Intermediate Econometrics

A Self-Contained Course Companion

Lecture course by **Prof. Xiaojun Song**
Guanghua School of Management, Peking University
Notes written and LaTeX-typeset by Rui Zhou

Spring 2023
(rewritten and unified, 2026)

How to Read This Book

This is a course companion for a one-semester **intermediate econometrics** course based on Wooldridge's *Introductory Econometrics*. It is written to be read straight through, like a short textbook, rather than skimmed like a formula sheet.

The first half of the course covers the mechanics of the linear regression model: how ordinary least squares (OLS) is defined, when it is unbiased, what its variance is, and how to test hypotheses about its coefficients. The second half relaxes the convenient assumptions one at a time — nonlinear functional forms, qualitative regressors, heteroskedasticity, problematic data, panel structure, and endogeneity — and shows what tool repairs each broken assumption. Every later chapter is, at heart, a story of the form "*assumption X fails; here is what goes wrong, how we detect it, and how we fix it.*"

How Each Chapter Is Organized

Each chapter opens with a short *story* in plain English: what problem are we solving, and why are we about to introduce a particular tool? We then formalize the idea, state the main results, and work through the algebra. Definitions, theorems, and assumptions are boxed so you can find them quickly when reviewing.

Box Color Code

- **Green: Definition**

  — the objects and concepts we build on.

- **Blue: Theorem**

  — the main results, usually with a proof or sketch.

- **Pink: Assumption**

  — the conditions we impose (e.g. SLR.1–SLR.5); we cite them by name.

- **Purple: Lemma**

  — intermediate results used in a proof.

- Worked **examples** and **remarks** appear in lightly ruled boxes throughout.

A Note on Notation

Throughout, n is the sample size, k the number of regressors (excluding the intercept), u the population error, and $\widehat{u}_i$ the OLS residual. We write $\mathbb{E}(\cdot)$ for expectation, $\mathrm{Var}(\cdot)$ for variance, $\mathrm{Cov}(\cdot,\cdot)$ for covariance, and plim for the probability limit. A hat denotes an estimator ($\widehat{\beta}_j$); a tilde denotes an estimator from a *misspecified* (e.g. short) regression ($\widetilde{\beta}_j$).

Contents

Chapter 1

The Simple Regression Model

Econometrics begins with a deceptively simple question: how does one variable move with another, *on average*, once we admit that the relationship is not exact? Suppose we believe that wages rise with education, that consumption rises with income, or that a crop yield rises with rainfall. In each case the relationship is real but noisy — two people with the same education earn different wages. The simple regression model is the smallest formal object that captures this idea: a straight line for the average, plus an error term for everything else.

This chapter introduces that model, defines the ordinary least squares (OLS) estimator that fits it, and asks the two questions we will ask of *every* estimator in this course: is it correct on average (unbiased), and how precise is it (variance)? The answers here, in the one-regressor case, set the template for the multiple regression model in Chapter 2.

1.1 The Population Model

Definition 1.1: Simple Linear Regression Model

The *simple linear regression (SLR) model* relates a dependent variable y to a single explanatory variable x through

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

where β_0 is the intercept, β_1 the slope, and u the error (or disturbance) term collecting all factors other than x that affect y.

The vocabulary is worth fixing once, because different fields use different names for the same objects:

- y: *dependent* variable, explained variable, response, or regressand;
- x: *independent* variable, explanatory variable, control, or regressor;
- u: *error* term or disturbance.

The slope β_1 is the object of interest: it is the change in y associated with a one-unit change in x, *holding the error fixed*, i.e. $\Delta y = \beta_1 \, \Delta x$ when $\Delta u = 0$. For this to have a causal reading we need the error to be unrelated to x, which we now make precise.

Normalizing $\mathbb{E}(u) = 0$ is free: any nonzero mean of u can be absorbed into the intercept β_0. The substantive assumption is the next one.

Definition 1.2: Zero Conditional Mean

The error has *zero conditional mean* if

$$\mathbb{E}(u \,|\, x) = 0 \qquad \text{for every value of } x.$$

Equivalently, the average of the unobserved factors does not vary with x.

Under zero conditional mean, taking the expectation of the model conditional on x gives the *population regression function (PRF)*,

$$\mathbb{E}(y \,|\, x) = \beta_0 + \beta_1 x,$$

which says that the conditional mean of y is linear in x. This is what OLS will try to recover from a sample.

1.2 Ordinary Least Squares

Zero conditional mean implies two *population moment conditions*: $\mathbb{E}(u) = 0$ and $\mathbb{E}(xu) = 0$ (the latter follows because $\mathbb{E}(u \,|\, x) = 0 \Rightarrow \mathrm{Cov}(x, u) = 0$). OLS is the method that imposes the sample analogues of these two conditions and solves for the two unknowns β_0, β_1.

Definition 1.3: OLS Estimators

Given a sample $\{(x_i, y_i) : i = 1, \dots, n\}$, the OLS estimators are

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})\, y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}, \qquad \widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x}.$$

The fitted (sample regression) function is $\hat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$, and the residual is $\hat{u}_i := y_i - \hat{y}_i$.

The two equalities for $\widehat{\beta}_1$ coincide because $\sum_{i=1}^n (x_i - \bar{x})\bar{y} = \bar{y} \sum_{i=1}^n (x_i - \bar{x}) = 0$; more generally $\sum_{i=1}^n c\,(x_i - \bar{x}) = 0$ for any constant c, a trick we will reuse constantly.

The slope can also be written in terms of moments,

$$\widehat{\beta}_1 = \frac{\mathrm{Cov}(x, y)}{\mathrm{Var}(x)} = \widehat{\rho}_{xy} \cdot \frac{\widehat{\sigma}_y}{\widehat{\sigma}_x},$$

which shows that the sign of the slope is the sign of the sample correlation between x and y.

Theorem 1.4: Algebraic Properties of OLS

The OLS fit satisfies three identities by construction:

1. $\sum_{i=1}^{n} \hat{u}_i = 0$ (residuals sum to zero, hence have zero sample mean);
2. $\sum_{i=1}^{n} x_i \hat{u}_i = 0$ (regressor and residuals have zero sample covariance);
3. $\bar{y} = \widehat{\beta}_0 + \widehat{\beta}_1 \bar{x}$ (the point of means lies on the regression line).

These are not assumptions about the world; they hold in *every* sample because they are exactly the first-order conditions of the least-squares problem $\min_{b_0,b_1} \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$.

1.3 Goodness of Fit

How much of the variation in y does the line explain? Decompose the total variation into an explained part and a residual part.

Definition 1.5: Sums of Squares and R^2

$$\text{SST} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \quad \text{(total)}$$

$$\text{SSE} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \quad \text{(explained)}$$

$$\text{SSR} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad \text{(residual)}$$

With $\text{SST} = \text{SSE} + \text{SSR}$, the *coefficient of determination* is

$$R^2 = \frac{\text{SSE}}{\text{SST}} = 1 - \frac{\text{SSR}}{\text{SST}} \in [0, 1].$$

R^2 is the fraction of the sample variation in y explained by the model. Two cautions worth internalizing now, because they recur in the multiple regression model (Chapter 2):

- R^2 equals the squared sample correlation between y_i and its fitted value $\hat{y}_i$; it is a measure of fit, *not* of causality.
- A low R^2 does not invalidate the regression. We may estimate the slope β_1 very precisely even when most of the variation in y is left to the error.

1.4 Units of Measurement and Functional Form

The linear model is more flexible than it looks: by taking logs of y, of x, or of both, the *same* OLS machinery delivers different and often more natural interpretations of the slope. The four standard cases are summarized below; $\%\Delta$ denotes a percentage change.

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

The log–log slope is an elasticity, and logged slopes are invariant to rescaling the units of the logged variable — two reasons logs are ubiquitous in applied work. We return to functional form systematically in Chapter 5.

1.5 Expected Value and Variance of the OLS Estimators

We now ask the two questions that organize the rest of the course. To answer them we need a list of assumptions; the first four deliver unbiasedness, and the fifth (homoskedasticity) delivers the simple variance formula.

Assumption 1.6: SLR.1–SLR.5 (Gauss–Markov Assumptions, Simple Regression)

SLR.1 — Linear in parameters. The population model is $y = \beta_0 + \beta_1 x + u$.

SLR.2 — Random sampling. $\{(x_i, y_i) : i = 1, \dots, n\}$ is a random sample from the population, so $y_i = \beta_0 + \beta_1 x_i + u_i$.

SLR.3 — Sample variation in x. The x_i are not all equal: $\mathrm{SST}_x := \sum_{i=1}^{n} (x_i - \bar{x})^2 > 0$.

SLR.4 — Zero conditional mean. $\mathbb{E}(u \mid x) = 0$.

SLR.5 — Homoskedasticity. $\mathrm{Var}(u \mid x) = \sigma^2$ (the error variance does not depend on x).

Theorem 1.7: Unbiasedness of OLS

Under SLR.1–SLR.4, the OLS estimators are unbiased:

$$\mathbb{E}\left(\widehat{\beta}_1\right) = \beta_1, \qquad \mathbb{E}\left(\widehat{\beta}_0\right) = \beta_0.$$

Proof. Write the slope as a function of the errors. Substituting $y_i = \beta_0 + \beta_1 x_i + u_i$ into the

numerator and using $\sum_{i=1}^n (x_i - \bar{x}) = 0$,

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})\, y_i}{\mathrm{SST}_x} = \frac{\sum_{i=1}^n (x_i - \bar{x})(\beta_0 + \beta_1 x_i + u_i)}{\mathrm{SST}_x}$$
$$= \frac{\beta_1 \sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x}) + \sum_{i=1}^n (x_i - \bar{x}) u_i}{\mathrm{SST}_x} = \beta_1 + \frac{\sum_{i=1}^n (x_i - \bar{x}) u_i}{\mathrm{SST}_x}.$$

Conditioning on the sample of x's and applying SLR.4 ($\mathbb{E}(u_i \mid x) = 0$),

$$\mathbb{E}\left(\widehat{\beta}_1\right) = \beta_1 + \frac{1}{\mathrm{SST}_x} \sum_{i=1}^n (x_i - \bar{x})\, \mathbb{E}(u_i \mid x) = \beta_1.$$

For the intercept, $\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x} = \beta_0 + (\beta_1 - \widehat{\beta}_1)\bar{x} + \frac{1}{n}\sum_{i=1}^n u_i$, so $\mathbb{E}\left(\widehat{\beta}_0\right) = \beta_0$ by the result just shown and $\mathbb{E}(u_i) = 0$. □

If any of SLR.1–SLR.4 fails, unbiasedness generally fails too. The fragile one is SLR.4: whenever an omitted factor inside u is correlated with x, the slope is biased — the central theme of Chapter 2.

Theorem 1.8: Sampling Variance of OLS

Under SLR.1–SLR.5,

$$\mathrm{Var}\left(\widehat{\beta}_1\right) = \frac{\sigma^2}{\mathrm{SST}_x}, \qquad \mathrm{Var}\left(\widehat{\beta}_0\right) = \frac{\sigma^2\, n^{-1} \sum_{i=1}^n x_i^2}{\mathrm{SST}_x}.$$

Proof. From $\widehat{\beta}_1 = \beta_1 + \dfrac{\sum_{i=1}^n (x_i - \bar{x}) u_i}{\mathrm{SST}_x}$, treating the x's as fixed and using independence across i together with homoskedasticity $\mathrm{Var}(u_i \mid x) = \sigma^2$,

$$\mathrm{Var}\left(\widehat{\beta}_1\right) = \frac{1}{(\mathrm{SST}_x)^2} \sum_{i=1}^n (x_i - \bar{x})^2\, \mathrm{Var}(u_i \mid x) = \frac{\sigma^2}{\mathrm{SST}_x}.$$

□

Two messages are already visible in $\mathrm{Var}\left(\widehat{\beta}_1\right) = \sigma^2/\mathrm{SST}_x$: the slope is estimated more precisely when the error is small (σ^2 low) and when the regressor varies a lot (SST_x large). More data raises SST_x and so shrinks the variance.

The error variance σ^2 is unknown and must itself be estimated.

Definition 1.9: Error Variance Estimator and Standard Error

The unbiased estimator of σ^2 and the implied standard error of the slope are

$$\widehat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n \hat{u}_i^2, \qquad \widehat{\sigma} = \sqrt{\widehat{\sigma}^2}, \qquad \mathrm{se}(\widehat{\beta}_1) = \frac{\widehat{\sigma}}{\sqrt{\mathrm{SST}_x}}.$$

The divisor $n - 2$ corrects for the two parameters (β_0, β_1) already used to form the residuals; $\widehat{\sigma}$ is called the *standard error of the regression* (SER).

Remark (Where this is heading).

> We have, for one regressor, found the OLS estimator, shown it is unbiased under SLR.1–SLR.4, and derived its variance under the extra homoskedasticity assumption SLR.5. Chapter 2 repeats this exact program with many regressors, where the new and important phenomenon is omitted-variable bias; Chapter 7 returns to drop SLR.5 and asks what survives under heteroskedasticity.

Chapter 2

Multiple Regression: Estimation

The simple regression model of Chapter 1 rested on one fragile assumption: that the single regressor x is uncorrelated with everything else that affects y. In practice this almost never holds. Wages rise with education, but more-educated people also tend to have more able parents, more innate ability, and better schools — all of which raise wages on their own and all of which are correlated with education. A simple regression of wage on education cannot tell the effect of an extra year of school apart from the effect of the things that travel with it. The slope we estimate is contaminated.

Multiple regression is the tool that lets us hold those other factors fixed. By bringing the confounders into the model as additional regressors, we can ask the question we actually care about: what is the effect of x_1 on y, *comparing observations that are identical in* $x_2, \dots, x_k$? This is the meaning of "other things equal" in an empirical setting, and it is what gives a regression coefficient its causal reading.

This chapter develops the multiple linear regression model and its OLS estimator. The mechanics will look familiar — we minimize a sum of squared residuals, the estimators are unbiased under an analogous list of assumptions, and we again derive a sampling variance. What is genuinely new, and what occupies the heart of the chapter, is *omitted-variable bias*: a precise account of exactly what goes wrong, and in which direction, when we leave a relevant variable out. That formula is the engine behind most of the rest of the book; nearly every later technique exists to defuse some version of it.

2.1 The Multiple Regression Model

Definition 2.1: Multiple Linear Regression Model

The *multiple linear regression (MLR) model* relates a dependent variable y to k explanatory variables through

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u,$$

where β_0 is the intercept, $\beta_1, \dots, \beta_k$ are the slope parameters, and u is the error term collecting all factors other than $x_1, \dots, x_k$ that affect y.

The interpretation of each slope is the key idea of the entire chapter. The coefficient β_j is a *partial effect*: it measures the change in y associated with a one-unit change in x_j, *holding all other regressors and the error fixed*,

$$\Delta y = \beta_j \, \Delta x_j \qquad \text{when } \Delta x_\ell = 0 \text{ for all } \ell \neq j \text{ and } \Delta u = 0.$$

This is the "other things equal" or *ceteris paribus* effect that simple regression could not isolate. Note that the intercept β_0 — the predicted value of y when every regressor is zero — frequently has no sensible real-world meaning, and we rarely interpret it directly.

2.2 Ordinary Least Squares

Exactly as in the simple case, OLS is defined by imposing the sample analogues of two population moment conditions. Zero conditional mean of the error will deliver these, and OLS solves for $\beta_0, \beta_1, \ldots, \beta_k$ by setting the sample versions to zero.

The two moment conditions behind OLS

$$\mathbb{E}(u) = 0, \qquad \mathbb{E}(x_j u) = 0 \quad \text{for every } j = 1, \ldots, k.$$

The first fixes the intercept; the k orthogonality conditions $\mathbb{E}(x_j u) = 0$ fix the slopes. Together with the model equation, these $k+1$ conditions have sample analogues that pin down the $k+1$ OLS estimates uniquely.

Given a sample $\{(x_{i1}, \ldots, x_{ik}, y_i) : i = 1, \ldots, n\}$, the OLS estimators $\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_k$ are the values $b_0, \ldots, b_k$ that minimize the sum of squared residuals,

$$\min_{b_0, \ldots, b_k} \sum_{i=1}^{n} \left(y_i - b_0 - b_1 x_{i1} - \cdots - b_k x_{ik} \right)^2 .$$

The fitted values are $\hat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \cdots + \widehat{\beta}_k x_{ik}$ and the residuals are $\hat{u}_i := y_i - \hat{y}_i$. The first-order conditions of this problem are precisely the sample analogues of the moment conditions above.

2.2.1 The Partialling-Out Interpretation

With several regressors the closed-form expression for a single slope is no longer a one-line ratio. There is, however, a beautiful and revealing formula for it, due to Frisch and Waugh, that makes the "other things equal" interpretation literal.

Theorem 2.2: Partialling Out

The OLS slope $\widehat{\beta}_1$ in the regression of y on $x_1, \dots, x_k$ can be written as

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} \hat{r}_{i1}\, y_i}{\sum_{i=1}^{n} \hat{r}_{i1}^{2}},$$

where $\hat{r}_{i1}$ is the residual from regressing x_1 on all the *other* regressors $x_2, \dots, x_k$ (with an intercept). The same statement holds for each $\widehat{\beta}_j$, regressing x_j on the remaining regressors.

The construction has two steps. First, regress x_1 on $x_2, \dots, x_k$ and save the residuals $\hat{r}_{i1}$; these are the part of x_1 that cannot be explained by the other regressors — x_1 with the linear influence of $x_2, \dots, x_k$ *partialled out*. Second, regress y on those residuals. Because $\hat{r}_{i1}$ is by construction uncorrelated with $x_2, \dots, x_k$, the resulting slope captures only the relationship between y and the part of x_1 that is unique to x_1.

What $\widehat{\beta}_1$ measures

$\widehat{\beta}_1$ measures the sample relationship between y and x_1 *after the effects of all other regressors have been removed from* x_1. This is the algebraic content of "holding $x_2, \dots, x_k$ fixed." If x_1 were already uncorrelated with the other regressors, $\hat{r}_{i1}$ would equal $x_{i1} - \bar{x}_1$ and $\widehat{\beta}_1$ would coincide with the simple-regression slope; in general it does not.

The intercept does not require a separate minimization: like the simple-regression case, it follows from the fact that the OLS fit passes through the sample means,

$$\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x}_1 - \widehat{\beta}_2 \bar{x}_2 - \cdots - \widehat{\beta}_k \bar{x}_k.$$

2.2.2 Algebraic Properties

The following identities hold in *every* sample, regardless of whether any assumption about the population is true, because they are exactly the first-order conditions of the least-squares problem.

Theorem 2.3: Algebraic Properties of OLS

The multiple-regression OLS fit satisfies:

1. The residuals have zero sample mean: $\sum_{i=1}^{n} \hat{u}_i = 0$, hence $\bar{\hat{y}} = \bar{y}$.
2. Each regressor has zero sample covariance with the residuals: $\sum_{i=1}^{n} x_{ij}\,\hat{u}_i = 0$ for every $j = 1, \dots, k$. Consequently the fitted values are also uncorrelated with the residuals, $\sum_{i=1}^{n} \hat{y}_i\,\hat{u}_i = 0$.
3. The point of means $(\bar{x}_1, \bar{x}_2, \dots, \bar{x}_k, \bar{y})$ lies on the OLS regression surface: $\bar{y} = \widehat{\beta}_0 + \widehat{\beta}_1 \bar{x}_1 + \cdots + \widehat{\beta}_k \bar{x}_k$.

2.3 Expected Value and Unbiasedness

We now state the assumptions under which OLS recovers the population parameters on average. The first four parallel SLR.1–SLR.4 of Chapter 1; the only genuinely new one is MLR.3, which has no real bite in the one-regressor case.

Assumption 2.4: MLR.1–MLR.5 (Gauss–Markov Assumptions, Multiple Regression)

MLR.1 — Linear in parameters. The population model is $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$.

MLR.2 — Random sampling. $\{(x_{i1}, \dots, x_{ik}, y_i) : i = 1, \dots, n\}$ is a random sample from the population, so $y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + u_i$.

MLR.3 — No perfect collinearity. In the sample no regressor is constant, and there is no *exact* linear relationship among the regressors.

MLR.4 — Zero conditional mean. $\mathbb{E}(u \mid x_1, \dots, x_k) = 0$.

MLR.5 — Homoskedasticity. $\mathrm{Var}(u \mid x_1, \dots, x_k) = \sigma^2$, constant in the regressors.

