Econometrics (ECO 350)

Lecture Notes

Feda Mohammadi

Professor: Dr. Caryn Vazzana

Berea College

Fall 2025

These notes are based on *Introductory Econometrics: A Modern Approach, 6th ed.* by Jeffrey M. Wooldridge (Michigan State University).

Contents

Chapter 1: The Nature of Econometrics and Economic Data

1.1 What is Econometrics?

Econometrics is based on the development of statistical methods for estimating economic relationships, testing economic theories, and evaluating and implementing government and business policy.

1.1.1 Steps in Empirical Economic Analysis

1. **Formulate the question of interest.** This might involve testing an aspect of an economic theory or evaluating a policy.
2. **Find a suitable economic model.** This model will specify the relationship between variables (e.g., a model of wage determination).
3. **Turn the economic model into an econometric model.** The econometric model specifies the functional form and the nature of the unobserved error term. For example:

$$\text{wage} = \beta_0 + \beta_1 \text{educ} + \beta_2 \text{exper} + u$$

4. **Collect data and use it to estimate the model's parameters.** This step involves using statistical software to obtain estimates like $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$.
5. **Perform statistical inference.** Use the estimates to test hypotheses about the economic theory and draw conclusions.

1.2 Types of Data in Economics

Econometric analysis can be based on several different data structures. Understanding the type of data helps determine the appropriate model, interpretation, and estimation method.

1.2.1 Cross-Sectional Data

Definition: Observations on many individuals, households, firms, or regions at a single point in time.

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

Example: A 2023 household survey recording income, education, and employment status for 1,000 individuals. **Intuition:** We compare differences across entities at one moment in time. Each observation represents a distinct individual or unit.

1.2.2 Time-Series Data

Definition: Observations on a single entity collected over multiple time periods.

$$y_t = \beta_0 + \beta_1 x_t + u_t$$

Example: Quarterly U.S. GDP from 1990 to 2024. **Intuition:** We follow one unit over time to examine trends, cycles, and dynamic effects of policy or shocks.

1.2.3 Pooled Cross-Section Data

Definition: Two or more cross-sectional datasets from different time periods or groups, combined (pooled) into one dataset. The individuals in each period are not necessarily the same.

$$y_{it} = \beta_0 + \beta_1 x_{it} + \beta_2 D_t + u_{it}$$

Example: Combining income surveys from 2010 and 2020 to study how the effect of education on earnings has changed. **Intuition:** We "stack" separate snapshots to observe how relationships evolve across time or populations.

1.2.4 Panel (Longitudinal) Data

Definition: Observations on multiple entities over several time periods, combining cross-sectional and time-series dimensions.

$$y_{it} = \beta_0 + \beta_1 x_{it} + a_i + u_{it}$$

where a_i represents unobserved, time-invariant characteristics of each entity. **Example:** Annual financial data on 500 firms from 2010 to 2024. **Intuition:** Panel data allow us to control for unobserved heterogeneity (e.g., ability, culture) by observing the same entities over time. This strengthens causal interpretation compared to simple cross-sectional analysis.

Summary Table

Type of Data	Units Observed	Time Dimension	Example	Key Use
Cross Section	Many entities, one time	No	2023 worker survey	Compare individuals at one time
Time Series	One entity over time	Yes	U.S. GDP 1990–2024	Study trends or cycles
Pooled Cross Section	Different entities, several times	Yes (different groups)	Combine 2010+2020 income surveys	Examine change across periods
Panel Data	Many entities over time	Yes	500 firms 2010–2024	Control for unobserved effects

Chapter 2: The Simple Regression Model

2.1 Model Setup

$$y_i = \beta_0 + \beta_1 x_i + u_i, \qquad i = 1, \ldots, n.$$

Here y_i is the outcome, x_i the regressor, and u_i the error term.

2.1.1 Fitted values and residuals

Given OLS estimates $\hat{\beta}_0, \hat{\beta}_1$,

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \qquad \hat{u}_i = y_i - \hat{y}_i.$$

If $\hat{u}_i > 0$, then $y_i > \hat{y}_i$: the actual value is higher than predicted (the regression *under-predicts*).

If $\hat{u}_i < 0$, then $y_i < \hat{y}_i$: the actual value is lower than predicted (the regression *over-predicts*).

2.1.2 Slope Intuition (Where the Formula Comes From)

OLS chooses the slope and intercept to minimize the sum of squared errors:

$$\min_{\beta_0, \beta_1} \sum_i (y_i - (\beta_0 + \beta_1 x_i))^2.$$

Solving this yields:

$$\hat{\beta}_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}, \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

Breaking down the slope formula.

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}.$$

- $(x_i - \bar{x})$ means: how far this x_i is from the average x.
- $(y_i - \bar{y})$ means: how far this y_i is from the average y.
- Multiplying them, $(x_i - \bar{x})(y_i - \bar{y})$, checks whether x and y move together:
  - Positive if both are above or both below their means.
  - Negative if one is above while the other is below.

  Summing across i gives the covariance — the overall "joint movement" of x and y.
- $(x_i - \bar{x})^2$ is always nonnegative and measures the spread of x around its mean.

- If x values are close together, the sum is small.
- If x values are very spread out, the sum is large.

Intuition: The slope $\hat{\beta}_1$ is the average "rise in y" per unit "run in x." It is computed as

$$\hat{\beta}_1 = \frac{\text{How } x \text{ and } y \text{ move together (covariance)}}{\text{How spread out } x \text{ is (variance)}}.$$

Breaking down the intercept formula.

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

- The intercept is determined after we compute the slope $\hat{\beta}_1$.
- It guarantees that the regression line passes through the averages: $(\bar{x}, \bar{y})$.
- This means that when $x = \bar{x}$, the predicted value is $\hat{y} = \bar{y}$.

Intuition: $\hat{\beta}_0$ is the starting level of the regression line. It adjusts the line so that the average predicted value equals the average actual value.

2.1.3 Algebraic OLS Properties (SLR, explained in plain words)

When we estimate a regression line with an intercept, three key things always happen:

1. Residuals sum to zero.

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

Intuition: A residual is the gap between the actual value and the predicted value. Some gaps are positive (line predicts too low), others are negative (line predicts too high). OLS chooses the line so that the total "upward misses" exactly cancel the total "downward misses." *Analogy:* Like a seesaw balancing — the line sits in the middle so errors on one side offset errors on the other.

2. Residuals have no relation with x.

$$\sum_{i=1}^{n} x_i \hat{u}_i = 0 \quad \Rightarrow \quad \text{Cov}(x, \hat{u}) = 0.$$

Intuition: If residuals were systematically higher when x is bigger (or smaller), we could tilt the line differently to capture that. But OLS already tilts the line in the best way, so the leftover errors have no pattern with x. *Analogy:* After fitting, the scatter of residuals against x looks like random "noise," not a hidden slope.

3. The line passes through the averages.

$(\bar{x}, \bar{y})$ is always on the fitted regression line.

Intuition: The regression line always goes through the point of averages. This guarantees that the mean of the fitted values equals the mean of the actual values: $\bar{y} = \overline{\hat{y}}$.

2.1.4 Decomposition of Variation: SST, SSE, SSR

We can split the total variation in y into two parts:

$$\text{SST} = \text{SSE} + \text{SSR}.$$

- SST = Total variation in y (how much the y's bounce around overall).
- SSE = Explained variation (how much of the bounce is captured by the fitted line).
- SSR = Unexplained variation (leftover scatter around the line).

Intuition: Think of explaining test score differences.

- Part of the variation is explained by study hours (students who study more tend to score higher).
- The rest is unexplained scatter from sleep, stress, luck, etc.

So:

$$\text{Total variation in } y = \text{Explained by regression} + \text{Unexplained (errors)}.$$

2.1.5 Goodness-of-fit: R^2

Assuming an intercept is included,

$$R^2 \equiv \frac{\text{SSE}}{\text{SST}} = 1 - \frac{\text{SSR}}{\text{SST}}.$$

Interpretation: Fraction of the sample variation in y explained by the fitted line $\hat{y}$. It lies in $[0, 1]$. A low R^2 can still accompany a meaningful slope $\hat{\beta}_1$ (ceteris paribus effect) in cross-sectional data.

2.1.6 Regression Through the Origin (No Intercept)

Sometimes a theory suggests $\beta_0 = 0$, so we fit y on x without an intercept. Be careful:

- The OLS residuals no longer have zero mean.
- The usual $R^2 = 1 - \text{SSR}/\text{SST}$ can be *negative*; using the intercept model's R^2 intuition is misleading here.

Takeaway: Include an intercept unless you have a compelling theoretical reason not to. If you must omit it, interpret R^2 with caution or use the correlation-squared definition between y and $\hat{y}$ instead. (See discussion in your text.)

2.1.7 Functional forms with logarithms (interpretation of β_1)

Model	Dependent	Independent	Interpretation of β_1
Level–level	y	x	$\Delta y = \beta_1 \, \Delta x$
Level–log	y	$\log x$	$\Delta y \approx (\beta_1/100) \, \%\Delta x$
Log–level	$\log y$	x	$\%\Delta y \approx (100\beta_1) \, \Delta x$
Log–log	$\log y$	$\log x$	$\%\Delta y \approx \beta_1 \, \%\Delta x$ (elasticity)

Notes: The "$\approx$" uses small-change (differential) approximations; for larger changes, use exact log differences.

2.1.8 Standard Error of Regression (SER)

Typical vertical spread of points around the fitted line:

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

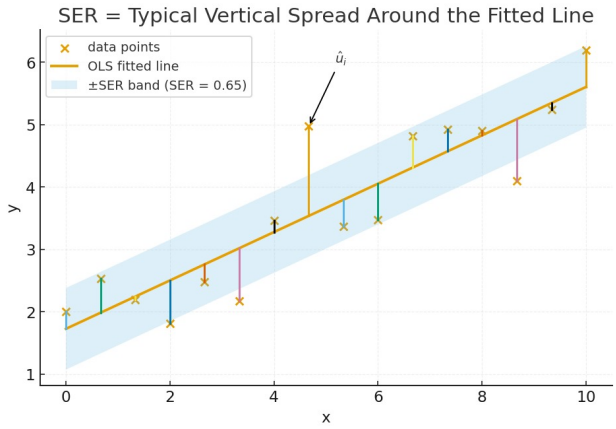

Figure 2.1: Standard Error of Regression (SER): typical vertical spread of points around the fitted line.

2.1.9 Units of Measurement (Rescaling Invariance)

Changing units rescales coefficients in predictable ways but leaves substantive conclusions intact.

- If y is multiplied by c, then both $\hat{\beta}_0$ and $\hat{\beta}_1$ are multiplied by c.
- If x is divided by c, then $\hat{\beta}_1$ is multiplied by c (intercept unchanged).
- Standard errors rescale with the coefficients so that t-statistics, F-tests, and p-values are *unchanged*.

Log models: When y or x is in logs, unit changes only shift the intercept; slopes (elasticities) are invariant.

2.2 The Gauss–Markov Assumptions for Simple Linear Regression

To study the statistical properties of the OLS estimators $\hat{\beta}_0$ and $\hat{\beta}_1$, we make the following assumptions about the population model:

$$y_i = \beta_0 + \beta_1 x_i + u_i, \qquad i = 1, 2, \ldots, n.$$

SLR.1 (Linearity in Parameters)

The model is linear in the coefficients:

$$y_i = \beta_0 + \beta_1 x_i + u_i.$$

This means β_0 and β_1 enter the equation linearly (not squared or multiplied together). The model can still include nonlinear functions of x (like $\ln x$ or x^2), as long as it remains linear in the parameters.

