EC 282: Introduction to Econometrics

Lecture Notes

Spring 2026

Contents

Part I

Probability & Statistics Review

1 Course Introduction

1.1 What is Econometrics?

Econometrics combines economic theory, mathematics, and statistical methods to analyze economic data. The term literally means "economic measurement"—the quantification of economic relationships.

Key Point

Econometrics sits at the intersection of several related fields:

- **Data Science**: Extracting insights from data
- **Statistical Learning**: Building predictive models
- **Machine Learning**: Automated pattern recognition (supervised and unsupervised)
- **Regression Analysis**: Modeling relationships between variables

1.2 The Big Data Revolution

Two important technological changes have transformed how we work with data:

1. **Smartphones and IoT Devices**: We became capable of collecting vastly more digital information than ever before.
2. **Cloud Computing and Servers**: We developed the infrastructure to store, manage, and process massive datasets using technologies like:
   - SQL databases (relational data)
   - Graph databases (network/relationship data)
   - Cloud storage and computing platforms

1.3 Prediction vs. Causal Inference

1.3.1 Machine Learning and Prediction

Machine learning approaches focus on **prediction**—forecasting outcomes based on patterns in data. These methods can be applied in virtually any market or domain.

Example 1.1. Real estate price estimation algorithms (e.g., Zillow's "Zestimate" or Redfin's estimates) use machine learning to predict home values based on property characteristics, location, and market conditions. Note that these predictions are **never perfect**—there is always some prediction error.

Machine learning can be categorized as:

- **Supervised Learning**: The algorithm learns from labeled training data (input-output pairs)
- **Unsupervised Learning**: The algorithm finds patterns in unlabeled data

1.3.2 Classical Econometrics and Causation

Classical econometrics focuses on understanding **causal relationships**—not just whether X and Y are correlated, but whether X actually *causes* changes in Y.

Key Point

The fundamental question in causal research: Does X cause Y, or are they merely correlated due to some other factor?
This requires careful **research design**, not just sophisticated statistical techniques.

Example 1.2. COVID-19 vaccine efficacy could not be established simply by observing that vaccinated people had lower infection rates (correlation). The FDA required **large-scale randomized controlled trials (RCTs)** to establish that vaccines actually *caused* reduced infection rates before granting approval.

1.4 Course Objectives

This course provides an introduction to both predictive and causal methods:

1. Learn the basic tools of regression analysis
2. Understand the critical difference between **correlation** and **causation**
3. Develop practical programming skills for data analysis

1.5 Course Logistics

1.5.1 Software and Tools

- **R and RStudio**: Industry-standard statistical programming environment
- **Stack Overflow**: Q&A platform for programming questions
- **GitHub**: Version control and code sharing platform
- Supplementary resource: "Econometrics Using R"

Note

While spreadsheet programs like Excel and Access remain useful, more complex data analysis increasingly requires programming languages like Python and R, which have become the industry standard. Traditional statistical software (Stata, SPSS) is less commonly used in modern data science workflows.

1.5.2 Grading Structure

High-Stakes Assessments (70%):

- 2 Midterm Exams
- 1 Final Exam
- Note: Exams do **not** require programming skills

Low-Stakes Assessments (30%):

- 7 Homework Assignments (submitted through Blackboard)
- Graded Pass/Fail
- Group work permitted (maximum 2 students per group)
- Each student receives different datasets

2 Review of Probability Theory

2.1 Random Variables

Definition 2.1 (Random Variable)**.** A **random variable** is a numerical summary of a random outcome. We typically denote random variables with capital letters (X, Y, Z) and their specific realized values with lowercase letters (x, y, z).

Random outcomes contain two components:

1. **Random component**: Inherent uncertainty (e.g., coin flip)
2. **Deterministic component**: Systematic patterns that can be modeled

Example 2.1. Purely random: A coin flip resulting in heads or tails.
Mixed: COVID-19 infection status (Yes/No) has both random elements (chance exposure) and deterministic elements (vaccination status, mask usage, etc.).

2.2 Types of Random Variables

2.2.1 Discrete Random Variables

Definition 2.2 (Discrete Random Variable)**.** A random variable is **discrete** if it can take only a finite or countably infinite number of distinct values.

Example 2.2.
- Binary outcome: $Y \in \{0, 1\}$ (e.g., infected or not)
- Grade points: $Y \in \{0, 0.7, 1, 1.3, 1.7, 2, 2.3, 2.7, 3, 3.3, 3.7, 4\}$
- Count data: Number of accidents per day

2.2.2 Continuous Random Variables

Definition 2.3 (Continuous Random Variable)**.** A random variable is **continuous** if it can take any numerical value within an interval or collection of intervals.

Example 2.3. Height, weight, income, temperature, time—any measurement that can take infinitely many values within a range.

Note

In this course, we will primarily work with discrete random variables, though many concepts extend naturally to the continuous case.

2.3 Probability Distributions

For a discrete random variable Y with possible outcomes $\{y_1, y_2, \ldots, y_k\}$, the **probability distribution** assigns a probability to each outcome:

$$\Pr(Y = y_1), \quad \Pr(Y = y_2), \quad \ldots, \quad \Pr(Y = y_k)$$

Property 2.1 (Properties of Probability Distributions)**.** For any valid probability distribution:

1. $0 \leq \Pr(Y = y_i) \leq 1$ for all i
2. $\sum_{i=1}^{k} \Pr(Y = y_i) = 1$

2.4 Expected Value

Definition 2.4 (Expected Value)**.** The **expected value** (or **mean**) of a discrete random variable Y is the long-run average value, defined as:

$$E[Y] = \mu_Y = \sum_{i=1}^{k} y_i \cdot \Pr(Y = y_i) = y_1 p_1 + y_2 p_2 + \cdots + y_k p_k$$

where $p_i = \Pr(Y = y_i)$.

Key Point

The expected value is a **weighted average** of all possible outcomes, where the weights are the probabilities of each outcome occurring.

2.4.1 The Bernoulli Distribution

Definition 2.5 (Bernoulli Random Variable)**.** A **Bernoulli** (or **binary/dummy**) random variable takes only two values:

$$Y \in \{0, 1\}$$

with probabilities:

$$\Pr(Y = 1) = p$$
$$\Pr(Y = 0) = 1 - p$$

Theorem 2.1 (Expected Value of Bernoulli)**.** For a Bernoulli random variable:

$$E[Y] = 0 \cdot (1 - p) + 1 \cdot p = p$$

The expected value equals the probability of "success" (Y = 1).

Example 2.4. Let Y indicate COVID-19 infection status, where $Y = 1$ means infected.
If $\Pr(Y = 1) = 0.01$ and $\Pr(Y = 0) = 0.99$, then:

$$E[Y] = 0 \times 0.99 + 1 \times 0.01 = 0.01 = p$$

The expected value represents the infection rate in the population.

2.5 Variance and Standard Deviation

Definition 2.6 (Variance)**.** The **variance** of a random variable Y measures the weighted spread of outcomes around the mean μ_Y:

$$\text{Var}(Y) = \sigma_Y^2 = E[(Y - \mu_Y)^2] = \sum_{i=1}^{k} (y_i - \mu_Y)^2 \cdot \Pr(Y = y_i)$$

Definition 2.7 (Standard Deviation)**.** The **standard deviation** is the square root of the variance:

$$\sigma_Y = \sqrt{\sigma_Y^2} = \sqrt{\text{Var}(Y)}$$

Note

We often prefer the standard deviation because it has the **same unit of measurement** as the original variable Y, making it more interpretable.

Property 2.2. The variance is always non-negative: $\sigma_Y^2 \geq 0$, and equals zero only when Y is constant (not random).

Theorem 2.2 (Variance of Bernoulli)**.** For a Bernoulli random variable with $\Pr(Y = 1) = p$:

$$\begin{aligned}\sigma_Y^2 &= (0-p)^2(1-p) + (1-p)^2 p \\ &= p^2(1-p) + (1-p)^2 p \\ &= p(1-p)[p + (1-p)] \\ &= p(1-p)\end{aligned}$$

Example 2.5. For COVID-19 infection with $p = 0.01$:

$$\sigma_Y^2 = 0.01 \times 0.99 = 0.0099$$

$$\sigma_Y = \sqrt{0.0099} \approx 0.0995$$

2.6 Two Random Variables: Joint and Marginal Distributions

When working with two discrete random variables X and Y, we need to understand how they relate to each other.

Definition 2.8 (Marginal Distribution)**.** The **marginal distribution** of X (or Y) describes the probability distribution of that variable alone, ignoring the other:

$$\Pr(X = x) \quad \text{and} \quad \Pr(Y = y)$$

Definition 2.9 (Joint Distribution)**.** The **joint distribution** describes the probability that X and Y simultaneously take specific values:

$$\Pr(X = x, Y = y)$$

Property 2.3 (Relationship Between Joint and Marginal)**.** The marginal distribution can be obtained from the joint distribution by summing over all values of the other variable:

$$\Pr(Y = y) = \sum_i \Pr(X = x_i, Y = y)$$

Example 2.6 (Commute Time and Rain)**.** Let:

- $Y \in \{0, 1\}$ where $Y = 0$ is long commute, $Y = 1$ is short commute
- $X \in \{0, 1\}$ where $X = 0$ is rain, $X = 1$ is no rain

Joint Distribution:

	$X = 0$ (Rain)	$X = 1$ (No Rain)	Marginal of Y
$Y = 0$ (Long)	0.15	0.07	0.22
$Y = 1$ (Short)	0.15	0.63	0.78
Marginal of X	0.30	0.70	1.00

Calculations:

$$\Pr(Y = 0) = \Pr(X = 0, Y = 0) + \Pr(X = 1, Y = 0) = 0.15 + 0.07 = 0.22$$
$$\Pr(Y = 1) = \Pr(X = 0, Y = 1) + \Pr(X = 1, Y = 1) = 0.15 + 0.63 = 0.78$$

Property 2.4. All joint probabilities must sum to 1:

$$\sum_i \sum_j \Pr(X = x_i, Y = y_j) = 1$$

3 Conditional Probability and Related Concepts

3.1 Bayes' Theorem and Conditional Probability

Definition 3.1 (Conditional Probability)**.** The **conditional probability** of $Y = y$ given that $X = x$ is:

$$\Pr(Y = y \mid X = x) = \frac{\Pr(Y = y, X = x)}{\Pr(X = x)}$$

This represents the probability that Y equals y, **conditional on** knowing that X equals x.

Theorem 3.1 (Bayes' Theorem)**.**

$$\Pr(Y = y \mid X = x) = \frac{\Pr(X = x \mid Y = y) \cdot \Pr(Y = y)}{\Pr(X = x)}$$

Example 3.1 (Continued: Commute and Rain)**.** What is the probability of a short commute given that it's raining?

$$\Pr(Y = 1 \mid X = 0) = \frac{\Pr(X = 0, Y = 1)}{\Pr(X = 0)} = \frac{0.15}{0.30} = 0.50$$

When it rains, there's a 50% chance of a short commute.

3.2 Conditional Expected Value

Definition 3.2 (Conditional Expected Value)**.** The **conditional expected value** of Y given $X = x$ is:

$$E[Y \mid X = x] = \sum_i y_i \cdot \Pr(Y = y_i \mid X = x)$$

Example 3.2 (Rolling a Die)**.** Consider rolling a fair six-sided die. Define:

- $Y \in \{1, 2, 3, 4, 5, 6\}$ with $\Pr(Y = y_i) = 1/6$ for all i
- $X \in \{0, 1\}$ where $X = 0$ if Y is even, $X = 1$ if Y is odd

Unconditional Expected Value:

$$E[Y] = \sum_{i=1}^{6} y_i \cdot \Pr(Y = y_i) = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + \cdots + 6 \cdot \frac{1}{6} = \frac{21}{6} = 3.5$$

Conditional Expected Value Given Odd ($X = 1$):

$$E[Y \mid X = 1] = 1 \cdot \frac{1}{3} + 3 \cdot \frac{1}{3} + 5 \cdot \frac{1}{3} = \frac{9}{3} = 3$$

Conditional Expected Value Given Even ($X = 0$):

$$E[Y \mid X = 0] = 2 \cdot \frac{1}{3} + 4 \cdot \frac{1}{3} + 6 \cdot \frac{1}{3} = \frac{12}{3} = 4$$

3.3 Law of Iterated Expectations

Theorem 3.2 (Law of Iterated Expectations (LIE))**.** The unconditional expected value equals the weighted average of conditional expected values:

$$E[Y] = \sum_i E[Y \mid X = x_i] \cdot \Pr(X = x_i)$$

In compact notation:

$$E[Y] = E\big[E[Y \mid X]\big]$$

Key Point

The Law of Iterated Expectations states that we can compute $E[Y]$ by:

1. Computing $E[Y \mid X = x]$ for each possible value of X
2. Taking the weighted average, using $\Pr(X = x)$ as weights

Example 3.3 (Verification with Die Example)**.** Using our die rolling example:

$$\begin{aligned} E[Y] &= E[Y \mid X = 0] \cdot \Pr(X = 0) + E[Y \mid X = 1] \cdot \Pr(X = 1) \\ &= 4 \cdot \frac{1}{2} + 3 \cdot \frac{1}{2} \\ &= 2 + 1.5 = 3.5\checkmark \end{aligned}$$

This matches our direct calculation of $E[Y] = 3.5$.

3.4 Independence

Definition 3.3 (Statistical Independence)**.** Two random variables X and Y are **independent** if knowing the value of one provides no information about the other:

$$\Pr(Y = y \mid X = x) = \Pr(Y = y) \quad \text{for all } x, y$$

Property 3.1 (Equivalent Characterization)**.** X and Y are independent if and only if:

$$\Pr(X = x, Y = y) = \Pr(X = x) \cdot \Pr(Y = y) \quad \text{for all } x, y$$

The joint probability equals the product of the marginal probabilities.

Note

Independence is a strong assumption. In our commute example, X (rain) and Y (commute time) are likely **not** independent—rain probably affects commute time!