MLR.3 deserves a word, because it is the one new assumption. It rules out an *exact* linear dependence among regressors, not a strong one; high-but-imperfect correlation (multicollinearity) is allowed and is treated separately below. The danger is subtle because collinearity can hide inside transformed variables: including both $\log(w)$ and $\log(w^2)$ violates MLR.3, since $\log(w^2) = 2\log(w)$ is an exact linear function of $\log(w)$. When MLR.3 fails, the OLS estimates are not even defined — there is no unique way to allocate a common movement between two regressors that always move together.

Theorem 2.5: Unbiasedness of OLS

Under MLR.1–MLR.4, the OLS estimators are unbiased for the population parameters:

$$\mathbb{E}\left(\widehat{\beta}_j\right) = \beta_j \qquad \text{for every } j = 0, 1, \ldots, k.$$

The proof generalizes the simple-regression argument: writing each $\widehat{\beta}_j$ via the partialling-out formula as β_j plus a linear combination of the errors, and applying MLR.4 conditional on the regressors, makes the error term vanish in expectation. As in Chapter 1, the fragile assumption is MLR.4: if any factor inside u is correlated with a regressor, MLR.4 fails and unbiasedness is generally lost. Quantifying exactly that failure is the subject of the next section.

2.4 Misspecified Models: Omitted-Variable Bias

This is the conceptual centerpiece of the chapter, and arguably of the first half of the course. We study two opposite mistakes — including a variable that does not belong, and excluding one that does — and find that they have very different consequences. Including an irrelevant variable costs us precision but not unbiasedness; omitting a relevant one biases our estimates, sometimes severely.

2.4.1 Including an Irrelevant Variable

Suppose the true model is $y = \beta_0 + \beta_1 x_1 + u$ but we estimate

$$\hat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2,$$

where x_2 has no effect on y — that is, its true coefficient is zero. Over-specifying the model in this way does *not* bias any OLS estimator: MLR.1–MLR.4 still hold (with the true $\beta_2 = 0$), so $\mathbb{E}\left(\widehat{\beta}_1\right) = \beta_1$ and $\mathbb{E}\left(\widehat{\beta}_2\right) = 0$. The price is paid in *variance*. As we will see in the variance formula below, adding x_2 can only raise (and generally does raise) the sampling variance of $\widehat{\beta}_1$ by introducing collinearity between x_1 and the new regressor. We therefore prefer not to include regressors that we have good reason to believe are irrelevant — not because they bias us, but because they make us less precise.

2.4.2 Omitting a Relevant Variable

Now the dangerous mistake. Suppose the correct model is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u, \qquad \beta_2 \neq 0,$$

but we leave x_2 out and run the *short* (misspecified) regression

$$\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1,$$

where we write $\tilde{\beta}_0, \tilde{\beta}_1$ for the short-regression estimator to distinguish it from the correct $\widehat{\beta}$ from the long regression. Let $\tilde{\delta}_1$ be the slope from regressing the omitted variable x_2 on the

included variable x_1,

$$\tilde{x}_2 = \tilde{\delta}_0 + \tilde{\delta}_1 x_1.$$

Then the short and long slope estimates are linked by an exact algebraic identity.

Theorem 2.6: Omitted-Variable Bias

With the notation above, the short-regression slope satisfies

$$\tilde{\beta}_1 = \widehat{\beta}_1 + \widehat{\beta}_2 \, \tilde{\delta}_1,$$

and therefore, taking expectations under the correct (long) model,

$$\text{Bias}(\tilde{\beta}_1) = \mathbb{E}\left(\tilde{\beta}_1\right) - \beta_1 = \beta_2 \, \tilde{\delta}_1.$$

Proof. By the partialling-out formula in the long regression, $\widehat{\beta}_1$ and $\widehat{\beta}_2$ are the OLS coefficients on x_1 and x_2. Regress x_2 on x_1 to get fitted values $\tilde{x}_{i2} = \tilde{\delta}_0 + \tilde{\delta}_1 x_{i1}$. Plugging the long fitted equation $\hat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \widehat{\beta}_2 x_{i2}$ into the short regression of y on x_1 and using the algebra of OLS gives

$$\tilde{\beta}_1 = \widehat{\beta}_1 + \widehat{\beta}_2 \, \tilde{\delta}_1.$$

Taking expectations conditional on the regressors and using $\mathbb{E}\left(\widehat{\beta}_1\right) = \beta_1$, $\mathbb{E}\left(\widehat{\beta}_2\right) = \beta_2$ from unbiasedness of the long regression yields $\mathbb{E}\left(\tilde{\beta}_1\right) = \beta_1 + \beta_2 \, \tilde{\delta}_1$, so $\text{Bias}(\tilde{\beta}_1) = \beta_2 \tilde{\delta}_1$. □

This little formula is worth memorizing, because it tells us not only that the short regression is biased but *in which direction*. The bias is the product of two pieces:

- β_2 — how the omitted variable affects y;
- $\tilde{\delta}_1$ — how the omitted variable correlates with the included variable, which has the same sign as $\text{Cov}(x_1, x_2)$.

The sign of the bias is the product of these two signs.

	$\text{Cov}(x_1, x_2) > 0$	$\text{Cov}(x_1, x_2) < 0$
$\beta_2 > 0$	Positive bias (upward)	Negative bias (downward)
$\beta_2 < 0$	Negative bias (downward)	Positive bias (upward)

A positive (*upward*) bias means $\mathbb{E}\left(\tilde{\beta}_1\right) > \beta_1$, so the short regression tends to overstate the effect; a negative (*downward*) bias means it understates the effect. When $\beta_1 > 0$ and the bias is negative — or $\beta_1 < 0$ and the bias is positive — the estimate is pulled *toward zero*, a case so common it earns its own name, *attenuation toward zero*. Note carefully that "upward" refers to the algebraic value of the estimate, not its magnitude: an upward bias on a negative true coefficient makes it less negative, i.e. moves it toward zero.

Example (The ability bias in a wage regression).

We want the return to education β_1 in $\log(\text{wage}) = \beta_0 + \beta_1 \,\text{educ} + \beta_2 \,\text{abil} + u$, but ability is unobserved and we run the short regression of log(wage) on education alone. Ability raises wages, so $\beta_2 > 0$; more able people tend to acquire more schooling, so $\text{Cov}(\text{educ}, \text{abil}) > 0$ and $\tilde{\delta}_1 > 0$.

Solution.

The bias is $\beta_2\tilde{\delta}_1 > 0$: the short regression has an *upward* (positive) bias. The naive estimate of the return to education overstates the true causal return, because it credits education with the wage gains that are really due to the higher ability that accompanies it. This is the classic "ability bias," and it is exactly why economists work so hard to control for or instrument out ability.

2.4.3 The General Lesson and Two Subtleties

The two-variable formula generalizes. In a model with many regressors, *if any included regressor is correlated with an omitted variable that belongs in the model (and hence is sitting in the error term), then in general all of the OLS slope estimates are biased*, not just the coefficient on the correlated regressor. The clean "only $\widehat{\beta}_1$ is biased" conclusion holds only in the special case where x_1 is the lone regressor correlated with the omission. This is why omitted-variable bias is so corrosive: a single confounder can contaminate an entire equation.

Two subtleties are worth stating precisely.

Remark (The bias is conditional on the sample).

The bias formula $\beta_2\tilde{\delta}_1$ is derived conditional on the sample values of the regressors, and $\tilde{\delta}_1$ is a *sample* regression coefficient. Even if x_1 and x_2 are *uncorrelated in the population* (so that the population analogue $\delta_1 = 0$), in any particular finite sample $\tilde{\delta}_1$ will generally not be exactly zero. Hence the short estimator can still be biased in a given sample. Population uncorrelatedness buys unbiasedness only in expectation over repeated sampling; it does not zero out the sample-specific tilt. (We return to the population version — inconsistency — in Chapter 4.)

The second subtlety is the bias–variance trade-off implicit in the choice of whether to include x_2:

- If $\beta_2 = 0$ (truly irrelevant), both $\tilde{\beta}_1$ and $\widehat{\beta}_1$ are unbiased, but

$$\text{Var}\left(\tilde{\beta}_1\right) = \frac{\sigma^2}{\text{SST}_1}, \qquad \text{Var}\left(\widehat{\beta}_1\right) = \frac{\sigma^2}{\text{SST}_1\,(1 - R_1^2)},$$

  so $\text{Var}\left(\tilde{\beta}_1\right) \leq \text{Var}\left(\widehat{\beta}_1\right)$ whenever x_1 has any sample correlation with x_2 (i.e. $R_1^2 > 0$). Including an irrelevant variable thus needlessly inflates the variance through multicollinearity — the precision cost flagged earlier.

- If $\beta_2 \neq 0$ (truly relevant), $\tilde{\beta}_1$ is biased while $\widehat{\beta}_1$ is not. The long regression's variance disadvantage can be overcome by collecting more data, but no amount of data removes the short regression's bias. We therefore generally prefer to include x_2.

2.5 Variance of OLS and Multicollinearity

Having established unbiasedness, we ask the second standard question: how precise is OLS? Adding the homoskedasticity assumption MLR.5 yields a clean formula.

Theorem 2.7: Sampling Variance of OLS

Under MLR.1–MLR.5, conditional on the sample values of the regressors,

$$\mathrm{Var}\left(\widehat{\beta}_j\right) = \frac{\sigma^2}{\mathrm{SST}_j\,(1 - R_j^2)} \qquad \text{for each } j = 1, \ldots, k,$$

where $\mathrm{SST}_j = \sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$ is the total sample variation in x_j, and R_j^2 is the R-squared from regressing x_j on all the other regressors.

Three forces drive the variance of $\widehat{\beta}_j$, all visible in the formula:

1. the error variance σ^2 — more unexplained noise in y makes every coefficient harder to pin down (numerator);
2. the total variation SST_j in the regressor — more spread in x_j sharpens the estimate, and more data raises SST_j (denominator);
3. the term $1 - R_j^2$ — how much of x_j is *linearly explained by the other regressors*. This is the genuinely new term relative to simple regression.

2.5.1 The Variance Inflation Factor

The third force has a name. As R_j^2 rises toward 1 — meaning x_j is nearly a linear combination of the other regressors — the factor $1/(1 - R_j^2)$ blows up and the variance of $\widehat{\beta}_j$ explodes. This is *multicollinearity*.

Definition 2.8: Variance Inflation Factor (VIF)

The *variance inflation factor* for regressor j is

$$\mathrm{VIF}_j = \frac{1}{1 - R_j^2},$$

the factor by which $\mathrm{Var}\left(\widehat{\beta}_j\right)$ is multiplied due to the linear association of x_j with the other regressors. A common (informal) rule of thumb flags $\mathrm{VIF}_j > 10$ (equivalently $R_j^2 > 0.9$) as a sign of serious multicollinearity.

It is important to be precise about what multicollinearity does and does not do.

What multicollinearity actually means

- It *inflates standard errors*: collinear coefficients are estimated imprecisely, are individually hard to make statistically significant, and can swing wildly across samples.

- It does *not* bias OLS or violate any Gauss–Markov assumption (MLR.3 forbids only *perfect* collinearity). OLS remains BLUE.
- It *cannot always be cured by more data* the way ordinary imprecision can: if two regressors are intrinsically near-duplicates, additional observations may add little independent variation.
- It afflicts the *individual* t-tests but not the *joint* F-test: an F-test of several collinear variables together is immune to the problem and can be highly significant even when no single coefficient is (see Chapter 3).

A coefficient that is imprecise because its regressor barely moves on its own is telling us something real — the data simply do not contain enough independent variation in x_j to separate its effect from the others. The remedy, when one exists, is more or better-targeted variation, not a statistical trick.

2.5.2 Estimating the Error Variance and Standard Errors

The variance formula depends on the unknown σ^2, which we must estimate from the residuals.

Definition 2.9: Error Variance Estimator and Standard Error

Under MLR.1–MLR.5, the unbiased estimator of σ^2 is

$$\widehat{\sigma}^2 = \frac{1}{n-k-1}\sum_{i=1}^{n}\hat{u}_i^2, \qquad \mathbb{E}\big(\widehat{\sigma}^2\big) = \sigma^2,$$

and $\widehat{\sigma} = \sqrt{\widehat{\sigma}^2}$ is the *standard error of the regression* (SER). The standard error of a slope coefficient is

$$\operatorname{se}(\widehat{\beta}_j) = \frac{\widehat{\sigma}}{\sqrt{\operatorname{SST}_j\,(1-R_j^2)}} = \frac{\widehat{\sigma}}{\sqrt{n}\,\operatorname{sd}(x_j)\sqrt{1-R_j^2}}.$$

The divisor $n-k-1$ is the degrees of freedom: we estimate $k+1$ parameters (the k slopes plus the intercept) before forming the residuals, and subtracting them makes $\widehat{\sigma}^2$ unbiased. This generalizes the $n-2$ of the one-regressor case in Chapter 1, which is exactly $n-k-1$ with $k=1$.

2.6 Efficiency: The Gauss–Markov Theorem

Unbiasedness alone does not single out OLS — many unbiased estimators exist. The case for OLS is that, among a natural and broad class, it is the most precise.

Theorem 2.10: Gauss–Markov Theorem

Under MLR.1–MLR.5, the OLS estimators $\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_k$ are the *best linear unbiased estimators* (BLUE) of $\beta_0, \beta_1, \ldots, \beta_k$: among all estimators that are linear in y and unbiased, OLS has the smallest variance.

Each word in "BLUE" carries content:

Linear. An estimator $\widehat{\theta}_j$ is linear if it can be written as a linear function of the sample values of the dependent variable, $\widehat{\theta}_j = \sum_{i=1}^{n} w_i y_i$, where the weights w_i may depend on the sample values of the regressors but not on y.

Unbiased. $\mathbb{E}\left(\widehat{\theta}_j\right) = \beta_j$ for every value of the parameters.

Best. Smallest sampling variance within the linear-unbiased class: $\operatorname{Var}\left(\widehat{\beta}_j\right) \leq \operatorname{Var}\left(\widehat{\theta}_j\right)$ for any other linear unbiased $\widehat{\theta}_j$.

The assumptions MLR.1–MLR.5 are collectively called the *Gauss–Markov assumptions*. The theorem is the formal justification for using OLS: as long as the model is correctly specified, sampling is random, there is no perfect collinearity, the error is mean-independent of the regressors, and the error variance is constant, no linear unbiased rule beats OLS. When homoskedasticity (MLR.5) fails, OLS loses the "best" property — a thread we pick up in Chapter 7.

2.7 Goodness of Fit

Finally, as in Chapter 1, we summarize how well the model fits with the coefficient of determination.

Definition 2.11: R^2 in Multiple Regression

With $\mathrm{SST} = \sum_{i=1}^{n}(y_i - \bar{y})^2$, $\mathrm{SSE} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$, and $\mathrm{SSR} = \sum_{i=1}^{n} \hat{u}_i^2$,

$$R^2 = \frac{\mathrm{SSE}}{\mathrm{SST}} = 1 - \frac{\mathrm{SSR}}{\mathrm{SST}} = \frac{\left[\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{y})\right]^2}{\left[\sum_{i=1}^{n}(y_i - \bar{y})^2\right]\left[\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2\right]}.$$

The last expression shows that R^2 equals the squared sample correlation between y_i and its fitted value $\hat{y}_i$.

Two cautions, both consequential, attach to R^2 in the multiple-regression setting.

Remark (R^2 never decreases when you add a regressor).

Adding any regressor to a model can only *increase* (or leave unchanged) the R^2, even if the new variable is pure noise. The reason is mechanical: OLS minimizes SSR over a larger parameter space, so the minimized SSR can only fall, and $R^2 = 1 - \mathrm{SSR}/\mathrm{SST}$ can only rise. It follows that a rise in R^2 is *not* a valid reason to keep a variable. The right criterion is whether the variable's partial effect is nonzero (a question of statistical significance, Chapter 3), not whether it nudges R^2 upward.

Because of this monotonicity, R^2 is a poor tool for comparing models with different numbers of regressors. The standard remedy is the *adjusted* R^2, which penalizes the addition of regressors via the degrees of freedom; we defer it to Chapter 5, where model comparison is treated systematically.

Remark (Where this is heading).

> We have built the multiple regression model, interpreted each coefficient as a partial effect via partialling out, shown OLS is unbiased under MLR.1–MLR.4 and BLUE under MLR.1–MLR.5, and — most importantly — derived the omitted-variable bias formula $\tilde{\beta}_1 = \widehat{\beta}_1 + \widehat{\beta}_2\tilde{\delta}_1$ that governs what goes wrong when MLR.4 fails. Chapter 3 adds the normality assumption MLR.6 to test hypotheses about these coefficients; Chapter 4 shows that the bias formula has a large-sample twin (inconsistency) that survives even without finite-sample assumptions; and the back half of the book is, in large part, a catalogue of repairs for a failing MLR.4.

Chapter 3

Multiple Regression: Inference

So far we have learned how to *estimate* the parameters of a multiple regression model and what we can say about the estimators across repeated samples: under the Gauss–Markov assumptions OLS is unbiased and, among linear unbiased estimators, has the smallest variance. But estimation is only half of empirical work. Having computed a number like $\widehat{\beta}_1 = 0.083$ for the return to a year of schooling, the practitioner wants to know whether that number is reliably different from zero, whether it is compatible with a theoretical value like 0.10, and how confident we should be in any of this. These are questions of *inference*: statements about the unknown population parameters that come with an explicit probability of error.

To make such statements precise we need to know the full *sampling distribution* of the OLS estimators, not just their first two moments. The Gauss–Markov assumptions pin down the mean and variance of $\widehat{\beta}_j$, but they say nothing about its shape; without a known distribution we cannot compute the probability that $\widehat{\beta}_j$ lands within a given distance of β_j. This chapter adds one final assumption — that the error is normally distributed — which delivers an exact normal sampling distribution and, through it, the t and F statistics that dominate applied econometrics. We develop the t test for a single coefficient and its equivalence with confidence intervals, the trick that turns a test of a linear combination of parameters into an ordinary t test, and the F test for several restrictions at once. We close by separating two ideas that are constantly confused: *statistical* significance, which is about precision, and *economic* significance, which is about magnitude.

3.1 The Normal Sampling Distribution

Conditional on the sample values of the regressors, the sampling distribution of each $\widehat{\beta}_j$ is inherited entirely from the distribution of the unobserved errors. In Chapter 2 we saw that $\widehat{\beta}_j$ is a linear function of the y_i, hence a linear function of the errors u_i; its mean and variance therefore followed from the mean and variance of u. To know its full distribution we must say something about the *shape* of the error distribution. The simplest tractable choice, and the one that makes the finite-sample theory exact, is normality.

Assumption 3.1: MLR.6 (Normality)

The population error u is independent of the explanatory variables $x_1, x_2, \ldots, x_k$ and is normally distributed with zero mean and variance σ^2:

$$u \sim N(0, \sigma^2).$$

This is a strong assumption. Independence of u from the regressors is more than the zero conditional mean MLR.4 and the constant conditional variance MLR.5; it fixes the *entire* conditional distribution of u, not just its first two moments. Consequently MLR.6 *implies* both MLR.4 ($\mathbb{E}(u \mid \mathbf{x}) = 0$) and MLR.5 ($\mathrm{Var}(u \mid \mathbf{x}) = \sigma^2$), and adds the assumption of a normal shape on top of them.

Definition 3.2: Classical Linear Model (CLM)

The assumptions MLR.1 through MLR.6 are called the *classical linear model (CLM)* assumptions. Under the CLM, the conditional distribution of the dependent variable given the regressors is

$$y \mid \mathbf{x} \;\sim\; N\big(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k,\; \sigma^2\big)\,.$$

In words, the CLM says that for each fixed configuration of the regressors, y is normally distributed about the population regression function $\beta_0+\beta_1 x_1+\cdots+\beta_k x_k$ with a variance σ^2 that is the same at every configuration. The population regression function fixes the mean of this normal distribution; homoskedasticity fixes its spread; normality fixes its shape.

Remark (Why normality is plausible — and when it fails).

> The error u is the sum of the many unobserved factors that influence y. If these factors are numerous, have comparable individual effects, and are roughly independent, a central limit argument suggests that their sum is approximately normal. The approximation is only as good as those premises: a single dominant factor, strongly skewed components, or strong dependence among them can all spoil it. Variables that are bounded, count-valued, or sharply skewed (wages, firm sizes) often have decidedly non-normal errors. This is exactly why Chapter 4 develops large-sample inference, which dispenses with MLR.6 entirely: as $n \to \infty$ the OLS estimators are approximately normal regardless of the error distribution.