SLR.2 (Random Sampling)
We have a random sample of size n, $\{(x_i, y_i) : i = 1, \ldots, n\}$, from the population. This ensures each observation follows the same population model and is independent of others.

SLR.3 (Sample Variation in x)
The explanatory variable varies in the sample:

$$x_1, x_2, \ldots, x_n \text{ are not all the same.}$$

Without variation in x, we cannot estimate a slope — the regression line would be undefined.

SLR.4 (Zero Conditional Mean)
The expected value of the error, given x, is zero:

$$E(u_i \mid x_i) = 0.$$

This means x and u are unrelated — there are no omitted factors correlated with x. It is the key condition for $\hat{\beta}_0$ and $\hat{\beta}_1$ to be *unbiased*.

Unbiasedness of OLS

Under assumptions SLR.1 through SLR.4, the OLS estimators are unbiased. This means that if we could repeatedly draw samples from the population and compute the estimates each time, the average of these estimates would be equal to the true population parameters.

$$E(\hat{\beta}_0) = \beta_0 \quad \text{and} \quad E(\hat{\beta}_1) = \beta_1$$

Unbiasedness is a property of the sampling distributions of the estimators, not a statement about a single estimate from one sample.

Assumption SLR.5 (Homoskedasticity)

The error term u has the same variance for all values of the explanatory variable x:

$$\text{Var}(u|x) = \sigma^2.$$

Meaning. This assumption states that the spread (variance) of the unobserved factors affecting y is the same for every value of x. In other words, the randomness or "noise" in y does not systematically increase or decrease as x changes. The errors are evenly scattered around the regression line, regardless of the level of x.

Relation to SLR.4. Homoskedasticity is different from the zero conditional mean assumption.

- **SLR.4 (Zero Conditional Mean):** $E(u|x) = 0$ deals with the *average value* of u.
- **SLR.5 (Homoskedasticity):** $\text{Var}(u|x) = \sigma^2$ deals with the *spread or variance* of u.

Why it matters.

- The homoskedasticity assumption is *not* needed for unbiasedness of OLS. Even if the variance of u changes with x (heteroskedasticity), the OLS estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ remain unbiased as long as $E(u|x) = 0$ holds.
- We add SLR.5 because it simplifies the mathematical formulas for the variances of $\hat{\beta}_0$ and $\hat{\beta}_1$, and because it gives OLS the desirable property of being the **Best Linear Unbiased Estimator (BLUE)**—that is, the most efficient among all linear unbiased estimators.

Interpretation. If $\text{Var}(u|x) = \sigma^2$, then the conditional variance of u does not depend on x:

$$E(u^2|x) = \sigma^2.$$

Because $E(u) = 0$, this σ^2 is also the unconditional variance of u:

$$\sigma^2 = E(u^2) = \text{Var}(u).$$

The standard deviation σ represents the typical size of the random error. A larger σ means the unobservables affecting y are more spread out around the regression line.

If violated (heteroskedasticity). When $\text{Var}(u|x)$ is not constant, OLS is still unbiased but no longer efficient, and the usual formulas for standard errors become invalid. In that case, we can use **heteroskedasticity-robust** standard errors to fix the problem.

Table 2.1: Population Error u_i vs. OLS Residual $\hat{u}_i$

Feature	**Population Error (u_i)**	**OLS Residual ($\hat{u}_i$)**
Definition	$u_i = y_i - (\beta_0 + \beta_1 x_i)$	$\hat{u}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$
Known?	Unobservable (true model)	Observable (sample estimate)
Used in	Assumption $E(u_i\|x_i) = 0$	OLS properties $\sum \hat{u}_i = 0$, $\text{Cov}(x, \hat{u}) = 0$
Average value	$E(u_i) = 0$ (theory)	$\bar{\hat{u}} = 0$ (by construction)
Purpose	Represents unobserved influences on y_i	Measures actual deviation from fitted line

Note on R^2. In cross sections, a low R^2 is common and does not imply the slope is meaningless. You can have a precisely estimated ceteris paribus effect even when overall fit is modest. Avoid over-weighting R^2 when judging a model.

Chapter 3: Multiple Regression Analysis: Estimation

3.1 Model Setup and Interpretation

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + u_i$$

3.1.1 Partial Effect Interpretation

β_j is the predicted change in y for a one-unit increase in x_j, *holding all other regressors fixed*. Multiple regression mimics the all-else-equal comparison statistically.

3.1.2 Partialling Out (Frisch–Waugh–Lovell)

With $k = 2$ regressors, the OLS slope on x_1 can be computed in two steps:

1. Regress x_1 on x_2 and keep residuals $\hat{r}_1$ (the part of x_1 orthogonal to x_2).
2. Regress y on $\hat{r}_1$; the slope equals $\hat{\beta}_1$ from the full regression of y on x_1, x_2.

This shows $\hat{\beta}_1$ measures the effect of x_1 on y *holding* x_2 *fixed* by literally stripping x_2 out of x_1. In general with many regressors, residualize x_j on the others and regress y on that residual to recover $\hat{\beta}_j$.

3.2 Gauss–Markov Assumptions for Multiple Regression

The population model is written as:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + u_i, \qquad i = 1, 2, \ldots, n.$$

These assumptions guarantee that OLS provides unbiased and efficient estimators under standard conditions.

Assumption MLR.1: Linearity in Parameters The model is linear in the coefficients:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$$

Intuition: The parameters β_j appear linearly (not squared or inside logarithms). OLS relies on this algebraic linearity to compute estimates. *Example:* $y = \beta_0 + \beta_1 x^2 + u$ is fine, but $y = e^{\beta_1 x} + u$ is not.

Assumption MLR.2: Random Sampling We have a random sample of observations:

$$\{(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i)\}_{i=1}^{n}$$

Intuition: Each observation gives independent information about the population relationship. Random sampling ensures our sample is representative and avoids selection bias.

Assumption MLR.3: No Perfect Collinearity None of the independent variables is an exact linear combination of the others.

$$\text{No } x_j = a_1 x_1 + a_2 x_2 + \cdots + a_k x_k.$$

Intuition: Each variable must provide unique information. If two regressors move together perfectly (like "income in dollars" and "income in thousands"), OLS cannot separate their effects.

Assumption MLR.4: Zero Conditional Mean The expected value of the error, given the explanatory variables, is zero:

$$E(u_i \mid x_{i1}, x_{i2}, \ldots, x_{ik}) = 0$$

Intuition: After accounting for all x's, nothing systematic remains in u. The unobserved factors are uncorrelated with the regressors. This is the key assumption that makes OLS **unbiased**.

Assumption MLR.5: Homoskedasticity The variance of the error term is constant across all observations:

$$\text{Var}(u_i \mid x_{i1}, x_{i2}, \ldots, x_{ik}) = \sigma^2$$

Intuition: The "noise" in y has the same average spread for every level of the x's. If the spread changes (larger for some groups), we have heteroskedasticity. Homoskedasticity ensures OLS is **efficient** (BLUE) and that standard errors and t-tests are valid.
Note: This assumption concerns the *variance of the error term*, not the coefficients themselves. However, it directly affects how we compute $\text{Var}(\hat{\beta}_j)$:

$$\text{Var}(\hat{\beta}_j) = \frac{\sigma^2}{SST_{x_j}(1 - R_j^2)}.$$

If homoskedasticity fails, OLS remains unbiased but these variance formulas no longer hold.

Summary:

Assumption	Why It Matters
MLR.1 Linearity	Ensures OLS can estimate parameters algebraically.
MLR.2 Random Sampling	Prevents sample-selection bias.
MLR.3 No Perfect Collinearity	Each variable adds unique information to the model.
MLR.4 Zero Conditional Mean	Guarantees OLS is unbiased.
MLR.5 Homoskedasticity	Ensures OLS is efficient and standard errors are correct.

Matrix remark (why residuals sum to zero). With an intercept, the normal equations imply $X^\top \hat{u} = 0$, so residuals sum to zero and are orthogonal to each regressor. This also underlies SST = SSE + SSR.

3.3 Understanding the Variance of OLS Coefficients

Notation and Subscripts (read this first!)

- x_{ij} = the actual value of regressor j for observation i.
- X_j = the entire variable (the full column of regressor j).
- $\bar{x}_j$ = the mean of regressor j across all observations.
- Y_i = actual outcome for observation i.
- $\hat{Y}_i$ = predicted outcome for observation i.
- $\sigma^2 = Var(u_i)$ = variance of the error term (baseline noise in Y).

—

3.3.1 The Key Formula

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2 \cdot (1 - R_j^2)}.$$

This formula tells us: the uncertainty in $\hat{\beta}_j$ (slope on X_j) equals the baseline noise in Y, divided by the amount of independent variation in X_j.

—

3.3.2 Error term and σ^2

The variance of the errors is denoted $\sigma^2 = Var(u_i)$.
Intuition: σ^2 measures how much Y bounces around even after taking the X's into account. This is often called the *baseline noise in Y* because it represents the starting level of unpredictability in the outcome.

Example: Suppose we predict test scores (Y) from study hours (X). If two students both study 10 hours and score 80 and 82, the errors are small and σ^2 is low. But if two students both study 10 hours and score 60 and 95, the errors are large and σ^2 is high. This baseline noise directly raises the variance of slope estimates.

3.3.3 $\sum(x_{ij} - \bar{x}_j)^2$: Raw variation in X_j

Subtracting the mean centers the regressor; squaring and summing gives its total spread. If this sum is large, X_j has a wide range. If it is small, everyone has nearly the same X_j, and the regression has little leverage to estimate its effect.

—

3.3.4 R_j^2: Multicollinearity (uniqueness of X_j)

R_j^2 is obtained by regressing X_j on all the other regressors. It measures how much of X_j's variation can be predicted from them.

- If $R_j^2 = 0.95$, then 95% of X_j's variation is redundant, leaving almost no unique variation. This makes $\hat{\beta}_j$ unstable.
- If R_j^2 is small, most of X_j is independent of the others, and the estimate of $\hat{\beta}_j$ is more precise.

—

3.3.5 Big Picture

$$Var(\hat{\beta}_j) = \frac{\text{Noise in Y } (\sigma^2)}{\text{Raw spread of } X_j \times \text{Uniqueness of } X_j \ (1 - R_j^2)}.$$

Interpretation: Variance of $\hat{\beta}_j$ is bigger if Y is noisy, if X_j has little spread, or if X_j is highly correlated with the other regressors.

Mini Cheat Sheet (At-a-Glance)

Three levers that control precision of $\hat{\beta}_j$:

- **Noise (σ^2):** Higher noise $\Rightarrow$ larger variance.
- **Spread in X_j ($\sum(x_{ij} - \bar{x}_j)^2$):** More spread $\Rightarrow$ smaller variance.
- **Uniqueness ($1 - R_j^2$):** Less collinearity $\Rightarrow$ smaller variance.

$$\boxed{Var(\hat{\beta}_j) \ = \ \frac{\text{Noise}}{\text{Spread} \times \text{Uniqueness}}}$$

Quick Intuition Diagram (Simple +/– Version)

3.4 Model Specification Issues

3.4.1 Including Irrelevant Variables (Overspecifying the Model)

Suppose the true population model is $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$, but we estimate $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$, where the true $\beta_3 = 0$.