3.5 Covariance

Definition 3.4 (Covariance)**.** The **covariance** between two random variables X and Y measures how they vary together:

$$\text{Cov}(X, Y) = \sigma_{XY} = E\big[(X - \mu_X)(Y - \mu_Y)\big]$$

For discrete random variables:

$$\sigma_{XY} = \sum_i \sum_j (x_i - \mu_X)(y_j - \mu_Y) \cdot \Pr(X = x_i, Y = y_j)$$

Property 3.2 (Interpretation of Covariance)**.**

- $\sigma_{XY} > 0$: X and Y tend to move in the same direction
- $\sigma_{XY} < 0$: X and Y tend to move in opposite directions
- $\sigma_{XY} = 0$: No linear relationship (but not necessarily independent!)

Theorem 3.3 (Covariance of Independent Variables)**.** If X and Y are independent, then $\text{Cov}(X, Y) = 0$.

Warning: The converse is **not** true! Zero covariance does not imply independence.

3.6 Correlation

Definition 3.5 (Correlation Coefficient)**.** The **correlation** between X and Y is the standardized covariance:

$$\text{Corr}(X, Y) = \rho_{XY} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

Property 3.3 (Properties of Correlation)**.**

1. $-1 \leq \rho_{XY} \leq 1$
2. $\rho_{XY} = 1$: Perfect positive linear relationship
3. $\rho_{XY} = -1$: Perfect negative linear relationship
4. $\rho_{XY} = 0$: No linear relationship
5. Correlation is **unitless** (unlike covariance)

Key Point

Correlation measures the strength and direction of the **linear** relationship between two variables. A strong nonlinear relationship might have correlation near zero!

Note

With correlation, **only the direction matters, not the scale**. A correlation of $\rho = 0.7$ indicates a positive relationship; whether the variables are measured in dollars or thousands of dollars doesn't change the correlation.

4 Random Sampling and the Sample Average

4.1 Population vs. Sample

In most real-world applications, we cannot observe the entire population. Instead, we work with samples to learn about population parameters.

Definition 4.1 (Population Parameters)**.** The true characteristics of the population distribution that we want to learn about:

- Population mean: $E[Y] = \mu_Y$
- Population variance: $\text{Var}(Y) = \sigma_Y^2$
- Population standard deviation: σ_Y
- Covariance: $\text{Cov}(X, Y)$
- Correlation: $\text{Corr}(X, Y)$

These parameters are **unknown** and must be estimated from sample data.

4.2 Random Sampling

Definition 4.2 (Random Sample)**.** A **random sample** $\{Y_1, Y_2, \ldots, Y_n\}$ consists of n observations drawn from a population such that:

1. Each Y_i is equally likely to be drawn
2. Each Y_i is drawn from the same probability distribution

Definition 4.3 (IID)**.** Random variables $Y_1, Y_2, \ldots, Y_n$ are **independently and identically distributed (i.i.d.)** if:

1. **Identically distributed**: Each Y_i comes from the same probability distribution
2. **Independent**: The value of any Y_i provides no information about any other Y_j

Random sampling ensures the i.i.d. property.

4.3 The Sample Mean

Definition 4.4 (Sample Mean)**.** The **sample mean** (or sample average) of n randomly drawn observations is:

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \frac{Y_1 + Y_2 + \cdots + Y_n}{n}$$

Key Point

The sample mean $\bar{Y}$ is itself a **random variable**. Every time we draw a new random sample, we get a different value of $\bar{Y}$. This means $\bar{Y}$ has its own probability distribution, expected value, and variance.

4.4 Properties of the Sample Mean

Theorem 4.1 (Expected Value of Sample Mean)**.** If $Y_1, Y_2, \ldots, Y_n$ are i.i.d. with $E[Y_i] = \mu_Y$, then:

$$E[\bar{Y}] = \mu_Y$$

Proof.

$$\begin{aligned} E[\bar{Y}] &= E\left[\frac{1}{n}\sum_{i=1}^{n} Y_i\right] = \frac{1}{n}E\left[\sum_{i=1}^{n} Y_i\right] \\ &= \frac{1}{n}\left(E[Y_1] + E[Y_2] + \cdots + E[Y_n]\right) \\ &= \frac{1}{n}\left(\mu_Y + \mu_Y + \cdots + \mu_Y\right) = \frac{1}{n}(n \cdot \mu_Y) = \mu_Y \end{aligned}$$

□

Theorem 4.2 (Variance of Sample Mean)**.** If $Y_1, Y_2, \ldots, Y_n$ are i.i.d. with $\text{Var}(Y_i) = \sigma_Y^2$, then:

$$\text{Var}(\bar{Y}) = \frac{\sigma_Y^2}{n}$$

The standard deviation of $\bar{Y}$ (called the **standard error**) is:

$$\sigma_{\bar{Y}} = \frac{\sigma_Y}{\sqrt{n}}$$

Proof. Because Y_i and Y_j are independent for $i \neq j$:

$$\begin{aligned} \text{Var}(\bar{Y}) &= \text{Var}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n^2}\text{Var}\left(\sum_{i=1}^{n} Y_i\right) \\ &= \frac{1}{n^2}\left(\text{Var}(Y_1) + \text{Var}(Y_2) + \cdots + \text{Var}(Y_n)\right) \\ &= \frac{1}{n^2}(n \cdot \sigma_Y^2) = \frac{\sigma_Y^2}{n} \end{aligned}$$

□

Note

As the sample size n increases, the variance of $\bar{Y}$ decreases. This means larger samples give us more precise estimates of the population mean.

5 Large Sample Approximations

5.1 Law of Large Numbers

Theorem 5.1 (Law of Large Numbers (LLN))**.** Let $Y_1, Y_2, \ldots, Y_n$ be i.i.d. random variables with $E[Y_i] = \mu_Y$ and $\text{Var}(Y_i) = \sigma_Y^2 < \infty$. Then as $n \to \infty$:

$$\bar{Y} \xrightarrow{p} \mu_Y$$

In words: the sample mean converges in probability to the population mean as the sample size grows.

Key Point

The Law of Large Numbers tells us that $\bar{Y}$ is a good approximation for μ_Y when the sample size n is large. The larger the sample, the closer $\bar{Y}$ tends to be to μ_Y.

5.2 Central Limit Theorem

Theorem 5.2 (Central Limit Theorem (CLT))**.** Let $Y_1, Y_2, \ldots, Y_n$ be i.i.d. random variables with $E[Y_i] = \mu_Y$ and $\text{Var}(Y_i) = \sigma_Y^2 < \infty$. Then as $n \to \infty$:

$$\bar{Y} \overset{a}{\sim} N\left(\mu_Y, \frac{\sigma_Y^2}{n}\right)$$

Or equivalently, the standardized sample mean converges to a standard normal:

$$\frac{\bar{Y} - \mu_Y}{\sigma_Y/\sqrt{n}} \xrightarrow{d} N(0,1)$$

Key Point

The Central Limit Theorem is remarkable: **regardless of the original distribution of** Y, the sampling distribution of $\bar{Y}$ is approximately normal for large n. This is why the normal distribution is so important in statistics!

5.3 The Normal Distribution

Definition 5.1 (Normal Distribution)**.** A random variable Y follows a **normal distribution** with mean μ_Y and variance σ_Y^2, written $Y \sim N(\mu_Y, \sigma_Y^2)$, if its probability density function is:

$$f(y) = \frac{1}{\sigma_Y\sqrt{2\pi}} \exp\left(-\frac{(y-\mu_Y)^2}{2\sigma_Y^2}\right)$$

Definition 5.2 (Standard Normal Distribution)**.** The **standard normal distribution** is a normal distribution with mean 0 and variance 1:

$$Z \sim N(0,1)$$

where $E[Z] = 0$ and $\text{Var}(Z) = 1$.

Property 5.1 (Standardization)**.** If $Y \sim N(\mu_Y, \sigma_Y^2)$, then the standardized variable:

$$Z = \frac{Y - \mu_Y}{\sigma_Y} \sim N(0, 1)$$

Property 5.2 (95% Interval for Normal Distribution)**.** For a standard normal variable $Z \sim N(0, 1)$:

$$\Pr(-1.96 \leq Z \leq 1.96) \approx 0.95$$

This means approximately 95% of the probability mass lies within 1.96 standard deviations of the mean.

Example 5.1. Suppose $Y \sim N(1, 4)$, so $\mu_Y = 1$ and $\sigma_Y = 2$. Find $\Pr(Y \leq 2)$.

Solution: Standardize to convert to the standard normal:

$$\Pr(Y \leq 2) = \Pr\left(\frac{Y - 1}{2} \leq \frac{2 - 1}{2}\right) = \Pr(Z \leq 0.5)$$

Using the standard normal table: $\Pr(Z \leq 0.5) = 0.691$.

For $\Pr(1 \leq Y \leq 2)$:

$$\Pr(1 \leq Y \leq 2) = \Pr(Y \leq 2) - \Pr(Y \leq 1) = \Pr(Z \leq 0.5) - \Pr(Z \leq 0)$$

$$= 0.691 - 0.50 = 0.191$$

6 Review of Statistics: Estimation

6.1 Estimators and Estimates

Definition 6.1 (Estimator)**.** An **estimator** is a function of sample data used to estimate an unknown population parameter. Since it depends on random sample data, an estimator is itself a random variable.

Common estimators include:

- Sample mean: $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$
- Sample median
- Sample variance

Definition 6.2 (Estimate)**.** An **estimate** is the numerical value obtained when we plug actual sample data into an estimator. While an estimator is a random variable, an estimate is a specific number.

Example 6.1. Suppose we want to estimate the average hourly earnings of college graduates. Let Y be hourly earnings at the population level, with unknown mean μ_Y.

We draw a random sample $\{Y_1, Y_2, \ldots, Y_n\}$ and compute:

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

Here $\bar{Y}$ is the **estimator** (a formula), and the computed value (say, \$25.50) is the **estimate**.

6.2 Properties of Good Estimators

In general, an estimator of μ_Y is denoted $\hat{\mu}_Y$. What makes a good estimator?

Definition 6.3 (Unbiasedness)**.** An estimator $\hat{\mu}_Y$ is **unbiased** if:

$$E[\hat{\mu}_Y] = \mu_Y$$

On average, the estimator equals the true parameter value.

Definition 6.4 (Consistency)**.** An estimator $\hat{\mu}_Y$ is **consistent** if:

$$\hat{\mu}_Y \xrightarrow{p} \mu_Y \quad \text{as } n \to \infty$$

As the sample size grows, the estimator converges to the true parameter.

Definition 6.5 (Efficiency)**.** Between two unbiased estimators $\hat{\mu}_Y$ and $\tilde{\mu}_Y$, we prefer the one with smaller variance. An estimator is **efficient** if it has the smallest variance among all unbiased estimators.

$$\text{Var}(\hat{\mu}_Y) < \text{Var}(\tilde{\mu}_Y) \implies \hat{\mu}_Y \text{ is more efficient}$$

6.3 BLUE: Best Linear Unbiased Estimator

Theorem 6.1 (Sample Mean is BLUE)**.** Under random sampling, the sample mean $\bar{Y}$ is the **Best Linear Unbiased Estimator (BLUE)** of μ_Y:

- **Linear**: $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ is a linear function of the observations
- **Unbiased**: $E[\bar{Y}] = \mu_Y$
- **Best**: $\bar{Y}$ has the smallest variance among all linear unbiased estimators

Note

The sample mean minimizes the sum of squared deviations. To see this, consider minimizing:

$$\sum_{i=1}^{n}(Y_i - m)^2$$

Taking the derivative with respect to m and setting equal to zero:

$$\frac{d}{dm}\sum_{i=1}^{n}(Y_i - m)^2 = -2\sum_{i=1}^{n}(Y_i - m) = 0$$

Solving: $\sum_{i=1}^{n} Y_i = nm$, so $m = \frac{1}{n}\sum_{i=1}^{n} Y_i = \bar{Y}$.

6.4 Sampling Distribution Examples

Example 6.2. By the CLT, $\bar{Y} \sim N\left(\mu_Y, \frac{\sigma_Y^2}{n}\right)$. Given $\mu_Y = 100$, $\sigma_Y^2 = 43$, and $n = 100$:

$$\bar{Y} \sim N\left(100, \frac{43}{100}\right) = N(100, 0.43)$$

(a) Find $\Pr(\bar{Y} < 101)$:

$$\Pr(\bar{Y} < 101) = \Pr\left(Z < \frac{101 - 100}{\sqrt{0.43}}\right) = \Pr(Z < 1.525) \approx 0.936$$

(b) With $n = 64$: $\sigma_{\bar{Y}}^2 = 43/64 = 0.672$

$$\Pr(101 < \bar{Y} < 103) = \Pr\left(\frac{101 - 100}{\sqrt{0.672}} < Z < \frac{103 - 100}{\sqrt{0.672}}\right)$$

$$= \Pr(1.22 < Z < 3.66) \approx 0.111$$

(c) With $n = 165$: $\sigma_{\bar{Y}}^2 = 43/165 = 0.26$

$$\Pr(\bar{Y} > 98) = 1 - \Pr\left(Z < \frac{98 - 100}{\sqrt{0.26}}\right) \approx 1.00$$

6.5 Non-Random Sampling and Selection Bias

Note

Random sampling with i.i.d. observations $(Y_1, Y_2, \ldots, Y_n)$ is crucial for valid inference. **Non-random sampling** can lead to **sample selection bias**:
Examples of selection bias:

- Surveying unemployment on Sundays (employed people may be less available)
- Studying cancer rates without accounting for age (survivorship bias)
- Online surveys (exclude those without internet access)

Selection bias means our sample is not representative of the population, and our estimates may be systematically wrong.

7 Hypothesis Testing

7.1 Introduction to Hypothesis Testing

Hypothesis testing provides a framework for making decisions about population parameters based on sample data.

Example 7.1 (Motivating Questions)**.**
- Do vaccines work? (Is the effect different from zero?)
- Do masks reduce transmission?
- Is there a gender wage gap? A racial gap in hiring?