Under the CLM we can strengthen what we know about OLS in two ways. First, its efficiency improves: under the Gauss–Markov assumptions OLS is best only within the class of *linear* unbiased estimators (the BLUE property of Chapter 2); under the additional normality of MLR.6, OLS is the *minimum variance unbiased estimator*, best among *all* unbiased estimators, linear or not. Second, and central to this chapter, we obtain its exact sampling distribution.

Theorem 3.3: Normal Sampling Distribution

Under the CLM assumptions MLR.1–MLR.6, conditional on the sample values of the independent variables,

$$\widehat{\beta}_j \;\sim\; N\Big(\beta_j,\; \mathrm{Var}\Big(\widehat{\beta}_j\Big)\Big), \qquad j = 0, 1, \ldots, k,$$

and therefore the standardized estimator is standard normal:

$$\frac{\widehat{\beta}_j - \beta_j}{\mathrm{sd}(\widehat{\beta}_j)} \;\sim\; N(0,1), \qquad \mathrm{sd}(\widehat{\beta}_j) = \sqrt{\mathrm{Var}\Big(\widehat{\beta}_j\Big)}.$$

Proof. Each OLS estimator can be written as a linear combination of the errors,

$$\widehat{\beta}_j = \beta_j + \sum_{i=1}^{n} w_{ij}\, u_i,$$

where the weights w_{ij} are (nonrandom) functions of the sample values of the regressors alone — for the slopes, $w_{ij} = \hat{r}_{ij} / \sum_i \hat{r}_{ij}^2$ in the partialling-out notation of Chapter 2. Conditional on the regressors the weights are constants. Under MLR.6 the u_i are independent $N(0, \sigma^2)$ variables, and any linear combination of independent normal random variables is itself normal. Hence $\widehat{\beta}_j$ is normal; its mean β_j and variance $\mathrm{Var}\Big(\widehat{\beta}_j\Big)$ are exactly the Gauss–Markov values derived in Chapter 2. Standardizing gives the $N(0,1)$ result. □

The same argument shows more: *any* linear combination of the $\widehat{\beta}_j$ is also exactly normal, because it too is a linear combination of the independent normal errors. We will exploit this repeatedly below.

3.2 Testing a Single Parameter: The t Test

The standard normal result above is not yet usable, because $\mathrm{sd}(\widehat{\beta}_j)$ contains the unknown error variance σ^2. Replacing σ by its estimator $\widehat{\sigma}$ (and hence $\mathrm{sd}(\widehat{\beta}_j)$ by the standard error $\mathrm{se}(\widehat{\beta}_j)$ from Chapter 2) changes the distribution from normal to Student's t, because we have introduced extra sampling noise through $\widehat{\sigma}$.

Theorem 3.4: t Distribution of the Standardized Estimator

Under the CLM assumptions MLR.1–MLR.6,

$$\frac{\widehat{\beta}_j - \beta_j}{\mathrm{se}(\widehat{\beta}_j)} \;\sim\; t_{\,n-k-1},$$

where $n-k-1$ is the residual degrees of freedom and $k+1$ is the number of parameters (intercept plus k slopes) estimated in the model.

The degrees of freedom $n - k - 1$ are exactly the divisor in the unbiased variance estimator $\widehat{\sigma}^2 = \frac{1}{n-k-1}\sum_{i=1}^{n} \hat{u}_i^2$: we lose one degree of freedom for each of the $k+1$ estimated

parameters. As $n-k-1$ grows the t distribution approaches the standard normal, which is why large-sample t tests are conducted using normal critical values.

3.2.1 Carrying out the test

To test a hypothesis about a single coefficient we form the t statistic by plugging the hypothesized value into the standardized ratio. The most common null is that β_j equals zero,

$$H_0: \ \beta_j = 0,$$

under which x_j has no *partial* effect on y once the other regressors are controlled for. The corresponding statistic is simply the coefficient divided by its standard error,

$$t_{\widehat{\beta}_j} = \frac{\widehat{\beta}_j}{\operatorname{se}(\widehat{\beta}_j)},$$

which under H_0 follows t_{n-k-1}. We reject H_0 in favor of the alternative when $|t_{\widehat{\beta}_j}|$ (for a two-sided alternative) or $t_{\widehat{\beta}_j}$ (for a one-sided alternative) exceeds the relevant critical value, or equivalently when the p-value falls below the chosen significance level.

Remark (Reading the t test correctly).

Several points trip up beginners and deserve to be stated once and for all.

- The null need not be zero. To test whether the coefficient equals some theoretical value a_j, use $H_0: \ \beta_j = a_j$ and the statistic $t = (\widehat{\beta}_j - a_j)/\operatorname{se}(\widehat{\beta}_j)$.
- Match the tails. A two-sided alternative $H_1: \ \beta_j \neq a_j$ uses the absolute value of t and a two-tailed critical value; a one-sided alternative ($H_1: \ \beta_j > a_j$ or $H_1: \ \beta_j < a_j$) uses the signed statistic and a one-tailed critical value. The p-value and critical value must be computed for the alternative actually being tested.
- In large samples the t test may be carried out using standard normal critical values.
- "Fail to reject H_0" is not the same as "accept H_0." Failing to reject means the data are consistent with H_0, not that H_0 is true; many other values of β_j would also fail to be rejected.

3.2.2 Confidence intervals and the duality with testing

A confidence interval inverts the test: instead of fixing one null value and asking whether the data reject it, it reports the entire set of null values the data would *not* reject. Under the CLM the $(1-\alpha)$ confidence interval for β_j is

$$\widehat{\beta}_j \ \pm \ c_{\alpha/2} \cdot \operatorname{se}(\widehat{\beta}_j),$$

where $c_{\alpha/2}$ is the two-tailed critical value from t_{n-k-1} (for the usual 95% interval and moderate degrees of freedom, $c_{\alpha/2} \approx 1.96$).

Testing and confidence intervals are equivalent

A value a_j lies *outside* the $(1-\alpha)$ confidence interval for β_j if and only if the two-sided t test of $H_0:\ \beta_j = a_j$ rejects at level α. The interval is therefore a complete summary of every two-sided test at once: one can read off acceptance or rejection of any null directly, without recomputing a statistic.

This duality is the reason confidence intervals are often preferred for reporting: a single interval communicates both the point estimate and the full range of values compatible with the data, and it answers every two-sided hypothesis test simultaneously.

3.3 Testing a Linear Combination of Parameters

Often the hypothesis of interest involves *several* coefficients tied together by one linear relation — for example, that the return to a year at a junior college equals the return to a year at a university, or that two elasticities are equal. The obstacle is that the relevant standard error is not printed by the regression software. A clever reparametrization sidesteps the difficulty by turning the question into an ordinary single-coefficient t test.

3.3.1 The reparametrization trick

Consider the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u,$$

and suppose we wish to test a linear restriction relating the two slopes,

$$H_0:\ \beta_1 = p + q\,\beta_2,$$

for known constants p and q. The idea is to express the restriction with a single new parameter θ that is zero exactly when H_0 holds. Write, with full generality,

$$\beta_1 = p + q\,\beta_2 + \theta, \qquad \text{so that} \qquad H_0:\ \theta = 0.$$

Substituting this into the regression equation and collecting terms,

$$\begin{aligned} y &= \beta_0 + (p + q\,\beta_2 + \theta)\,x_1 + \beta_2 x_2 + u \\ &= \beta_0 + p\,x_1 + \theta\,x_1 + \beta_2(q\,x_1 + x_2) + u, \end{aligned}$$

which rearranges to

$$\underbrace{y - p\,x_1}_{\text{new dependent variable}} = \beta_0 + \theta\,x_1 + \beta_2\,\underbrace{(q\,x_1 + x_2)}_{\text{new regressor}} + u.$$

Now run OLS of the transformed dependent variable $y - p\,x_1$ on x_1 and the transformed regressor $q\,x_1 + x_2$. The coefficient on x_1 in this regression *is* θ, and the original null $H_0:\ \beta_1 = p + q\beta_2$ is now the elementary null $H_0:\ \theta = 0$. We test it with the standard t statistic $\widehat{\theta}/\operatorname{se}(\widehat{\theta})$ that the software reports directly.

The payoff is twofold. The regression delivers $\widehat{\theta} = \widehat{\beta}_1 - p - q\,\widehat{\beta}_2$ (a point estimate of the linear combination) and, crucially, its standard error $\mathrm{se}(\widehat{\theta}) = \mathrm{se}(\widehat{\beta}_1 - q\,\widehat{\beta}_2)$ *automatically* — a quantity that would otherwise be awkward to compute by hand, because it requires the covariance between two estimated coefficients.

Example (Equal returns to two-year and four-year college).

> Let $\log(\text{wage}) = \beta_0 + \beta_1\,\text{jc} + \beta_2\,\text{univ} + \beta_3\,\text{exper} + u$, where jc and univ are years at a junior college and at a university. To test whether a year of each yields the same return, $H_0:\ \beta_1 = \beta_2$, take $p = 0$ and $q = 1$: define $\theta = \beta_1 - \beta_2$, regress log(wage) on jc, the combined variable jc + univ, and exper, and read the t statistic on jc. A negative, significant $\widehat{\theta}$ would say a junior-college year is worth less than a university year.

3.3.2 The covariance of two coefficients

To see why the standard error of $\widehat{\beta}_1 - q\widehat{\beta}_2$ requires the covariance $\mathrm{Cov}\left(\widehat{\beta}_1, \widehat{\beta}_2\right)$, expand the variance of the linear combination:

$$\mathrm{Var}\left(\widehat{\beta}_1 - q\,\widehat{\beta}_2\right) = \mathrm{Var}\left(\widehat{\beta}_1\right) + q^2\ \mathrm{Var}\left(\widehat{\beta}_2\right) - 2q\ \mathrm{Cov}\left(\widehat{\beta}_1, \widehat{\beta}_2\right).$$

The two individual variances are printed by the software, but the cross term is not. We compute it directly, using the fact (from Chapter 2) that each slope estimator is a linear function of the y_i written through the partialling-out residuals $\hat{r}_{ij}$ — the residuals from regressing x_j on all the other regressors:

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n \hat{r}_{i1}\,y_i}{\sum_{i=1}^n \hat{r}_{i1}^2}, \qquad \widehat{\beta}_2 = \frac{\sum_{i=1}^n \hat{r}_{i2}\,y_i}{\sum_{i=1}^n \hat{r}_{i2}^2}.$$

Because the observations are independently sampled, the y_i are mutually independent (conditional on the regressors), each with variance $\mathrm{Var}(y_i \mid \mathbf{x}) = \sigma^2$. Bilinearity of covariance then gives

$$\begin{aligned}
\mathrm{Cov}\left(\widehat{\beta}_1, \widehat{\beta}_2\right) &= \mathrm{Cov}\left(\frac{\sum_{i=1}^n \hat{r}_{i1} y_i}{\sum_{i=1}^n \hat{r}_{i1}^2}, \frac{\sum_{i=1}^n \hat{r}_{i2} y_i}{\sum_{i=1}^n \hat{r}_{i2}^2}\right) \\
&= \frac{1}{\left(\sum_{i=1}^n \hat{r}_{i1}^2\right)\left(\sum_{i=1}^n \hat{r}_{i2}^2\right)}\ \mathrm{Cov}\left(\sum_{i=1}^n \hat{r}_{i1} y_i, \sum_{i=1}^n \hat{r}_{i2} y_i\right) \\
&= \frac{1}{\left(\sum_{i=1}^n \hat{r}_{i1}^2\right)\left(\sum_{i=1}^n \hat{r}_{i2}^2\right)}\ \sum_{i=1}^n \hat{r}_{i1}\,\hat{r}_{i2}\ \mathrm{Var}(y_i) \\
&= \sigma^2\,\frac{\sum_{i=1}^n \hat{r}_{i1}\,\hat{r}_{i2}}{\left(\sum_{i=1}^n \hat{r}_{i1}^2\right)\left(\sum_{i=1}^n \hat{r}_{i2}^2\right)},
\end{aligned}$$

where the cross terms $\mathrm{Cov}(y_i, y_j) = 0$ for $i \neq j$ vanish by independence. The same idea gives a single tidy formula for any weighted combination of the two slopes. Writing $a_{ij} := \hat{r}_{ij} / \sum_l \hat{r}_{lj}^2$, so that $\widehat{\beta}_j = \sum_i a_{ij}\,y_i$, the combination $w_1\widehat{\beta}_1 + w_2\widehat{\beta}_2 = \sum_{i=1}^n (w_1 a_{i1} + w_2 a_{i2})\,y_i$

is again a linear function of the independent y_i, hence

$$\operatorname{Var}\left(w_1\widehat{\beta}_1 + w_2\widehat{\beta}_2\right) = \sigma^2 \sum_{i=1}^{n} \left(w_1\, a_{i1} + w_2\, a_{i2}\right)^2 .$$

Expanding the square reproduces $w_1^2 \operatorname{Var}\left(\widehat{\beta}_1\right) + w_2^2 \operatorname{Var}\left(\widehat{\beta}_2\right) + 2w_1w_2 \operatorname{Cov}\left(\widehat{\beta}_1, \widehat{\beta}_2\right)$ with the variances and covariance just derived. In practice one never plugs into this by hand; the reparametrization of the previous subsection lets OLS compute $\operatorname{se}(\widehat{\theta})$ for us. The point of the derivation is conceptual: the standard error of a *difference* of coefficients depends on how those coefficients covary, and ignoring the covariance term — treating the two estimates as independent — generally gives the wrong answer.

Remark (General linear combinations).

> For a general linear combination $\sum_j w_j\widehat{\beta}_j$ of several coefficients, the variance is
>
> $$\operatorname{Var}\left(\textstyle\sum_j w_j\widehat{\beta}_j\right) = \displaystyle\sum_j w_j^2 \operatorname{Var}\left(\widehat{\beta}_j\right) + \sum_{j\neq l} w_j w_l \operatorname{Cov}\left(\widehat{\beta}_j, \widehat{\beta}_l\right),$$
>
> with each covariance computed exactly as above. Equivalently, in matrix form $\operatorname{Var}\left(\mathbf{w}'\widehat{\beta}\right) = \mathbf{w}' \operatorname{Var}\left(\widehat{\beta}\right) \mathbf{w}$, where $\operatorname{Var}\left(\widehat{\beta}\right) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ is the full covariance matrix of the coefficient vector. The reparametrization trick remains the practical route whenever the combination can be folded into a single redefined regressor.

3.4 Testing Multiple Linear Restrictions: The F Test

A single t statistic tests one restriction. To test *several* restrictions simultaneously — "do these three variables jointly belong in the model?" — we need a statistic that evaluates all the restrictions at once. This is the F test.

3.4.1 Restricted and unrestricted models

A null that sets one population parameter to zero, say $H_0 : \beta_k = 0$, is an *exclusion restriction*: it excludes the corresponding variable. A null that constrains several parameters at once is a set of *multiple restrictions*, and the test of such a null is a *joint* (or multiple) hypothesis test. The leading case is a joint exclusion null,

$$H_0 : \beta_{k-q+1} = \beta_{k-q+2} = \cdots = \beta_k = 0,$$

which asserts that a designated group of q regressors has no partial effect on y as a group.

Imposing the restrictions gives the *restricted model*, which drops the q variables; the full model with all regressors is the *unrestricted model*. The restricted model is *nested* inside the unrestricted one — it is the special case obtained by forcing q coefficients to zero — and the difference in the number of estimated parameters between the two models is exactly q, the number of restrictions being tested.

3.4.2 The F statistic in SSR form

Dropping variables can never lower the sum of squared residuals: a restricted model fits no better than the model that is free to use the extra variables, so $\text{SSR}_r \geq \text{SSR}_{ur}$ always. The question is whether the *increase* $\text{SSR}_r - \text{SSR}_{ur}$ is large enough — relative to how well the unrestricted model fits — to be incompatible with the restrictions being true. The F statistic scales this increase appropriately.

Theorem 3.5: The F Statistic (SSR Form)

Under the CLM assumptions, with q restrictions and k regressors in the unrestricted model,

$$F = \frac{(\text{SSR}_r - \text{SSR}_{ur})/q}{\text{SSR}_{ur}/(n-k-1)} \;\sim\; F_{q,\ n-k-1} \qquad \text{under } H_0.$$

The numerator is the increase in unexplained variation per restriction; the denominator is $\widehat{\sigma}^2$ from the unrestricted model, i.e. the unbiased estimate of the unrestricted error variance. The numerator degrees of freedom q is the number of restrictions (equivalently, the difference in degrees of freedom between the two models), and the denominator degrees of freedom $n-k-1$ is the residual degrees of freedom of the unrestricted model. Because $\text{SSR}_r \geq \text{SSR}_{ur}$, the statistic is always nonnegative, and large values are evidence against H_0. We reject when F exceeds the upper-tail critical value of the $F_{q,\,n-k-1}$ distribution.

What rejection means

If F exceeds its critical value we reject H_0 and conclude that the excluded variables are *jointly statistically significant* at the chosen level. If we cannot reject, the group is jointly insignificant — the data do not establish that those variables, taken together, belong in the model. The SSR form is completely general: it is valid whether or not the restricted and unrestricted models share the same dependent variable, since it is built only from sums of squared residuals.

3.4.3 The F statistic in R^2 form

When the restricted and unrestricted models have the *same* dependent variable — as in any joint exclusion test — the total sum of squares SST is identical across the two models, so we may divide numerator and denominator by it and rewrite the statistic in terms of the R^2's. Using $\text{SSR} = \text{SST}\,(1 - R^2)$,

$$F = \frac{(R^2_{ur} - R^2_r)/q}{(1 - R^2_{ur})/(n-k-1)} \;\sim\; F_{q,\ n-k-1}.$$

This form is convenient because regression output reports R^2 directly. Note carefully that the denominator is $(1-R^2_{ur})$, the unexplained fraction of variation in the unrestricted model, divided by its degrees of freedom — *not* $(1-R_{ur})^2$. The R^2 form is restricted to the equal-dependent-variable case; whenever the dependent variable changes (e.g. a log transformation under one model), only the SSR form applies.

3.4.4 Overall significance of the regression

A special case worth singling out is the test that *no* regressor matters,

$$H_0: \ \beta_1 = \beta_2 = \cdots = \beta_k = 0,$$

the test of overall significance of the regression. Here the restricted model contains only the intercept, so $R_r^2 = 0$ and $q = k$. The R^2 form collapses to

$$F = \frac{R^2/k}{(1-R^2)/(n-k-1)} \ \sim \ F_{k,\ n-k-1}.$$

This statistic appears in nearly every regression printout. Rejecting H_0 says only that *at least one* regressor has explanatory power — it does not identify which.

3.4.5 F versus t, and joint versus individual significance

The t and F tests are intimately related but not interchangeable.

Joint significance can survive when individual significance does not

When two or more regressors are highly correlated (multicollinearity), each individual coefficient is estimated imprecisely, so the individual t statistics may all be insignificant even though the variables clearly matter as a group. The F test, which examines the variables *jointly*, is essentially immune to this: it asks whether the group as a whole improves the fit, a question multicollinearity does not obscure. It is entirely possible — and common — for an F test to reject the joint null while no single t statistic is significant.

The reason is geometric. Multicollinearity inflates the variance of each $\widehat{\beta}_j$ (recall $\mathrm{Var}\left(\widehat{\beta}_j\right) = \sigma^2/[\mathrm{SST}_j(1-R_j^2)]$ from Chapter 2, which blows up as $R_j^2 \to 1$) and induces a strong negative covariance between the correlated coefficients, so that each is poorly determined alone while their joint contribution is well determined. The F statistic accounts for this covariance automatically; the separate t statistics do not.

For a *single* restriction the two tests coincide: the square of the t statistic equals the F statistic ($t^2 = F$ with $q = 1$), and the two-sided t test is exactly equivalent to the corresponding F test. The t statistic retains two practical advantages, however: it handles *one-sided* alternatives, which the (inherently two-sided, nonnegative) F statistic cannot, and it is more directly available, being printed automatically for every coefficient.