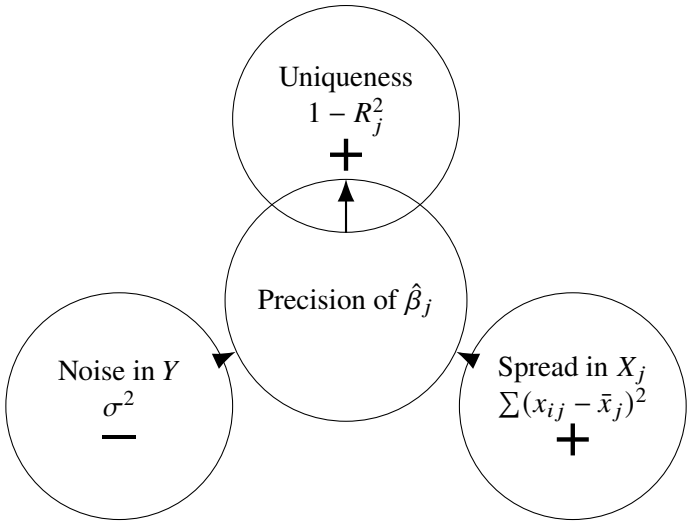

Figure 3.1: Precision of $\hat{\beta}_j$ decreases with noise (–) and increases with spread and uniqueness (+).

Effect on OLS Estimates.

- **Unbiasedness:** OLS estimates remain unbiased ($E[\hat{\beta}_j] = \beta_j$).
- **Efficiency:** The estimates become less precise. Adding irrelevant variables increases the variance of the other coefficients.

Summary. Including irrelevant variables does not bias coefficients but makes them noisier: $\text{Var}(\hat{\beta}_j) \uparrow$.

3.4.2 Omitting a Relevant Variable (Underspecifying the Model)

Now suppose the true model is $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$, but we estimate $y = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 + v$.

Effect on OLS Estimates.

- **Bias:** The estimator $\tilde{\beta}_1$ is generally biased.
- **Formula for Bias:** $\text{Bias}(\tilde{\beta}_1) = \beta_2 \cdot \delta_1$, where δ_1 is the slope from regressing x_2 on x_1.
- **Direction of Bias:**

β_2	$\text{Corr}(x_1, x_2)$	Direction of Bias in $\tilde{\beta}_1$
+	+	Upward (too large)
+	−	Downward (too small)
−	+	Downward (too small)
−	−	Upward (too large)

Intuition. Omitting a relevant variable is harmful because the included regressors "pick up" part of the missing effect.

Comparison Summary:

Case	Bias	Efficiency	Interpretation
Include Irrelevant Variable	Unbiased	Less efficient	Coefficients are correct on average but noisier.
Omit Relevant Variable	Biased	(Biased)	Coefficients become misleading.

Chapter 4: Multiple Regression Analysis: Inference

Inference involves using sample estimates to test hypotheses about population parameters. To do this, we need one more assumption to get the exact sampling distributions.

> Assumption MLR.6: Normality: The population error u is independent of the explanatory variables $(x_1, \ldots, x_k)$ and is normally distributed with zero mean and variance σ^2: $u \sim \text{Normal}(0, \sigma^2)$.

Under MLR.1-MLR.6 (the Classical Linear Model assumptions), the OLS estimators are normally distributed. When we standardize them, we get the t distribution.

4.1 t-Distribution for the Standardized OLS Estimator

Theorem (t-Distribution of the Estimator)

Under the Classical Linear Model assumptions (MLR.1 through MLR.6),

$$\frac{\hat{\beta}_j - \beta_j}{\text{se}(\hat{\beta}_j)} \sim t_{n-k-1},$$

where:

- $k + 1$ = number of estimated parameters (including the intercept),
- $n - k - 1$ = degrees of freedom.

4.1.1 Intuition Behind the t-Distribution for $\hat{\beta}_j$

Step 1. The Ideal Case: σ^2 Known. If we *knew* the true error variance σ^2, then the OLS estimator $\hat{\beta}_j$ would have an exact normal sampling distribution:

$$\frac{\hat{\beta}_j - \beta_j}{\text{sd}(\hat{\beta}_j)} \sim N(0, 1).$$

This is like a z-score: it measures how many true standard deviations $\hat{\beta}_j$ is away from the true β_j.

Step 2. The Problem: σ^2 is Unknown. In reality, σ^2 is unknown. We must estimate it using the residuals:

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

Because $\hat{\sigma}$ is computed from the sample, it is a random variable. When we substitute $\hat{\sigma}$ for σ to get the standard error, the standardized estimator becomes

$$\frac{\hat{\beta}_j - \beta_j}{\text{se}(\hat{\beta}_j)} = \frac{\text{Normal}(0,1)}{\sqrt{\chi^2_{n-k-1}/(n-k-1)}} \sim t_{n-k-1}$$

This is the definition of a t variable: a ratio of a standard normal and the square root of an independent chi-squared variable divided by its degrees of freedom.

Step 3. The Meaning: Extra Uncertainty.

- The **numerator** tells us how far $\hat{\beta}_j$ is from the true β_j.
- The **denominator** adds extra uncertainty because σ^2 is estimated, not known.

This extra randomness makes the t distribution have *fatter tails* than the standard normal. It reflects the idea that:

> "We are less certain because the noise level is estimated rather than known."

As the sample size n increases and $\hat{\sigma} \to \sigma$, the t distribution approaches the standard normal.

Step 4. Summary Table:

Case	What You Divide By	Distribution	Why
σ^2 known	True $\text{sd}(\hat{\beta}_j)$	$N(0,1)$	No extra uncertainty
σ^2 unknown	Estimated $\text{se}(\hat{\beta}_j)$	t_{n-k-1}	Adds variability

4.2 Hypothesis Testing: The t-Test

The most common hypothesis is that a population parameter is zero, meaning a variable has no partial effect on the outcome.

4.2.1 The Five Steps of a t-test

1. **State the null (H_0) and alternative (H_1) hypotheses.**
   - Two-sided: $H_0 : \beta_j = 0$ vs. $H_1 : \beta_j \neq 0$
   - One-sided (right): $H_0 : \beta_j = 0$ vs. $H_1 : \beta_j > 0$
   - One-sided (left): $H_0 : \beta_j = 0$ vs. $H_1 : \beta_j < 0$
2. **Choose a significance level (α).** This is the probability of a Type I error (rejecting H_0 when it's true). Common levels are 5% ($\alpha = 0.05$) and 1% ($\alpha = 0.01$).
3. **Compute the t statistic (or t ratio):**

$$t = \frac{\text{estimate} - \text{hypothesized value}}{\text{standard error}} = \frac{\hat{\beta}_j - 0}{\text{se}(\hat{\beta}_j)}$$

4. **Find the critical value or the p-value.**

5. **State your conclusion.** "We reject the null hypothesis at the 5% significance level" or "We fail to reject the null hypothesis."

4.2.2 Interpreting One-Sided vs. Two-Sided Tests

When testing hypotheses, the direction of the *alternative hypothesis* determines which side of the t distribution contains the rejection region.

Example: School Enrollment From the regression

$$\widehat{\text{math10}} = 2.274 + 0.00046\,\text{totcomp} + 0.048\,\text{staff} - 0.00020\,\text{enroll},$$

we want to test whether larger school enrollment lowers performance:

$$H_0 : \beta_{\text{enroll}} = 0 \quad \text{versus} \quad H_1 : \beta_{\text{enroll}} < 0.$$

The estimated coefficient and standard error are:

$$\hat{\beta}_{\text{enroll}} = -0.00020, \quad \text{se}(\hat{\beta}_{\text{enroll}}) = 0.00022,$$

so the t statistic is

$$t_{\text{enroll}} = \frac{-0.00020}{0.00022} \approx -0.91.$$

Because the alternative hypothesis is $H_1 : \beta_{\text{enroll}} < 0$, this is a **left-tailed test**. The rejection region lies entirely in the left tail. For $\alpha = 0.05$, we reject H_0 if $t < -t_{0.05}$ (e.g., $t < -1.65$). Here, the critical value is negative because we are only looking for *sufficiently large negative evidence* against H_0.

We obtained $t_{\text{enroll}} = -0.91$, which is not less than -1.65. Therefore, we **fail to reject** H_0.

Summary Table for Test Types:

Alternative Hypothesis	Tail	Rejection Rule	Example of Rejection
$H_1 : \beta_j > 0$	Right-tailed	$t > +t_{0.05}$	$t = 2.3 > 1.65$
$H_1 : \beta_j < 0$	Left-tailed	$t < -t_{0.05}$	$t = -2.3 < -1.65$
$H_1 : \beta_j \neq 0$	Two-tailed	$\lvert t\rvert > t_{0.025}$	$\lvert t\rvert = 2.3 > 1.96$

4.2.3 Computing and Interpreting p-Values

Instead of picking α first, the **p-value** asks:

"If the null hypothesis were true, what is the probability of observing a t as extreme as the one we obtained?"

Decision Rule:

$$\text{Reject } H_0 \text{ if } p\text{-value} < \alpha; \quad \text{otherwise, fail to reject.}$$

Example: Two-Sided Test. Suppose $t = 1.85$ with 40 degrees of freedom. For a two-sided alternative ($H_1 : \beta_j \neq 0$), the p-value is

$$p = P(|T| > 1.85) = 2P(T > 1.85) \approx 0.0718$$

Since $0.0718 > 0.05$, we fail to reject H_0 at the 5% level.

One-Sided p-Values.

- If $H_1 : \beta_j > 0$, then $p = P(T > t_{\text{obs}})$
- If $H_1 : \beta_j < 0$, then $p = P(T < t_{\text{obs}})$

Interpreting p-Values:

p-Value Range	Evidence Against H_0	Decision (at $\alpha = 0.05$)
$p < 0.01$	Very strong	Reject H_0
$0.01 \leq p < 0.05$	Moderate	Reject H_0
$0.05 \leq p < 0.10$	Weak	Marginal (Fail to reject)
$p \geq 0.10$	None	Fail to reject H_0

4.3 Confidence Intervals for Regression Coefficients

A confidence interval (CI) provides a range of plausible values for the population parameter β_j.

4.3.1 Formula and Interpretation

A $(1 - \alpha)100\%$ confidence interval for β_j is:

$$\boxed{\hat{\beta}_j \pm c \times \text{se}(\hat{\beta}_j)},$$

where c is the critical value from the t_{n-k-1} distribution for a two-tailed test.

Interpretation: If we repeatedly drew random samples, then 95% of the intervals we compute would contain the true, unknown value of β_j.

4.3.2 Example: Wage Regression

Consider the wage regression:

$$\widehat{\log(\text{wage})} = 0.284 + 0.092\,\text{educ} + 0.0041\,\text{exper} + 0.022\,\text{tenure}.$$

The standard error for *exper* is $\text{se}(\hat{\beta}_{\text{exper}}) = 0.0017$, and $df = 522$. For a 95% confidence level, $c \approx 1.96$.

$$\text{CI}_{95\%}(\beta_{\text{exper}}) = 0.0041 \pm 1.96(0.0017) = [0.0008,\ 0.0074].$$

Interpretation: We are 95% confident that the true effect of an additional year of experience on log(wage) is between 0.0008 and 0.0074. In percentage terms, this is an increase of 0.08% to 0.74%.

4.3.3 Relation to Hypothesis Testing

A 95% CI provides the same information as a 5% two-tailed t-test:

- If 0 is **outside** the CI (as in the example above), we **reject** $H_0 : \beta_j = 0$.
- If 0 is **inside** the CI, we **fail to reject** $H_0 : \beta_j = 0$.

