Multiples of 40 (41, 81, etc.).
- **Mechanism:** A grade with 40 students is in one large class. A grade with 41 students is split into two small classes (average size 20.5). This rule creates a sharp, discontinuous drop in average class size at the cutoff.
- **Finding:** By comparing the test scores of students in schools just above and just below the enrollment cutoffs, the authors found a significant causal effect: smaller class sizes led to higher student achievement.

Part IV
ADVANCED TOPICS

13 Time Series Econometrics

We now shift gears to a distinct and highly specialized branch of econometrics: the analysis of **time series data**. Until this point, our primary focus has been on estimating causal effects while battling the problem of endogeneity. Time series econometrics, while also concerned with relationships between variables, often has a different set of primary objectives: **forecasting**, modeling **dynamic relationships**, and understanding the impact of **shocks over time**. The challenges are also different; issues like serial correlation and non-stationarity take center stage.

13.1 The Nature of Time Series Data

Time series data consists of observations on a variable or set of variables over time (e.g., quarterly GDP, monthly inflation, daily stock prices). A key feature of such data is that the observations are typically **not independent** across time. The value of GDP today is highly dependent on its value last quarter. This temporal dependence, known as **autocorrelation** or **serial correlation**, is not a problem to be fixed but a central feature to be modeled.

Autoregressive (AR) Models The simplest way to model this dependence is with an Autoregressive model. An AR(p) model specifies that the current value of a variable, Y_t, depends on its own past values up to p lags. An AR(1) model is:

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + u_t \tag{69}$$

Here, β_1 captures the degree of persistence in the series. This type of model is a fundamental building block for time series forecasting.

13.2 Stationarity: The Foundation of Time Series Analysis

The most critical concept in time series analysis is **stationarity**. A time series is stationary if its statistical properties—mean, variance, and autocorrelation—are all constant over time.

- A stationary series tends to revert to a constant long-run mean.
- The variance of a stationary series is finite and does not depend on time.

Many economic time series are clearly **non-stationary**. For example, nominal GDP trends upwards over time and does not have a constant mean. A regression involving non-stationary variables can lead to a **spurious regression**, where we find a statistically significant relationship between two variables that are in fact unrelated, simply because both are trending over time.

Unit Root Tests for Stationarity How do we formally check for non-stationarity? The most common cause of non-stationarity in economics is the presence of a **unit root**. In the AR(1) model above, if the coefficient $\beta_1 = 1$, the series has a unit root and is non-stationary. Such a series is also called a ”random walk.” Shocks to a unit root process are permanent; the series never reverts to a mean.

To test for this, we use **unit root tests**. The most common are:

- **Augmented Dickey-Fuller (ADF) test**
- **Phillips-Perron (PP) test**

For these tests, the null hypothesis is that the series has a unit root (it is non-stationary). If we find a small p-value, we reject the null and conclude the series is stationary. If a series is found to be non-stationary, it is common to transform it by taking the first difference ($\Delta Y_t = Y_t - Y_{t-1}$) to make it stationary.

13.3 Advanced Topics in Time Series

While the details of advanced time series models are beyond the scope of these introductory notes, it is useful to be aware of the key concepts that practitioners use.

Cointegration: Long-Run Relationships What if we have two or more variables that are themselves non-stationary, but we believe they share a stable, long-run relationship? For example, consumption and income both trend upwards over time, but economic theory suggests they should not drift infinitely far apart. If a linear combination of these non-stationary variables is stationary, they are said to be **cointegrated**. The **Johansen test** is a common method for testing for cointegration. This concept is crucial for building valid long-run economic models.

Granger Causality In the context of stationary time series, we can test for a specific, predictive form of causality called **Granger causality**. The concept is simple: a variable X is said to ”Granger-cause” a variable Y if past values of X contain information that helps predict future values of Y, over and above the information contained in past values of Y alone. This is tested with an F-test on the lags of X in a regression of Y on its own lags and the lags of X. While it is a test of predictive power rather than deep structural causality, it is a valuable tool for understanding dynamic relationships.

Real-World Application: Macroeconomic Forecasting and Policy Analysis Central banks, financial institutions, and governments rely heavily on time series models, particularly **Vector Autoregressions (VARs)**, which model multiple time series variables as a function of their own and each other’s past values.

- **Forecasting:** VARs are used to produce forecasts for key macroeconomic variables like inflation, GDP growth, and unemployment.
- **Impulse Response Functions (IRFs):** A key output of VAR models is the IRF, which traces out the dynamic effect of a one-time ”shock” to one variable on all other variables in the system over time. For example, a central bank might use an IRF to estimate the effect of a 1% interest rate hike on inflation and unemployment over the next three years.

> This highlights the different focus of time series analysis: understanding the dynamic, system-wide interplay of variables over time, which is a different but equally important goal to the microeconometric focus on identifying the causal effect of a single program or variable.

Part V

UNIFYING FRAMEWORKS AND STRUCTURAL ESTIMATION

Part VI

MODERN FRONTIERS IN ECONOMETRICS

17 Analyzing Heterogeneity: Quantile Regression and Treatment Effects

18 Econometrics and Machine Learning

19 Nonparametric and Semiparametric Methods