7.2 Setting Up Hypotheses

Definition 7.1 (Null and Alternative Hypotheses)**.**
- **Null hypothesis** (H_0): The hypothesis we are trying to reject. Typically states "no effect" or "no difference."

$$H_0 : \mu_Y = \mu_{Y,0}$$

where $\mu_{Y,0}$ is a specific hypothesized value (often 0).

- **Alternative hypothesis** (H_A): What we believe if we reject H_0.

$$H_A : \mu_Y \neq \mu_{Y,0} \quad \text{(two-sided alternative)}$$

Or one-sided: $H_A : \mu_Y > \mu_{Y,0}$ or $H_A : \mu_Y < \mu_{Y,0}$

7.3 The Testing Procedure

1. **State the hypotheses**: Define H_0 and H_A.
2. **Collect data**: Draw a random sample $\{Y_1, Y_2, \ldots, Y_n\}$ and compute the sample mean $\bar{Y}$.
3. **Acknowledge sampling variation**: Due to randomness, $\bar{Y}$ will almost never exactly equal $\mu_{Y,0}$, even if H_0 is true.
4. **Assume H_0 is true**: Under H_0, by the CLT:

$$\bar{Y} \sim N\left(\mu_{Y,0}, \frac{\sigma_Y^2}{n}\right)$$

5. **Calculate the p-value**: The probability of observing a sample mean at least as extreme as what we observed, assuming H_0 is true.

Definition 7.2 (P-value)**.** The **p-value** is the probability of obtaining a test statistic at least as extreme as the one observed, assuming the null hypothesis is true.

A small p-value indicates that the observed result would be unlikely if H_0 were true, providing evidence against H_0.

Key Point

The p-value answers: "If the null hypothesis were true, how likely would we be to see results this extreme (or more extreme) just by chance?"

- Small p-value (e.g., < 0.05): Evidence against H_0; reject H_0
- Large p-value: Insufficient evidence to reject H_0

7.4 Sample Variance and Standard Error

When the population variance σ_Y^2 is unknown (the typical case), we estimate it using the **sample variance**.

Definition 7.3 (Sample Variance)**.** The **sample variance** is:

$$S_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

The sample standard deviation is $S_Y = \sqrt{S_Y^2}$.

Note

We divide by $n-1$ (not n) to obtain an unbiased estimator of σ_Y^2. This is called **Bessel's correction**.

Definition 7.4 (Standard Error)**.** The **standard error** of $\bar{Y}$ is the estimated standard deviation of the sampling distribution:

$$SE[\bar{Y}] = \hat{\sigma}_{\bar{Y}} = \frac{S_Y}{\sqrt{n}}$$

This serves as a proxy for the true (unknown) $\sigma_{\bar{Y}} = \sigma_Y/\sqrt{n}$.

7.5 The Test Statistic

Definition 7.5 (Z-statistic (variance known))**.** When σ_Y^2 is known (rare case):

$$Z = \frac{\bar{Y} - \mu_{Y,0}}{\sigma_{\bar{Y}}} = \frac{\bar{Y} - \mu_{Y,0}}{\sigma_Y/\sqrt{n}}$$

Under H_0: $Z \sim N(0,1)$

Definition 7.6 (t-statistic (variance unknown))**.** When σ_Y^2 is unknown (typical case):

$$t\text{-stat} = \frac{\bar{Y} - \mu_{Y,0}}{SE[\bar{Y}]} = \frac{\bar{Y} - \mu_{Y,0}}{S_Y/\sqrt{n}}$$

Under H_0 and for large n: t-stat $\overset{a}{\sim} N(0,1)$

Definition 7.7 (P-value Calculation)**.** For a two-sided test with test statistic t:

$$\text{p-value} = 2 \times \Phi(-|t\text{-stat}|)$$

where Φ is the standard normal CDF.

7.6 Complete Hypothesis Testing Procedure

1. **State** H_0 and H_A
2. **Use CLT** to predict the distribution of $\bar{Y}$ under H_0: $\bar{Y} \sim N(\mu_{Y,0}, \sigma_Y^2/n)$
3. **Calculate** the sample mean $\bar{Y} = \frac{1}{n}\sum_i Y_i$
4. **Compute** the test statistic:
$$t\text{-stat} = \frac{\bar{Y} - \mu_{Y,0}}{SE[\bar{Y}]}$$
5. **Calculate** the p-value: p-value $= 2\Phi(-|t\text{-stat}|)$
6. **Compare** p-value to significance level ($\alpha = 0.01, 0.05, 0.10$)
7. **Decision**: If p-value $< \alpha$, reject H_0

7.7 Types of Errors

Definition 7.8 (Type I and Type II Errors).

- **Type I Error** (False Positive): H_0 is true, but you incorrectly reject it.
- **Type II Error** (False Negative): H_0 is false, but you fail to reject it.

	Decision	
Reality	Fail to Reject H_0	Reject H_0
H_0 True	Correct $(1-\alpha)$	Type I Error (α)
H_0 False	Type II Error (β)	Correct (Power $= 1-\beta$)

Definition 7.9 (Key Terminology).

- **Significance Level** (α): Pre-specified probability of Type I error (commonly 0.01, 0.05, or 0.10)
- **Critical Value**: Value of the test statistic at which the test rejects H_0
- **Rejection Region**: Area where we reject H_0
- **Size of Test**: Probability of incorrectly rejecting H_0 (equals α)
- **Power of Test**: Probability of correctly rejecting H_0 when it is false $(1-\beta)$

8 Confidence Intervals

8.1 Definition and Construction

Definition 8.1 (Confidence Interval)**.** A **confidence interval** provides a range of plausible values for the unknown population parameter, based on sample data.

Theorem 8.1 (Confidence Interval for the Mean)**.** A $(1-\alpha)\times 100\%$ confidence interval for μ_Y is:

$$\bar{Y} \pm z_{\alpha/2} \times SE[\bar{Y}]$$

where $z_{\alpha/2}$ is the critical value from the standard normal distribution.

Common confidence intervals:

- 90% CI: $\bar{Y} \pm 1.65 \times SE[\bar{Y}]$
- 95% CI: $\bar{Y} \pm 1.96 \times SE[\bar{Y}]$
- 99% CI: $\bar{Y} \pm 2.576 \times SE[\bar{Y}]$

Example 8.1. With $\bar{Y} = 0.61$ and $SE[\bar{Y}] = 0.049$:

$$95\%\ \text{CI} = 0.61 \pm 1.96 \times 0.049 = [0.51, 0.71]$$

Key Point

Interpretation: We are 95% confident that the true population mean μ_Y lies within the confidence interval.
Note: As we increase the confidence level, the interval becomes wider (more conservative but less precise).

9 Comparing Means from Two Populations

9.1 Setup

Often we want to compare means from two different populations (e.g., men vs. women, treatment vs. control).

Let:

- μ_M = population mean for group M (e.g., men)
- μ_W = population mean for group W (e.g., women)

Definition 9.1 (Hypotheses for Two-Sample Test)**.**

$$H_0 : \mu_M - \mu_W = d_0 \quad (\text{often } d_0 = 0)$$
$$H_A : \mu_M - \mu_W \neq d_0$$

9.2 Test Statistic for Difference in Means

Draw independent random samples:

- Sample from population M: $\bar{Y}_M$ with n_M observations
- Sample from population W: $\bar{Y}_W$ with n_W observations

By the CLT:

$$\bar{Y}_M \sim N\left(\mu_M, \frac{\sigma_M^2}{n_M}\right), \quad \bar{Y}_W \sim N\left(\mu_W, \frac{\sigma_W^2}{n_W}\right)$$

Therefore:

$$\bar{Y}_M - \bar{Y}_W \sim N\left(\mu_M - \mu_W, \frac{\sigma_M^2}{n_M} + \frac{\sigma_W^2}{n_W}\right)$$

Definition 9.2 (Standard Error for Difference in Means)**.**

$$SE[\bar{Y}_M - \bar{Y}_W] = \sqrt{\frac{S_M^2}{n_M} + \frac{S_W^2}{n_W}}$$

Definition 9.3 (t-statistic for Two-Sample Test)**.**

$$t\text{-stat} = \frac{(\bar{Y}_M - \bar{Y}_W) - d_0}{SE[\bar{Y}_M - \bar{Y}_W]}$$

Under H_0 and large n: t-stat $\overset{a}{\sim} N(0,1)$

Definition 9.4 (Confidence Interval for Difference in Means)**.**

$$(\bar{Y}_M - \bar{Y}_W) \pm 1.96 \times SE[\bar{Y}_M - \bar{Y}_W]$$

10 Practice Problems: Confidence Intervals and Two-Sample Tests

Example 10.1 (Single Sample Confidence Interval)**.** Given: $n = 420$, $\bar{Y} = 646.2$, $S_Y = 19.5$

(a) Construct a 95% confidence interval for μ_Y.

Solution: First, compute the standard error:

$$SE[\bar{Y}] = \frac{S_Y}{\sqrt{n}} = \frac{19.5}{\sqrt{420}} \approx 0.95$$

The 95% confidence interval is:

$$\bar{Y} \pm 1.96 \times SE[\bar{Y}] = 646.2 \pm 1.96 \times 0.95 = [644.34, 648.06]$$

Example 10.2 (Two-Sample Test for Class Size Effect)**. (b)** Compare test scores between districts with different class sizes.

Given:

- Group 1 (small classes): $\bar{Y}_1 = 657.4$, $S^2_{Y_1} = 19.4$, $n_1 = 238$
- Group 2 (large classes): $\bar{Y}_2 = 650$, $S^2_{Y_2} = 17.9$, $n_2 = 182$

Difference in means:

$$\bar{Y}_1 - \bar{Y}_2 = 657.4 - 650 = 7.4$$

Standard error of the difference:

$$SE[\bar{Y}_1 - \bar{Y}_2] = \sqrt{\frac{S^2_{Y_1}}{n_1} + \frac{S^2_{Y_2}}{n_2}} = \sqrt{\frac{19.4}{238} + \frac{17.9}{182}} = 1.828$$

95% Confidence Interval:

$$7.4 \pm 1.96 \times 1.828 = [3.82, 10.98]$$

Hypothesis Test:

$$H_0 : \mu_1 - \mu_2 = 0$$
$$H_A : \mu_1 - \mu_2 > 0$$

t-statistic:

$$t\text{-stat} = \frac{7.4 - 0}{1.828} = 4.05$$

Conclusion: Reject the null hypothesis. Districts with smaller classes have significantly better outcomes.

Example 10.3 (Another Two-Sample Comparison)**.** Given:

- $\bar{Y}_1 = 3178.832$, $S_{Y_1} = 580.0068$
- $\bar{Y}_2 = 3432.06$, $S_{Y_2} = 584.622$
- $\bar{Y}_1 - \bar{Y}_2 = -253.2284$

Standard error:

$$SE[\bar{Y}_1 - \bar{Y}_2] = \sqrt{\frac{S^2_{Y_1}}{n_1} + \frac{S^2_{Y_2}}{n_2}} = 28.82106$$

t-statistic:

$$t\text{-stat} = \frac{-253.2284}{28.82106} = -8.79$$

The p-value ≈ 0, so we reject the null hypothesis of no difference.

Part II

Simple Linear Regression

11 Introduction to Linear Regression

11.1 From Correlation to Regression

We have established that the sample correlation r_{XY} measures the strength of the **linear association** between X and Y. However, correlation has limitations:

- Correlation does **not** imply causation
- Correlation only shows the strength of association, not the nature of the relationship

Key Point

Regression analysis allows us to:

1. Quantify the relationship between variables
2. Make predictions
3. (Under certain conditions) Make causal inferences

11.2 Sample Statistics Review

Before diving into regression, let's review the sample statistics we'll need.

Definition 11.1 (Sample Covariance)**.**

$$S_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$$

Definition 11.2 (Sample Variance)**.**

$$S_X^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

Definition 11.3 (Sample Correlation)**.**

$$r_{XY} = \frac{S_{XY}}{S_X S_Y}$$

where $-1 \leq r_{XY} \leq 1$. This tells us how much X and Y are related.

11.3 The Population Regression Function

Definition 11.4 (Population Regression Model)**.** The **population regression function** describes the relationship between Y and X in the entire population:

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

where:

- Y_i = **dependent variable** (outcome, response, left-hand-side variable)
- X_i = **independent variable** (regressor, right-hand-side variable)
- β_0 = **intercept** (population parameter)
- β_1 = **slope** (population parameter)

- u_i = **error term** (unobserved factors affecting Y)
- $i = 1, 2, \ldots, n$ indexes observations

Note

The term $\beta_0 + \beta_1 X_i$ is called the **population regression line**. It represents the systematic (predictable) component of Y, while u_i captures everything else—the "leftover" or unexplained variation.

11.4 Two Main Challenges

When working with regression, we face two fundamental problems:

1. **We don't observe the population**: We only have access to random samples, not the entire population. The population parameters β_0 and β_1 are **unknown**.
2. **Which line should we fit?**: Given sample data, how do we choose the "best" line to estimate the population regression?

11.5 From Population to Sample

Definition 11.5 (Sample Regression Model)**.** Given a random sample of n observations $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n,$ we estimate:

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{u}_i$$

where:

- $\hat{\beta}_0, \hat{\beta}_1$ are **estimated coefficients** (from the sample)
- $\hat{u}_i$ is the **residual** (sample analog of the error term)

Key Point

Error term (u_i) vs. **Residual** ($\hat{u}_i$):

- u_i = population error (unobservable)
- $\hat{u}_i$ = residual (observable, computed from sample)

Definition 11.6 (Predicted (Fitted) Value)**.** The **predicted value** of Y for observation i is:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

This is the value of Y predicted by our estimated regression line.