4.4 Testing Multiple Linear Restrictions: The F-Test

4.4.1 Purpose and Intuition

The t-statistic tests whether a *single* coefficient is zero. The **F-test** checks whether a *group* of variables has no joint effect.

For example, in a salary model:

$$\log(\text{salary}) = \beta_0 + \beta_1 \text{years} + \beta_2 \text{gamesyr} + \beta_3 \text{bavg} + \beta_4 \text{hrunsyr} + \beta_5 \text{rbisyr} + u,$$

We might want to test if the three performance variables have no joint impact. This is a **null hypothesis of exclusion restrictions**:

$$H_0 : \beta_3 = \beta_4 = \beta_5 = 0,$$

versus

$$H_a : \text{At least one of } \beta_3, \beta_4, \beta_5 \text{ is nonzero.}$$

4.4.2 Restricted vs. Unrestricted Models

To test this, we compare two regressions:

- The **Unrestricted Model (UR)** includes all variables.
- The **Restricted Model (R)** omits the variables being tested (imposes H_0).

Because OLS minimizes the Sum of Squared Residuals (SSR), adding variables can *never* make the fit worse. Therefore:

$$SSR_{\text{Restricted}} \geq SSR_{\text{Unrestricted}}.$$

If removing the variables makes the model fit *much* worse (a large increase in SSR), this suggests H_0 is false and the variables do matter.

4.4.3 Computing the F-Statistic

The test statistic is:

$$\boxed{F = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)}}$$

where

- q = number of restrictions (number of variables dropped).
- n = sample size.
- $k + 1$ = number of parameters in the **unrestricted** model.

- $n - k - 1$ = degrees of freedom in the **unrestricted** model.

Under H_0, this statistic follows an F distribution with $(q, n - k - 1)$ degrees of freedom. We reject H_0 if F is larger than the critical value.

4.4.4 Example: MLB Salary Regression

From the data:

$$SSR_{ur} = 183.186 \quad (k + 1 = 6)$$
$$SSR_r = 198.311$$
$$n = 353, \quad q = 3$$

Then:

$$F = \frac{(198.311 - 183.186)/3}{183.186/(353 - 6)} = \frac{15.125/3}{183.186/347} \approx 9.55.$$

The 1% critical value for an $F(3, 347)$ distribution is 3.78. Since $9.55 > 3.78$, we **strongly reject** H_0.

Interpretation: The joint effect of batting average, home runs per year, and RBIs per year on salary is statistically significant. Even if their individual t-statistics are small (due to multicollinearity), as a group they have strong explanatory power.

4.4.5 The R^2 Form of the F-Statistic

For exclusion restrictions, an equivalent formula using R-squared is:

$$\boxed{F = \frac{(R^2_{ur} - R^2_r)/q}{(1 - R^2_{ur})/(n - k - 1)}}$$

Note: The unrestricted R^2 is in the denominator.

4.4.6 The F-Statistic for Overall Significance

To test whether *all* slope coefficients are zero ($H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$), we can use a simplified F-statistic that depends only on the unrestricted R^2:

$$\boxed{F = \frac{R^2/k}{(1 - R^2)/(n - k - 1)}}$$

This is a standard part of all regression output. A significant F-stat means the model has *some* explanatory power.

4.4.7 Testing General Linear Restrictions

Sometimes restrictions are not simple exclusions. For example, testing for constant returns to scale might be $H_0 : \beta_1 + \beta_2 = 1$. The F-statistic (SSR form) can handle this, but the R^2 form cannot. We must run the restricted and unrestricted models, find the SSR for each, and plug them into the main F-statistic formula.

4.4.8 Comparing the t-Test and the F-Test

The t-test and F-test are closely related but answer different questions.

Feature	t**-Test**	F**-Test**
Purpose	Tests **one** coefficient	Tests **a group** of coefficients jointly
Null hypothesis	$H_0 : \beta_j = 0$	$H_0 : \beta_{j_1} = \beta_{j_2} = \cdots = 0$
Use Case	"Does this one variable matter?"	"Does this group of variables matter?"
Multicollinearity	High correlation can inflate standard errors, making t-stats small.	Can detect joint significance even when t-stats are small.
Connection	For a single restriction ($q = 1$), $F = t^2$	Generalizes the t-test.

Chapter 5: Multiple Regression Analysis: OLS Asymptotics

This chapter introduces the **asymptotic properties** of OLS, which describe how the estimators behave as the sample size n grows infinitely large ($n \to \infty$). These "large sample" properties are important because they do not require the strong Classical Linear Model (CLM) assumptions, particularly the Normality of errors (MLR.6).

5.1 Consistency

What are Asymptotic Properties? The Gauss-Markov assumptions (MLR.1-MLR.5) give us the "finite sample" properties of OLS, such as unbiasedness and BLUE. Unbiasedness means $E(\hat{\beta}_j) = \beta_j$ for *any* sample size n. However, this relies on the strict Zero Conditional Mean assumption (MLR.4). Asymptotic properties are a weaker, more fundamental requirement.

Definition of Consistency: An estimator $\hat{\beta}_j$ is **consistent** for β_j if it converges in probability to the true parameter as n increases. We write this as:

$$\text{plim}(\hat{\beta}_j) = \beta_j$$

Intuition: This means that as we get more and more data, the estimator $\hat{\beta}_j$ gets closer and closer to the true population value β_j. The distribution of $\hat{\beta}_j$ collapses to a single point at β_j. Consistency is a minimum requirement for any good estimator.

5.1.1 Asymptotic Assumptions for OLS

To prove consistency, we can replace the strong MLR.4 assumption with a weaker one.

> Assumption MLR.4': Zero Mean and Zero Correlation: The error u has an expected value of zero, and it is uncorrelated with each explanatory variable:
>
> $$E(u) = 0 \quad \text{and} \quad \text{Cov}(x_j, u) = 0, \quad \text{for all } j = 1, \ldots, k$$

Comparing MLR.4 and MLR.4'

- **MLR.4 (Zero Conditional Mean):** $E(u|x_1, \ldots, x_k) = 0$. This is a very strong assumption. It implies u is uncorrelated with *any function* of the x variables.
- **MLR.4' (Zero Correlation):** $\text{Cov}(x_j, u) = 0$. This is much weaker. It only requires u to be uncorrelated with the x_j themselves, not all possible functions of them.

- **Key Point:** $E(u|x) = 0$ implies $\text{Cov}(x, u) = 0$, but the reverse is not true. Therefore, MLR.4' is a weaker assumption.

Theorem 5.1 (Consistency of OLS): Under assumptions MLR.1, MLR.2, MLR.3, and the weaker MLR.4', the OLS estimator $\hat{\beta}_j$ is consistent for β_j.

5.2 Deriving the Inconsistency in OLS (Omitted Variable Bias)

Consistency fails if MLR.4' is violated, i.e., if the error term u is correlated with an explanatory variable x_j. The most common reason for this is **omitted variable bias**.

The Setup

- **True Model:** $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + v$. (Here, v is the true error).
- **Estimated Model:** $y = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 + u$. (We omit x_2).

When we estimate the short model, the error term u actually contains the omitted variable: $u = \beta_2 x_2 + v$.

When does OLS fail? OLS in the short model will be inconsistent if the regressor x_1 is correlated with the new error term u.

$$\text{Cov}(x_1, u) = \text{Cov}(x_1, \beta_2 x_2 + v) = \beta_2 \text{Cov}(x_1, x_2) + \text{Cov}(x_1, v)$$

Assuming the true error v is uncorrelated with x_1 ($\text{Cov}(x_1, v) = 0$), the inconsistency is driven by $\beta_2 \text{Cov}(x_1, x_2)$.

The Two Conditions for Inconsistency OLS is inconsistent for β_1 if:

1. The omitted variable x_2 truly belongs in the model ($\beta_2 \neq 0$).
2. The omitted variable x_2 is correlated with the included variable x_1 ($\text{Cov}(x_1, x_2) \neq 0$).

The Inconsistency Formula The OLS estimator $\tilde{\beta}_1$ converges in probability to the wrong value:

$$\text{plim}(\tilde{\beta}_1) = \beta_1 + \beta_2 \delta_1$$

where $\delta_1 = \frac{\text{Cov}(x_1, x_2)}{\text{Var}(x_1)}$ (this is the slope from a simple regression of x_2 on x_1). The term $\beta_2 \delta_1$ is the **asymptotic bias** or **inconsistency**.

5.3 Asymptotic Normality and Large Sample Inference

Consistency is good, but it doesn't let us perform t-tests or F-tests. For that, we need to know the estimator's distribution. **Asymptotic normality** states that, in large samples, the OLS estimators are approximately normally distributed.

Theorem 5.2: Asymptotic Normality of OLS Under the Gauss-Markov assumptions (MLR.1 through MLR.5), the OLS estimators are asymptotically normal. This means:

$$\frac{\hat{\beta}_j - \beta_j}{\text{se}(\hat{\beta}_j)} \xrightarrow{a} \text{Normal}(0, 1) \quad \text{as } n \to \infty$$

(The $\xrightarrow{a}$ means "approaches in distribution").

Practical Implication: Why We Can Use t-tests **This is the most important result of the chapter.** Theorem 5.2 shows that even if the error terms u are **not** normally distributed (i.e., MLR.6 fails), the t-statistic still has a distribution that is *approximately* standard normal in large samples.

- This justifies using t-tests, F-tests, and confidence intervals in large samples, regardless of the distribution of the errors.
- This is why, with $n > 120$, we often just use the 1.96 critical value for a 5% test, as the t-distribution becomes indistinguishable from the normal distribution.

5.4 Other Large Sample Tests: The Lagrange Multiplier (LM) Statistic

The F-test requires running two regressions (restricted and unrestricted). The **LM statistic** is an alternative for testing exclusion restrictions that only requires estimating the **restricted** model.

5.4.1 The LM Statistic for q Exclusion Restrictions

Suppose we want to test $H_0 : \beta_{k-q+1} = 0, \ldots, \beta_k = 0$ (i.e., that q variables are jointly zero).

Procedure (The "Residual Regression")

1. **Run the restricted model:** Regress y on the variables that are *not* being tested (i.e., omit the q variables) and save the residuals, $\tilde{u}$.
2. **Run an auxiliary regression:** Regress these residuals $\tilde{u}$ on **all** of the original independent variables (both the restricted and the q excluded ones).
3. **Get the R^2:** Save the R-squared from this auxiliary regression, call it R_u^2.
4. **Calculate the LM statistic:**
$$LM = n \cdot R_u^2$$
where n is the sample size.

Decision Rule Under the null hypothesis, the LM statistic follows a chi-squared distribution with q degrees of freedom.

$$LM \xrightarrow{a} \chi_q^2$$

We reject H_0 if LM is greater than the critical value from the χ_q^2 distribution. Software will typically report a p-value for this test.

5.5 Asymptotic Efficiency of OLS

Theorem 5.3: Asymptotic Efficiency of OLS Under the Gauss-Markov assumptions (MLR.1 through MLR.5), the OLS estimators are **asymptotically efficient**.

- **Intuition:** This means that in large samples, OLS has the smallest asymptotic variance among a large class of estimators.
- This is the large-sample equivalent of the BLUE property.

What if Homoskedasticity Fails? If MLR.5 (Homoskedasticity) fails, OLS is:

- Still unbiased (if MLR.1-4 hold).
- Still consistent (if MLR.1-4' hold).
- Still asymptotically normal.
- **BUT...** The standard errors, t-stats, and F-stats are **no longer valid**, even in large samples.
- **AND...** OLS is **no longer BLUE** and **no longer asymptotically efficient**. Other estimators (like Weighted Least Squares) are more efficient.