3.4.6 General linear restrictions

The F machinery is not limited to exclusion restrictions. Any set of linear restrictions can be reduced to exclusion restrictions by the same reparametrization used for the single-restriction t test. A general linear restriction has the form $\beta_1 = f(\boldsymbol{\beta})$, where f is a linear function of the other parameters. Writing it most generally as $\beta_1 = f(\boldsymbol{\beta}) + \theta$ and substituting into the equation, the restriction H_0 becomes $\theta = 0$:

- additive constants in f get absorbed into a redefined dependent variable;
- the linear coefficients in f get absorbed into redefined regressors (combinations of the original ones).

After this transformation the joint null is a set of plain exclusion restrictions on the new parameters, and the ordinary SSR-form F test applies. Thus exclusion restrictions are the canonical case to which all linear restrictions reduce.

3.5 Economic versus Statistical Significance

A coefficient can be statistically significant and economically trivial, or economically large and statistically uncertain. Keeping the two notions distinct is one of the most important habits in applied work.

Two different questions

The *economic* (or *practical*) significance of a variable is determined entirely by the *size and sign* of its coefficient: how much does y move, in units that matter, for a realistic change in x_j? The *statistical* significance is determined entirely by the *size of the t statistic*: is the coefficient precisely enough estimated to rule out zero? A large coefficient with a small t statistic is economically important but statistically shaky; a tiny coefficient with a huge t statistic (common in very large samples) is statistically significant but may be economically negligible.

Several practical guidelines follow.

- As the sample grows, smaller significance levels are warranted. Standard errors shrink with n, so even economically trivial effects eventually become statistically significant; using a stricter level offsets this and makes economic and statistical significance more likely to agree.
- Check statistical significance first, but interpret economic significance in light of how the variable enters the equation — its units, and whether it appears in levels, logs, or interactions, all change what "the size of the coefficient" means.
- A variable with the wrong sign or an unexpected magnitude that is statistically insignificant can usually be set aside, but it may also be a symptom of a deeper specification problem — an omitted variable, a wrong functional form, or multicollinearity — and is worth investigating rather than ignoring.

Remark (Where this is heading).

> We have added normality (MLR.6) to complete the classical linear model, obtained the exact normal sampling distribution of OLS, and built the t and F tests on top of it, along with the reparametrization device that handles any linear hypothesis. The exactness of these results rests on MLR.6, which is often only approximately true. Chapter 4 shows that the same t and F procedures remain valid in large samples *without* normality, by appeal to the central limit theorem; Chapter 7 revisits inference once the homoskedasticity

assumption MLR.5 is dropped.

Chapter 4

Large-Sample Properties of OLS

So far we have judged the OLS estimator by its *finite-sample* behavior: under the Gauss–Markov assumptions it is unbiased (Chapter 2), it has the smallest variance among linear unbiased estimators, and — once we add normality of the errors — its t and F statistics have exact t and F distributions (Chapter 3). Those are strong guarantees, but they come at a price. Unbiasedness leans on the zero conditional mean assumption MLR.4, which is genuinely demanding; and the exact distributions of t and F lean on the normality assumption MLR.6, which real data rarely honor.

This chapter asks a different and more forgiving question: what can we say about OLS when the sample is large? Two ideas organize the answer. The first is *consistency* — as the sample grows, $\widehat{\beta}_j$ collapses onto the true β_j. We will see that consistency survives even when unbiasedness does not: it needs only a weaker, zero-*correlation* version of MLR.4, and it lets us define an asymptotic analogue of omitted-variable bias. The second is *asymptotic normality* — even when the errors are not normal, the central limit theorem makes $\widehat{\beta}_j$ approximately normal in large samples, so the usual t and F inference is approximately valid *without* assuming MLR.6. The cost is that all of these statements are approximations that improve as n grows, with the estimated variance of $\widehat{\beta}_j$ shrinking to zero at the rate $1/n$.

Throughout, keep one distinction firmly in mind, because it is the conceptual heart of the chapter: *bias* is a statement about averaging over many hypothetical samples of fixed size, while *inconsistency* is a statement about what happens to a *single* estimate as the one sample we have grows without bound. They usually point the same way, but they are not the same thing, and the algebra that describes them differs in a subtle way we will make precise.

4.1 Consistency of OLS

Consistency is the most basic requirement we can ask of an estimator. An estimator that is not even consistent is using more data to home in on the wrong answer, and no sample size can rescue it. We recall the definition in the form we will use.

Definition 4.1: Consistency

An estimator $\widehat{\theta}_n$ of a parameter θ (computed from a sample of size n) is *consistent* if it converges in probability to θ:

$$\widehat{\theta}_n \xrightarrow{p} \theta, \qquad \text{i.e.} \qquad \operatorname{plim}_{n\to\infty} \widehat{\theta}_n = \theta.$$

Equivalently, for every $\varepsilon > 0$, $\mathrm{P}\left(\left|\widehat{\theta}_n - \theta\right| > \varepsilon\right) \to 0$ as $n \to \infty$: with probability approaching one, $\widehat{\theta}_n$ lies arbitrarily close to θ.

The central result of this section is that OLS is consistent under exactly the assumptions that already delivered unbiasedness in Chapter 2.

Theorem 4.2: Consistency of OLS

Under assumptions MLR.1 through MLR.4 (linear in parameters, random sampling, no perfect collinearity, and zero conditional mean $\mathbb{E}(u \mid x_1, \dots, x_k) = 0$), the OLS estimators are consistent: for every $j = 0, 1, \dots, k$,

$$\operatorname{plim}_{n\to\infty} \widehat{\beta}_j = \beta_j.$$

It is instructive to prove this in the simple regression model, where the algebra is transparent and the mechanism — a law of large numbers acting on sample moments — is laid bare.

Proof (simple regression). Take the model $y = \beta_0 + \beta_1 x_1 + u$ with $\mathbb{E}(u) = 0$ and $\operatorname{Cov}(x_1, u) = 0$. As in Chapter 1, write the slope estimator in error form:

$$\widehat{\beta}_1 = \beta_1 + \frac{n^{-1}\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)\, u_i}{n^{-1}\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2}.$$

By the law of large numbers each sample average converges in probability to its population counterpart: the numerator $\xrightarrow{p} \operatorname{Cov}(x_1, u)$ and the denominator $\xrightarrow{p} \operatorname{Var}(x_1)$. Since $\operatorname{Var}(x_1) > 0$ (the population analogue of MLR.3), Slutsky's theorem lets us pass the plim through the ratio:

$$\operatorname{plim} \widehat{\beta}_1 = \beta_1 + \frac{\operatorname{Cov}(x_1, u)}{\operatorname{Var}(x_1)} = \beta_1 + \frac{0}{\operatorname{Var}(x_1)} = \beta_1.$$

The same argument applied to the general k-regressor case (using the partialled-out residual $\hat{r}_{ij}$ in place of $x_{i1} - \bar{x}_1$) gives $\operatorname{plim} \widehat{\beta}_j = \beta_j$ for every j. □

The proof exposes the engine of consistency: OLS replaces the *population* moment condition $\operatorname{Cov}(x_1, u) = 0$ with its *sample* analogue, and the law of large numbers guarantees that the sample analogue converges to the truth. Nothing in this argument required the errors to be normal, homoskedastic, or even to have a zero conditional mean at every value of x — only that, in the population, the regressors are uncorrelated with the error.

4.1.1 The Weaker Condition for Consistency: MLR.4-prime

The proof above used much less than the full force of MLR.4. It never needed $\mathbb{E}(u \mid x_1, \ldots, x_k) = 0$ for every configuration of the regressors; it needed only that u has mean zero and is *uncorrelated* with each regressor. This motivates a deliberately weaker assumption.

Assumption 4.3: MLR.4′ (Zero Mean and Zero Correlation)

The error has mean zero and is uncorrelated with each regressor:

$$\mathbb{E}(u) = 0 \qquad \text{and} \qquad \operatorname{Cov}(x_j, u) = 0 \quad \text{for all } j = 1, 2, \ldots, k.$$

Assumption MLR.4′ is strictly weaker than MLR.4: zero conditional mean implies both $\mathbb{E}(u) = 0$ and $\operatorname{Cov}(x_j, u) = 0$, but not conversely. The two are worth comparing carefully, because the difference is exactly the difference between consistency and unbiasedness.

MLR.4 versus MLR.4′

- **MLR.4′ is the natural condition for consistency.** It is precisely the population moment condition that OLS imposes in the sample, so it is the minimal requirement under which $\operatorname{plim} \widehat{\beta}_j = \beta_j$. Under MLR.4′ alone, however, OLS may be *biased in finite samples* even though it is consistent.
- **MLR.4 is stronger, and buys finite-sample properties.** Zero conditional mean is what we need to establish exact, sample-size-n properties such as unbiasedness (Chapter 2) and the Gauss–Markov optimality of OLS.
- **MLR.4 says the model is correctly specified as a conditional mean.** It implies
$$\mathbb{E}(y \mid x_1, \ldots, x_k) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k,$$
so the coefficients are partial effects of the regressors on the *conditional expectation* of y. Under MLR.4′ this need not hold: the linear index $\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$ need not be the population regression function, because some nonlinear function of the regressors could still be correlated with u even though the regressors themselves are not.

The practical upshot is symmetric. If u is correlated with *any* of $x_1, \ldots, x_k$, then MLR.4′ fails, and OLS generally loses *both* unbiasedness and consistency at once — the worst case, because no amount of data fixes it.

4.2 Inconsistency and Asymptotic Bias

When MLR.4′ fails, the plim of $\widehat{\beta}_j$ no longer equals β_j. The gap is the asymptotic counterpart of bias, and it has a clean closed form.

Definition 4.4: Inconsistency (Asymptotic Bias)

The *inconsistency* or *asymptotic bias* of an estimator $\widehat{\theta}$ for θ is

$$\operatorname{plim} \widehat{\theta} - \theta.$$

A consistent estimator has zero inconsistency by definition. A nonzero inconsistency is the error that *remains* no matter how large the sample.

In the simple regression $y = \beta_0 + \beta_1 x_1 + u$, retracing the consistency proof but now *without* assuming $\operatorname{Cov}(x_1, u) = 0$ gives the inconsistency of the slope directly.

Theorem 4.5: Inconsistency of the Slope

In the simple regression model, the inconsistency of the OLS slope is

$$\operatorname{plim} \widehat{\beta}_1 - \beta_1 = \frac{\operatorname{Cov}(x_1, u)}{\operatorname{Var}(x_1)}.$$

Hence $\widehat{\beta}_1$ is consistent if and only if x_1 is uncorrelated with u in the population.

The sign of the asymptotic bias is the sign of $\operatorname{Cov}(x_1, u)$: if the regressor is positively correlated with the omitted factors in u, OLS overstates β_1 asymptotically, and conversely.

4.2.1 The Asymptotic Analogue of Omitted-Variable Bias

The most common reason $\operatorname{Cov}(x_1, u) \neq 0$ is that the error absorbs a relevant variable that happens to be correlated with x_1. Suppose the correctly specified model is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + v, \qquad \mathbb{E}(v \mid x_1, x_2) = 0,$$

but we estimate the short regression of y on x_1 alone, so the error we are actually working with is $u = \beta_2 x_2 + v$. Then $\operatorname{Cov}(x_1, u) = \beta_2 \operatorname{Cov}(x_1, x_2)$, and substituting into the inconsistency formula yields the asymptotic version of omitted-variable bias.

Theorem 4.6: Asymptotic Omitted-Variable Bias

If x_2 is omitted from the regression of y on x_1, the OLS slope $\tilde{\beta}_1$ from the short regression satisfies

$$\operatorname{plim} \tilde{\beta}_1 = \beta_1 + \beta_2 \, \delta_1, \qquad \delta_1 = \frac{\operatorname{Cov}(x_1, x_2)}{\operatorname{Var}(x_1)},$$

where δ_1 is the slope of the *population* regression of x_2 on x_1. The asymptotic bias is therefore $\beta_2 \delta_1$, whose sign is given by the table below.

	$\mathrm{Cov}(x_1, x_2) > 0$	$\mathrm{Cov}(x_1, x_2) < 0$
$\beta_2 > 0$	positive bias	negative bias
$\beta_2 < 0$	negative bias	positive bias

This is exactly the table you met for finite-sample omitted-variable bias in Chapter 2, and the parallel is the point. But the parallel is not an identity, and the difference is worth stating precisely.

Bias and inconsistency are related but not the same

The two formulas look almost identical, yet they describe different objects:

- **Inconsistency** is written with *population* moments: $\delta_1 = \mathrm{Cov}(x_1, x_2) \,/\, \mathrm{Var}(x_1)$. It is the error that survives as $n \to \infty$.
- **Bias** is written with the *sample* regression slope $\hat{\delta}_1 = \widehat{\mathrm{Cov}}(x_1, x_2)/\widehat{\mathrm{Var}}(x_1)$, because finite-sample bias is computed conditional on the particular sample of regressors we hold.

The sample quantity $\hat{\delta}_1$ converges in probability to the population quantity δ_1, which is why the two notions coincide in large samples — but they can diverge in any given finite sample.

The cleanest way to feel the distinction is to look at a case where the population correlation is exactly zero.

Example (Zero population correlation: consistent but possibly biased).

Let the true model be $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$, and suppose x_1 and x_2 are *uncorrelated in the population*, so $\mathrm{Cov}(x_1, x_2) = 0$ and hence $\delta_1 = 0$. For a given sample, run the three regressions

$$\begin{aligned} &(1) \quad \hat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 \quad \text{(long, both regressors)}, \\ &(2) \quad \hat{y} = \tilde{\alpha}_0 + \tilde{\alpha}_1 x_1 \quad \text{(short, omit } x_2\text{)}, \\ &(3) \quad \hat{y} = \tilde{\gamma}_0 + \tilde{\gamma}_2 x_2 \quad \text{(short, omit } x_1\text{)}. \end{aligned}$$

What can we say about the short-regression slopes $\tilde{\alpha}_1$ and $\tilde{\gamma}_2$?

Solution.

Because $\delta_1 = \mathrm{Cov}(x_1, x_2) \,/\, \mathrm{Var}(x_1) = 0$, the asymptotic bias term $\beta_2 \delta_1$ vanishes, so

$$\operatorname{plim} \tilde{\alpha}_1 = \beta_1 \qquad \text{and similarly} \qquad \operatorname{plim} \tilde{\gamma}_2 = \beta_2.$$

Both short-regression slopes are therefore *consistent* for the corresponding long-regression coefficients.

Unbiasedness, however, can fail. Finite-sample bias depends on the *sample* regression slope $\hat{\delta}_1 = \widehat{\mathrm{Cov}}(x_1, x_2)/\widehat{\mathrm{Var}}(x_1)$, and in any particular sample the sample covariance between x_1 and x_2 is essentially never exactly zero, so $\hat{\delta}_1 \neq 0$ generically. Hence the

omitted-variable bias term $\beta_2\hat{\delta}_1$ is nonzero in that sample, and $\tilde{\alpha}_1$ need not be unbiased for β_1 (likewise $\tilde{\gamma}_2$ for β_2). The population correlation being zero kills the *asymptotic* bias but not the *finite-sample* bias — a crisp illustration that consistency and unbiasedness are logically distinct.

Remark (Why the asymptotics is the more robust statement).

In practice we never average over infinitely many samples; we have one sample, and we want to know whether using more of it would push the estimate toward the truth. That is exactly what consistency answers. This is why much of modern econometrics is built on plim arguments: they require weaker assumptions (MLR.4′ rather than MLR.4) and they speak directly to the question an applied researcher actually faces.

4.3 Asymptotic Normality and Large-Sample Inference

Consistency tells us that $\widehat{\beta}_j$ converges to β_j, but it says nothing about the *distribution* of the estimate around the truth, and inference requires a distribution. In Chapter 3 we obtained exact t and F distributions by adding the normality assumption MLR.6 (the errors are normally distributed and independent of the regressors). That assumption is convenient but often implausible. The remarkable fact of this section is that we can drop it: in large samples the central limit theorem manufactures approximate normality for us.

Normality is not needed for the core results

The normality assumption MLR.6 plays *no* role in the unbiasedness of OLS, and it plays no role in the Gauss–Markov conclusion that OLS is the best linear unbiased estimator under MLR.1–MLR.5. Normality enters only to make the t and F statistics *exactly* t- and F-distributed in finite samples. Once we are content with large-sample approximations, even that role disappears.

The intuition is the one behind every appearance of the normal distribution. From the error representation, $\widehat{\beta}_j - \beta_j$ is (after scaling) an average of the form $n^{-1}\sum_{i=1}^{n} \hat{r}_{ij}u_i$, and an appropriately normalized average of independent terms is asymptotically normal even when the individual y_i are drawn from a decidedly non-normal distribution. The price is the factor $\sqrt{n}$: it is $\sqrt{n}(\widehat{\beta}_j - \beta_j)$, not $\widehat{\beta}_j - \beta_j$ itself, that settles down to a nondegenerate normal limit.

Theorem 4.7: Asymptotic Normality of OLS

Under the Gauss–Markov assumptions MLR.1 through MLR.5 (i.e. *without* the normality assumption MLR.6):

1. For each j,

$$\sqrt{n}\,(\widehat{\beta}_j - \beta_j) \xrightarrow{d} N\big(0,\ \sigma^2/a_j^2\big), \qquad a_j^2 = \operatorname{plim}\left(\frac{1}{n}\sum_{i=1}^{n}\hat{r}_{ij}^2\right),$$

   where $\hat{r}_{ij}$ is the residual from regressing x_j on all the other regressors. Equivalently, $\widehat{\beta}_j$ is *asymptotically normally distributed*, with asymptotic variance σ^2/a_j^2.

2. The error-variance estimator $\widehat{\sigma}^2 = \frac{1}{n-k-1}\sum_{i=1}^{n}\hat{u}_i^2$ is a consistent estimator of $\sigma^2 = \operatorname{Var}(u)$: $\widehat{\sigma}^2 \xrightarrow{p} \sigma^2$.

3. For each j, the standardized estimator is asymptotically standard normal, whether we standardize by the (unknown) standard deviation or by the (estimated) standard error:

$$\frac{\widehat{\beta}_j - \beta_j}{\operatorname{sd}(\widehat{\beta}_j)} \xrightarrow{d} N(0,1), \qquad \frac{\widehat{\beta}_j - \beta_j}{\operatorname{se}(\widehat{\beta}_j)} \xrightarrow{d} N(0,1).$$

Remark (Why se and sd are asymptotically interchangeable).

> The two standardizations in part (3) give the same limit because $\widehat{\sigma} \xrightarrow{p} \sigma$ (part 2): from an asymptotic point of view $\widehat{\sigma}$ and σ are "equivalent," so replacing the unknown $\operatorname{sd}(\widehat{\beta}_j)$ by the estimated $\operatorname{se}(\widehat{\beta}_j)$ does not change the limiting distribution. This is exactly why, in large samples, we may treat the usual t ratio as standard normal even though we never know σ.

4.3.1 Large-Sample t and F Inference

Part (3) of the theorem is the license for everyday inference. The t statistic $t_{\widehat{\beta}_j} = (\widehat{\beta}_j - \beta_j)/\operatorname{se}(\widehat{\beta}_j)$, which under MLR.6 had an exact t_{n-k-1} distribution, now has an *approximate* standard normal distribution in large samples even when the errors are non-normal. Because $t_{n-k-1} \to N(0,1)$ as $n \to \infty$, this is no contradiction: for large n the t distribution and the standard normal are indistinguishable, so we may keep using the same critical values and p-values. The same logic applies to multiple restrictions: the F statistic from Chapter 3 retains its (approximate) $F_{q,\,n-k-1}$ distribution without normality, and $q \cdot F$ is asymptotically χ_q^2.

Fact 4.8: What changes and what does not

With MLR.6 dropped and only MLR.1–MLR.5 maintained, the mechanics of inference are unchanged: compute the same $\widehat{\beta}_j$, the same $\text{se}(\widehat{\beta}_j)$, the same t and F statistics, and compare to the same critical values. Only the *justification* changes — from "exact in finite samples under normality" to "approximately valid in large samples by the central limit theorem."