Property 11.1 (Decomposition of Observed Value)**.** Each observed Y_i can be decomposed as:

$$Y_i = \hat{Y}_i + \hat{u}_i = \underbrace{\hat{\beta}_0 + \hat{\beta}_1 X_i}_{\text{predicted}} + \underbrace{\hat{u}_i}_{\text{residual}}$$

12 Ordinary Least Squares (OLS)

12.1 The OLS Problem

How do we choose $\hat{\beta}_0$ and $\hat{\beta}_1$? We want to minimize the prediction errors.

Definition 12.1 (OLS Criterion)**. Ordinary Least Squares (OLS)** chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the **sum of squared residuals (SSR)**:

$$\min_{\hat{\beta}_0,\hat{\beta}_1} \sum_{i=1}^{n} \hat{u}_i^2 = \min_{\hat{\beta}_0,\hat{\beta}_1} \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$

Note

Why minimize squared residuals?

- Squaring ensures all errors are positive (large negative errors are as bad as large positive ones)
- Squaring penalizes larger errors more heavily
- The math works out nicely (differentiable, unique solution)

12.2 OLS Formulas

Taking derivatives and setting them to zero yields the OLS estimators:

Theorem 12.1 (OLS Estimators)**.**

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{S_{XY}}{S_X^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

Key Point

Everything in these formulas is **observable**—we can compute $\hat{\beta}_0$ and $\hat{\beta}_1$ directly from our sample data.

12.3 Interpreting the Coefficients

Property 12.1 (Interpretation of $\hat{\beta}_1$)**.** Consider two observations with X values differing by ΔX:

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{u}_i$$
$$Y_i + \Delta Y = \hat{\beta}_0 + \hat{\beta}_1 (X_i + \Delta X) + \hat{u}_i$$

Subtracting:

$$\Delta Y = \hat{\beta}_1 \Delta X \quad \Rightarrow \quad \hat{\beta}_1 = \frac{\Delta Y}{\Delta X}$$

Interpretation: $\hat{\beta}_1$ is the predicted change in Y associated with a one-unit increase in X.

Property 12.2 (Interpretation of $\hat{\beta}_0$)**.** From $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$:

$$\hat{\beta}_0 = \bar{Y} \quad \text{when } \bar{X} = 0$$

$\hat{\beta}_0$ is the predicted value of Y when $X = 0$.

Caution: Sometimes this interpretation makes sense (e.g., baseline value), but often $X = 0$ is outside the range of the data or meaningless.

Example 12.1 (Test Scores and Class Size)**.** Suppose we estimate:

$$\widehat{\text{TestScore}} = 698.9 - 2.28 \times \text{ClassSize}$$

Interpretation of $\hat{\beta}_1 = -2.28$: A one-student increase in class size is associated with a 2.28-point decrease in test scores.

Is this effect large or small?

Compare to the outcome's scale:

- As percentage of mean: $\frac{2.28}{698.9} \approx 0.33\%$
- In standard deviation units: If $S_Y = 19$, then $\frac{2.28}{19} \approx 0.12$ SD

Prediction: If class size is 24 students:

$$\hat{Y} = 698.9 - 2.28(24) = 644.18$$

13 Measures of Fit

How well does our regression line fit the data? We need measures to assess the "goodness of fit."

13.1 Decomposition of Variance

Definition 13.1 (Total Sum of Squares (TSS))**.** The **total sum of squares** measures the total variation in Y:

$$TSS = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

Definition 13.2 (Residual Sum of Squares (RSS/SSR))**.** The **residual sum of squares** measures the unexplained variation:

$$RSS = \sum_{i=1}^{n}\hat{u}_i^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

This is what OLS minimizes.

Definition 13.3 (Explained Sum of Squares (ESS))**.** The **explained sum of squares** measures variation explained by the regression:

$$ESS = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$$

Theorem 13.1 (Variance Decomposition)**.**

$$TSS = ESS + RSS$$

Total variation = Explained variation + Unexplained variation

13.2 The Coefficient of Determination (R^2)

Definition 13.4 (R^2)**.** The **coefficient of determination** is the fraction of variance in Y explained by X:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

Property 13.1 (Properties of R^2)**.**

- $0 \leq R^2 \leq 1$
- $R^2 = 0$: X explains none of the variation in Y
- $R^2 = 1$: X explains all the variation in Y (perfect fit)
- In simple regression: $R^2 = r_{XY}^2$ (squared correlation)

Key Point

R^2 tells us what **share** of the variation in Y is explained by the variation in X.

13.3 Standard Error of the Regression (SER)

Definition 13.5 (Standard Error of the Regression)**.** The **SER** measures the typical size of the residuals:

$$SER = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}\hat{u}_i^2} = \sqrt{\frac{RSS}{n-2}}$$

We divide by $n-2$ because we estimated two parameters ($\hat{\beta}_0$ and $\hat{\beta}_1$).

14 Properties of OLS Residuals

OLS has several important algebraic properties that hold by construction.

Property 14.1 (Property 1: Mean of Residuals is Zero)**.**

$$\frac{1}{n}\sum_{i=1}^{n}\hat{u}_i = 0$$

The sample average of OLS residuals is always zero.

Proof. Recall $\hat{u}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$ and $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$.

Substituting:

$$\hat{u}_i = Y_i - \bar{Y} + \hat{\beta}_1\bar{X} - \hat{\beta}_1 X_i = (Y_i - \bar{Y}) - \hat{\beta}_1(X_i - \bar{X})$$

Summing over all observations:

$$\sum_{i=1}^{n}\hat{u}_i = \underbrace{\sum_{i=1}^{n}(Y_i - \bar{Y})}_{=0} - \hat{\beta}_1 \underbrace{\sum_{i=1}^{n}(X_i - \bar{X})}_{=0} = 0$$

□

Property 14.2 (Property 2: Mean of Predicted Values Equals Mean of Y)**.**

$$\frac{1}{n}\sum_{i=1}^{n}\hat{Y}_i = \bar{Y}$$

Proof. Since $Y_i = \hat{Y}_i + \hat{u}_i$:

$$\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n}\hat{Y}_i + \sum_{i=1}^{n}\hat{u}_i = \sum_{i=1}^{n}\hat{Y}_i + 0$$

Therefore $n\bar{Y} = \sum_{i=1}^{n}\hat{Y}_i$, so $\frac{1}{n}\sum_{i=1}^{n}\hat{Y}_i = \bar{Y}$. □

Property 14.3 (Property 3: Residuals are Uncorrelated with X)**.**

$$\sum_{i=1}^{n}\hat{u}_i X_i = \sum_{i=1}^{n}\hat{u}_i(X_i - \bar{X}) = 0$$

Proof.

$$\begin{aligned}\sum_{i=1}^{n}\hat{u}_i X_i &= \sum_{i=1}^{n}\left[(Y_i - \bar{Y}) - \hat{\beta}_1(X_i - \bar{X})\right](X_i - \bar{X}) \\ &= \sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X}) - \hat{\beta}_1\sum_{i=1}^{n}(X_i - \bar{X})^2\end{aligned}$$

By the definition of $\hat{\beta}_1$:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

Substituting:

$$\sum_{i=1}^{n}\hat{u}_i X_i = \sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X}) - \sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X}) = 0$$

□

Property 14.4 (Property 4: TSS = ESS + RSS)**.**

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n}\hat{u}_i^2$$

Proof. Write $Y_i - \bar{Y} = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y}) = \hat{u}_i + (\hat{Y}_i - \bar{Y})$.

Squaring and summing:

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}\hat{u}_i^2 + \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + 2\sum_{i=1}^{n}\hat{u}_i(\hat{Y}_i - \bar{Y})$$

The cross-term vanishes:

$$\begin{aligned}\sum_{i=1}^{n}\hat{u}_i\hat{Y}_i &= \sum_{i=1}^{n}\hat{u}_i(\hat{\beta}_0 + \hat{\beta}_1 X_i) \\ &= \hat{\beta}_0 \underbrace{\sum_{i=1}^{n}\hat{u}_i}_{=0} + \hat{\beta}_1 \underbrace{\sum_{i=1}^{n}\hat{u}_i X_i}_{=0} = 0\end{aligned}$$

Therefore $TSS = RSS + ESS$. □

Key Point

These properties are **algebraic facts** that hold for any OLS regression—they follow directly from how OLS is constructed, not from any assumptions about the data.

15 The Least Squares Assumptions

When are $\hat{\beta}_0$ and $\hat{\beta}_1$ "good" estimators of the population parameters β_0 and β_1? We need certain assumptions to hold.

15.1 Assumption 1: Conditional Mean Zero

Definition 15.1 (Assumption 1: Conditional Mean Independence)**.** The conditional distribution of u_i given X_i has mean zero:

$$E[u_i \mid X_i] = 0$$

This is equivalent to:

$$\text{Corr}(u_i, X_i) = 0$$

Key Point

This assumption says: at any given value of X, the errors u average out to zero. The error term is **not systematically related** to the independent variable.
If $E[u_i \mid X_i] = 0$, then the OLS estimators are **unbiased**:

$$E[\hat{\beta}_0] = \beta_0 \quad \text{and} \quad E[\hat{\beta}_1] = \beta_1$$

Note

Graphical interpretation: At any value of X, the distribution of Y is centered on the population regression line $\beta_0 + \beta_1 X$. Sometimes we overpredict, sometimes we underpredict, but on average the error is zero.

15.2 What Happens When Assumption 1 Fails?

Example 15.1 (Test Scores and Class Size)**.** Consider regressing test scores on student-to-teacher ratio (STR):

$$\text{TestScore}_i = \beta_0 + \beta_1 \times \text{STR}_i + u_i$$

What's in the error term u_i? Everything else that affects test scores:

- Poverty level
- Parental education
- School funding
- Teacher quality
- etc.

Problem: If $\text{Corr}(\text{Poverty}, \text{STR}) > 0$ (poorer districts have larger class sizes), then:

$$\text{Corr}(u_i, X_i) \neq 0$$

and Assumption 1 is violated!

Definition 15.2 (Omitted Variable Bias)**.** When a variable that affects Y is:

1. Omitted from the regression, AND

2. Correlated with the included variable X

the OLS estimator is **biased**. This is called **omitted variable bias**.

Key Point

When Assumption 1 fails:

- The estimated regression line is **biased**
- We systematically over- or under-predict
- The slope $\hat{\beta}_1$ does NOT have a causal interpretation
- We can only interpret the relationship as **association**, not causation

15.3 Proof: $E[u_i|X_i] = 0 \Rightarrow \mathbf{Cov}(u_i, X_i) = 0$

Proof. Recall the definition of covariance:

$$\text{Cov}(X, u) = E[(X - E[X])(u - E[u])]$$

Expanding:

$$\begin{aligned}\text{Cov}(X, u) &= E[Xu - XE[u] - E[X]u + E[X]E[u]] \\ &= E[Xu] - E[X]E[u] - E[X]E[u] + E[X]E[u] \\ &= E[Xu] - E[X]E[u]\end{aligned}$$

Using the Law of Iterated Expectations:

$$E[Xu] = E[E[Xu \mid X]] = E[X \cdot E[u \mid X]]$$

If $E[u \mid X] = 0$:

$$E[Xu] = E[X \cdot 0] = 0$$

Also, by the Law of Iterated Expectations:

$$E[u] = E[E[u \mid X]] = E[0] = 0$$

Therefore:

$$\text{Cov}(X, u) = E[Xu] - E[X]E[u] = 0 - E[X] \cdot 0 = 0$$

□

15.4 Assumption 2: Independent and Identically Distributed (i.i.d.)

Definition 15.3 (Assumption 2: i.i.d. Sampling)**.** The observations (X_i, Y_i) for $i = 1, 2, \ldots, n$ are independently and identically distributed (i.i.d.).

Note

This assumption is ensured by **random sampling**:

- **Identically distributed**: All observations come from the same population/probability distribution
- **Independent**: Draws have no memory—knowing one observation tells you nothing about another

15.5 Assumption 3: No Large Outliers

Definition 15.4 (Assumption 3: Finite Fourth Moments)**.** Large outliers are unlikely. Technically: X and Y have finite fourth moments (kurtosis exists).

$$E[X^4] < \infty \quad \text{and} \quad E[Y^4] < \infty$$

Note

This is a technical assumption needed for:

- The Law of Large Numbers to apply
- The Central Limit Theorem to work
- OLS to be consistent

OLS can be misleading if there are large outliers in the data.

16 Sampling Distribution of OLS Estimators

16.1 OLS Estimators as Random Variables

Just like the sample mean $\bar{Y}$, the OLS estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are **random variables**. Each time we draw a new random sample, we get different estimates.

16.2 Properties Under the Three Assumptions

Theorem 16.1 (Unbiasedness of OLS)**.** If Assumptions 1–3 hold, then the OLS estimators are **unbiased**:

$$E[\hat{\beta}_0] = \beta_0 \quad \text{and} \quad E[\hat{\beta}_1] = \beta_1$$

Theorem 16.2 (Consistency of OLS)**.** If Assumptions 1–3 hold, then as $n \to \infty$:

$$\hat{\beta}_0 \xrightarrow{p} \beta_0 \quad \text{and} \quad \hat{\beta}_1 \xrightarrow{p} \beta_1$$

The OLS estimators are **consistent**.

16.3 Large Sample Distribution

Theorem 16.3 (Large Sample Distribution of OLS)**.** If Assumptions 1–3 hold, then for large samples, by the Central Limit Theorem:

$$\hat{\beta}_1 \overset{a}{\sim} N\left(\beta_1, \sigma^2_{\hat{\beta}_1}\right)$$

where the variance of $\hat{\beta}_1$ is:

$$\sigma^2_{\hat{\beta}_1} = \frac{1}{n} \cdot \frac{\text{Var}[(X_i - \mu_X)u_i]}{[\text{Var}(X_i)]^2}$$

Similarly:

$$\hat{\beta}_0 \overset{a}{\sim} N\left(\beta_0, \sigma^2_{\hat{\beta}_0}\right)$$

Key Point

What affects the precision of $\hat{\beta}_1$?
The variance $\sigma^2_{\hat{\beta}_1}$ is:

- **Smaller** when n is larger (more data = more precision)
- **Smaller** when $\text{Var}(X_i)$ is larger (more spread in X = better estimates)
- **Larger** when $\text{Var}(u_i)$ is larger (more noise = less precision)

Note

Intuition for Var(X): If all your X values are clustered together, it's hard to estimate the slope. You need variation in X to trace out the regression line.