This is why we must test for heteroskedasticity (Chapter 8) and use robust standard errors if it is present.

Chapter 6: Multiple Regression Analysis: Further Issues

This chapter talks about how rescaling data affects OLS statistics, standardized (beta) coefficients, choosing and interpreting functional forms (logs, quadratics, interactions), computing average partial effects, model selection ideas (including adjusted R^2), and prediction issues, including when the dependent variable is in logs.

6.1 Effects of Data Scaling on OLS Statistics

Changing units (rescaling) makes output easier to read (e.g., dollars to thousands) without changing economic content. If we divide the dependent variable by a constant (e.g., ounces $\rightarrow$ pounds), *all* slope estimates and their standard errors scale by the same constant, so t-statistics, F-statistics, and hypothesis test outcomes are unchanged; R^2 is also unchanged. SSR and SER rescale mechanically with the residuals.

Worked example (birth weight). Starting from

$$\widehat{\text{bwght}} = \hat{\beta}_0 + \hat{\beta}_1 \text{ cigs} + \hat{\beta}_2 \text{ faminc},$$

measuring birth weight in pounds instead of ounces divides each coefficient and its s.e. by 16; t-ratios remain identical and R^2 is the same.

Logs are unit-invariant. If y (or some x_j) appears in logs, changing units only shifts the intercept; slopes on log variables are unchanged (elasticities and percentage effects are unit free).

6.2 Standardized (Beta) Coefficients

Sometimes scales are arbitrary (e.g., test scores). Standardize each variable to a z-score and regress z_y on $z_{x_1}, \ldots, z_{x_k}$:

$$z_y = \tilde{\beta}_1 z_{x_1} + \cdots + \tilde{\beta}_k z_{x_k} + \text{error},$$

where $\tilde{\beta}_j = \frac{\hat{\sigma}_{x_j}}{\hat{\sigma}_y} \hat{\beta}_j$. These are the *beta coefficients*. Interpretation: a one standard deviation increase in x_j changes $\hat{y}$ by $\tilde{\beta}_j$ standard deviations. With a single regressor, the beta equals the sample correlation and lies in $[-1, 1]$. Statistical significance does not change when moving to betas.

Use case. Betas put regressors on equal footing when units differ; they can also complement elasticities when a given percent change spans very different parts of variables' ranges.

6.3 More on Functional Form

Functional form determines how we interpret coefficients, how effects vary across the range of x, and whether assumptions (e.g., homoskedasticity, linearity in parameters) are reasonable. Below are the **four canonical forms** plus **quadratics** and **interactions**, each with: model, interpretation, exact vs. approximate changes, when to use, and a worked example.

6.3.1 Level–Level (Linear in levels)

Model:

$$y = \beta_0 + \beta_1 x + u.$$

Interpretation (marginal effect). A one-unit increase in x changes y by *approximately exactly* β_1 units, *holding other regressors fixed*. The effect is *constant* across all x.

Exact vs. Approx. No approximation needed: $\Delta\hat{y} = \beta_1 \, \Delta x$.

When to use.

- Units are meaningful (e.g., "one year of education adds *some dollars* to wages" may be too rigid; but "one additional tutoring hour raises score by 2 points" is reasonable).
- The data scatter suggests a linear trend with roughly constant vertical spread across x.

Worked example (test scores). Suppose $\widehat{\text{score}} = 58.0 + 1.95 \cdot \text{hours}$; then +5 hours $\Rightarrow$ +9.75 points no matter if you go from 0 to 5 hours or from 50 to 55 hours. The marginal effect does not depend on the starting level of x.

6.3.2 Log–Level (Semi-log with log y)

Model:

$$\log y = \beta_0 + \beta_1 x + u.$$

Interpretation (semi-elasticity). A one-unit increase in x changes y by *approximately* $100 \cdot \beta_1$ percent. More precisely, for a change Δx:

$$\%\Delta y \approx 100 \cdot \beta_1 \, \Delta x, \qquad \text{Exact:} \quad \%\Delta y = 100 \cdot \big[\exp(\beta_1 \Delta x) - 1\big].$$

For small Δx, the approximation is very accurate; for larger changes, report the exact percentage.

When to use.

- When the effect of x on y is proportional rather than additive (e.g., each extra year of education produces a *percentage* increase in wages).
- When the variance of y grows with its level (logs can stabilize variance).

Worked example (returns to education). If $\widehat{\log(\text{wage})} = 1.2 + 0.07 \cdot \text{educ}$, then one more year of education increases wage by about 7% (exact: $100[\exp(0.07) - 1] \approx 7.25\%$). Five extra years $\Rightarrow 100[\exp(0.35) - 1] \approx 41.9\%$ increase.

6.3.3 Level–Log (Semi-log with log x)

Model:

$$y = \beta_0 + \beta_1 \log x + u, \quad x > 0.$$

Interpretation (per-% change in x). A 1% increase in x changes y by about $\beta_1/100$ units. For a p% increase in x:

$$\Delta\hat{y} \approx \beta_1 \cdot \log\Big(1 + \frac{p}{100}\Big) \approx \beta_1 \cdot \frac{p}{100} \quad \text{(small } p\text{)}.$$

Thus β_1 maps proportional changes in x into absolute changes in y.

When to use.

- When x has a wide positive range and proportional changes in x (rather than absolute changes) are the relevant shock.
- When we believe y responds in *levels* to *percent changes* in x (e.g., \$ change in consumption from a % change in income).

Worked example (consumption vs. income). If $\widehat{C} = 400 + 1200 \cdot \log(Y)$ and income rises by 10% ($\Delta \log Y \approx 0.0953$), then $\Delta\widehat{C} \approx 1200 \times 0.0953 \approx 114.4$ currency units.

6.3.4 Log–Log (Elasticity model)

Model:

$$\log y = \beta_0 + \beta_1 \log x + u, \quad x > 0,\ y > 0.$$

Interpretation (elasticity). β_1 is the elasticity of y with respect to x: a 1% increase in x changes y by β_1% (constant elasticity across the range of x).

Exact vs. Approx. Exact for small changes; for larger proportional changes, use:

$$\Delta \log y = \beta_1 \Delta \log x \ \Rightarrow\ \%\Delta y = 100 \cdot \big[\exp(\beta_1 \Delta \log x) - 1\big].$$

When to use.

- When x and y are strictly positive and proportional responses are natural (e.g., demand elasticity w.r.t. price, Engel curves).
- When unit-free comparisons across samples/contexts are useful (elasticities are unitless).

Worked example (demand elasticity). If $\widehat{\log Q} = 5.1 - 1.2 \log P$, then the price elasticity of demand is -1.2: a 10% price increase reduces Q by about 12% (exact: $100[\exp(-1.2\times0.0953)-1] \approx -11.0\%$).

Comparing the four forms (cheat table)

Form	Equation	Slope meaning	Typical use
Level–Level	y on x	$\Delta y = \beta_1 \Delta x$	Additive effects, constant ME
Log–Level	$\log y$ on x	$\%\Delta y \approx 100\,\beta_1 \Delta x$	Percent change in y per unit x
Level–Log	y on $\log x$	$\Delta y \approx (\beta_1/100)\ \%\Delta x$	Level change from % change in x
Log–Log	$\log y$ on $\log x$	$\%\Delta y \approx \beta_1\ \%\Delta x$	Elasticity; unit-free comparison

Important cautions.

- **Zeros/negatives:** Logs require strictly positive values. Common fixes (adding a small constant, $\log(1+y)$) change interpretations, use with care and disclose.
- **Percent vs. percentage points:** If x is itself a percentage (e.g., unemployment rate), a one-unit change is *one percentage point*, not 1%.
- R^2 **comparability:** Do not compare R^2 across models with different dependent variables (e.g., y vs. $\log y$).

6.3.5 Quadratic Models (Curvature)

Model (single regressor):

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + u.$$

Marginal effect:

$$\frac{\partial \hat{y}}{\partial x} = \hat{\beta}_1 + 2\hat{\beta}_2 x,$$

so the effect *changes with* x. The turning point (vertex) occurs at

$$x^\star = -\frac{\hat{\beta}_1}{2\hat{\beta}_2} \quad (\hat{\beta}_2 \neq 0),$$

with concavity determined by $\hat{\beta}_2$ (< 0 concave, > 0 convex).

When to use. Diminishing/increasing returns (e.g., experience on wages), inverted-U relationships (e.g., age and productivity), or any curvature suggested by residual plots.

Worked example (returns to tenure). $\widehat{\log(\text{wage})} = 1.6 + 0.09 \cdot \text{tenure} - 0.002 \cdot \text{tenure}^2$. The marginal % effect per year is $100(0.09 - 0.004 \cdot \text{tenure})$. The peak occurs at $x^\star = 0.09/(2 \cdot 0.002) = 22.5$ years.

6.3.6 Interaction Terms (Effect Moderation)

Interactions allow the effect of one variable to depend on another.

Continuous × Continuous

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 (x_1 x_2) + u.$$

Marginal effects:

$$\frac{\partial \hat{y}}{\partial x_1} = \hat{\beta}_1 + \hat{\beta}_3 x_2, \qquad \frac{\partial \hat{y}}{\partial x_2} = \hat{\beta}_2 + \hat{\beta}_3 x_1.$$

Interpretation. β_3 captures how the slope of x_1 varies with x_2 (and vice versa). Test $H_0 : \beta_3 = 0$ to assess whether the slope truly depends on the other variable.

Example (class size × teacher experience). If the negative effect of class size on test scores attenuates with teacher experience, $\hat{\beta}_3 > 0$ (the class-size penalty is smaller when experience is higher).

Dummy × Continuous

Let $D \in \{0, 1\}$.

$$y = \beta_0 + \beta_1 x + \delta_0 D + \delta_1 (D \cdot x) + u.$$

Interpretation. When $D = 0$: intercept β_0, slope β_1. When $D = 1$: intercept $\beta_0 + \delta_0$, slope $\beta_1 + \delta_1$. Thus δ_0 is the intercept shift, δ_1 is the slope shift for the $D = 1$ group.

Example (policy effect heterogeneity). Suppose D indicates treated districts. Then $\delta_1 > 0$ means the return to x (e.g., funding) is larger in treated districts.

Dummy × Dummy

$$y = \beta_0 + \delta_A A + \delta_B B + \delta_{AB}(A \cdot B) + u, \quad A, B \in \{0, 1\}.$$

The interaction δ_{AB} allows the joint presence of A and B to have an additional effect beyond their separate contributions.

6.3.7 Discrete Changes and Dummies in Log Models

If the dependent variable is in logs and you have a dummy D:

$$\log y = \beta_0 + \delta D + \text{controls} + u,$$

then the *percentage difference* for $D = 1$ vs. $D = 0$ is *exactly*

$$100 \cdot \left[\exp(\delta) - 1 \right] \%.$$

For small $|\delta|$, the approximation $100\delta\%$ is fine, but for larger $|\delta|$ report the exact value above.

6.3.8 Average Partial Effects (APE) vs. Effect at the Mean (EAM)

When effects depend on x (quadratics/interactions), summarize as:

$$\text{EAM: } \left. \frac{\partial \hat{y}}{\partial x_j} \right|_{x=\bar{x}}, \qquad \text{APE: } \frac{1}{n} \sum_{i=1}^{n} \left. \frac{\partial \hat{y}}{\partial x_j} \right|_{x=x_i}.$$

Practice. Report both if they differ; APE is typically more representative in skewed samples or when heterogeneity is strong.