4.3.2 The Estimated Variance and Its Rate of Decay

The estimated sampling variance of $\widehat{\beta}_j$ has the same algebraic form as in the finite-sample theory of Chapter 2:

$$\widehat{\text{Var}}(\widehat{\beta}_j) = \frac{\widehat{\sigma}^2}{\text{SST}_j\,(1 - R_j^2)},$$

where $\text{SST}_j = \sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$ is the total sample variation in x_j and R_j^2 is the R^2 from regressing x_j on the remaining regressors. The asymptotic theory tells us how fast this quantity shrinks. As $n \to \infty$ each ingredient behaves predictably: $\widehat{\sigma}^2 \xrightarrow{p} \sigma^2$, a finite constant; while $\text{SST}_j \approx n\,\sigma_j^2$ grows linearly in n, where $\sigma_j^2 = \text{Var}(x_j)$ is the population variance of x_j. Holding R_j^2 at its limiting value, the denominator therefore grows like n while the numerator settles to a constant, so

$$\widehat{\text{Var}}(\widehat{\beta}_j) \;\approx\; \frac{\sigma^2}{n\,\sigma_j^2\,(1 - R_j^2)} \;=\; O_p(1/n).$$

The $1/n$ shrinkage of the variance

The estimated variance of $\widehat{\beta}_j$ shrinks to zero at the rate $1/n$; equivalently, the standard error $\text{se}(\widehat{\beta}_j)$ shrinks at the rate $1/\sqrt{n}$. Two consequences follow:

- This $1/n$ rate is the quantitative content of consistency: as the variance vanishes and $\widehat{\beta}_j$ stays centered at β_j, the estimator concentrates on the truth.
- Because $\text{se}(\widehat{\beta}_j) \propto 1/\sqrt{n}$, cutting the standard error in half requires *quadrupling* the sample. Precision improves with more data, but only at the square-root rate — a useful sense of how much data "buys" how much precision.

Remark (Where this is heading).

We now have the two large-sample pillars: OLS is consistent under the weak condition MLR.4′, with a clean formula for its inconsistency when that fails; and OLS is asymptotically normal under MLR.1–MLR.5, so the familiar t and F procedures remain (approximately) valid without the normality assumption MLR.6. Both pillars matter for what follows. Chapter 7 drops homoskedasticity (MLR.5) and rebuilds the standard errors so that this same asymptotic inference survives heteroskedasticity; Chapter 10 confronts the case where MLR.4′ itself fails — a regressor is correlated with the error — and

introduces instrumental variables, an estimator that is biased but, crucially, consistent.

Chapter 5

Further Issues in Multiple Regression

By now the multiple regression model is a finished machine: Chapter 2 built the OLS estimator, Chapter 3 taught us to test its coefficients, and Chapter 4 showed that the same estimator behaves well even when the errors are not normal. In every one of those chapters the model was written as $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$, linear in the parameters *and* linear in the variables, with the slope β_j read off as "the change in y for a one-unit change in x_j." That reading is correct, but it is also limiting. Many of the relationships we care about are not straight lines: returns to experience flatten out, a tax cut helps the rich differently than the poor, wages respond to a *percentage* change in firm sales rather than a dollar change.

This chapter shows how far the linear model stretches once we are willing to transform the variables. Logarithms turn slopes into elasticities and tame skewed data; quadratics let a relationship rise and then fall; interaction terms let one variable's effect depend on the level of another. None of this requires a new estimator — OLS still does all the work, because the model remains linear *in the parameters*. What changes is the bookkeeping: how we read a coefficient, where the "turning point" of a quadratic lies, and how to summarize an effect that is no longer a single number. We close with the question that always follows a richer specification: did the extra flexibility actually buy us a better model? The honest answer requires a goodness-of-fit measure that charges for complexity — the adjusted R^2 — and an understanding of exactly what it can and cannot compare.

5.1 More on Functional Form

5.1.1 Which Variables to Include

The first decision in any specification is which regressors to put in. There is a genuine tension here, and it is worth stating plainly before we get to functional form.

Adding a regressor that genuinely belongs in the model can only help on one front: it removes a piece of variation from the error term, lowering the error variance σ^2 and, all else equal, tightening every standard error. But "all else equal" is exactly what we cannot take for granted, because a new regressor that is correlated with the ones already present *raises*

the multicollinearity term R_j^2 in the variance formula

$$\operatorname{Var}\left(\widehat{\beta}_j\right) = \frac{\sigma^2}{\mathrm{SST}_j\,(1 - R_j^2)}.$$

The net effect on $\operatorname{Var}\left(\widehat{\beta}_j\right)$ is a contest between a smaller numerator σ^2 and a larger denominator factor $(1 - R_j^2)$.

Add what is orthogonal, be wary of what is collinear

A regressor that is *uncorrelated* with the existing regressors is close to a free lunch: it can lower σ^2 without touching any R_j^2, so it shrinks the standard errors of the coefficients you already care about. The catch is that such variables are hard to find in practice — in observational data, the interesting controls tend to be correlated with everything. And there is a deeper warning: if you "hold fixed" too many things, you may hold fixed the very channel you set out to study. Controlling for occupation when estimating the return to education, for instance, partials out exactly the mechanism (people get better jobs) through which education raises wages.

5.1.2 Logarithmic Functional Forms

We met the four level/log combinations in Chapter 1; here we collect the practical reasons logs are so common in applied work, and one important pitfall when $\log y$ is the dependent variable.

The interpretive advantages are familiar. A logged independent variable lets the slope speak in percentages; a log–log slope is an elasticity; and because $\log(c\,x) = \log c + \log x$, rescaling the units of a logged variable changes only the intercept, leaving the slope of interest untouched. To these we add three more pragmatic benefits. Taking logs of a strictly positive, right-skewed variable (income, firm size, population) typically compresses the long upper tail, so a few enormous observations no longer dominate the fit — logs *dampen the influence of outliers*. The same compression often makes the conditional distribution of the error look more symmetric and bell-shaped, helping the *normality* approximation, and it frequently stabilizes the spread of the error, helping *homoskedasticity* (a theme we return to in Chapter 7).

When *not* to take logs

Logs are not automatic. Three cautions:

- A variable already measured in *percentage points* or as a *rate* — an unemployment rate, a literacy rate, an interest rate — should usually be left in levels, because a change in such a variable is already naturally read as a percentage-point change, and logging it obscures that.
- A variable measured in a small number of natural *units* or *years* (years of schooling, number of children) is often kept in levels, both for interpretability and because the log transform exaggerates differences at the low end.

- Logs require strictly *positive* values: $\log 0$ and log(negative) are undefined, so variables that can be zero or negative (net profit, change in inventory) cannot be logged directly.

5.1.3 Predicting y When the Dependent Variable Is $\log y$

Suppose we have estimated a model in which the dependent variable is $\log y$:

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

This is ideal for interpreting how the regressors move y in percentage terms. But often we want an actual prediction of y itself — a predicted wage in dollars, not a predicted log-wage. Exponentiating both sides,

$$y = \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k) \cdot \exp(u),$$

and the naive temptation is to estimate y by simply exponentiating the fitted log-value, $\widehat{\log y} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \cdots + \widehat{\beta}_k x_k$. This is *systematically too low*, and understanding why is the point of this subsection.

The clean way to see the problem is to compute the conditional mean of y. Add the assumption that u is *independent* of $(x_1, \ldots, x_k)$ — stronger than the zero-conditional-mean assumption we usually impose, but it lets the error term factor out cleanly. Then

$$\mathbb{E}(y \mid x) = \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k) \cdot \mathbb{E}(\exp(u)).$$

The first factor is what we would get from exponentiating the population regression function; the second factor, $\mathbb{E}(\exp(u))$, is a constant multiplier that the naive predictor silently sets to 1. It is not 1.

Theorem 5.1: Jensen's Inequality Makes the Naive Predictor Biased Downward

Because $\exp(\cdot)$ is a strictly convex function, Jensen's inequality gives

$$\mathbb{E}(\exp(u)) \;>\; \exp\big(\mathbb{E}(u)\big) = \exp(0) = 1,$$

using $\mathbb{E}(u) = 0$. Hence $\mathbb{E}(y \mid x)$ is the naive exponentiated prediction scaled *up* by a factor strictly greater than 1: simply exponentiating $\widehat{\log y}$ underpredicts y on average.

The practical fix is to estimate the missing multiplier from the residuals. Since $\mathbb{E}(\exp(u))$ is a population mean, replace it by its sample analogue using the OLS residuals $\hat{u}_i = \log y_i - \widehat{\log y}_i$, giving the corrected prediction

$$\hat{y} = \widehat{\alpha}_0 \cdot \exp\big(\widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \cdots + \widehat{\beta}_k x_k\big), \qquad \widehat{\alpha}_0 = \frac{1}{n}\sum_{i=1}^{n} \exp(\hat{u}_i).$$

This is Duan's *smearing* estimate of the retransformation factor.

Remark (Consistent but biased: a recurring distinction).

The corrected predictor $\hat{y}$ is *not* an unbiased estimator of $\mathbb{E}(y \mid x)$ in finite samples: $\frac{1}{n}\sum_{i=1}^{n}\exp(\hat{u}_i)$ is a nonlinear function of estimated residuals, and nonlinear functions of unbiased inputs are not unbiased. What we *can* guarantee is *consistency*. As $n \to \infty$, $\widehat{\beta}_j \xrightarrow{p} \beta_j$ and $\frac{1}{n}\sum_{i=1}^{n}\exp(\hat{u}_i) \xrightarrow{p} \mathbb{E}(\exp(u))$, so by the continuous mapping theorem $\operatorname{plim}\hat{y} = \mathbb{E}(y \mid x)$. "Consistent but biased" is a pattern you will see again — most prominently for the IV estimator in Chapter 10 — and it is the right standard to hold a nonlinear predictor to.

Example (Retransforming a log-wage prediction).

A wage equation is estimated as $\widehat{\log(\text{wage})} = 1.20 + 0.090\,\text{educ}$, with smearing factor $\widehat{\alpha}_0 = \frac{1}{n}\sum_{i=1}^{n}\exp(\hat{u}_i) = 1.08$. Predict the wage for a worker with educ $= 12$.

Solution.

The fitted log-wage is $\widehat{\log(\text{wage})} = 1.20 + 0.090(12) = 2.28$. The naive prediction would be $\exp(2.28) \approx 9.78$. Applying the smearing correction,

$$\hat{y} = 1.08 \cdot \exp(2.28) \approx 1.08 \cdot 9.78 \approx 10.56.$$

Ignoring the factor of 1.08 would understate the predicted wage by about eight percent — exactly the Jensen gap.

5.1.4 Quadratics and the Turning Point

When y bends with x — rising quickly at first and then leveling off, or falling and then rising — a straight line is the wrong shape. Adding the square of a regressor lets the model curve while staying linear in the parameters:

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

The crucial change is that the partial effect of x is no longer a constant. Differentiating,

$$\frac{\partial y}{\partial x} = \beta_1 + 2\beta_2 x,$$

so the marginal effect of x depends on the level of x itself. A useful approximation for a one-unit change is $\Delta y \approx (\beta_1 + 2\beta_2 x)\,\Delta x$. The coefficient β_1 alone is the slope only at $x = 0$ and is meaningless to report in isolation; you must always pair it with β_2.

The two coefficients typically have opposite signs, producing a parabola with a single *turning point* (a maximum if $\beta_2 < 0$, a minimum if $\beta_2 > 0$). Setting the partial effect to zero,

$$\beta_1 + 2\beta_2 x^* = 0 \quad \Longrightarrow \quad x^* = -\frac{\beta_1}{2\beta_2}.$$

On one side of x^* the relationship between y and x slopes up; on the other side it slopes down. In the typical case of opposite-signed coefficients ($\beta_1 > 0$, $\beta_2 < 0$, giving a downward-opening parabola), this turning point is positive and equals $-\beta_1/(2\beta_2)$.

Always interpret *both* sides — and count how many observations lie past the turn

A quadratic forces a turnaround on the data whether or not the turnaround is economically real. If a wage–experience profile peaks at $x^* \approx 24.4$ years and then formally "declines," you must ask: does that downturn describe a meaningful feature of the data, or is it an artifact? The diagnostic is to count how many sample observations lie on the "unwanted" side of x^*. If almost no one in the sample has more than 24 years of experience, the implied decline rests on a handful of points (or none) and should be read with great caution — effectively, the fitted relationship is increasing over the entire range that matters. If, on the other hand, a non-negligible fraction of the sample sits on a side where the sign of the effect makes no economic sense, that is a signal that the quadratic is the wrong functional form and the model must be refined.

Example (Locating and reading a turning point).

A regression of hourly wage on experience gives $\widehat{\text{wage}} = 3.73 + 0.298\,\text{exper} - 0.0061\,\text{exper}^2$. Find the experience level at which predicted wage is maximized, and describe the shape on each side.

Solution.

Here $\beta_1 = 0.298 > 0$ and $\beta_2 = -0.0061 < 0$, so the parabola opens downward and has a maximum. The turning point is

$$\text{exper}^* = -\frac{0.298}{2(-0.0061)} = \frac{0.298}{0.0122} \approx 24.4 \text{ years.}$$

For exper < 24.4, the marginal effect $0.298 - 2(0.0061)\,\text{exper}$ is positive, so wage rises with experience; for exper > 24.4 the effect turns negative and predicted wage falls. Whether to take the post-24.4 decline seriously depends on how many workers in the sample have more than 24 years of experience; if few do, the model is effectively saying wages rise (at a diminishing rate) throughout the relevant range.

5.1.5 Interaction Terms

Sometimes the effect of one regressor depends on the *level* of another — the value of a bedroom depends on the size of the house, the wage premium for experience differs by education. An *interaction term* captures this by entering the product of two regressors:

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

The partial effect of x_2 is now

$$\frac{\partial y}{\partial x_2} = \beta_2 + \beta_3 x_1,$$

which depends on x_1. The interaction coefficient β_3 measures how the effect of x_2 changes as x_1 increases (equivalently, how the effect of x_1 changes with x_2 — the relationship is symmetric).

This flexibility comes at a cost to interpretation. The coefficient β_2 is *not* the average effect of x_2; it is the effect of x_2 *only when* $x_1 = 0$. If $x_1 = 0$ is an unusual value — or, worse, an impossible one, as when x_1 is firm size or a test score that is never zero — then β_2 describes a counterfactual no one in the data occupies, and reporting it as "the effect of x_2" is misleading.

Reparametrization by centering. The clean remedy is to center the interacting variables at meaningful values, typically their means $\mu_1 = \mathbb{E}(x_1)$ and $\mu_2 = \mathbb{E}(x_2)$ (replaced in practice by the sample means $\bar{x}_1, \bar{x}_2$). Rewrite the model as

$$y = \alpha_0 + \delta_1 x_1 + \delta_2 x_2 + \beta_3 (x_1 - \mu_1)(x_2 - \mu_2) + u.$$

This is the *same* model algebraically — expanding the product and collecting terms recovers the original equation, with β_3 unchanged — but the level coefficients now have a far more useful meaning. In the centered form,

$$\frac{\partial y}{\partial x_2} = \delta_2 + \beta_3 (x_1 - \mu_1),$$

so δ_2 is the partial effect of x_2 evaluated *at the mean of* x_1, the effect for a typical observation rather than for the (possibly fictional) $x_1 = 0$ case. Centering carries a second, equally important benefit: the standard error of the partial effect at the mean is the standard error of the single coefficient δ_2, which the regression reports directly. In the uncentered model, $\widehat{\partial y/\partial x_2} = \widehat{\beta}_2 + \widehat{\beta}_3 \bar{x}_1$ is a *linear combination* of two estimates, and its standard error requires the covariance $\mathrm{Cov}\left(\widehat{\beta}_2, \widehat{\beta}_3\right)$ — extra work the centering sidesteps. If some value other than the mean is of interest, center there instead.

5.1.6 The Average Partial Effect

Quadratics and interactions share a feature that breaks the old habit of reading a single slope: the partial effect of a regressor is no longer one number but a *function* of the data. In

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + u, \qquad \frac{\partial y}{\partial x} = \beta_1 + 2\beta_2 x,$$

the effect varies observation by observation. To report a single representative figure we use the *average partial effect*.

Definition 5.2: Average Partial Effect (APE)

The *average partial effect* of x_j is the partial effect $\partial y/\partial x_j$, evaluated at the estimated coefficients and averaged across the sample:

$$\widehat{\mathrm{APE}}_j = \frac{1}{n} \sum_{i=1}^{n} \left. \frac{\partial \hat{y}}{\partial x_j} \right|_{\text{obs } i}.$$

The computation is mechanical: write down the partial-effect expression, plug in the

OLS estimates, evaluate it at each observation, and average. For the quadratic above,

$$\widehat{\text{APE}}_x = \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{\beta}_1 + 2\widehat{\beta}_2 x_i\right) = \widehat{\beta}_1 + 2\widehat{\beta}_2\,\bar{x},$$

which equals the partial effect evaluated at the sample mean $\bar{x}$ — here the "average of the effect" and the "effect at the average" coincide because the partial effect is linear in x. This gives a clean connection to centering.

Centering makes a coefficient equal its own APE

If you center the regressor before squaring or interacting — replacing x by $x - \bar{x}$ — the coefficient on the linear term *becomes* the average partial effect, with the correct standard error reported automatically. For the quadratic, writing $y = \gamma_0 + \gamma_1(x - \bar{x}) + \gamma_2(x - \bar{x})^2 + u$ makes $\widehat{\gamma}_1 = \widehat{\beta}_1 + 2\widehat{\beta}_2\bar{x} = \widehat{\text{APE}}_x$. This is the practical reason centering is the default for nonlinear specifications: it puts the number you want to report, with its standard error, directly on the regression output.

5.2 Goodness of Fit Revisited

A richer specification almost always raises the ordinary R^2. That alone is not evidence that the richer model is better, so we need to revisit what R^2 measures and build a version that charges for added complexity.

5.2.1 What R^2 Does and Does Not Tell You

Two cautions carry over from Chapter 1 and are worth restating now that the model is rich:

- A *high* R^2 does not imply a causal interpretation. Fit measures association between y and its fitted value, nothing more; an omitted variable can inflate R^2 while ruining every coefficient.
- A *low* R^2 does not preclude precise estimation of a partial effect. We may pin down a slope β_j tightly even when the regressors leave most of the variation in y to the error.

There is also a population target lurking behind R^2. Writing it with the sample sizes made explicit,

$$R^2 = 1 - \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSR}/n}{\text{SST}/n} \xrightarrow{p} 1 - \frac{\sigma_u^2}{\sigma_y^2},$$

because $\text{SSR}/n \xrightarrow{p} \sigma_u^2 = \text{Var}(u)$ and $\text{SST}/n \xrightarrow{p} \sigma_y^2 = \text{Var}(y)$. So R^2 is a (slightly biased) estimate of the population quantity $1 - \sigma_u^2/\sigma_y^2$, the fraction of the variance of y that is *not* error. The bias is the entry point for the adjustment that follows.

5.2.2 The Adjusted R^2

The estimator R^2 uses SSR/n and SST/n, but neither SSR/n nor SST/n is an unbiased estimator of the variance it targets. Correcting each by its proper degrees of freedom —

$n-k-1$ for the residual sum of squares (we estimated $k+1$ parameters to form it) and $n-1$ for the total sum of squares (one parameter, the mean, was used) — gives the adjusted R^2.

Definition 5.3: Adjusted R^2

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

Equivalently, in terms of the ordinary R^2,

$$\bar{R}^2 = 1 - \frac{n-1}{n-k-1}\,(1-R^2).$$

The two forms are identical: substitute $\text{SSR}/\text{SST} = 1-R^2$ into the first and factor out. The fraction $\frac{n-1}{n-k-1}$ exceeds 1 whenever $k \geq 1$, so $\bar{R}^2 < R^2$ in any model with at least one regressor; the gap is the *penalty*.

The penalty and the $|t| > 1$ rule

Because the $\frac{n-1}{n-k-1}$ factor grows with k, $\bar{R}^2$ does not automatically rise when a regressor is added. There is a sharp characterization of when it rises:

Adding one regressor raises $\bar{R}^2$ if and only if the absolute t statistic of that regressor exceeds 1.

A new variable lifts the penalized fit only if it clears a $|t| > 1$ bar — a threshold notably *lower* than the usual significance cutoffs near 2. So a variable can raise $\bar{R}^2$ while still being statistically insignificant at conventional levels. The adjusted R^2 is therefore a lenient model-selection device, not a hypothesis test.

Remark ($\bar{R}^2$ can be negative).

Unlike $R^2 \in [0,1]$, the adjusted R^2 has no nonnegativity guarantee. If the regressors explain very little and k is large relative to n, the penalty term can drive $\bar{R}^2$ below zero. A negative $\bar{R}^2$ is a blunt warning that the model fits worse, after accounting for its complexity, than a model with no regressors at all (which predicts y by its mean).

Example (When an extra regressor lowers the adjusted fit).