17 Practice Problems: Regression

Example 17.1 (Birth Weight Regression)**.** Regression of birth weight (Y) on number of cigarettes smoked during pregnancy (X):

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

Given: $\hat{\beta}_0 = 509.384$, $\hat{\beta}_1 = -5.614$, and $X_i = 22$ cigarettes.

(a) Predict birth weight when mother smokes 22 cigarettes:

$$\hat{Y}_i = 509.384 - 5.614 \times 22 = 509.384 - 123.508 = 385.9 \text{ grams}$$

(b) If cigarette consumption changes by $\Delta X = 23 - 19 = 4$:

$$\Delta\hat{Y} = \hat{\beta}_1 \times \Delta X = -5.614 \times 4 = -22.5 \text{ grams}$$

A 4-cigarette increase is associated with a 22.5 gram decrease in birth weight.

(c) Find the average outcome $\bar{Y}$ if $\bar{X} = 21$: Using $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$:

$$\bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{X} = 509.384 - 5.614 \times 21 = 391.5 \text{ grams}$$

18 Hypothesis Testing for Regression Coefficients

18.1 Testing β_1

We often want to test whether the slope coefficient is statistically significant—that is, whether X has a real effect on Y.

Definition 18.1 (Hypotheses About β_1)**. Most common case** (two-sided test):

$$H_0 : \beta_1 = 0 \quad \text{(no relationship between } X \text{ and } Y\text{)}$$
$$H_A : \beta_1 \neq 0 \quad \text{(there is a relationship)}$$

General case:

$$H_0 : \beta_1 = \beta_{1,0}$$
$$H_A : \beta_1 \neq \beta_{1,0}$$

where $\beta_{1,0}$ is some hypothesized value.

18.2 Three Steps for Hypothesis Testing

1. **Step 1: Compute the Standard Error**

   $$SE(\hat{\beta}_1) = \sqrt{\hat{\sigma}^2_{\hat{\beta}_1}}$$

   where the estimated variance is:

   $$\hat{\sigma}^2_{\hat{\beta}_1} = \frac{1}{n} \times \frac{\frac{1}{n-2}\sum_{i=1}^{n}(X_i - \bar{X})^2 \hat{u}_i^2}{\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right]^2}$$

2. **Step 2: Calculate the t-statistic**

   $$t\text{-stat} = \frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)}$$

3. **Step 3: Compute the p-value**

   For a **two-sided test**:

   $$\text{p-value} = \Pr_{H_0}\left(|t| > |t^{\text{act}}|\right) = 2\Phi(-|t^{\text{act}}|)$$

   For a **one-sided test** ($H_A : \beta_1 < 0$):

   $$\text{p-value} = \Phi(t^{\text{act}})$$

Key Point

Decision Rule: Reject H_0 if the p-value is less than the pre-specified significance level α (typically 1%, 5%, or 10%).
Critical values for two-sided tests:

- 1% level: $|t| > 2.576$
- 5% level: $|t| > 1.96$
- 10% level: $|t| > 1.645$

Example 18.1 (Testing Significance of Birth Weight Regression)**.** Given: $\hat{\beta}_1 = -5.614$, $SE(\hat{\beta}_1) = 1.862$

Test $H_0 : \beta_1 = 0$ vs. $H_A : \beta_1 \neq 0$

Step 1: $SE(\hat{\beta}_1) = 1.862$ (given)

Step 2: t-statistic:

$$t\text{-stat} = \frac{-5.614 - 0}{1.862} = -3.015$$

Step 3: p-value (two-sided):

$$\text{p-value} = 2\Phi(-|-3.015|) = 2\Phi(-3.015) \approx 0.0026$$

Since p-value < 0.01, we reject H_0 at the 1% significance level. There is strong evidence that cigarette smoking affects birth weight.

Example 18.2 (One-Sided Test)**.** For the same regression, test $H_0 : \beta_1 = 0$ vs. $H_A : \beta_1 < 0$

The t-statistic is still -3.015.

For a one-sided test:

$$\text{p-value} = \Phi(-3.015) \approx 0.0013$$

We reject H_0—there is strong evidence that smoking *decreases* birth weight.

Note

Testing hypotheses about the intercept β_0 is rare and often not meaningful. Always include an intercept in your regression, even if you fail to reject $H_0 : \beta_0 = 0$.

19 Confidence Intervals for Regression Coefficients

19.1 Confidence Interval for β_1

Definition 19.1 (Two Interpretations)**.** A 95% confidence interval for β_1 is:

1. The set of values that cannot be rejected using a two-sided hypothesis test at the 5% level
2. In 95% of all possible samples, the interval will contain the true value of β_1

Theorem 19.1 (Confidence Interval Formula)**.** A $(1-\alpha)\times 100\%$ confidence interval for β_1 is:

$$\hat{\beta}_1 \pm z_{\alpha/2} \times SE(\hat{\beta}_1)$$

Common intervals:

- 90% CI: $\hat{\beta}_1 \pm 1.645 \times SE(\hat{\beta}_1)$
- 95% CI: $\hat{\beta}_1 \pm 1.96 \times SE(\hat{\beta}_1)$
- 99% CI: $\hat{\beta}_1 \pm 2.576 \times SE(\hat{\beta}_1)$

Example 19.1. With $\hat{\beta}_1 = -5.614$ and $SE(\hat{\beta}_1) = 1.862$:
99% Confidence Interval:

$$-5.614 \pm 2.576 \times 1.862 = [-10.41, -0.82]$$

Interpretation: We are 99% confident that the true effect of one additional cigarette on birth weight is between -10.41 and -0.82 grams.

19.2 Confidence Interval for Predicted Change

Definition 19.2 (Confidence Interval for $\beta_1 \cdot \Delta X$)**.** When X changes by ΔX, the predicted change in Y is $\Delta Y = \beta_1 \cdot \Delta X$.

A 95% CI for this predicted change is:

$$\Delta X \times \left[\hat{\beta}_1 \pm 1.96 \times SE(\hat{\beta}_1)\right]$$

Example 19.2. If a mother reduces smoking by 2 cigarettes ($\Delta X = -2$):
99% CI for the effect on birth weight:

$$(-2) \times [-10.41, -0.82] = [1.64, 20.82] \text{ grams}$$

We are 99% confident that reducing smoking by 2 cigarettes increases birth weight by between 1.64 and 20.82 grams.

20 Regression with Binary Variables

20.1 Binary (Dummy) Variables

One of the most common cases in regression is when the independent variable is binary.

Definition 20.1 (Binary/Dummy Variable)**.** A **binary** (or **dummy**, **indicator**) variable takes only two values:

$$D_i \in \{0, 1\}$$

Examples:

- Gender: Male = 1, Female = 0
- Education: BA degree = 1, No BA = 0
- Treatment: Treated = 1, Control = 0
- Class size: Small (STR < 20) = 1, Large (STR ≥ 20) = 0

20.2 Interpreting Binary Regression

Consider the regression:

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

Property 20.1 (Interpretation of Coefficients)**.** When $D_i = 0$:

$$Y_i = \beta_0 + u_i \quad \Rightarrow \quad E[Y_i|D_i = 0] = \beta_0$$

When $D_i = 1$:

$$Y_i = \beta_0 + \beta_1 + u_i \quad \Rightarrow \quad E[Y_i|D_i = 1] = \beta_0 + \beta_1$$

Therefore:

$$\beta_1 = E[Y_i|D_i = 1] - E[Y_i|D_i = 0]$$

β_1 **is the difference in population means between the two groups!**

Key Point

In OLS with a binary regressor:

- $\hat{\beta}_0$ = sample mean of Y for the group where $D = 0$
- $\hat{\beta}_1$ = difference in sample means: $\bar{Y}_{D=1} - \bar{Y}_{D=0}$

This is exactly the two-sample comparison we studied earlier!

Example 20.1 (Test Scores and Class Size)**.** Let $D_i = 1$ if district i has small classes (STR < 20), and $D_i = 0$ otherwise.

Regression result:

$$\widehat{\text{TestScore}}_i = 369.92 + 44.45 \times D_i$$

Interpretation:

- $\hat{\beta}_0 = 369.92$: Average test score in districts with large classes
- $\hat{\beta}_0 + \hat{\beta}_1 = 369.92 + 44.45 = 414.37$: Average test score in districts with small classes
- $\hat{\beta}_1 = 44.45$: Difference in average scores (small − large)

95% Confidence Interval (given $SE(\hat{\beta}_1) = 22.19$):

$$44.45 \pm 1.96 \times 22.19 = [0.96, 87.94]$$

Since the CI does not include zero, the difference is statistically significant at the 5% level.
t-statistic:

$$t = \frac{44.45 - 0}{22.19} = 2.003$$

Part III

Multiple Regression

21 Introduction to Multiple Regression

21.1 Why Multiple Regressors?

So far, we've studied **simple regression** with one independent variable:

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

But what if other factors affect Y? We need **multiple regression**:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i$$

21.2 The Problem of Omitted Variable Bias

Example 21.1 (Class Size and Test Scores in California)**.** Consider regressing test scores on student-to-teacher ratio (STR):

$$\text{TestScore}_i = \beta_0 + \beta_1 \times \text{STR}_i + u_i$$

California has a large immigrant population. Districts with more English learners may:

1. Perform worse on English tests (direct effect)
2. Have larger class sizes (correlation with STR)

If we omit "% English Learners" from the regression, we attribute its effect to STR!

Definition 21.1 (Omitted Variable Bias (Formal))**. Omitted variable bias** occurs when:

1. An omitted variable **affects the outcome** Y
2. The omitted variable is **correlated with an included regressor** X

Both conditions must hold for bias to occur.

Key Point

When omitted variable bias is present:

- $E[\hat{\beta}_1] \neq \beta_1$ (the estimator is biased)
- The estimated relationship may be driven by the omitted factor
- The true relationship might be weaker, stronger, or even opposite in sign
- We cannot give $\hat{\beta}_1$ a causal interpretation

Solution: Include the omitted variable in the regression (if possible).

Note

In the California schools example:

- If districts with large classes have more English learners
- And English learners score lower on tests
- Then $\hat{\beta}_1$ will be more negative than the true effect of class size
- We're incorrectly attributing the "English learner effect" to class size

22 Summary of Key Formulas

22.1 Single Random Variable

Concept	Formula
Expected Value	$E[Y] = \mu_Y = \sum_i y_i \cdot \Pr(Y = y_i)$
Variance	$\text{Var}(Y) = \sigma_Y^2 = \sum_i (y_i - \mu_Y)^2 \cdot \Pr(Y = y_i)$
Standard Deviation	$\sigma_Y = \sqrt{\sigma_Y^2}$
Bernoulli Mean	$E[Y] = p$
Bernoulli Variance	$\text{Var}(Y) = p(1-p)$

22.2 Two Random Variables

Concept	Formula
Conditional Probability	$\Pr(Y = y \mid X = x) = \dfrac{\Pr(X = x, Y = y)}{\Pr(X = x)}$
Conditional Expectation	$E[Y \mid X = x] = \sum_i y_i \cdot \Pr(Y = y_i \mid X = x)$
Law of Iterated Expectations	$E[Y] = E[E[Y \mid X]]$
Independence	$\Pr(X = x, Y = y) = \Pr(X = x) \cdot \Pr(Y = y)$
Covariance	$\text{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$
Correlation	$\rho_{XY} = \dfrac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$

22.3 Sample Mean and Large Sample Results

Concept	Formula
Sample Mean	$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$
Expected Value of $\bar{Y}$	$E[\bar{Y}] = \mu_Y$
Variance of $\bar{Y}$	$\text{Var}(\bar{Y}) = \frac{\sigma_Y^2}{n}$
Standard Error	$\sigma_{\bar{Y}} = \frac{\sigma_Y}{\sqrt{n}}$
Law of Large Numbers	$\bar{Y} \xrightarrow{p} \mu_Y$ as $n \to \infty$
Central Limit Theorem	$\bar{Y} \overset{a}{\sim} N\left(\mu_Y, \frac{\sigma_Y^2}{n}\right)$
Standardization	$Z = \frac{Y - \mu_Y}{\sigma_Y} \sim N(0,1)$

22.4 Estimator Properties

Property	Definition
Unbiased	$E[\hat{\mu}_Y] = \mu_Y$
Consistent	$\hat{\mu}_Y \xrightarrow{p} \mu_Y$ as $n \to \infty$
Efficient	Smallest variance among unbiased estimators
BLUE	Best Linear Unbiased Estimator

22.5 OLS Regression

Concept	Formula
Population Regression	$Y_i = \beta_0 + \beta_1 X_i + u_i$
Sample Regression	$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{u}_i$
OLS Slope	$\hat{\beta}_1 = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_i (X_i - \bar{X})^2} = \frac{S_{XY}}{S_X^2}$
OLS Intercept	$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$
Predicted Value	$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$
Residual	$\hat{u}_i = Y_i - \hat{Y}_i$

22.6 Measures of Fit

Concept	Formula
Total Sum of Squares	$TSS = \sum_i (Y_i - \bar{Y})^2$
Residual Sum of Squares	$RSS = \sum_i \hat{u}_i^2$
Explained Sum of Squares	$ESS = \sum_i (\hat{Y}_i - \bar{Y})^2$
Variance Decomposition	$TSS = ESS + RSS$
R^2	$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$
Standard Error of Regression	$SER = \sqrt{\frac{RSS}{n-2}}$

23 Multiple Regression: Detailed Treatment

23.1 Omitted Variable Bias Revisited

When the conditional mean independence assumption $E[u_i|X_i] = 0$ fails, we have:

$$E[u_i|X_i] \neq 0 \quad \Rightarrow \quad \text{Corr}(u_i, X_i) \neq 0$$

The correlation between the regressor and the error term is denoted:

$$\text{Corr}(X_i, u_i) = \rho_{Xu}$$

Theorem 23.1 (Expected Value of $\hat{\beta}_1$ with OVB)**.** When there is correlation between the regressor and the error:

$$E[\hat{\beta}_1] = \beta_1 + \rho_{Xu} \cdot \frac{\sigma_u}{\sigma_X}$$

The term $\rho_{Xu} \cdot \frac{\sigma_u}{\sigma_X}$ represents the **bias due to the omitted variable**.