6.3.9 Choosing Among Functional Forms (Practical Guidance)

- **Think economics first.** Is the response proportional (logs), additive (levels), or elastic (log–log)?
- **Look at the data.** Plots of y vs. x, residuals vs. fits, binned scatter plots: linearity, curvature, spread.
- **Stability of variance.** Logs often reduce right skew and multiplicative heteroskedasticity.
- **Comparability.** Do not compare R^2 across different dependent variables (e.g., y vs. $\log y$).
- **Transparency.** State clearly whether your reported effect is a level change, a percent change, or a percentage-point change.

6.4 Goodness-of-Fit and Choosing Regressors

6.4.1 Adjusted R^2

The usual R^2 weakly increases as we add regressors, even if they have no true effect. The adjusted R^2 penalizes for extra regressors:

$$\bar{R}^2 = 1 - \frac{\text{SSR}/(n-k-1)}{\text{SST}/(n-1)}.$$

Use $\bar{R}^2$ to compare models with the *same* dependent variable; do *not* compare R^2 (or $\bar{R}^2$) across y vs. $\log y$.

6.4.2 Using adjusted R^2 for non-nested choices

When theory does not dictate a unique functional form (e.g. level vs. log of a regressor), $\bar{R}^2$ comparisons can be informative if the dependent variable is the same across models. Remember that economic interpretability and plausibility trump tiny differences in $\bar{R}^2$.

6.4.3 Controlling for many factors

Adding regressors can reduce residual variance (improve precision) and mitigate omitted-variable bias, but at the cost of degrees of freedom and potential multicollinearity. Prefer parsimonious, theory-driven models; add controls that are plausibly related to y and correlated with our regressor of interest.

6.5 Prediction and Residual Analysis

6.5.1 Prediction and confidence intervals

For a new x^*, the *mean* prediction is $\hat{y}(x^*)$ with a CI that uses the standard error of the fitted mean; a *forecast* (for a new individual) adds the irreducible error variance and is wider. Always state which interval you are reporting.

6.5.2 Residual analysis

Plot residuals vs. fitted values (and key regressors) to look for:

- Systematic patterns (misspecification, omitted nonlinearities/interactions),
- Changing spread (heteroskedasticity),
- Outliers/high leverage points (influence diagnostics).

6.5.3 Predicting y when $\log(y)$ was modeled

We estimate a model for $\log(y)$ but often need predictions of y itself. Simply exponentiating $\widehat{\log y}$ understates $E[y|x]$; a retransformation adjustment is needed. The text flags this issue and shows how to convert predictions for $\log(y)$ into predictions for y (and discusses comparable fit measures

when the dependent variable differs across specifications). Use an appropriate retransformation (e.g., a smearing or parametric correction) when reporting predictions in levels.

6.6 Some Reference Tables

(A) Interpreting common functional forms

Model	Slope meaning (approx.)	Notes
Level–Level: y on x	$\Delta y \approx \beta_1 \Delta x$	Unit-specific
Log–Level: $\log y$ on x	$\%\Delta y \approx 100\,\beta_1 \Delta x$	Use exact $100[\exp(\beta_1 \Delta x) - 1]$ for large Δx
Level–Log: y on $\log x$	$\Delta y \approx (\beta_1/100)\,\%\Delta x$	Unit of y; x must be > 0
Log–Log: $\log y$ on $\log x$	$\%\Delta y \approx \beta_1\,\%\Delta x$	Elasticity; unit-free

(B) Quadratic and interaction effects

Specification	Marginal effect
$y = \beta_0 + \beta_1 x + \beta_2 x^2 + u$	$\partial \hat{y}/\partial x \approx \hat{\beta}_1 + 2\hat{\beta}_2 x$
$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + u$	$\partial \hat{y}/\partial x_1 = \hat{\beta}_1 + \hat{\beta}_3 x_2$
$y = \beta_0 + \beta_1 x + \delta_0 D + \delta_1 D \cdot x + u$	Slope shift by δ_1 when $D = 1$

(C) Standardized (beta) coefficients

Definition	$\tilde{\beta}_j = \dfrac{\hat{\sigma}_{x_j}}{\hat{\sigma}_y}\hat{\beta}_j$ (regress z_y on z_x's)
Interpretation	$+1$ sd in $x_j \Rightarrow \tilde{\beta}_j$ sd change in $\hat{y}$
Significance	t-stats identical to unstandardized model

6.7 Some tips

- Use rescaling to make coefficients readable, but remember it *never* changes t/F decisions or R^2.
- Prefer logs for strictly positive monetary and size variables; be cautious with zeros and with proportion variables near zero; keep p.p. vs. % straight.
- For curvature, include x *and* x^2; for interactions, include the main effects with the product term.
- Do not compare R^2 across y vs. $\log y$; consider adjusted R^2 only when the dependent variable is the same.
- When predicting levels after modeling logs, apply a re-transformation adjustment rather than just exponentiating fitted log values.

6.8 Nested vs. Non-nested Models

Nested vs. Non-nested Models in Regression

Purpose: To understand when one model is a special case (subset) of another and what comparison tools are appropriate.

Type	Definition / Relationship	Comparison and Use
Nested Models	One model is a special (restricted) case of another. Formally, Model A imposes constraints on Model B's parameters. Example: Model A omits a variable that Model B includes.	**Formal test possible:** use t- or F-tests. If Model B adds extra variables, test H_0 : "added coefficients = 0." If H_0 is rejected $\Rightarrow$ larger model improves fit significantly. $\bar{R}^2$ and F-tests are valid because both models explain the same y.
Non-nested Models	Models are not special cases of each other; they differ in functional form or dependent variable transformation. Examples: y vs. $\log y$ as the dependent variable, or $\log(x)$ vs. x^2 as regressors.	**No exact F-test exists.** R^2 cannot be compared directly if y differs (e.g., y vs. $\log y$). Model choice must rely on *economic reasoning*, *predictive accuracy*, or information criteria (AIC, BIC), not formal nesting tests.

Example 1 (nested):

$$\text{Model A: } y = \beta_0 + \beta_1 x_1 + u,$$
$$\text{Model B: } y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u.$$

Model A is nested in Model B; test $H_0 : \beta_2 = 0$ using an F-test.

Example 2 (non-nested):

$$\text{Model 1: } y = \beta_0 + \beta_1 x + u, \qquad \text{Model 2: } \log y = \alpha_0 + \alpha_1 x + v.$$

Model 2 is not a restricted form of Model 1 because the dependent variable differs. Use adjusted R^2, AIC/BIC, or predictive performance - not F-tests - to compare.

Summary:

- Nested models $\Rightarrow$ hypothesis testing possible (restrictions on parameters of the same y).
- Non-nested models $\Rightarrow$ rely on fit measures or theory, not formal F tests.
- Only compare $\bar{R}^2$ across models with the same dependent variable.

Chapter 7: Multiple Regression Analysis with Qualitative Information: Binary (or Dummy) Variables

Quantitative variables (like income, price, or age) measure "how much." **Qualitative variables** measure "which group" or "what category." Dummy variables (also called *indicator* or *binary* variables) allow regression models to include such categorical information.

7.1 Describing Qualitative Information

A dummy variable takes only two values:

$$D_i = \begin{cases} 1, & \text{if observation } i \text{ belongs to a category,} \\ 0, & \text{otherwise.} \end{cases}$$

Examples:

- $D_{\text{female}} = 1$ if female, 0 if male.
- $D_{\text{union}} = 1$ if union member, 0 otherwise.
- $D_{\text{south}} = 1$ if lives in the South, 0 otherwise.

Including dummy variables lets us compare mean outcomes across groups while controlling for other variables.

Intuition: Dummies "turn on" or "off" an effect. The coefficient attached to a dummy tells you how much higher or lower the average outcome is for that group compared to the base group, all else equal.

7.2 A Single Dummy Independent Variable

Model:

$$y_i = \beta_0 + \delta_0 D_i + u_i.$$

When $D_i = 0$, $E(y|D = 0) = \beta_0$. When $D_i = 1$, $E(y|D = 1) = \beta_0 + \delta_0$.
Interpretation: δ_0 measures the difference in the mean of y between the two groups.
Example (Gender wage gap):

$$\text{wage} = \beta_0 + \delta_0 \text{female} + u.$$

If $\hat{\delta}_0 = -2.45$, women earn on average \$2.45 less per hour than men.

Adding more variables:

$$\text{wage} = \beta_0 + \delta_0 \text{female} + \beta_1 \text{educ} + \beta_2 \text{exper} + u.$$

Now δ_0 measures the *ceteris paribus* difference in wages between men and women, holding education and experience constant.

Intuition: Without controls, δ_0 just compares group averages. With controls, it isolates how much of the gap remains after accounting for differences in education, experience, etc. It's a way to "hold everything else fixed."

7.3 When the Dependent Variable Is $\log(y)$

Model:

$$\log(y_i) = \beta_0 + \delta_0 D_i + u_i.$$

Interpretation:

$$100 \times \delta_0$$

is the *approximate* percentage difference in y between the two groups. For more accuracy, use:

$$\%\text{difference} = 100 \times [\exp(\delta_0) - 1].$$

Example: If $\hat{\delta}_0 = -0.20$, then $\exp(-0.20) - 1 = -0.1813$. $\Rightarrow$ Women earn about 18% less than men, holding other factors constant.

Intuition: Taking logs makes comparisons multiplicative. If the model is in logs, differences between groups are naturally expressed as *percentage* changes, not dollar differences. This is useful when variability increases with income or size.

7.4 Using Dummy Variables for Multiple Categories

For a categorical variable with c groups (e.g., race: White, Black, Hispanic, Asian), create $c - 1$ dummy variables. The omitted category becomes the **base group.**

Example (Four regions):

$$\text{wage} = \beta_0 + \delta_1 D_{\text{NE}} + \delta_2 D_{\text{MW}} + \delta_3 D_{\text{S}} + u,$$

where the West is omitted as the base group.

Interpretation:

- δ_1: mean difference between Northeast and West, holding others constant.
- δ_2: mean difference between Midwest and West, etc.
- β_0: expected wage for the base (West) group when other covariates are zero.

Avoiding the Dummy Variable Trap

Rule: Never include all category dummies and an intercept together. Why? Because they perfectly add up to one, causing **perfect multicollinearity.**
Correct setup: If there are c categories, include only $c - 1$ dummies and an intercept. The omitted category automatically becomes the baseline against which all other categories are compared.
Tip: The choice of base group doesn't change overall fit or R^2—it only changes the interpretation of coefficients.

Intuition: Leaving one category out gives the regression a reference point. All other coefficients tell you how each group's mean differs from that baseline.

7.5 Incorporating Ordinal Information

If a categorical variable has a natural order (e.g., education level: High School, College, Graduate), create dummy variables for each level except one.

Example:

$$\text{wage} = \beta_0 + \delta_1 D_{\text{college}} + \delta_2 D_{\text{grad}} + u,$$

where High School is the base.

Interpretation:

- δ_1: difference in mean wage between College and High School.
- δ_2: difference in mean wage between Graduate and High School.

Intuition: Ordinal dummies treat each category as a step up from the base. The coefficients show how outcomes improve (or worsen) as we move up the ordered scale.

7.6 Interactions among Dummy Variables

Interacting dummy variables allows for combined group effects.