A model with k regressors on $n = 50$ observations has $R^2 = 0.400$. Adding one more regressor (so $k+1$ in total) raises the ordinary fit to $R^2 = 0.405$. Did the adjusted R^2 go up? Take $k = 4$ before the addition.

Solution.

Before: $\bar{R}^2 = 1 - \frac{49}{45}(1-0.400) = 1 - \frac{49}{45}(0.600) \approx 1 - 0.653 = 0.347$.

After (now $k = 5$, so denominator $n-k-1 = 44$): $\bar{R}^2 = 1 - \frac{49}{44}(1-0.405) = 1 - \frac{49}{44}(0.595) \approx 1 - 0.663 = 0.337$.

The ordinary R^2 rose, but the adjusted R^2 *fell* from 0.347 to 0.337: the half-percentage-point gain in raw fit did not justify the lost degree of freedom. Equivalently,

the new regressor's t statistic must have been below 1 in absolute value.

5.2.3 Using $\bar{R}^2$ to Compare Nonnested Models

Two models are *nested* when one is a special case of the other (obtained by setting some coefficients to zero); they are *nonnested* when neither contains the other — for example, one uses x and x^2 while the other uses $\log x$, or one has three regressors and a competitor has three entirely different regressors. Nested models are compared with an F test; nonnested models cannot be, because there is no set of restrictions taking one to the other.

$\bar{R}^2$ levels the playing field across nonnested models

The adjusted R^2 is well suited to choosing between nonnested specifications precisely because it penalizes for the number of parameters. Comparing ordinary R^2 values here would be *unfair*: the model with more regressors (or with more flexible terms) has a built-in advantage in raw fit that has nothing to do with being a better model. Because $\bar{R}^2$ docks each model for its complexity, a higher $\bar{R}^2$ is a defensible reason to prefer one nonnested model over another.

There is one hard limit, and it is not a matter of fairness but of arithmetic.

Neither R^2 measure can compare models with different dependent variables

If two models use different definitions of the dependent variable — most commonly y in one and $\log y$ in the other — then *neither* R^2 nor $\bar{R}^2$ can be used to choose between them. The reason is that R^2 measures explained variation *relative to* the total variation SST in the dependent variable, and SST for y and SST for $\log y$ are variations in different quantities, on different scales. A model explaining 70% of the variation in log(wage) and a model explaining 65% of the variation in wage are simply not on the same ruler. Choosing between a level and a log dependent variable requires a different tool — for instance, comparing how well each predicts the *same* target after retransforming, as in the smearing prediction of Section 5.1.3 — not a contest of R^2 values.

Remark (Where this leaves us).

We have stretched the linear model to fit curves, percentages, and effects that depend on context, all without leaving OLS. The recurring lesson is that flexibility shifts the work from estimation to *interpretation*: a logged dependent variable demands a retransformation, a quadratic demands a turning point and a count of observations past it, an interaction demands centering, and a richer model demands the adjusted R^2 to keep the comparison honest. Chapter 6 adds the last piece of functional-form vocabulary — qualitative regressors — by letting dummy variables shift intercepts and, through interactions, slopes.

Chapter 6

Qualitative Information: Dummy Variables

So far every variable in our regressions has been quantitative — wages, years of schooling, prices, interest rates — things that arrive already attached to a number. But much of the information an economist cares about is qualitative: a worker is male or female, a firm is unionized or not, a household lives in the city or the country, an applicant is approved or denied. These attributes have no natural numerical scale, yet they plainly belong in the model. The device that lets us fold them in is the *dummy variable*: a variable that takes only the values 0 and 1, recording the *presence* or *absence* of a qualitative trait.

The beauty of the dummy variable is that, once coded, it is just another regressor. All the OLS machinery of Chapters 2–3 applies unchanged — the estimators, their unbiasedness, the t and F tests. What changes is the *interpretation*. A dummy entered on its own shifts the intercept, splitting one population into a base group and a comparison group; a dummy *interacted* with a continuous regressor lets the two groups have different slopes as well. We will see how to combine several dummies to encode ordered categories, how to test whether two groups obey the same regression at all (the Chow test), what happens when the dependent variable is itself a dummy (the linear probability model), and finally how dummies organize the modern language of treatment effects and self-selection.

A single recurring caution ties the chapter together. Because a dummy and its complement are perfectly correlated by construction — knowing you are male tells you exactly that you are not female — one must always leave a base group *out* of the regression, or drop the intercept. Forgetting this is the *dummy variable trap*, and it is the first thing to internalize.

6.1 Different Intercepts: Dummies as Regressors

A dummy variable — also called a *binary* or 0–1 *variable* — encodes a two-way qualitative distinction. The choice of which state gets coded 0 and which gets coded 1 is arbitrary and does not affect the substance of the analysis; what it *does* fix is the *base group* (or *benchmark group*), the state coded 0, against which everything else is measured. Keeping clear sight of the base group is the single most important habit in reading dummy-variable regressions.

Definition 6.1: Dummy (Binary) Variable

A *dummy variable* d takes the value

$$d = \begin{cases} 1 & \text{if the observation has the attribute,} \\ 0 & \text{if it does not (the base group).} \end{cases}$$

It carries qualitative information into an otherwise quantitative regression.

Consider a population split into two groups by a single dummy d, alongside one continuous regressor x_1. The model is

$$y = \beta_0 + \delta_0\, d + \beta_1 x_1 + u.$$

To read off what δ_0 does, take the conditional mean of y within each group. For the base group ($d = 0$),

$$\mathbb{E}(y \mid x_1, d = 0) = \beta_0 + \beta_1 x_1,$$

while for the comparison group ($d = 1$),

$$\mathbb{E}(y \mid x_1, d = 1) = (\beta_0 + \delta_0) + \beta_1 x_1.$$

The two population lines are *parallel* — both have slope β_1 — but the second sits a vertical distance δ_0 above (or below, if $\delta_0 < 0$) the first. The base group's intercept is β_0; the comparison group's intercept is $\beta_0 + \delta_0$. Thus δ_0 is an *intercept shift*.

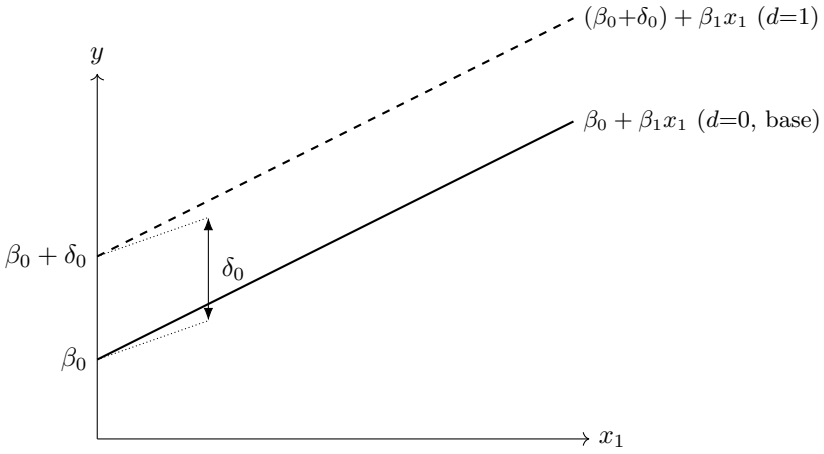

Figure 6.1: An intercept shift. The dummy d moves the whole line up by δ_0 without changing the slope β_1; the two groups differ by the constant vertical gap δ_0 at every value of x_1.

The interpretation generalizes immediately. *The coefficient on a dummy variable is the difference in the (conditional) mean of y between the two states, holding all other regressors fixed.* Here, at any fixed x_1, the average of y in the $d = 1$ group exceeds that in the base group by exactly δ_0.

Example (Wage and gender).

Let wage be hourly wage, educ years of schooling, and female a dummy equal to 1 for women, so men are the base group. The model

$$\text{wage} = \beta_0 + \delta_0 \,\text{female} + \beta_1 \,\text{educ} + u$$

makes δ_0 the average wage gap between a woman and a man with the *same* education. A negative estimate $\widehat{\delta}_0 < 0$ says that, at every level of schooling, women earn on average $|\widehat{\delta}_0|$ less per hour; the return to a year of education, β_1, is assumed common to both.

6.1.1 The Dummy Variable Trap

If we tried to include a dummy for *every* level of a qualitative variable *and* kept the intercept, we would create perfect multicollinearity. With two states and dummies d_0 (for the base) and d_1 (for the other), every observation has $d_0 + d_1 = 1$, which is exactly the constant column already supplied by the intercept. OLS cannot then separate the intercept from the dummies, and the design matrix loses full rank — a violation of the no-perfect-collinearity assumption MLR.3 of Chapter 2. This is the *dummy variable trap*.

The dummy variable trap

To represent a qualitative variable with m mutually exclusive categories in a model *with an intercept*, include exactly $m-1$ dummies and omit one category as the base group. Including all m dummies alongside the intercept produces perfect collinearity, and OLS breaks down.

There is one legitimate alternative: include all m dummies but *drop the intercept*. With two groups,

$$y = \gamma_0 \, d_0 + \gamma_1 \, d_1 + \beta_1 x_1 + u,$$

where now d_0 is the base-group indicator. Here γ_0 and γ_1 are the two group intercepts *directly*: $\mathbb{E}(y \,|\, x_1, \text{base}) = \gamma_0 + \beta_1 x_1$ and $\mathbb{E}(y \,|\, x_1, \text{other}) = \gamma_1 + \beta_1 x_1$. The estimated difference between the groups is $|\widehat{\gamma}_1 - \widehat{\gamma}_0|$. This parametrization is internally consistent and gives correct coefficient estimates, but it has two practical drawbacks:

- The quantity we usually care about — the *difference* between groups — is now a linear combination $\gamma_1 - \gamma_0$ rather than a single coefficient, so testing whether the groups differ requires a test on that combination instead of a simple t test on one estimate. The intercept-plus-$(m-1)$-dummies parametrization is more convenient precisely because the gap is read off a single coefficient (δ_0 above).
- With no intercept in the model, the usual R^2 loses its standard interpretation (and reported software values can even be misleading), because the total-sum-of-squares decomposition that defines R^2 relies on an intercept being present.

For these reasons the textbook default is to keep the intercept and omit one category.

6.2 Incorporating Ordinal Information

Some qualitative variables are not merely categorical but *ordered*: they rank observations without supplying a meaningful scale of distances. A leading example is a city's credit rating CR, an integer running $0, 1, 2, 3, 4$, which we wish to relate to the interest rate MBR the city pays on its municipal bonds.

The naive approach treats CR as if it were an ordinary continuous regressor,

$$\text{MBR} = \beta_0 + \beta_1 \,\text{CR} + (\text{other factors}),$$

which we call the *fixed-partial-effect model*. Its single slope β_1 forces the effect of moving up one notch in the rating to be *identical* at every notch: the gap between ratings 0 and 1 is assumed equal to the gap between ratings 3 and 4. But an ordinal variable carries only ordinal meaning, not interval meaning; there is no reason the rungs of the ladder should be equally spaced in their effect on MBR.

A more flexible specification defines a separate dummy for each level (minding the dummy variable trap by omitting one as the base):

$$\text{MBR} = \beta_0 + \delta_1 \,\text{CR}_1 + \delta_2 \,\text{CR}_2 + \delta_3 \,\text{CR}_3 + \delta_4 \,\text{CR}_4 + (\text{other factors}),$$

where $\text{CR}_j = 1$ if the rating equals j and 0 otherwise, and $\text{CR} = 0$ is the base group. Now each δ_j is the effect of having rating j *relative to* rating 0, free to differ across levels. This lets the data, rather than the functional form, decide how the steps are spaced.

6.2.1 The Fixed-Partial-Effect Model as a Restricted Special Case

The continuous-CR model is nested inside the dummy model: it is the dummy model with the rungs forced to be evenly spaced. Indeed, if the per-notch effect is the constant δ_1, then rating j should be worth $j\,\delta_1$ relative to the base, i.e.

$$\delta_2 = 2\delta_1, \qquad \delta_3 = 3\delta_1, \qquad \delta_4 = 4\delta_1.$$

Imposing these three restrictions on the dummy model collapses it back to the fixed-partial-effect model:

$$\begin{aligned} \text{MBR} &= \beta_0 + \delta_1\big(\text{CR}_1 + 2\,\text{CR}_2 + 3\,\text{CR}_3 + 4\,\text{CR}_4\big) + (\text{other factors}) \\ &= \beta_0 + \delta_1 \,\text{CR} + (\text{other factors}), \end{aligned}$$

because $\text{CR}_1 + 2\,\text{CR}_2 + 3\,\text{CR}_3 + 4\,\text{CR}_4$ is exactly the numerical rating CR (only one of the dummies is nonzero, and it equals the rating). The restriction

$$H_0: \quad \delta_2 = 2\delta_1, \;\; \delta_3 = 3\delta_1, \;\; \delta_4 = 4\delta_1$$

imposes three linear constraints and can be tested with the usual F test (Chapter 3): estimate the unrestricted dummy model and the restricted continuous-CR model, and compare their sums of squared residuals. Rejecting H_0 says the equal-spacing shortcut is too crude.

Remark (When there are too many levels).

If the ordinal variable takes a great many values, a separate dummy for each is impractical (and burns degrees of freedom). The usual remedy is to *bin* the variable into a handful of ranges — for instance, group credit scores into "low," "medium," and "high" bands — and enter a dummy for each band but one. This trades some resolution for parsimony while still relaxing the equal-spacing assumption.

6.3 Different Slopes: Interaction with a Dummy

An intercept-shift dummy lets two groups differ by a constant. But often we suspect the groups respond *differently* to a continuous regressor — that the *return* to education differs by gender, say, not merely the level of wages. To allow this we interact the dummy with the continuous regressor:

$$y = \beta_0 + \beta_1 x_1 + \delta_0\, d + \delta_1\,(d \cdot x_1) + u.$$

Reading the two group lines off this equation:

$$\begin{aligned} \mathbb{E}(y \mid x_1, d = 0) &= \beta_0 + \beta_1 x_1, \\ \mathbb{E}(y \mid x_1, d = 1) &= (\beta_0 + \delta_0) + (\beta_1 + \delta_1)\, x_1. \end{aligned}$$

Now the two groups differ in *both* intercept and slope. As before, δ_0 shifts the intercept. The new coefficient δ_1 is the *difference in slopes*: the partial effect of x_1 on y is β_1 in the base group and $\beta_1 + \delta_1$ in the comparison group, so the two partial effects differ by exactly δ_1.

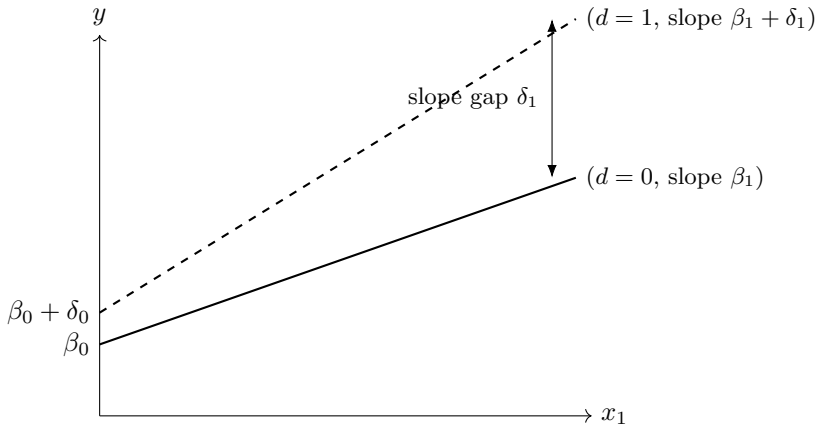

Figure 6.2: Different slopes. The interaction term $d \cdot x_1$ lets the comparison line ($d = 1$) have slope $\beta_1 + \delta_1$, so the vertical gap between the groups widens (or narrows) with x_1 instead of staying constant.

Several special cases are worth naming. If $\delta_1 = 0$, the slopes coincide and we are back to the parallel-lines model of Figure 6.1: x_1 has the same partial effect in both groups. If $\delta_0 = \delta_1 = 0$, the two groups obey *exactly the same* regression — the dummy is irrelevant. These restrictions are testable, and testing them jointly is the subject of the Chow test below.

Interpreting a coefficient once its variable is interacted

When x_1 enters both alone and through an interaction, β_1 is *not* the partial effect of x_1 in general — it is the partial effect of x_1 only in the base group, i.e. only when $d = 0$. Symmetrically, δ_0 is the group gap only at $x_1 = 0$. If $x_1 = 0$ is not a meaningful or observed value, δ_0 as reported is not interpretable on its own. The standard fix is to *center* x_1 at its sample mean (replace x_1 by $x_1 - \bar{x}_1$) before forming the interaction, so that δ_0 becomes the group gap *at the average* x_1, which is a quantity one can actually speak about.

6.3.1 The Chow Test

The interaction approach extends naturally to many regressors: to let *every* slope and the intercept differ between two groups, interact the dummy with each regressor (and include the dummy itself for the intercept). Testing whether the dummy classification matters at all then amounts to a joint F test that *all* the interaction coefficients are zero. When there are only a few regressors this "direct method" is easy. But with many regressors it is cumbersome to build and name all the interaction terms. The *Chow test* is an algebraically equivalent shortcut that computes the same F statistic from three separate regressions.

The null hypothesis of the Chow test is that the *entire regression function is identical* across the two groups — same intercept and same slopes. Let there be k regressors (so $k+1$ parameters including the intercept), and split the n observations into group 1 (size n_1) and group 2 (size n_2), $n = n_1 + n_2$.

Chow Test — procedure

1. **Unrestricted fit.** Run *separate* regressions on each group and record their residual sums of squares. The unrestricted SSR is their sum,

$$\mathrm{SSR}_{ur} = \mathrm{SSR}_1 + \mathrm{SSR}_2.$$

   (This is exactly the SSR one would get from the fully-interacted single regression, which allows all coefficients to differ by group.)

2. **Restricted (pooled) fit.** Run one regression on the *pooled* data, ignoring the group distinction, and record the restricted (pooled) residual sum of squares SSR_p.

3. **Form the statistic.**

$$F = \frac{(\mathrm{SSR}_p - \mathrm{SSR}_{ur})/(k+1)}{\mathrm{SSR}_{ur}/\big(n-2(k+1)\big)} = \frac{\mathrm{SSR}_p - (\mathrm{SSR}_1 + \mathrm{SSR}_2)}{\mathrm{SSR}_1 + \mathrm{SSR}_2} \cdot \frac{n-2(k+1)}{k+1}.$$

   Under H_0, $F \sim F_{k+1,\ n-2(k+1)}$.

The numerator degrees of freedom, $k+1$, count the restrictions: forcing all k slopes *and* the intercept to be equal across groups is $k+1$ constraints. The denominator degrees of

freedom, $n-2(k+1)$, count the residual degrees of freedom of the unrestricted model, which fits $2(k+1)$ parameters in all ($k+1$ for each group).

Remark (Allowing the intercept to differ).

> The plain Chow test forces the intercept to be common as well, which is often too strong — two groups may share the same slopes yet sit at different baseline levels. A more useful variant allows the intercepts to differ and tests only the *slopes*. Operationally one keeps an intercept-shift dummy in both the restricted and unrestricted models, so the restriction count drops from $k+1$ to k, and the F statistic uses numerator degrees of freedom k accordingly.

6.3.2 Structural Change Over Time

The same idea detects *structural change*: whether the regression relationship is stable across T time periods, or whether the coefficients (intercept included) shift from one era to the next. We now use dummies for the time periods. Running a separate regression in each period $t = 1, \ldots, T$ gives

$$\mathrm{SSR}_{ur} = \mathrm{SSR}_1 + \mathrm{SSR}_2 + \cdots + \mathrm{SSR}_T.$$

With k regressors and T periods, the unrestricted model estimates $T(k+1)$ parameters — a full set of $k+1$ coefficients per period. Stability requires that all T periods share one set of coefficients, which imposes $(T-1)(k+1)$ restrictions (each of the $T-1$ non-base periods must match the base period in all $k+1$ coefficients). Hence, with $n = n_1 + \cdots + n_T$ total observations, the F statistic has

$$(T-1)(k+1) \quad \text{numerator and} \quad n - T(k+1) \quad \text{denominator degrees of freedom.}$$

Heteroskedasticity caveat

The Chow test — in either its group form or its structural-change form — *assumes the error variance is constant across groups (and over time).* The derivation of the pooled F statistic relies on a single common σ^2; if the error variance differs across groups, the statistic does not have the stated F distribution and the test is invalid. Under heteroskedasticity one should not use the pooled-SSR shortcut at all. Instead, build the full set of interaction terms explicitly and run one regression, then conduct a *heteroskedasticity-robust* joint test (Wald/F) on the interaction coefficients using robust standard errors (Chapter 7).