Meanwhile, the intercept estimator remains unbiased: $E[\hat{\beta}_0] = \beta_0$.

Key Point

Key facts about omitted variable bias:

1. **Larger sample size will NOT help** — OVB is a systematic bias that does not diminish with more data
2. The **magnitude of bias** depends on:
   - $\text{Corr}(u_i, X_i)$ — correlation between error and regressor
   - $\text{Corr}(X_i, Y_i)$ — correlation between regressor and outcome
3. The larger the correlations, the larger the bias
4. The **direction** of bias depends on the three-way relationship between X_{1i} (included), X_{2i} (omitted), and Y_i (outcome)

23.2 Addressing Omitted Variable Bias

Strategy: Measure the impact of X_1 (e.g., student-teacher ratio) on the outcome Y while holding X_2 (e.g., % of English learners) constant.

Idea: Compare outcomes among observations with similar values of the potentially omitted variable.

Example 23.1 (Class Size and Test Scores)**.** To estimate the effect of class size on test scores without bias from English learner proportions:

- Compare small vs. large classes **among districts with similar % of English learners**
- This "controls for" the confounding variable

This motivates the **multiple regression model**, which allows us to estimate the impact of X_1 on Y while holding X_2 constant.

23.3 The Population Multiple Regression Function

Definition 23.1 (Multiple Regression Model with Two Regressors)**.** The population regression model with two explanatory variables is:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

where:

- Y_i is the dependent variable (outcome)

- X_{1i} is the primary regressor of interest
- X_{2i} is the control variable
- u_i is the error term — the part of Y_i that cannot be explained by X_{1i} and X_{2i}

Definition 23.2 (Conditional Expectation Function)**.** The conditional expectation of Y given specific values of both regressors:

$$E[Y_i|X_{1i} = x_1, X_{2i} = x_2] = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

This gives the **population average of test scores** for districts with student-teacher ratio x_1 and % of English learners x_2.

23.4 Interpretation of Coefficients

Definition 23.3 (Coefficient Interpretation)**.** In the model $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$:

- $\beta_0 =$ **intercept**: $E[Y|X_1 = 0, X_2 = 0]$ — population average when both regressors equal zero
- $\beta_1 =$ **slope coefficient for** X_1: the change in Y induced by a one-unit change in X_1, **holding** X_2 **constant**
- $\beta_2 =$ **slope coefficient for** X_2: the change in Y induced by a one-unit change in X_2, **holding** X_1 **constant**

Theorem 23.2 (Partial Effect / Ceteris Paribus Interpretation)**.** The coefficient β_1 represents the **partial effect** of X_1 on Y:

$$\beta_1 = \frac{\Delta Y}{\Delta X_1} \quad \text{(holding } X_2 \text{ constant)}$$

Derivation:

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 \\ Y + \Delta Y &= \beta_0 + \beta_1(X_1 + \Delta X_1) + \beta_2 X_2 \\ \Delta Y &= \beta_1 \Delta X_1 \end{aligned}$$

Therefore: $\beta_1 = \frac{\Delta Y}{\Delta X_1}$

Equivalent phrases for "holding X_2 constant":

- "After accounting for X_2"
- "After controlling for X_2"
- "Ceteris paribus" (all else equal)

23.5 General Multiple Regression Model

Definition 23.4 (Multiple Regression with k Regressors)**.** The general population multiple regression model:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i$$

for $i = 1, 2, \ldots, n$

Alternative notation using a constant regressor:

$$Y_i = \beta_0 X_{0i} + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i$$

where $X_{0i} \equiv 1$ for all i (constant term/regressor).

The conditional expectation:

$$E[Y_i | X_{1i} = x_1, X_{2i} = x_2, \ldots, X_{ki} = x_k] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

Definition 23.5 (Interpretation of β_j in General Model)**.** The coefficient β_j represents:

$$\beta_j = \frac{\Delta Y}{\Delta X_j} \quad \text{(holding all other } X\text{'s constant)}$$

This is the change in Y induced by a one-unit change in X_j, holding **all else constant**.

23.6 The OLS Estimator for Multiple Regression

Definition 23.6 (OLS Objective in Multiple Regression)**.** Given the model:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i$$

Objective: Find $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$ (a total of $k+1$ parameters) that minimize the sum of squared residuals.

The fitted values and residuals:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_k X_{ki}$$
$$\hat{u}_i = Y_i - \hat{Y}_i$$

Theorem 23.3 (OLS Minimization Problem)**.** The OLS estimators are found by minimizing:

$$\min_{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k} \sum_{i=1}^{n} \hat{u}_i^2 = \min \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

This minimizes the Sum of Squared Residuals (SSR), also called Residual Sum of Squares (RSS).

Note

The OLS estimates are found by solving $k+1$ simultaneous equations (the first-order conditions). While we derived explicit formulas for simple regression, in multiple regression the solution typically requires matrix algebra.

23.7 Omitted Variable Bias Formula

Example 23.2 (Numerical Example of OVB)**.** Consider the true model:

$$\text{Model 1 (True): } Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

But we estimate the misspecified model (omitting X_2):

$$\text{Model 2 (Estimated): } Y_i = \tilde{\beta}_0 + \tilde{\beta}_1 X_{1i} + \tilde{u}_i$$

Also consider the auxiliary regression:

$$X_{2i} = \alpha_0 + \alpha_1 X_{1i} + \varepsilon_i$$

Suppose we obtain these estimates:

$$\begin{aligned}
\tilde{\beta}_1 &= -2.6210 \quad \text{(from Model 2 — biased)} \\
\hat{\beta}_1 &= -1.28970 \quad \text{(from Model 1 — unbiased)} \\
\hat{\beta}_2 &= -0.73403 \\
\hat{\alpha}_1 &= 1.8137
\end{aligned}$$

OVB Formula:

$$\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2 \times \hat{\alpha}_1$$

Verification:

$$-2.6210 = -1.28970 + (-0.73403)(1.8137) = -1.28970 - 1.33 = -2.6210 \quad \checkmark$$

The bias is $\hat{\beta}_2 \times \hat{\alpha}_1 \approx -1.33$.

Key Point

Intuition: Districts with a high % of English learners tend to have not only lower test scores but also a high student-teacher ratio. When we omit English learners from the regression, the estimated coefficient on class size captures both effects, leading to a larger (in absolute value) estimated coefficient.

24 Measures of Fit in Multiple Regression

24.1 Standard Error of the Regression (SER) and RMSE

Both SER and RMSE measure the spread of Y_i around $\hat{Y}_i$ (the average distance between observed values and the regression line predictions).

Definition 24.1 (Standard Error of the Regression)**.**

$$SER = \sqrt{\frac{1}{n-k-1}\sum_{i=1}^{n}\hat{u}_i^2} = \sqrt{\frac{SSR}{n-k-1}}$$

where k is the number of independent variables (regressors, not counting the constant).

The denominator $n-k-1$ represents the **degrees of freedom**.

Definition 24.2 (Root Mean Squared Error)**.**

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\hat{u}_i^2} = \sqrt{\frac{SSR}{n}}$$

Note

Key properties of SER and RMSE:

- Both are measured in the **same units as the dependent variable** Y
- They represent the "average prediction error" — the typical distance between observed and fitted values
- When n is large, SER and RMSE are close to each other
- Since $Y_i = \hat{Y}_i + \hat{u}_i$, the residual $\hat{u}_i$ represents the prediction error

24.2 The Problem with R^2

Recall the definition:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS}$$

where:

$$ESS = \sum_i (\hat{Y}_i - \bar{Y})^2 \quad \text{(Explained Sum of Squares)}$$

$$TSS = \sum_i (Y_i - \bar{Y})^2 \quad \text{(Total Sum of Squares)}$$

$$SSR = \sum_i \hat{u}_i^2 \quad \text{(Sum of Squared Residuals)}$$

Problem: Every time you add a new variable, R^2 will **increase** (or at worst stay the same), regardless of whether that variable is actually useful for explaining Y.

Decomposition: $TSS = ESS + SSR$

24.3 Adjusted R^2

Definition 24.3 (Adjusted R^2)**.** The adjusted R^2 penalizes for adding regressors:

$$\bar{R}^2 = 1 - \frac{n-1}{n-k-1} \cdot \frac{SSR}{TSS}$$

Compare to the regular R^2:

$$R^2 = 1 - \frac{SSR}{TSS}$$

The factor $\frac{n-1}{n-k-1}$ **penalizes for each additional regressor**.

Theorem 24.1 (Properties of Adjusted R^2)**.**

1. Since $\frac{n-1}{n-k-1} > 1$ (when $k \geq 1$), we always have $\bar{R}^2 \leq R^2$

2. As $n \to \infty$, $\bar{R}^2 \to R^2$ (they converge for large samples)
3. Unlike R^2, the adjusted $\bar{R}^2$ can be negative (theoretically)
4. $\bar{R}^2$ can decrease when adding a variable that doesn't improve fit enough to justify the penalty

25 Least Squares Assumptions for Multiple Regression

For the model:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i, \quad i = 1, \ldots, n$$

Definition 25.1 (LS Assumption 1: Conditional Mean Zero)**.**

$$E[u_i | X_{1i} = x_1, X_{2i} = x_2, \ldots, X_{ki} = x_k] = 0$$

The conditional distribution of u_i given all regressors has mean zero.

Implications:

- No omitted variable bias
- Two conditions that would cause OVB:
  1. $\text{Corr}(u_i, X_j) \neq 0$ for some included regressor
  2. The omitted variable affects Y_i

Definition 25.2 (LS Assumption 2: Random Sampling (i.i.d.))**.**

$$(X_{1i}, X_{2i}, \ldots, X_{ki}, Y_i) \text{ are i.i.d.}$$

The observations are **independently and identically distributed**.

This assumption is automatically satisfied with a random sample from the population.

Definition 25.3 (LS Assumption 3: No Large Outliers)**.** Large outliers are unlikely.

How to check: Examine the min, max, mean, and median of each variable. The gap between max and mean (or min and mean) will be larger when there are outliers.

Definition 25.4 (LS Assumption 4: No Perfect Multicollinearity)**.** No regressor is an exact linear function of another regressor.

Perfect multicollinearity occurs when one regressor can be written as a perfect linear combination of others:

$$X_{2i} = c \cdot X_{1i} \quad \text{for some constant } c$$

Example 25.1 (Perfect Multicollinearity)**.** Consider the wage regression:

$$Y_i = \beta_0 + \beta_1 X_{1i} + u_i$$

where Y_i is wage and X_{1i} is years of education.

If we try to add X_{2i} where $X_{2i} = 2 \times X_{1i}$ (e.g., semesters instead of years):

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

This model cannot be estimated because X_2 is a perfect linear function of X_1. The OLS algorithm will return "NA" for one of the coefficients (as shown in R output: "1 not defined because of singularities").

Note

[Homoskedasticity Assumption (Optional)] An additional assumption sometimes made:

$$\text{Var}(u_i|X_{1i}, X_{2i}, \ldots, X_{ki}) = \sigma_u^2 \quad \text{(constant)}$$

The variance of the error term is constant across all values of the regressors. When this fails, we have **heteroskedasticity**.

26 The Dummy Variable Trap

26.1 Including Indicator Variables

Suppose you want to include an indicator (dummy) variable D_i in your regression.

Example 26.1 (Gender and Wages)**.** Define two dummy variables:

$$D_{1i} = \begin{cases} 0 & \text{if female} \\ 1 & \text{if male} \end{cases}$$

$$D_{2i} = \begin{cases} 1 & \text{if female} \\ 0 & \text{if male} \end{cases}$$

With Y_i = wage as the outcome variable, consider the regression:

$$Y_i = \beta_0 + \beta_1 D_{1i} + u_i$$

Here $\hat{\beta}_1$ represents the **average wage difference between males and females**.
Alternatively:

$$Y_i = \alpha_0 + \alpha_1 D_{2i} + u_i$$

In this case, $\hat{\alpha}_1$ also represents the average wage difference between males and females, but with opposite sign: $\hat{\alpha}_1 = -\hat{\beta}_1$.

26.2 The Trap: Including All Categories

Definition 26.1 (Dummy Variable Trap)**.** If you include **both** dummy variables representing all categories of a categorical variable, you create **perfect multicollinearity**:

$$D_{1i} + D_{2i} = 1 \quad \Rightarrow \quad D_{1i} = 1 - D_{2i}$$

This is a perfect linear function!

Theorem 26.1 (Why the Model Fails)**.** Consider the model:

$$Y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + u_i$$

Substituting $D_{2i} = 1 - D_{1i}$:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 D_{1i} + \beta_2(1 - D_{1i}) + u_i \\ &= \beta_0 + \beta_1 D_{1i} + \beta_2 - \beta_2 D_{1i} + u_i \\ &= (\beta_0 + \beta_2) + (\beta_1 - \beta_2) D_{1i} + u_i \end{aligned}$$

This collapses to a **single regressor** with **two parameters** — impossible to estimate separately!

The software will report "NA" for one coefficient ("not defined because of singularities").

Key Point

Rule: When including dummy variables for a categorical variable with k categories, include only $k-1$ dummies. The omitted category becomes the **reference group** or **baseline category**.