Example (gender $\times$ union):

$$\text{wage} = \beta_0 + \delta_1 \text{female} + \delta_2 \text{union} + \delta_3 (\text{female} \times \text{union}) + u.$$

- δ_3 measures how the union wage premium differs between men and women.
- The union effect for women: $\delta_2 + \delta_3$.
- The union effect for men: δ_2.

Intuition: Interaction terms show that "the effect of one factor depends on another." In the above example, unions might raise wages for both men and women, but perhaps the increase is larger (or smaller) for women—δ_3 captures that difference.

7.7 Allowing for Different Slopes across Groups

You can let slopes (not just intercepts) vary by group using interactions with continuous variables.

Model:

$$y = \beta_0 + \delta_0 D + \beta_1 x + \delta_1 (D \times x) + u.$$

Interpretation:

- For group $D = 0$: $E(y|D = 0) = \beta_0 + \beta_1 x$.
- For group $D = 1$: $E(y|D = 1) = (\beta_0 + \delta_0) + (\beta_1 + \delta_1)x$.

So both intercept and slope differ by δ_0 and δ_1 respectively.

Example: If $\delta_1 > 0$, the return to education is higher for women than for men.

Intuition: This setup allows the regression line to have a different shape for each group. Instead of assuming one overall slope for everyone, we allow each group's relationship to tilt differently—like giving each group its own trend line.

7.8 Testing for Differences Across Groups (Expanded)

Sometimes we want to know whether *the entire regression function* is the same across two groups (e.g., women vs. men), not just whether the intercept differs. The natural way is to allow the intercept *and* all slopes to differ across groups and then test those differences jointly.

Unified interaction specification

Let D be a group dummy (e.g., D = female). With regressors $x_1, \ldots, x_k$, write the *fully interacted* model:

$$y = \beta_0 + \delta_0 D + \beta_1 x_1 + \cdots + \beta_k x_k + \delta_1 (D \cdot x_1) + \cdots + \delta_k (D \cdot x_k) + u.$$

Here, δ_0 is the intercept difference; δ_j is the slope difference on x_j (group $D = 1$ minus group $D = 0$). The null that *all* function parameters are the same across groups is

$$H_0 : \ \delta_0 = \delta_1 = \cdots = \delta_k = 0 \quad \text{(jointly } q = k + 1 \text{ restrictions).}$$

This is a standard **linear restrictions** test: run the fully interacted regression (unrestricted), then test whether the q interaction/dummy coefficients are jointly zero (an F-test).

Intuition: The unrestricted model gives each group its own line (own intercept and slopes). Under the null, both groups share the same line. If the restrictions cause a meaningful loss of fit, we reject the idea that one line suffices.

Worked template (athlete GPA example)

Suppose

$$\text{cumgpa} = \beta_0 + \delta_0\,\text{female} + \beta_1\,\text{sat} + \delta_1\,(\text{female}{\cdot}\text{sat}) + \beta_2\,\text{hsperc} + \delta_2\,(\text{female}{\cdot}\text{hsperc}) + \beta_3\,\text{tothrs} + \delta_3\,(\text{female}{\cdot}\text{tothrs}) + u.$$

To test "same GPA model for men and women," use the joint null

$$H_0: \ \delta_0 = \delta_1 = \delta_2 = \delta_3 = 0 \quad (q = 4).$$

Do not look at interaction t-tests one-by-one; use the *joint* F-test.

How to compute the joint F-test

There are two equivalent ways (with the same data and dependent variable):

(a) R^2 form (single-regression approach). Let R^2_{ur} be from the fully interacted model (unrestricted), and R^2_r from the restricted model (drop D and all $D \cdot x_j$ terms). If k_{ur} is the number of regressors (including the intercept) in the unrestricted model, then

$$F = \frac{(R^2_{ur} - R^2_r)/q}{(1 - R^2_{ur})/(n - k_{ur})} \quad \sim F_{q,\, n-k_{ur}} \quad \text{(large-sample approximation if using robust SEs).}$$

(b) SSR/Chow form (two-regression approach). Let k be the number of *slopes* in x (not counting the intercept). Estimate separate regressions by group:

- Group 1 (size n_1): get SSR_1 from $y = \beta_{1,0} + \beta_{1,1}x_1 + \cdots + \beta_{1,k}x_k + u$.
- Group 2 (size n_2): get SSR_2 analogously.

Then $SSR_{ur} = SSR_1 + SSR_2$ comes from the "unrestricted" fit that allows *all* coefficients to differ. Next, pool the two groups and estimate a single common-parameters regression to obtain SSR_P (the "restricted" fit). The **Chow** F-statistic is

$$F = \frac{(SSR_P - SSR_{ur})/\,(k+1)}{SSR_{ur}/\,(n - 2(k+1))} \quad \sim F_{k+1,\, n-2(k+1)}.$$

Caution: The classic Chow test assumes **homoskedastic** errors (and equal error variances across groups). If heteroskedasticity is a concern, prefer the single-regression approach and compute a *robust* Wald/F for the interactions.

Intuition. The Chow logic compares "one line for everyone" vs. "two separate lines." If letting each group have its own line reduces the residual sum of squares a lot (relative to sampling noise), then the groups do not share the same regression function.

Testing only for slope differences (allowing intercept shift under H_0)

Often we want to allow different intercepts under the null and test *only* whether slopes are equal. There are two equivalent implementations:

(i) **Single-regression Wald F:** Include D and *all* interactions, but test only

$$H_0: \ \delta_1 = \cdots = \delta_k = 0 \quad (q = k),$$

leaving δ_0 unrestricted.

(ii) **Chow/SSR form:** Let $SSR_{ur} = SSR_1 + SSR_2$ as before. For the restricted SSR, estimate the *pooled* model that includes an intercept shift D but *no* slope interactions (so the two groups have different intercepts but common slopes) and call its residual sum SSR_P. Then

$$F = \frac{(SSR_P - SSR_{ur})/\,k}{SSR_{ur}/\,(n-2(k+1))} \quad \sim F_{k,\,n-2(k+1)}.$$

Intuition: This version asks a narrower question: once we allow groups to start at different baselines (different intercepts), do they respond to $x_1, \ldots, x_k$ in the same way (same slopes)? If not, the interactions matter.

Interpreting group differences at realistic covariate values

With interactions, the group difference in the *predicted* outcome depends on x:

$$\widehat{y}(D{=}1) - \widehat{y}(D{=}0) = \hat{\delta}_0 + \hat{\delta}_1 x_1 + \cdots + \hat{\delta}_k x_k.$$

Always compute differences at meaningful values of x (e.g., typical SAT, % rank, hours). A large negative $\hat{\delta}_0$ alone *does not* imply a lower outcome for the $D{=}1$ group; the interaction terms may dominate at realistic x values.

Intuition: Dummies with interactions do not have a single "gap." The gap varies with x. Summarize at reference points (means, quartiles) or report an average partial effect of the group dummy.

A checklist

- Decide whether you want to test *everything* (intercept and slopes) or *only slopes*.
- Prefer a **single interacted regression** and a **robust** Wald/F test (handles heteroskedasticity).
- Chow (SSR) formulas are convenient and intuitive but rely on homoskedasticity for exact F.
- Report the joint F and also give *interpretable* group differences at realistic x values.

More on Chow Test

---

1. What the Chow Test Really Checks

The Chow Test answers a very simple question:

> *Should we use one regression for everyone, or should we split the sample and run two separate regressions because the groups behave differently?*

This usually arises when we compare two groups, such as:

- men vs. women,
- native vs. foreign-born,
- before vs. after a policy change,

- rural vs. urban areas.

If the relationship between the variables is the same in both groups, one regression is enough. If not, using two separate regressions gives a better fit. The Chow Test checks this formally.

2. What the Test Actually Compares

Suppose our model is

$$y = \alpha + \beta x + u.$$

We want to know if both groups share the same coefficients. This means testing:

$$H_0 : \ \alpha_1 = \alpha_2, \quad \beta_1 = \beta_2.$$

If H_0 is true, the model is the same for both groups. If H_0 is false, the groups differ and should be modeled separately.

3. The Three Regressions You Must Run

To perform a Chow Test, we estimate three regressions:

1. **Pooled regression (restricted model):** This uses the entire sample. Let its sum of squared residuals be SSR_R.
2. **Group 1 regression (unrestricted):** Run the same model on Group 1 only. Let its SSR be SSR_1.
3. **Group 2 regression (unrestricted):** Run the same model on Group 2 only. Let its SSR be SSR_2.

The unrestricted SSR is:

$$SSR_{UR} = SSR_1 + SSR_2.$$

The idea is simple: If the two separate regressions reduce SSR a lot, the groups must differ.

4. Chow Test Statistic

Let k be the number of parameters in the model (intercept + slopes).

Let n_1 and n_2 be the sample sizes for the two groups. Total sample size:

$$n = n_1 + n_2.$$

The Chow Test statistic is:

$$F = \frac{(SSR_R - SSR_{UR})/k}{SSR_{UR}/(n_1 + n_2 - 2k)}.$$

Under the null hypothesis:

$$F \sim F(k,\ n_1 + n_2 - 2k).$$

Decision Rule: Reject H_0 at the 5% level if

$$F > F_{0.95}(k,\ n_1 + n_2 - 2k).$$

5. A Simple Numerical Example

Suppose we study the effect of education on wage:

$$wage = \alpha + \beta\, educ + u.$$

There are $k = 2$ coefficients.
We compare:

- **Group 1:** Men ($n_1 = 30$)
- **Group 2:** Women ($n_2 = 35$)

Assume we obtain:

$$SSR_R = 500, \qquad SSR_1 = 200, \qquad SSR_2 = 180.$$

Then

$$SSR_{UR} = SSR_1 + SSR_2 = 380.$$

Compute the Chow statistic:

$$F = \frac{(500 - 380)/2}{380/(65 - 4)} = \frac{120/2}{380/61} = \frac{60}{6.23} \approx 9.63.$$

Degrees of freedom:

$$df_1 = k = 2, \qquad df_2 = n_1 + n_2 - 2k = 61.$$

Critical value:

$$F_{0.95}(2, 61) \approx 3.15.$$

Since $9.63 > 3.15$, we reject the null hypothesis.
Interpretation: The wage–education relationship is different for men and women. We should estimate separate regressions for the two groups.

6. Why the Chow Test Makes Sense

The pooled model forces the same coefficients on both groups. Running separate regressions gives the model more freedom.
If these two separate regressions reduce the SSR dramatically, it means:

> *The two groups follow different relationships, and the pooled model is hiding important differences.*

The Chow Test quantifies this difference and tells us whether it is large enough to matter statistically.

7.9 Binary Dependent Variable: The Linear Probability Model (LPM)

When y itself is binary (0 or 1), we can still use OLS:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + u_i.$$

Interpretation: $E(y|x) = P(y = 1|x)$, so coefficients represent *changes in probability*.
Example: If vote = 1 if a person votes, 0 otherwise,

$$\text{vote} = \beta_0 + \beta_1 \text{income} + \beta_2 \text{age} + u.$$

Then $\hat{\beta}_1 = 0.04$ means: an extra \$1,000 in income increases the probability of voting by 4 percentage points.

Problems:

- Fitted probabilities can be below 0 or above 1.
- Errors are heteroskedastic.

Use **robust standard errors** or consider nonlinear models (logit or probit).

Intuition: Even though OLS isn't designed for probabilities, it gives a quick, easy-to-read estimate of how x changes the likelihood of an event. It's a first step, simple and interpretable, but not perfect because probabilities can stray outside $[0, 1]$.