6.4 A Binary Regressand: The Linear Probability Model

Dummies need not be confined to the right-hand side. When the *dependent* variable is itself binary — loan approved or denied, employed or not, in the labor force or out — regressing it on a set of explanatory variables produces the *linear probability model* (LPM).

Write the binary outcome $y \in \{0, 1\}$ as

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

Because y takes only the values 0 and 1, its conditional expectation *is* a probability:

$$\mathbb{E}(y \,|\, \mathbf{x}) = 1 \cdot \mathrm{P}(y = 1 \,|\, \mathbf{x}) + 0 \cdot \mathrm{P}(y = 0 \,|\, \mathbf{x}) = \mathrm{P}(y = 1 \,|\, \mathbf{x}) \,.$$

Combining this with the zero-conditional-mean assumption $\mathbb{E}(u \,|\, \mathbf{x}) = 0$ gives the defining identity of the LPM:

$$\mathrm{P}(y = 1 \,|\, \mathbf{x}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k.$$

The response probability is modeled as linear in the parameters, and consequently each slope has a clean probability interpretation:

$$\beta_j = \frac{\Delta \, \mathrm{P}(y = 1 \,|\, \mathbf{x})}{\Delta x_j}.$$

A one-unit increase in x_j changes the probability that $y = 1$ by β_j (in percentage-point terms, holding the other regressors fixed).

Definition 6.2: Linear Probability Model

A regression $y = \beta_0 + \sum_{j=1}^{k} \beta_j x_j + u$ with a binary dependent variable $y \in \{0, 1\}$ is a *linear probability model*. The fitted value $\widehat{y}$ estimates the conditional probability $\mathrm{P}(y = 1 \,|\, \mathbf{x})$, and each β_j measures the change in that probability per unit of x_j.

The LPM is attractive for its transparency — it is estimated by ordinary OLS and read like any other regression — but it has genuine drawbacks, all stemming from forcing a probability to be linear.

- **Out-of-range predictions.** A linear function is unbounded, so fitted probabilities can fall below 0 or rise above 1, which is nonsensical for a probability.
- **Constant (and sometimes impossible) marginal effects.** A constant β_j says a unit change in x_j moves the probability by the same amount no matter where one starts. Near the boundaries this can be logically impossible (you cannot raise a probability already at 0.98 by another 0.10), whereas in reality such effects are usually nonlinear, tapering off as the probability approaches 0 or 1.
- **Built-in heteroskedasticity.** This is unavoidable, not incidental. Since y given $\mathbf{x}$ is a Bernoulli random variable with success probability $p(\mathbf{x}) := \mathrm{P}(y = 1 \,|\, \mathbf{x})$, its conditional variance is

  $$\mathrm{Var}(y \,|\, \mathbf{x}) = p(\mathbf{x}) \left[1 - p(\mathbf{x})\right] = \big(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k\big)\big(1 - \beta_0 - \cdots - \beta_k x_k\big).$$

  This depends on $\mathbf{x}$, so the homoskedasticity assumption of the classical model necessarily fails. OLS remains unbiased and consistent, but its usual standard errors are wrong; one must use heteroskedasticity-robust (White/Eicker–Huber) standard errors, the subject of Chapter 7.

Remark (Why use it anyway).

> Despite these flaws, the LPM is widely used because its coefficients are directly interpretable as effects on a probability and because, for x values near the center of the data, its estimated partial effects are often close to those from the more elaborate logit and probit models. It is an honest first pass; the out-of-range and heteroskedasticity problems are managed (robust SEs) rather than ignored.

6.5 Self-Selection and Treatment Effects

Dummies are the natural language for program evaluation, where we ask whether a binary *treatment* w — a job-training program, a scholarship, a medical intervention — causes a change in an outcome y. The challenge is *self-selection*: the people who take a treatment usually differ systematically from those who do not, so a raw comparison of treated and untreated outcomes confounds the effect of treatment with the effect of those differences.

6.5.1 The Potential-Outcomes Setup

For each unit, imagine two potential outcomes: $y(0)$, the outcome it would realize *without* treatment, and $y(1)$, the outcome it would realize *with* treatment. We only ever observe one of them, the one corresponding to the treatment actually received:

$$y = (1 - w)\, y(0) + w\, y(1),$$

where $w = 1$ for the treated and $w = 0$ for controls. The individual causal effect is $y(1) - y(0)$, which is never directly observed for any single unit.

A regression model for the observed outcome, including controls $x_1, \ldots, x_k$, is

$$\mathbb{E}(y \mid w, \mathbf{x}) = \alpha + \tau\, w + \gamma_1 x_1 + \cdots + \gamma_k x_k.$$

The controls $x_1, \ldots, x_k$ are included precisely to address self-selection: they soak up the systematic differences between who gets treated and who does not, moving us closer to the as-good-as-random-assignment ideal. The coefficient τ on the treatment dummy is the treatment effect, *assumed constant across units* in this specification.

6.5.2 Regression Adjustment

For τ to recover the causal effect, we need the controls to render treatment ignorable. The required condition is *conditional independence*: given the covariates, treatment assignment is independent of the potential outcomes,

$$w \perp \big[\, y(0),\, y(1) \,\big] \;\Big|\; x_1, \ldots, x_k.$$

This is the *unconfoundedness* (or *selection-on-observables*) assumption, and the estimator built on it is called *regression adjustment*: we adjust for measured differences across units, treating treatment as random *within* cells defined by the covariates. It is a strong assumption — it asserts there are no unmeasured confounders — but it is what makes the regression coefficient τ interpretable as a causal effect.

6.5.3 Heterogeneous Effects and the Average Treatment Effect

The constant-τ model is restrictive: in reality the treatment may help some units more than others. We can relax it by letting the effect vary with the covariates, interacting the treatment dummy with each (mean-centered) control:

$$y_i = \alpha + \tau\, w_i + \sum_{j=1}^{k} \gamma_j x_{ij} + \sum_{j=1}^{k} \delta_j\, w_i\,(x_{ij} - \bar{x}_j) + u_i.$$

With the interactions *centered* at the covariate means, the coefficient τ on w_i equals the *average treatment effect* (ATE) — the mean of the individual effects $y(1) - y(0)$ over the population. The centering is what makes this work: it ensures the interaction terms average to zero in the sample, so τ captures the effect at the average covariate values, which coincides with the population-average effect.

Restricted vs. unrestricted regression adjustment

- *Unrestricted regression adjustment* (URA): the model with the $w_i\,(x_{ij} - \bar{x}_j)$ interactions, allowing the treatment effect to differ across units. It is closer to reality, since effects are rarely homogeneous, and its τ estimates the ATE.
- *Restricted regression adjustment* (RRA): the simpler model with no treatment–covariate interactions, forcing a single common treatment effect for everyone.

Remark (Which terms to center).

It does not matter whether the *main* control terms $\gamma_j x_{ij}$ are centered — centering them only relabels the intercept α. But it *does* matter that the *interaction* terms are centered: if $w_i\, x_{ij}$ is used uncentered, the coefficient τ no longer equals the ATE but instead the treatment effect evaluated at $x = 0$, which is generally not the quantity of interest.

6.5.4 An Equivalent Two-Regression Estimator of the ATE

The URA-based ATE can also be computed by fitting the covariate model separately within each treatment arm and then predicting both potential outcomes for every unit. Using only the control observations, estimate

$$\widehat{y}_i^{(0)} = \widehat{\alpha}^{(0)} + \widehat{\gamma}_{0,1} x_{i1} + \cdots + \widehat{\gamma}_{0,k} x_{ik},$$

and, using only the treated observations, estimate

$$\widehat{y}_i^{(1)} = \widehat{\alpha}^{(1)} + \widehat{\gamma}_{1,1} x_{i1} + \cdots + \widehat{\gamma}_{1,k} x_{ik}.$$

Then, for *every* unit in the full sample — regardless of which group it actually belongs to — form both predicted potential outcomes $\widehat{y}_i^{(0)}$ and $\widehat{y}_i^{(1)}$, and average their difference:

$$\widehat{\mathrm{ATE}} = \frac{1}{n} \sum_{i=1}^{n} \left(\widehat{y}_i^{(1)} - \widehat{y}_i^{(0)} \right).$$

This delivers the *same* point estimate as the centered-interaction (URA) regression. Its drawback is purely practical: computing a correct standard error by hand for this two-step predict-and-average estimator is awkward, whereas the single interacted regression reports the standard error of τ automatically.

Remark (Where this is heading).

> Dummy variables let qualitative information enter regression without any new estimation theory — only new interpretation: an isolated dummy shifts the intercept, an interacted dummy shifts the slope, and joint F (Chow) tests ask whether the groups differ at all. Two threads point forward. The linear probability model and the Chow test both founder on non-constant error variance, motivating the systematic treatment of heteroskedasticity in Chapter 7; and the treatment-effect framework, which here rests on selection-on-observables, is exactly where instrumental variables (Chapter 10) take over when unmeasured confounders make that assumption untenable.

Chapter 7

Heteroskedasticity

Every variance formula we have written so far rests on a single quiet assumption: that the error term has the same spread at every value of the regressors. This is homoskedasticity, assumption SLR.5 / MLR.5 of the Gauss–Markov family. It is a convenience, not a law of nature, and in real data it routinely fails. The dispersion of savings around its mean is wider for high-income households than for low-income ones; the spread of test scores is narrower in small classes than in large ones; the variance of a binary outcome is mechanically tied to its mean. When the conditional variance of the error moves with the regressors, we say the error is *heteroskedastic*.

The good news is that heteroskedasticity is a far milder ailment than the endogeneity of Chapter 2. It has nothing to do with whether OLS lands on the right answer: the slope estimators remain unbiased and consistent, and the unconditional-variance reading of R^2 is untouched. What heteroskedasticity breaks is the bookkeeping of *precision*. The textbook variance formula σ^2/SST_x is no longer correct, so every standard error, t statistic, F statistic, and confidence interval built from it is wrong — and OLS forfeits its claim to being the best linear unbiased estimator.

This chapter does three things, in order. First we sort out exactly what survives and what fails. Then we repair inference without touching the estimator, through heteroskedasticity-robust (White/Huber/Eicker) standard errors, and we develop two tests — Breusch–Pagan and White — for deciding whether the repair is needed. Finally we ask whether we can do better than OLS by reweighting the data, through weighted and feasible generalized least squares. Throughout, keep one slogan in mind: heteroskedasticity is a problem of *inference*, not of *estimation*.

7.1 Properties of OLS Under Heteroskedasticity

Recall the homoskedasticity assumption, here stated for the multiple regression model with regressor vector $\vec{x} = (x_1, \ldots, x_k)$.

Assumption 7.1: MLR.5 (Homoskedasticity)

The error has constant conditional variance,

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

the same value of σ^2 for every configuration of the regressors. When this fails — when $\mathrm{Var}(u \mid \vec{x})$ depends on $\vec{x}$ — the error is *heteroskedastic*.

7.1.1 Unbiasedness and Consistency Survive

It is worth being very precise about which assumptions each property needs, because the headline of this chapter is that heteroskedasticity touches none of them.

- **Unbiasedness** ($\mathbb{E}\left(\widehat{\beta}_j\right) = \beta_j$ for every sample size) follows from MLR.1–MLR.4. The binding assumption is the zero conditional mean MLR.4, $\mathbb{E}(u \mid \vec{x}) = 0$.
- **Consistency** ($\widehat{\beta}_j \xrightarrow{p} \beta_j$ as $n \to \infty$) needs much less. It is enough that the regressors be *uncorrelated* with the error, namely $\mathbb{E}(u) = 0$ and $\mathrm{Cov}(x_j, u) = 0$ for each j. This is strictly weaker than zero conditional mean.

What survives heteroskedasticity

Neither the unbiasedness condition (MLR.4) nor the weaker consistency condition (zero correlation of regressors with the error) mentions the conditional *variance* of u at all. Therefore:

Under heteroskedasticity, OLS is still unbiased and still consistent.

Heteroskedasticity is a statement about the second moment of u given $\vec{x}$; unbiasedness and consistency are statements about its first moment. The two never meet.

Remark (Do not conflate the two conditions).

A common slip is to write "consistency requires $\mathbb{E}(u \mid \vec{x}) = 0$." It does not. Consistency requires only the population orthogonality conditions $\mathbb{E}(u) = 0$ and $\mathrm{Cov}(x_j, u) = 0$, which the sample moment conditions of OLS impose by construction. Zero conditional mean is sufficient for these (it implies $\mathrm{Cov}(x_j, u) = 0$) but is much stronger than necessary. This distinction matters here: it is precisely because consistency asks so little that heteroskedasticity — a statement purely about variances — cannot disturb it.

7.1.2 The Interpretation of R^2 Is Unchanged

A pleasant corollary is that the population content of R^2 does not change either. Recall from Chapter 4 that the sample sums of squares, divided by n, converge to population variances:

$$\frac{1}{n}\mathrm{SSR} = \frac{1}{n}\sum_{i=1}^{n}\hat{u}_i^2 \xrightarrow{p} \sigma_u^2, \qquad \frac{1}{n}\mathrm{SST} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 \xrightarrow{p} \sigma_y^2,$$

where $\sigma_u^2 = \mathrm{Var}(u)$ and $\sigma_y^2 = \mathrm{Var}(y)$ are *unconditional* variances. Hence

$$R^2 = 1 - \frac{\mathrm{SSR}}{\mathrm{SST}} \xrightarrow{p} 1 - \frac{\sigma_u^2}{\sigma_y^2}.$$

The point is that $\sigma_u^2 = \mathrm{Var}(u)$ is the *unconditional* error variance, an average over all values of $\vec{x}$. Heteroskedasticity concerns how $\mathrm{Var}(u \mid \vec{x})$ varies *across* values of $\vec{x}$, but it leaves the overall average $\mathrm{Var}(u)$ in place. By the law of total variance, the unconditional variance is unaffected by how the conditional variance is distributed, so the probability limit of R^2 is exactly what it was under homoskedasticity. The familiar reading — "R^2 is the fraction of the variance of y explained by the regressors" — still holds.

7.1.3 What Breaks: The Usual Variance Formula and BLUE

Two things do fail, and they are the reason this chapter exists.

What heteroskedasticity breaks

1. **The usual variance formulas for OLS are no longer valid.** The formula $\mathrm{Var}\left(\widehat{\beta}_1\right) = \sigma^2/\mathrm{SST}_x$ (and its multiple-regression analogue) was derived using MLR.5. Without it, that formula is simply the wrong number, so the conventional standard errors, t statistics, F statistics, and confidence intervals are all invalid — regardless of sample size.
2. **OLS is no longer BLUE.** The Gauss–Markov theorem requires homoskedasticity. Drop MLR.5 and OLS, while still unbiased, is no longer the minimum-variance linear unbiased estimator. A reweighted estimator (Section 7.4) can do better.

7.1.4 The Correct Variance Under Heteroskedasticity

To see exactly how the variance formula changes, return to the simple regression $y_i = \beta_0 + \beta_1 x_i + u_i$ and let the conditional variance carry an index,

$$\mathrm{Var}(u_i \mid x_i) = \sigma_i^2,$$

the general form of heteroskedasticity (each observation may have its own error variance). As in Chapter 1, write the slope estimator as the truth plus a weighted sum of errors,

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})\, y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} = \beta_1 + \frac{\sum_{i=1}^n (x_i - \bar{x})\, u_i}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

Conditioning on the regressors and using independence across i,

$$\mathrm{Var}\left(\widehat{\beta}_1 \mid \vec{x}\right) = \frac{\sum_{i=1}^n (x_i - \bar{x})^2\, \mathrm{Var}(u_i \mid x_i)}{\left[\sum_{i=1}^n (x_i - \bar{x})^2\right]^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})^2\, \sigma_i^2}{\left[\sum_{i=1}^n (x_i - \bar{x})^2\right]^2}. \qquad (\star)$$

Remark (Which formula is general?).

Equation $(\star)$ is the *general* variance of the OLS slope: it holds whether or not the errors are homoskedastic. Only when every σ_i^2 equals the same σ^2 can we pull it out of the sum, $\sum (x_i - \bar{x})^2 \sigma^2 = \sigma^2 \, \mathrm{SST}_x$, and collapse $(\star)$ to the familiar σ^2/SST_x. So the textbook formula is the special case; $(\star)$ is the truth.

7.2 Heteroskedasticity-Robust Inference

Formula $(\star)$ involves the unknown variances σ_i^2, and there are as many of them as there are observations, so we cannot estimate each one separately. The breakthrough of White (1980), building on Eicker and Huber, is that we do not need to: to estimate the *sum* in $(\star)$ consistently it suffices to replace each unknown σ_i^2 by the squared OLS residual $\hat{u}_i^2$.

Definition 7.2: Heteroskedasticity-Robust Variance Estimator (simple regression)

A valid estimator of $\mathrm{Var}\left(\widehat{\beta}_1\right)$ under arbitrary heteroskedasticity is

$$\widehat{\mathrm{Var}\left(\widehat{\beta}_1\right)} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2 \, \hat{u}_i^2}{\left[\sum_{i=1}^{n} (x_i - \bar{x})^2\right]^2}, \qquad \mathrm{se}(\widehat{\beta}_1) = \sqrt{\widehat{\mathrm{Var}\left(\widehat{\beta}_1\right)}}.$$

The resulting $\mathrm{se}(\widehat{\beta}_1)$ is the *heteroskedasticity-robust* standard error.

The recipe is to take the general formula $(\star)$ and plug $\hat{u}_i^2$ in for the unobservable σ_i^2. The justification is asymptotic: the robust estimator does not converge to the true finite-sample variance term by term, but its scaled version has the same probability limit as the scaled true variance.

Theorem 7.3: Consistency of the Robust Variance Estimator

Under MLR.1–MLR.4 (no homoskedasticity required), the scaled robust estimator is consistent for the scaled true variance:

$$\mathrm{plim}_{n\to\infty} \; n \, \widehat{\mathrm{Var}\left(\widehat{\beta}_1\right)} = \mathrm{plim}_{n\to\infty} \; n \, \mathrm{Var}\left(\widehat{\beta}_1 \,|\, \vec{x}\right) = \frac{\mathbb{E}\left((x - \mu_x)^2 \, \sigma^2(x)\right)}{\left(\sigma_x^2\right)^2},$$

where $\mu_x = \mathbb{E}(x)$, $\sigma_x^2 = \mathrm{Var}(x)$, and $\sigma^2(x) = \mathbb{E}\left(u^2 \,|\, x\right)$ is the conditional-variance function.

Proof. Scale numerator and denominator of $(\star)$ by appropriate powers of n. The denominator obeys $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \xrightarrow{p} \sigma_x^2$ by the law of large numbers, so its square $\xrightarrow{p} (\sigma_x^2)^2$. For the numerator, replacing $\bar{x}$ by μ_x is asymptotically negligible, and the law of large numbers gives

$$\frac{1}{n}\sum_{i=1}^{n} (x_i - \mu_x)^2 \, \sigma_i^2 \;\xrightarrow{p}\; \mathbb{E}\left((x - \mu_x)^2 \, \sigma^2(x)\right).$$

This delivers $\mathrm{plim}\; n \, \mathrm{Var}\left(\widehat{\beta}_1 \,|\, \vec{x}\right) = \mathbb{E}\left((x - \mu_x)^2 \sigma^2(x)\right) / (\sigma_x^2)^2$. For the estimated version,

the key step is that replacing σ_i^2 by $\hat{u}_i^2$ does not change the limit:

$$\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu_x)^2\,\hat{u}_i^2 \;\xrightarrow{p}\; \mathbb{E}\big((x-\mu_x)^2\,\sigma^2(x)\big)\,.$$

To see why the same limit appears, use the law of iterated expectations and $\mathbb{E}\big(u^2\,|\,x\big) = \sigma^2(x)$:

$$\mathbb{E}\big((x-\mu_x)^2\,u^2\big) = \mathbb{E}\big(\,\mathbb{E}\big((x-\mu_x)^2\,u^2\,|\,x\big)\,\big) = \mathbb{E}\big((x-\mu_x)^2\,\mathbb{E}\big(u^2\,|\,x\big)\big) = \mathbb{E}\big((x-\mu_x)^2\,\sigma^2(x)\big)\,.$$

Since the two scaled quantities share a probability limit, their ratio converges to 1, which is what consistency of the robust standard error means. □

Remark (Robust standard errors are large-sample objects).