Example 26.2 (R Output with Dummy Variable Trap)**.** Correct specification (one dummy for "white"):

```
lm(formula = wage ~ education + white, data = CPS1988)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -123.2581    14.2508  -8.649  <2e-16 ***
education     46.2504     0.8882  52.070  <2e-16 ***
white        133.1458     9.5332  13.967  <2e-16 ***
```

Interpretation:

- $\hat{\beta}_0 = -123.26$: Expected wage when education = 0 and white = 0 (non-white)
- $\hat{\beta}_1 = 46.25$: Each additional year of education increases wage by \$46.25
- $\hat{\beta}_2 = 133.15$: White workers earn \$133.15 more on average than non-white workers, holding education constant

If we include both "white" and "non.white" dummies, R drops one automatically and shows "NA".

27 Imperfect Multicollinearity

27.1 Definition and Consequences

Definition 27.1 (Imperfect Multicollinearity)**. Imperfect multicollinearity** occurs when X_1 and X_2 are **highly correlated** but **not a perfect linear function** of each other.

In the model:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

OLS will work, but **poorly**:

- $\hat{\beta}_1$ and $\hat{\beta}_2$ will be very **imprecise**
- **Large standard errors** for the coefficients

Theorem 27.1 (Why Imperfect Multicollinearity Causes Problems)**.** Recall that $\hat{\beta}_1$ estimates the relationship between X_1 and Y **holding** X_2 **constant**.

If X_1 and X_2 are highly correlated, then:

- It is difficult to estimate the "net effect" of X_1 alone
- Very little variation in X_1 is left to exploit after accounting for X_2
- Result: $SE[\hat{\beta}_1]$ is large $\Rightarrow$ 95% CI for $\hat{\beta}_1$ is wide

27.2 How to Detect Multicollinearity

1. **Pairwise correlations**: If $|\text{Corr}(X_1, X_2)| > 0.7$ or 0.8, this is not a good sign
2. **High R^2 but insignificant coefficients**: If the overall model has high R^2 but individual coefficients are not statistically significant due to large standard errors
3. **Variance Inflation Factor (VIF)**: A formal diagnostic measure

27.3 Variance Inflation Factor (VIF)

Definition 27.2 (Variance Inflation Factor)**.** For the model:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i$$

To calculate $VIF(\hat{\beta}_1)$:

1. Run the auxiliary regression:

$$X_{1i} = \alpha_0 + \alpha_2 X_{2i} + \alpha_3 X_{3i} + \cdots + \alpha_k X_{ki} + \varepsilon_i$$

2. Calculate R_1^2 from this regression
3. Compute:

$$VIF(\hat{\beta}_1) = \frac{1}{1 - R_1^2}$$

Theorem 27.2 (Interpreting VIF)**.**

- $VIF(\hat{\beta}_j) \geq 5$ indicates **severe multicollinearity**
- Some sources use $VIF \geq 10$ as the threshold
- Higher VIF means the standard error of $\hat{\beta}_j$ is inflated by that factor

27.4 How to Fix Multicollinearity

1. **Do nothing**: Sometimes multicollinearity is unavoidable and doesn't prevent valid inference
2. **Drop redundant variables**: If two variables measure essentially the same thing
3. **Transform multicollinear variables**:
   - If GDP and Population are both correlated $\rightarrow$ use GDP per capita instead
   - Combine related variables into an index
4. **Larger sample size**: More data can help provide more variation to separate effects

Example 27.1 (R Output: Multiple Regression with Control Variables)**.** Model: Test scores on STR, English learners, free lunch eligibility, and expenditure per student.

```
lm(formula = read ~ str + english + lunch + expenditure)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 662.090395   9.071658  72.984  < 2e-16 ***
str          -0.203130   0.286034  -0.710    0.478
english      -0.210593   0.030448  -6.916 1.76e-11 ***
lunch        -0.550214   0.020324 -27.072  < 2e-16 ***
expenditure   0.004667   0.000841   5.551 5.08e-08 ***
```

95% CI for STR coefficient: $-0.203 \pm 1.96 \times 0.286 = [-0.76, +0.36]$

The correlation matrix and VIF values show no severe multicollinearity (all VIF < 2).

Key insight: Control variables do not need to have a causal interpretation to be useful for reducing omitted variable bias.

28 Hypothesis Testing in Multiple Regression

28.1 Testing Individual Coefficients

For the model:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i$$

Step 1: State hypotheses for coefficient β_j (where $\beta_{j,0}$ can be any number):

$$H_0 : \beta_j = \beta_{j,0} \quad (j = 1, \ldots, k)$$
$$H_A : \beta_j \neq \beta_{j,0}$$

Step 2: Compute the t-statistic:

$$t\text{-stat} = \frac{\hat{\beta}_j - \beta_{j,0}}{SE[\hat{\beta}_j]}$$

By CLT: $\frac{\hat{\beta}_j - E[\hat{\beta}_j]}{\sqrt{\text{Var}(\hat{\beta}_j)}} \overset{a}{\sim} N(0,1)$

Step 3: Calculate p-value and make decision:

$$p\text{-value} = 2 \times \Phi(-|t\text{-stat}|)$$

If p-value $<$ critical p $\Rightarrow$ Reject H_0

95% Confidence Interval:

$$\hat{\beta}_j \pm 1.96 \times SE[\hat{\beta}_j]$$

29 Test of Joint Hypotheses: The F-Test

29.1 Why Individual t-Tests Don't Work for Joint Hypotheses

Sometimes we need to test **multiple restrictions simultaneously** (joint hypothesis).

Example 29.1 (Testing if School Resources Matter)**.** Consider:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + u_i$$

where: Y = test score, X_1 = STR, X_2 = % English learners, X_3 = % eligible for free lunch, X_4 = expenditure per student.

Null hypothesis: School resources (STR and expenditure) don't matter

$$H_0 : \beta_1 = 0 \text{ AND } \beta_4 = 0$$

Alternative: They do matter

$$H_A : \text{Either } \beta_1 \neq 0 \text{ OR } \beta_4 \neq 0 \text{ (or both)}$$

Here $q = 2$ (number of restrictions in the joint hypothesis).

Theorem 29.1 (Why Individual t-Tests Fail for Joint Hypotheses)**.** Suppose $\beta_1 = 0$ and $\beta_4 = 0$ are both true. At 5% significance level:

- Probability of failing to reject $H_0 : \beta_1 = 0$ is 95%
- Probability of failing to reject $H_0 : \beta_4 = 0$ is 95%
- Probability of failing to reject **both**: $0.95 \times 0.95 = 0.9025$
- Probability of rejecting at least one (Type I error): $1 - 0.9025 = 9.75\%$

This is almost **double** the intended 5% significance level! We reject too often when using individual t-tests for joint hypotheses.

Additional problem: If $\hat{\beta}_1$ and $\hat{\beta}_4$ are **correlated** (which they typically are), the calculation is even more complicated.

29.2 The F-Statistic

Definition 29.1 (F-Test for Joint Hypotheses)**.** The F-statistic accounts for the correlation between coefficient estimates:

$$F = \frac{1}{2}\left[\frac{t_1^2 + t_4^2 - 2\hat{\rho}_{t_1,t_4} \cdot t_1 \cdot t_4}{1 - \hat{\rho}_{t_1,t_4}^2}\right]$$

where $\hat{\rho}_{t_1,t_4}$ is the estimated correlation between t_1 and t_4.

Special case: If t_1 and t_4 are independent:

$$F = \frac{1}{2}(t_1^2 + t_4^2)$$

Under H_0 and in large samples: $F \sim \chi_q^2/q$ (approximately $F_{q,\infty}$)

Theorem 29.2 (Decision Rule)**.** If F is large $\Rightarrow$ Reject the null hypothesis.

Critical values at 5% significance level:

- $q = 1$: Critical $F = 3.84$
- $q = 2$: Critical $F = 3.00$
- $q = 3$: Critical $F = 2.60$

29.3 F-Test Using R^2

Theorem 29.3 (F-Statistic Formula Using R^2)**.** Under homoskedasticity, there's a simpler formula comparing two regressions:

- **Restricted model**: Impose H_0 (e.g., $\beta_1 = 0, \beta_4 = 0$)
- **Unrestricted model**: Full model with all parameters

Calculate R^2 from both models:

$$F = \frac{(R^2_{unr} - R^2_{res})/q}{(1 - R^2_{unr})/(n - k - 1)}$$

where:

- q = number of restrictions
- k = number of regressors in unrestricted model
- n = sample size

Key Point

Intuition:

- If the difference in R^2 is large $\Rightarrow$ F is large $\Rightarrow$ more likely to reject H_0
- If the difference is not big, then maybe the coefficients jointly do not add much prediction power to the model

Example 29.2 (Numerical Example)**.** Given:

$$\begin{aligned} R^2_{unr} &= 0.8212 \quad (q = 2, \text{ testing } \beta_1 = 0, \beta_4 = 0) \\ R^2_{res} &= 0.7959 \\ n - k - 1 &= 420 - 4 - 1 = 415 \end{aligned}$$

Calculate:

$$F = \frac{(0.8212 - 0.7959)/2}{(1 - 0.8212)/415} = \frac{0.0253/2}{0.1788/415} = \frac{0.01265}{0.000431} = 29.36$$

Since $29.36 > 3.00$ (critical value for $q = 2$ at 5%), we **reject the null hypothesis**.

Conclusion: Classroom resources (STR and expenditure) are jointly statistically significant. They do matter for test scores.

Part IV

Extensions

30 Non-Linear Regression Models

30.1 Motivation

In many applications, the relationship between Y and X is **non-linear**:

- The impact of X on Y depends on the **level of** X
- β is not constant — it is a function of X
- Sometimes this is theoretically justified

Example 30.1 (Wage and Age)**.** Wages typically increase faster early in one's career, then the rate of increase slows down:

- The marginal effect of age on wage is larger at age 25 than at age 50
- This suggests a concave (diminishing returns) relationship

Definition 30.1 (General Non-Linear Population Regression Function)**.**

$$Y_i = f(X_{1i}, X_{2i}, \ldots, X_{ki}), \quad i = 1, 2, \ldots, n$$

where $f(\cdot)$ is a non-linear function.

Key insight: In some cases, we can still use OLS after appropriate transformations.

30.2 Two Main Approaches

1. **Polynomial regression**: Population regression function can be approximated by a quadratic, cubic, or higher-order polynomial
2. **Logarithmic transformation**: Transform X, Y, or both to logarithms, which makes interpretation easier

31 Polynomial Regression

Definition 31.1 (Polynomial Regression Model)**.**

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \cdots + \beta_k X_i^k + u_i$$

Key features:

- Single underlying variable X
- All regressors are powers of X
- $Y_i = f(X_i)$ — a polynomial function
- The model is **linear in parameters**, so we can use OLS
- Individual coefficients are hard to interpret directly

Example 31.1 (Common Polynomial Specifications)**.**

1. **Linear**: $Y_i = \beta_0 + \beta_1 X_i + u_i$
2. **Quadratic**: $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + u_i$
3. **Cubic**: $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + u_i$

31.1 Interpreting Polynomial Coefficients

For the quadratic model:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + u_i$$

The marginal effect of X on Y is found by taking the derivative:

$$\frac{dY_i}{dX_i} = \beta_1 + 2\beta_2 X_i$$

Key Point

The marginal effect **depends on the level of** X:

- If $\beta_1 > 0$ and $\beta_2 < 0$: positive but diminishing effect (concave)
- The effect gets smaller as X increases

Example 31.2 (Test Scores and Income (Quadratic))**.** Given estimates: $\hat{\beta}_1 = 3.473$ (SE $= 0.310$) and $\hat{\beta}_2 = -0.036$ (SE $= 0.006$)

Model: $\hat{Y}_i = \hat{\beta}_0 + 3.473 X_i - 0.036 X_i^2$

Marginal effect at different income levels:

$$\frac{d\hat{Y}_i}{dX_i} = 3.473 + 2(-0.036)X_i = 3.473 - 0.072 X_i$$

Evaluating at $\bar{X} = 15.32$ (mean income in \$1000s):

$$\left.\frac{d\hat{Y}_i}{dX_i}\right|_{X=15.32} = 3.473 - 0.072(15.32) = 3.473 - 1.103 = 2.37$$

Interpretation: A \$1,000 increase in income when income is around \$15,317 is associated with a 2.37 point increase in test scores.

At different income levels:

Income (X)	Marginal Effect (dY/dX)
\$5,000	3.11
\$15,000	2.39
\$30,000	1.31

Note

[Caution] **Never extrapolate** (make predictions/evaluate) outside the data range of X. Polynomial models can behave erratically outside the observed data.

31.2 Testing for Non-Linearity

Use F-tests to determine the appropriate polynomial degree:

Example 31.3 (Testing Linear vs. Cubic)**. Unrestricted model**: $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + u_i$
Restricted model: $Y_i = \beta_0 + \beta_1 X_i + u_i$
Test: $H_0 : \beta_2 = 0$ AND $\beta_3 = 0$ vs. H_A: Either $\beta_2 \neq 0$ OR $\beta_3 \neq 0$
If F-stat $= 15.702$, p-value $< 0.01 \Rightarrow$ Reject H_0. Model should be non-linear.

Example 31.4 (Testing Quadratic vs. Cubic)**.** $H_0 : \beta_3 = 0$ (quadratic is sufficient) vs. $H_A : \beta_3 \neq 0$
If F-stat $= 0.2768$, p-value $= 0.5997 > 0.01 \Rightarrow$ Fail to reject H_0.
Conclusion: X^3 should NOT be in the model (overfitting). Quadratic is sufficient.
Note: This is equivalent to an individual t-test on β_3.

Key Point

Summary for Polynomial Models:

- Can be estimated using OLS
- Individual coefficients are hard to interpret
- Best option: Take the derivative and evaluate the marginal effect at a specific X
- Decide on the appropriate form using F-tests or t-tests

32 Logarithmic Transformations

32.1 Properties of Logarithms

Definition 32.1 (Natural Logarithm)**.** $\log(X) = \ln(X)$ denotes the natural logarithm of X.
Logarithms are very useful for modeling **relative (percentage) changes**.