7.10 Policy Analysis and Program Evaluation

Dummy variables are powerful tools for estimating causal effects of programs or policies.

Example (training program):

$$\text{earnings} = \beta_0 + \delta_0 \text{training} + \beta_1 \text{educ} + \beta_2 \text{exper} + u.$$

Here, δ_0 estimates the **average treatment effect** of participating in the program, controlling for education and experience.

Difference-in-means (no controls): average earnings difference between trained and untrained.
Regression with controls: adjusts for other differences (e.g., education, experience).

Intuition. Think of δ_0 as "the estimated causal impact" of the program. By controlling for factors that might influence both training and earnings, we isolate how much of the difference is truly due to the program rather than preexisting differences.

7.11 Discrete Dependent Variables in General

When y is discrete (binary, counts, or categories):

- The linear model is easy to interpret but may predict impossible values.
- More advanced models, logit, probit, Poisson, handle limited outcomes better.

However, understanding the Linear Probability Model builds intuition for how regressors affect probabilities in nonlinear models.

Intuition: In real-world data, many dependent variables are yes/no or counts (vote, employed, default, number of children). Linear models can approximate these relationships surprisingly well for small probability ranges, but specialized nonlinear models are needed when predictions near 0 or 1 matter.

Summary: Dummy Variable Uses

Purpose	**Example / Interpretation**
Group mean differences	wage = β_0 + $\delta_0$female + u; δ_0 = mean difference (women–men)
Multiple categories	Include $c - 1$ dummies for c groups; one base group omitted
Ordinal variables	Create dummies by level; interpret relative to lowest category
Interaction among dummies	Gender × union effect differences
Different slopes by group	Add $D \times x$ to let slope vary
Binary dependent variable	Linear Probability Model: coefficients = Δ in probability
Policy evaluation	Dummy = treatment indicator; coefficient = program effect

Chapter 8: Heteroskedasticity

8.1 What Is Heteroskedasticity and Why Do We Care?

In the multiple regression model,

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u,$$

one key assumption under the classical linear model is **homoskedasticity**:

$$\text{Var}(u \mid x_1, \ldots, x_k) = \sigma^2.$$

This means the variance of the error term is constant across all values of the explanatory variables.

Heteroskedasticity occurs when this assumption fails:

$$\text{Var}(u \mid x_1, \ldots, x_k) = \sigma^2(x_1, \ldots, x_k),$$

so the error variance changes with the level of one or more regressors.

Intuition

Heteroskedasticity is common in cross-sectional data. Typical examples include:

- Income models where high-income individuals have more dispersed outcomes.
- Firm data where larger firms show more variability in profits.
- Models where the dependent variable is bounded, such as binary outcomes.

Ignoring heteroskedasticity does *not* usually bias coefficient estimates, but it *does* affect inference.

8.2 Consequences of Heteroskedasticity for OLS

8.2.1 What Still Works

Under heteroskedasticity, as long as the zero conditional mean assumption holds,

$$\text{E}(u \mid x_1, \ldots, x_k) = 0,$$

the OLS estimators remain:

- **Unbiased**
- **Consistent**

So heteroskedasticity does *not* invalidate point estimates.

8.2.2 What Breaks

The usual OLS variance formula,

$$\text{Var}(\hat{\beta}_j \mid X) = \frac{\sigma^2}{\sum (x_{ij} - \bar{x}_j)^2},$$

is no longer valid.

As a result:

- Standard errors are wrong.
- t statistics and F tests are invalid.
- Confidence intervals are unreliable.

OLS is also no longer efficient: there exist estimators with smaller variance.

8.3 Heteroskedasticity-Robust Inference After OLS

To fix inference without changing coefficient estimates, we use **heteroskedasticity-robust standard errors**.

8.3.1 Robust Variance Estimator

The robust variance estimator replaces the constant variance assumption with squared residuals:

$$\widehat{\text{Var}}(\hat{\beta}) = (X'X)^{-1} \left(\sum_{i=1}^{n} \hat{u}_i^2 x_i x_i' \right) (X'X)^{-1}.$$

This estimator is consistent even when heteroskedasticity is present.

8.3.2 Interpretation

- Coefficients stay the same.
- Standard errors change.
- Inference becomes valid asymptotically.

This is now the default in most applied work.

8.3.3 Robust LM Tests

Lagrange Multiplier tests can also be made robust by using the robust covariance matrix. These tests are valid under heteroskedasticity and are often simpler than Wald tests in large models.

8.4 Testing for Heteroskedasticity

While robust inference works regardless, testing can help diagnose model structure.

8.4.1 General Idea

Most tests examine whether squared residuals are systematically related to explanatory variables:

$$\hat{u}_i^2 = \delta_0 + \delta_1 z_{i1} + \cdots + \delta_m z_{im} + v_i.$$

If the regressors explain residual variance, heteroskedasticity is present.

8.4.2 The White Test

The White test regresses squared residuals on:

- Original regressors
- Their squares
- Their cross-products

The test statistic is:

$$nR^2 \sim \chi_q^2,$$

where q is the number of auxiliary regressors.

Pros and Cons

- Very general
- Does not require specifying a functional form
- Can suffer from low power in small samples

8.5 Weighted Least Squares (WLS)

When heteroskedasticity has a known structure, OLS can be improved.

8.5.1 Known up to a Multiplicative Constant

Suppose:

$$\text{Var}(u_i \mid x_i) = \sigma^2 h_i,$$

where h_i is known.
Divide the model by $\sqrt{h_i}$:

$$\frac{y_i}{\sqrt{h_i}} = \beta_0 \frac{1}{\sqrt{h_i}} + \beta_1 \frac{x_{i1}}{\sqrt{h_i}} + \cdots + \frac{u_i}{\sqrt{h_i}}.$$

OLS on the transformed model is efficient.

8.5.2 Feasible GLS (FGLS)

Often h_i is unknown and must be estimated:

1. Estimate the original model by OLS.
2. Regress $\hat{u}_i^2$ on suspected drivers of variance.

3. Use fitted values $\hat{h}_i$ as weights.
4. Re-estimate using WLS.

FGLS is consistent if the variance model is correct.

8.5.3 What If the Variance Model Is Wrong?

If the heteroskedasticity function is misspecified:

- FGLS may be inefficient.
- Inference may be unreliable.

In practice, OLS with robust standard errors is often safer.

8.6 Prediction with Heteroskedasticity

Heteroskedasticity affects prediction intervals because:

$$\text{Var}(y_{n+1} \mid x_{n+1}) = \text{Var}(u_{n+1} \mid x_{n+1}).$$

Prediction intervals should allow variance to depend on covariates. Ignoring this leads to intervals that are too narrow or too wide.

8.7 The Linear Probability Model Revisited

In the Linear Probability Model (LPM),

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

heteroskedasticity is guaranteed because:

$$\text{Var}(u \mid x) = p(x)(1 - p(x)).$$

Implications

- OLS coefficients remain consistent.
- Usual standard errors are wrong.
- Robust standard errors are mandatory.

This explains why heteroskedasticity-robust inference is standard in binary outcome models.

8.8 How to Handle Heteroskedasticity in Stata

Let's see how to (1) diagnose heteroskedasticity and (2) run valid inference when it is present. Throughout, assume a cross-sectional OLS regression:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u.$$

8.8.1 Step 0: Run the baseline OLS

In Stata:

```
reg y x1 x2 x3
```

This gives OLS coefficient estimates. If heteroskedasticity is present, the coefficients are still unbiased/consistent (under zero conditional mean), but the usual standard errors can be wrong.

8.8.2 Step 1: Use heteroskedasticity-robust standard errors (default fix)

The simplest correct response is to keep OLS coefficients and use robust standard errors:

```
reg y x1 x2 x3, vce(robust)
```

Why: `vce(robust)` replaces the homoskedastic variance formula with a heteroskedasticity-robust variance estimator, making t tests, F tests, and confidence intervals asymptotically valid.

8.8.3 Step 2: Breusch-Pagan / Cook-Weisberg test

After running `reg`, you can test for heteroskedasticity using:

```
estat hettest
```

By default, this is a Breusch-Pagan/Cook-Weisberg test of whether the error variance depends on the fitted values.

You can also test heteroskedasticity as a function of specific regressors:

```
estat hettest x1 x2 x3
```

Null hypothesis: $\text{Var}(u \mid X) = \sigma^2$ (homoskedasticity).
Alternative: variance changes with the included variables.

Interpretation:

- Small p-value (e.g., < 0.05): evidence of heteroskedasticity.
- Large p-value: no strong evidence, but this does not prove homoskedasticity.

8.8.4 Step 3: White test idea (general heteroskedasticity)

Stata has a built-in IM test that is commonly used as a White-type general test:

```
estat imtest, white
```

Why: This is a more flexible test that can detect many forms of heteroskedasticity without you choosing a specific variance function. It is useful when you do not have a clear theory about how variance changes.

Warning: Like many general tests, it can have low power in small samples and can be sensitive to outliers.

8.8.5 Step 4: Visual diagnosis (fast and often revealing)

Plot residual patterns against fitted values:

```
predict uhat, resid
predict yhat, xb
scatter uhat yhat
```

A "fan shape" (residual spread growing with fitted values) is a classic sign of heteroskedasticity. A cleaner plot uses squared residuals:

```
gen uhat2 = uhat^2
scatter uhat2 yhat
```

8.8.6 Step 5: Weighted Least Squares (WLS) and Feasible GLS (FGLS)

If you have a credible model for the variance, you can use weights.

Case A: Known form up to a constant

If theory suggests $\text{Var}(u_i \mid X_i) = \sigma^2 h_i$ and you know h_i (up to scale), use analytic weights:

```
reg y x1 x2 x3 [aw = 1/h]
```

Here h is your known function (must be a variable in your data).

Why: WLS downweights observations with higher error variance and can be more efficient than OLS.

Case B: FGLS when variance must be estimated

A common workflow is:

1. Run OLS and get residuals.
2. Model the variance using $\log(\hat{u}_i^2)$ or $\hat{u}_i^2$ on variables that might drive variance.
3. Convert fitted values to weights and re-run regression.

Example:

```
reg y x1 x2 x3
predict uhat, resid
gen lnu2 = ln(uhat^2)

reg lnu2 z1 z2 z3
predict ghat, xb
gen hhat = exp(ghat)
reg y x1 x2 x3 [aw = 1/hhat]
```

Why: This is a feasible version of GLS: you estimate the heteroskedasticity function, then weight accordingly.

Important caution: If your variance model is wrong, FGLS can misbehave. In practice, OLS with robust standard errors is often the safer default.

8.8.7 Step 6: Linear Probability Model (LPM) reminder

If *y* is binary (0/1), heteroskedasticity is guaranteed, so robust SEs are not optional:

```
reg y x1 x2 x3, vce(robust)
```

8.8.8 One-page decision rule

Situation	What to do in Stata
You only care about valid inference	`reg y X, vce(robust)`
You want a formal test	`estat hettest` and/or `estat imtest, white`
You have strong theory for variance form	WLS: `reg y X [aw=1/h]`
You want efficiency and can model variance well	FGLS (two-step weights)
Binary dependent variable (LPM)	Always robust SEs

Acknowledgment and Source Notice

These lecture notes were created for personal academic study and educational use in ECO 350: Econometrics. They are based on concepts, examples, and figures presented in *Introductory Econometrics: A Modern Approach, 6th Edition* by Jeffrey M. Wooldridge (Michigan State University).

All credit for the original content and figures belongs to the author and publisher. This document is a student-produced summary for learning purposes only and is not intended for commercial distribution.