Theorem 32.1 (Logarithm Approximation for Small Changes)**.** For small changes:

$$\ln(X + \Delta X) - \ln(X) = \ln\left(\frac{X + \Delta X}{X}\right) \approx \frac{\Delta X}{X}$$

This equals the **relative change** in X. Multiplying by 100 gives the **percentage change**:

$$\frac{\Delta X}{X} \times 100 = \%\text{ change in } X$$

32.2 Three Logarithmic Specifications

Definition 32.2 (Logarithmic Model Types)**.**

1. **Linear-Log**: $Y_i = \beta_0 + \beta_1 \ln(X_i) + u_i$
2. **Log-Linear**: $\ln(Y_i) = \beta_0 + \beta_1 X_i + u_i$
3. **Log-Log**: $\ln(Y_i) = \beta_0 + \beta_1 \ln(X_i) + u_i$

β_1 is interpreted **very differently** in each case!

32.3 Linear-Log Model

Definition 32.3 (Linear-Log Specification)**.**

$$Y_i = \beta_0 + \beta_1 \ln(X_i) + u_i$$

Derivation of interpretation:
Before: $Y_i = \beta_0 + \beta_1 \ln(X_i) + u_i$
After: $Y_i + \Delta Y = \beta_0 + \beta_1 \ln(X_i + \Delta X) + u_i$
Subtracting:

$$\Delta Y = \beta_1[\ln(X_i + \Delta X) - \ln(X_i)] \approx \beta_1 \cdot \frac{\Delta X}{X_i}$$

Therefore:

$$\beta_1 = \frac{\Delta Y}{\Delta X / X}$$

Key Point

Interpretation (Linear-Log): A 1% increase in X is associated with a $\beta_1/100$ unit change in Y.
Equivalently: A 1% increase in X is associated with a $0.01\beta_1$ change in Y.

32.4 Log-Linear Model

Definition 32.4 (Log-Linear Specification)**.**

$$\ln(Y_i) = \beta_0 + \beta_1 X_i + u_i$$

Derivation:
Before: $\ln(Y_i) = \beta_0 + \beta_1 X_i + u_i$
After: $\ln(Y_i + \Delta Y) = \beta_0 + \beta_1(X_i + \Delta X) + u_i$
Subtracting:

$$\ln(Y_i + \Delta Y) - \ln(Y_i) = \beta_1 \Delta X$$

$$\frac{\Delta Y}{Y} \approx \beta_1 \Delta X$$

Therefore:

$$\beta_1 = \frac{\Delta Y / Y}{\Delta X}$$

Key Point

Interpretation (Log-Linear): A 1-unit increase in X is associated with a $100 \times \beta_1\%$ change in Y.
Equivalently: β_1 represents the percentage change in Y (divided by 100) for a one-unit increase in X.

32.5 Log-Log Model

Definition 32.5 (Log-Log Specification)**.**

$$\ln(Y_i) = \beta_0 + \beta_1 \ln(X_i) + u_i$$

Derivation:
Before: $\ln(Y_i) = \beta_0 + \beta_1 \ln(X_i) + u_i$
After: $\ln(Y_i + \Delta Y) = \beta_0 + \beta_1 \ln(X_i + \Delta X) + u_i$
Subtracting:

$$\ln(Y_i + \Delta Y) - \ln(Y_i) = \beta_1[\ln(X_i + \Delta X) - \ln(X_i)]$$

$$\frac{\Delta Y}{Y} \approx \beta_1 \cdot \frac{\Delta X}{X}$$

Therefore:

$$\beta_1 = \frac{\Delta Y/Y}{\Delta X/X}$$

Key Point

Interpretation (Log-Log): A 1% increase in X is associated with a β_1% change in Y. β_1 is the **elasticity** of Y with respect to X.

32.6 Summary Table

Model	**Specification**	β_1 **Interpretation**
Linear	$Y = \beta_0 + \beta_1 X$	$\Delta X = 1 \Rightarrow \Delta Y = \beta_1$
Linear-Log	$Y = \beta_0 + \beta_1 \ln(X)$	$1\% \uparrow$ in $X \Rightarrow \Delta Y = \beta_1/100$
Log-Linear	$\ln(Y) = \beta_0 + \beta_1 X$	$\Delta X = 1 \Rightarrow \%\Delta Y = 100\beta_1$
Log-Log	$\ln(Y) = \beta_0 + \beta_1 \ln(X)$	$1\% \uparrow$ in $X \Rightarrow \%\Delta Y = \beta_1$

32.7 Practical Considerations

Note

All logarithmic models:

- Can be estimated using OLS
- Hypothesis tests and confidence intervals are interpreted as usual
- Standard errors and t-statistics apply to the transformed model

When to use log transformations:

- Income, wages: Often have skewed distributions — log transformation helps

- Plot the relationships to see if logs are appropriate
- Use diagnostic tests to compare model fits

Example 32.1 (Log-Linear with Dummy Variable)**.** Model: $\ln(Y_i) = \beta_0 + \beta_1 D_i + u_i$
where D_i is a dummy variable (0 or 1).
When $D_i = 0$: $\ln(Y_i|D_i = 0) = \beta_0 + u_i$
When $D_i = 1$: $\ln(Y_i|D_i = 1) = \beta_0 + \beta_1 + u_i$
Therefore:

$$\beta_1 = \ln\left(\frac{Y_i|D_i = 1}{Y_i|D_i = 0}\right)$$

Interpretation: β_1 represents the approximate percentage difference in Y between groups (when $D = 0$ vs. $D = 1$).

Note

[Warning] The approximation $\ln(1 + x) \approx x$ only works when $\Delta Y/Y$ is **small** (typically less than 10-15%).
For exact changes, use the exponential:

$$\ln(Y_i) = \beta_0 + \beta_1 X_i \quad \Rightarrow \quad Y_i = e^{\beta_0 + \beta_1 X_i}$$

33 Exact Percentage Change in Log Models

Recall the log-linear model:

$$\ln(Y_i) = \beta_0 + \beta_1 X_i + u_i$$

When β_1 is **small**, we use the approximation $\ln(1 + x) \approx x$ to interpret β_1 as the approximate percentage change in Y for a one-unit change in X. However, this approximation **only matters when β_1 is large**.

33.1 Deriving the Exact Formula

Starting from the regression output:

$$\ln\left(\frac{\Delta Y_i + Y_i}{Y_i}\right) = \beta_1 \Delta X_i$$

This simplifies to:

$$\ln\left(\frac{\Delta Y_i}{Y_i} + 1\right) = \beta_1 \Delta X_i$$

Using the approximation when changes are small:

$$\frac{\Delta Y_i}{Y_i} \approx \beta_1 \Delta X_i = \%\Delta Y_i$$

33.2 Exact Percentage Change

For the **exact** percentage change, consider two values X_{i0} and X_{i1}:

$$\ln Y_{i0} = \beta_0 + \beta_1 X_{i0} + u_i$$
$$\ln Y_{i1} = \beta_0 + \beta_1 X_{i1} + u_i$$

The percentage change in Y is:

$$\frac{\Delta Y_i}{Y_i} = \frac{Y_{i1} - Y_{i0}}{Y_{i0}} = \frac{Y_{i1}}{Y_{i0}} - 1$$

Taking the exponential:

$$\frac{\Delta Y_i}{Y_i} = \exp\left(\ln \frac{Y_{i1}}{Y_{i0}}\right) - 1 = \exp(\beta_1 X_{i1} - \beta_1 X_{i0}) - 1$$

Key Point

The **exact percentage change formula** for log-linear models is:

$$\boxed{\%\Delta Y = \exp(\beta_1 \Delta X) - 1}$$

Example 33.1 (Large Coefficient)**.** If $\hat{\beta}_1 = 0.3$ and $\Delta X = 1$:

- Approximate: $\%\Delta Y \approx 100 \times 0.3 = 30\%$
- Exact: $\%\Delta Y = \exp(0.3) - 1 = 1.35 - 1 = 0.35 = 35\%$

The difference of 5 percentage points matters when β_1 is large!

34 Linear Probability Model

34.1 Regression with a Binary Dependent Variable

So far we have considered cases where Y is continuous (e.g., test scores, traffic fatality rates). Now consider cases where Y is **binary**:

Outcome	Y **Values**
Getting into college	$\{0, 1\}$
Smoking status	$\{0, 1\}$
Obesity	$\{0, 1\}$
Mortgage approval/denial	$\{0, 1\}$

Question: How do we interpret $\hat{\beta}_1$ when X_i is continuous and Y_i is binary?

34.2 The Linear Probability Model (LPM)

Definition 34.1 (Linear Probability Model)**.** The **Linear Probability Model** is:

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

where $Y_i \in \{0, 1\}$.

34.3 Interpreting the Conditional Expectation

For a binary outcome:

$$E[Y_i|X_i] = 1 \times \Pr(Y_i = 1|X_i) + 0 \times \Pr(Y_i = 0|X_i) = \Pr(Y_i = 1|X_i)$$

Therefore:

$$E[Y_i|X_i] = \Pr(Y_i = 1|X_i)$$

Key Point

In the Linear Probability Model, the conditional expectation equals the **probability** that $Y = 1$ given X.

34.4 Derivation

Starting with:

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

Taking the conditional expectation:

$$\begin{aligned} E[Y_i|X_i] &= E[\beta_0 + \beta_1 X_i + u_i|X_i] \\ &= E[\beta_0|X_i] + E[\beta_1 X_i|X_i] + E[u_i|X_i] \\ &= \beta_0 + \beta_1 X_i + 0 \end{aligned}$$

Since $E[Y_i|X_i] = \Pr(Y_i = 1|X_i)$:

$$\boxed{\Pr(Y_i = 1|X_i) = \beta_0 + \beta_1 X_i}$$

The **sample** analog:

$$\widehat{\Pr}(Y_i = 1|X_i) = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

34.5 Interpretation of Coefficients

In the LPM:

$$\hat{\beta}_1 = \frac{\Delta Y}{\Delta X} = \frac{\Delta \Pr(Y = 1|X)}{\Delta X}$$

Key Point

$\hat{\beta}_1$ represents the **change in probability** (in percentage points) that $Y = 1$ for a one-unit change in X.

Example 34.1 (Mortgage Denial — Continuous Regressor)**.** Consider the probability of mortgage denial conditional on the payment-to-income (P/I) ratio:

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

where $Y_i = 1$ if mortgage is denied, and X_i is the P/I ratio.

Example calculations:

- $\widehat{\Pr}(Y_i = 1|X_i = 0.3) = 0.12$ (12% denial rate)
- $\widehat{\Pr}(Y_i = 1|X_i = 0.5) = 0.26$ (26% denial rate)

The slope coefficient:

$$\hat{\beta}_1 = \frac{\widehat{\Pr}(Y_i = 1|0.5) - \widehat{\Pr}(Y_i = 1|0.3)}{0.5 - 0.3} = \frac{0.26 - 0.12}{0.2} = \frac{0.14}{0.2} = 0.70$$

Interpretation: The probability of denial goes up by 0.14 (or **14 percentage points**) as the P/I ratio increases by 0.2.

Using regression coefficients $\hat{\beta}_0 = -0.07991$ and $\hat{\beta}_1 = 0.60353$:

$$\widehat{\Pr}(\text{denial}|X = 0.4) = -0.07991 + 0.4 \times 0.60353 = 16.2\%$$
$$\widehat{\Pr}(\text{denial}|X = 0.3) = -0.07991 + 0.3 \times 0.60353 = 10.1\%$$

The difference is $16.2\% - 10.1\% = 6.1\%$. If P/I increases by 0.1, the probability of denial goes up by approximately **6 percentage points**.

34.6 LPM with Binary Regressor

Consider:

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

where both $Y_i \in \{0, 1\}$ and $D_i \in \{0, 1\}$.

The conditional expectation:

$$E[Y_i|D_i] = \Pr(Y_i = 1|D_i) = \beta_0 + \beta_1 D_i$$

For $D_i = 0$:

$$\Pr(Y_i = 1|D_i = 0) = \beta_0 \quad \text{(probability when } D = 0)$$

For $D_i = 1$:

$$\Pr(Y_i = 1|D_i = 1) = \beta_0 + \beta_1 \quad \text{(probability when } D = 1)$$

Example 34.2 (Mortgage Denial by Race)**.** Let $Y_i = 1$ if mortgage denied, $D_i = 1$ if applicant is Black.

Results:

- $\widehat{\Pr}(Y_i = 1|D_i = 0) = \hat{\beta}_0 = 9.3\%$ (denial rate for non-Black applicants)
- $\widehat{\Pr}(Y_i = 1|D_i = 1) = \hat{\beta}_0 + \hat{\beta}_1 = 28.4\%$ (denial rate for Black applicants)

Therefore: $\hat{\beta}_1 = 28.4\% - 9.3\% = 19.1\%$

Interpretation: $\hat{\beta}_1 = 0.191$ means being Black is associated with a **19.1 percentage point** higher probability of mortgage denial.

With controls (3rd model): When holding P/I ratio constant, being Black is associated with a **17.7 percentage point** decline in approval probability.

34.7 Issues with the Linear Probability Model

Note

[Important Limitation] The LPM has a fundamental problem: predicted probabilities can fall **outside the** $[0, 1]$ **interval**.

Since $\Pr(Y = 1|X) = \beta_0 + \beta_1 X$ is a linear function of X:

- If $\beta_1 > 0$, for sufficiently large X: $\Pr(Y = 1|X) > 1$
- If $\beta_1 > 0$, for sufficiently small X: $\Pr(Y = 1|X) < 0$

This is **not realistic** since probabilities must be between 0 and 1.

Conceptually:

- The LPM assumes the effect of X on $\Pr(Y = 1|X)$ is constant (slope $= \beta_1$)
- In reality, the marginal effect should diminish as we approach probability bounds
- More sophisticated models (Probit, Logit) constrain predictions to $[0, 1]$

Key Point

Despite this limitation, the LPM remains useful for:

- **Estimation**: OLS provides consistent estimates
- **Interpretation**: Coefficients have straightforward interpretation
- **Inference**: Standard hypothesis tests and confidence intervals apply

The predicted probabilities may be problematic, but the estimated coefficients and their interpretations remain valid.