> Two caveats follow from the proof. First, the robust standard error is justified only *asymptotically*; in small samples it can be a poor approximation (degrees-of-freedom corrections such as the HC1–HC3 variants exist to mitigate this). Second, the residual $\hat{u}_i$ is not the error u_i, but in large samples $\hat{u}_i^2 - u_i^2 \xrightarrow{p} 0$ uniformly enough that the substitution is harmless.

7.2.1 The Multiple Regression Case

The same idea extends to any coefficient in a multiple regression $y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + u_i$. Let $\hat{r}_{ij}$ denote the i-th residual from regressing x_j on all the other regressors, and let $\mathrm{SSR}_j = \sum_{i=1}^{n} \hat{r}_{ij}^2$ be the corresponding residual sum of squares (the partialling-out machinery of Chapter 2).

Definition 7.4: Robust Variance Estimator (multiple regression)

A valid estimator of $\mathrm{Var}\left(\widehat{\beta}_j\right)$ under arbitrary heteroskedasticity is

$$\widehat{\mathrm{Var}\left(\widehat{\beta}_j\right)} = \frac{\sum_{i=1}^{n} \hat{r}_{ij}^2\,\hat{u}_i^2}{\left(\mathrm{SSR}_j\right)^2}.$$

The square roots of these quantities are the *White*, *Huber*, or *Eicker* standard errors, after the three statisticians who developed them.

The construction uses two sets of residuals: the $\hat{u}_i$ from the original regression, and the $\hat{r}_{ij}$ from the auxiliary regression of x_j on the other regressors. In the one-regressor case $\hat{r}_{i1} = x_i - \bar{x}$ and the formula reduces to the simple-regression version above.

7.2.2 Robust t Statistics and the Fate of the F Test

Once we have a robust standard error, inference proceeds almost as before. The robust t statistic is the usual ratio with the robust standard error in the denominator,

$$t = \frac{\widehat{\beta}_j - \beta_{j,0}}{\mathrm{se}(\widehat{\beta}_j)}, \qquad t_{\text{robust}} = \frac{\widehat{\beta}_j - \beta_{j,0}}{\mathrm{se}_{\text{robust}}(\widehat{\beta}_j)}.$$

Under heteroskedasticity the ordinary t statistic is invalid even in large samples, because its standard error estimates the wrong variance. The robust t statistic, by contrast, is valid *asymptotically*: it has an approximate standard normal (equivalently t) distribution under the null. The only thing that changes between the two formulas is the version of the standard error in the denominator.

Remark (Are robust standard errors always larger?).

> As a rough empirical regularity — not a theorem — heteroskedasticity-robust standard errors tend to be *somewhat larger* than the conventional ones. They can also be smaller. What is diagnostically useful is the *gap*: if the robust and conventional standard errors differ substantially, that is a sign of strong heteroskedasticity. If they are close, the homoskedasticity-based inference was probably fine.

The usual F statistic for joint hypotheses is built on the conventional variance estimates and therefore also fails under heteroskedasticity. Heteroskedasticity-robust versions of the F test (and its asymptotic cousin, the robust Wald statistic) exist and are produced automatically by modern software, but their algebra is involved and we do not reproduce it here. The practical message is simple: report robust standard errors and robust test statistics whenever heteroskedasticity is a live possibility, which in cross-sectional work is almost always.

7.3 Testing for Heteroskedasticity

Robust standard errors fix inference whether or not heteroskedasticity is present, so why test at all? Two reasons. First, if the errors are in fact homoskedastic, OLS is BLUE and we lose efficiency by switching to a reweighting scheme; a test tells us whether the extra machinery is warranted. Second, evidence of heteroskedasticity that depends on the regressors in a structured way can be a symptom of deeper misspecification (a wrong functional form), which is worth knowing about. Every test below shares one idea: the conditional variance of u equals the conditional mean of u^2, so we look for dependence of $\hat{u}_i^2$ on the regressors.

7.3.1 The Breusch–Pagan Test

Work under MLR.4, so that $\mathbb{E}(u \mid \vec{x}) = 0$. Then the conditional variance is just the conditional second moment,

$$\mathrm{Var}(u \mid \vec{x}) = \mathbb{E}\left(u^2 \mid \vec{x}\right) - \left[\mathbb{E}(u \mid \vec{x})\right]^2 = \mathbb{E}\left(u^2 \mid \vec{x}\right).$$

Homoskedasticity says $\operatorname{Var}(u \mid \vec{x}) = \sigma^2$, a constant, so the null hypothesis can be written as the statement that $\mathbb{E}\big(u^2 \mid \vec{x}\big)$ does not vary with the regressors:

$$H_0: \quad \mathbb{E}\big(u^2 \mid x_1, \dots, x_k\big) = \mathbb{E}\big(u^2\big) = \sigma^2.$$

The Breusch–Pagan idea is to test this by regressing u^2 linearly on the regressors,

$$u^2 = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \cdots + \delta_k x_k + v.$$

Under H_0 none of the regressors should help, i.e. $\delta_1 = \delta_2 = \cdots = \delta_k = 0$. Of course u^2 is unobservable, so we use the squared OLS residual $\hat{u}^2$ in its place,

$$\hat{u}^2 = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \cdots + \delta_k x_k + \text{error}.$$

The error in this feasible regression is $\text{error}_i = v_i + (\hat{u}_i^2 - u_i^2)$, and the substitution is asymptotically innocuous because $\hat{u}_i^2 - u_i^2 \xrightarrow{p} 0$ in large samples.

Let $R^2_{\hat{u}^2}$ be the R-squared from this auxiliary regression of $\hat{u}^2$ on $x_1, \dots, x_k$. There are two equivalent test statistics.

Theorem 7.5: Breusch–Pagan Test Statistics

To test $H_0 : \delta_1 = \cdots = \delta_k = 0$ in the auxiliary regression of $\hat{u}^2$ on $x_1, \dots, x_k$:

- the F **form** is the usual F statistic for joint significance of all k regressors,

$$F = \frac{R^2_{\hat{u}^2}/k}{\big(1 - R^2_{\hat{u}^2}\big)/(n-k-1)} \overset{a}{\sim} F_{k,\, n-k-1};$$

- the LM **form** (Lagrange multiplier, or "nR^2" form) is

$$LM = n \cdot R^2_{\hat{u}^2} \overset{a}{\sim} \chi^2_k.$$

A large value of either statistic rejects homoskedasticity.

Note carefully the denominator of the F statistic: it is $(1 - R^2_{\hat{u}^2})/(n-k-1)$ — one minus the auxiliary R-squared, divided by the residual degrees of freedom $n-k-1$ (k slopes plus an intercept). No quantity is squared here. The LM form multiplies the auxiliary R^2 by the full sample size n and compares it to a χ^2_k critical value.

Remark (A small R^2 need not mean a small statistic).

> The auxiliary $R^2_{\hat{u}^2}$ is typically small even under strong heteroskedasticity, but this does *not* make the LM statistic small: $LM = nR^2_{\hat{u}^2}$ is the product of R^2 with the sample size, and a large n can turn a modest R^2 into a decisive rejection. Read the statistic, not the R^2.

7.3.2 The Algebraic Link Between F and LM

The two forms are asymptotically equivalent, and it is worth seeing why, since the same pairing of an F test and an nR^2 Lagrange-multiplier test recurs throughout the book. As

$n \to \infty$ with k fixed, an $F_{k,\,n-k-1}$ random variable behaves like χ^2_k/k:

$$F_{k,\,n-k-1} \xrightarrow{n\to\infty} F_{k,\infty} \overset{a}{\sim} \chi^2_k/k.$$

The reason is that the denominator of an F statistic is itself a normalized chi-square that converges to a constant. If $X_i \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$, then $\frac{1}{n}\sum_{i=1}^{n} X_i^2 \sim \chi^2_n/n$, and by the law of large numbers

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 \xrightarrow{p} \mathbb{E}\big(X_i^2\big) = \mathrm{Var}(X_i) = 1, \qquad \text{so} \qquad \chi^2_n/n \xrightarrow{p} 1.$$

Hence the F denominator stabilizes at 1 and F collapses to χ^2_k/k. Multiplying through, $LM \overset{a}{\sim} k \cdot F$ as $n \to \infty$, which is exactly the relationship $LM = nR^2_{\hat{u}^2}$ versus $F = \frac{R^2/k}{(1-R^2)/(n-k-1)}$ encodes once the small terms are dropped.

Remark (Two practical notes).

> Two further points about the Breusch–Pagan test. First, it only detects heteroskedasticity that is a *linear* function of the regressors; a variance that depends on $\vec{x}$ in a purely nonlinear, mean-zero way can slip past it. Second, taking logarithms of variables often dampens heteroskedasticity, and this works chiefly through logging the *dependent* variable rather than the regressors, because logs compress the right tail where the largest errors usually live.

7.3.3 The White Test

Because the Breusch–Pagan auxiliary regression is linear in $\vec{x}$, it misses heteroskedasticity that operates through squares and cross-products of the regressors. White's test enlarges the auxiliary regression to include the regressors, their squares, and all their pairwise interactions. With three regressors x_1, x_2, x_3 the White auxiliary regression is

$$\hat{u}^2 = \delta_0 + \underbrace{\delta_1 x_1 + \delta_2 x_2 + \delta_3 x_3}_{\text{levels}} + \underbrace{\delta_4 x_1^2 + \delta_5 x_2^2 + \delta_6 x_3^2}_{\text{squares}} + \underbrace{\delta_7 x_1 x_2 + \delta_8 x_1 x_3 + \delta_9 x_2 x_3}_{\text{interactions}} + \text{error},$$

and we test the joint significance of all the included terms (everything but the intercept) via the same F or $LM = nR^2$ statistic as before, now with the larger number of regressors driving the degrees of freedom.

The drawback is combinatorial. With k regressors the number of squares and interactions grows roughly as k^2, so the auxiliary regression burns through degrees of freedom quickly; with many regressors the test estimates a large number of nuisance parameters and its power deteriorates.

7.3.4 The Reduced (Special-Case) White Test

A clever economization sidesteps the explosion of terms. The fitted value $\hat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \cdots + \widehat{\beta}_k x_{ik}$ is a linear combination of the regressors, so $\hat{y}_i$ and $\hat{y}_i^2$ are functions of exactly the levels, squares, and cross-products that the full White test enumerates. We can therefore proxy for that whole list with just two terms.

Definition 7.6: Reduced White Test (special case)

Run the auxiliary regression of the squared residuals on the fitted values and their squares,

$$\hat{u}^2 = \delta_0 + \delta_1 \hat{y} + \delta_2 \hat{y}^2 + \text{error},$$

and test $H_0 : \delta_1 = \delta_2 = 0$ with the F statistic ($\overset{a}{\sim} F_{2,\,n-3}$) or the LM statistic $nR^2_{\hat{u}^2} \overset{a}{\sim} \chi^2_2$.

Two cautions on the construction. First, it is the *fitted* value $\hat{y}$, not the observed y, that enters: only $\hat{y}$ is a function of the regressors, so only $\hat{y}$ implicitly carries the squares and interactions we are after. Second, because $\hat{y}$ and $\hat{y}^2$ together stand in for the full battery of nonlinear terms, this regression tests dependence of $\hat{u}^2$ on the regressors, their squares, and their interactions *indirectly*, using only two degrees of freedom regardless of how many regressors the original model contains. This is why it is called the special case of the White test.

7.4 Weighted Least Squares

Robust standard errors leave OLS in place and merely correct its standard errors. But if we actually know the shape of the heteroskedasticity, we can do strictly better: by reweighting the data to undo the unequal variances we recover an estimator that is BLUE again. This is *weighted least squares* (WLS), the leading special case of generalized least squares.

7.4.1 Known Variance Up to a Constant

Suppose the conditional variance is known up to an overall scale,

$$\operatorname{Var}(u_i \,|\, \vec{x}_i) = \sigma^2 \, h(\vec{x}_i), \qquad h(\vec{x}_i) = h_i > 0,$$

where the function $h(\cdot) > 0$ captures the *shape* of the heteroskedasticity and the unknown constant σ^2 its overall level. The positivity $h_i > 0$ is the primary condition to verify — a variance cannot be negative — and it is also what limits where WLS can be applied. The functional form $h(\cdot)$ is something we must *assume*; in practice that assumption deserves an argument grounded in how the data are generated.

7.4.2 The Model Transformation

The trick is to divide the entire equation by $\sqrt{h_i}$, which rescales each observation's error to a common variance. Indeed,

$$\operatorname{Var}\left(\frac{u_i}{\sqrt{h_i}} \;\middle|\; \vec{x}_i \right) = \frac{1}{h_i} \operatorname{Var}(u_i \,|\, \vec{x}_i) = \frac{\sigma^2 h_i}{h_i} = \sigma^2,$$

a constant. Applying the same division to every term of the model $y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + u_i$ produces a transformed model that is homoskedastic:

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

or compactly, with starred variables $y_i^* = y_i/\sqrt{h_i}$, $x_{ij}^* = x_{ij}/\sqrt{h_i}$, $x_{i0}^* = 1/\sqrt{h_i}$, and $u_i^* = u_i/\sqrt{h_i}$,

$$y_i^* = \beta_0\, x_{i0}^* + \beta_1\, x_{i1}^* + \cdots + \beta_k\, x_{ik}^* + u_i^*.$$

Remark (Two features of the transformed model).

> Two things to notice. First, the transformed model has *no* ordinary intercept: the constant term becomes the regressor $x_{i0}^* = 1/\sqrt{h_i}$, which now varies across i. Second, the weighting factor need not be exactly $1/\sqrt{h_i}$ — any constant multiple of it gives the same WLS estimates, since the overall scale cancels. This is the same freedom that let us leave σ^2 unspecified.

7.4.3 WLS as Weighted Minimization

Running OLS on the transformed (starred) model is, by construction, the same as minimizing a *weighted* sum of squared residuals in the original variables:

$$\min_{b_0,\ldots,b_k} \sum_{i=1}^{n} \left(\frac{y_i}{\sqrt{h_i}} - b_0 \frac{1}{\sqrt{h_i}} - \cdots - b_k \frac{x_{ik}}{\sqrt{h_i}} \right)^2 = \min_{b_0,\ldots,b_k} \sum_{i=1}^{n} \frac{1}{h_i} \left(y_i - b_0 - b_1 x_{i1} - \cdots - b_k x_{ik} \right)^2.$$

Each squared residual is weighted by $1/h_i$, which gives the method its name. The logic of the weights is exactly the logic of efficiency: an observation with a large variance h_i is noisy and uninformative, so it receives a small weight $1/h_i$; an observation with a small variance is reliable and gets a large weight. (If h_i is constant across i, every weight is the same and WLS collapses back to OLS.)

Theorem 7.7: Optimality of WLS

If $\mathrm{Var}(u_i \mid \vec{x}_i) = \sigma^2 h_i$ with h_i known, and the remaining Gauss–Markov assumptions hold, then OLS applied to the transformed model — equivalently, WLS with weights $1/h_i$ — is the *best linear unbiased estimator* (BLUE). Because the transformed errors u_i^* are homoskedastic, the Gauss–Markov theorem applies to the transformed model.

In practice WLS standard errors are often noticeably smaller than OLS standard errors, which is precisely the efficiency gain the theorem promises. Two reminders complete the picture. The goodness-of-fit measure for WLS is a *weighted* R^2, $R^2 = 1 - \mathrm{SSR_w}/\mathrm{SST_w}$, computed from the weighted sums of squares of the transformed model. And the *interpretation* of the coefficients always returns to the *original* model $y = \beta_0 + \cdots + \beta_k x_k + u$: the transformation is a device for efficient estimation, not a respecification, so β_j still measures the partial effect of x_j on y.

Remark (WLS and robust SEs as a cross-check).

Robust standard errors and WLS are two responses to the same disease. They use the data differently, so when both the OLS-with-robust-SE results and the WLS results agree — coefficients close in magnitude with comparable standard errors — you can be confident the inference is sound. Large discrepancies are a warning that something beyond heteroskedasticity (such as a violation of MLR.4) may be at work.

7.4.4 A Leading Special Case: Averaged (Grouped) Data

A common and important situation where h_i is known arises when the data are reported as *averages over groups of unequal size* — average outcomes by city, by firm, by school. Suppose the individual-level model holds for person e in unit i,

$$y_{i,e} = \beta_0 + \beta_1 x_{1,i,e} + \beta_2 x_{2,i,e} + \beta_3 x_{3,i} + u_{i,e},$$

where unit i contains m_i individuals and $x_{3,i}$ is a unit-level regressor (the same for everyone in the unit). Averaging over the m_i members of unit i gives a regression in the group means,

$$\bar{y}_i = \beta_0 + \beta_1 \bar{x}_{1,i} + \beta_2 \bar{x}_{2,i} + \beta_3 x_{3,i} + \bar{u}_i, \qquad \bar{u}_i = \frac{1}{m_i} \sum_{e=1}^{m_i} u_{i,e}.$$

Even if the individual errors $u_{i,e}$ are homoskedastic at the individual level — i.i.d. with variance σ^2 — the averaged error is not. Averaging shrinks variance in proportion to group size:

$$\operatorname{Var}(\bar{u}_i) = \operatorname{Var}\left(\frac{1}{m_i} \sum_{e=1}^{m_i} u_{i,e} \right) = \frac{1}{m_i^2} \sum_{e=1}^{m_i} \operatorname{Var}(u_{i,e}) = \frac{m_i \, \sigma^2}{m_i^2} = \frac{\sigma^2}{m_i}.$$

So the grouped-data error variance is σ^2 / m_i: large units have precise averages and small units have noisy ones. This is heteroskedasticity with a *known* form, $h_i = 1/m_i$, and the correct WLS weight is $1/h_i = m_i$.

Weighting grouped data

When the regression is run on group averages and the underlying individual errors are homoskedastic, use WLS with weights equal to the *group size* m_i. Larger groups give more reliable averages and so deserve more weight. If you doubt the individual-level homoskedasticity, run WLS and report *robust* standard errors for the transformed model, getting the best of both corrections.

7.5 Feasible GLS: Unknown Variance Function

The preceding section assumed $h(\vec{x})$ known. Usually it is not. *Feasible* generalized least squares (FGLS) estimates the variance function from the data and then performs WLS with the estimated weights. The challenge is to model $h(\vec{x})$ in a way that automatically respects the positivity $h(\vec{x}) > 0$. The standard device is to put the linear index inside an exponential, which is positive for any value of its argument.

Definition 7.8: Exponential Variance Model

Model the conditional variance as

$$\operatorname{Var}(u \mid \vec{x}) = \sigma^2 \exp(\delta_0 + \delta_1 x_1 + \cdots + \delta_k x_k) = \sigma^2 h(\vec{x}),$$

where the $\exp(\cdot)$ guarantees $h(\vec{x}) > 0$ regardless of the δ's.

To estimate the δ's, start from $\mathbb{E}\big(u^2 \mid \vec{x}\big) = \operatorname{Var}(u \mid \vec{x})$ and write the second moment as the variance times a multiplicative error,

$$u^2 = \sigma^2 \exp(\delta_0 + \delta_1 x_1 + \cdots + \delta_k x_k)\ v,$$

where $v > 0$ is assumed independent of the regressors with $\mathbb{E}(v) = 1$. Taking logs linearizes the model:

$$\log(u^2) = \alpha_0 + \delta_1 x_1 + \cdots + \delta_k x_k + e,$$

where α_0 absorbs $\log \sigma^2$, δ_0, and $\mathbb{E}(\log v)$, and $e = \log v - \mathbb{E}(\log v)$ has mean zero. Since u^2 is unobserved, replace it with $\hat{u}^2$ and run OLS:

$$\log(\hat{u}^2) = \hat{\alpha}_0 + \hat{\delta}_1 x_1 + \cdots + \hat{\delta}_k x_k + \text{error}.$$

Exponentiating the fitted values recovers estimated weights that are guaranteed positive,

$$\hat{h}_i = \exp\Big(\hat{\alpha}_0 + \hat{\delta}_1 x_{i1} + \cdots + \hat{\delta}_k x_{ik}\Big) > 0.$$

These $\hat{h}_i$ are then used as the WLS weights $1/\hat{h}_i$. As always, $\hat{h}_i$ need only be proportional to the true h_i; the overall scale is irrelevant.

FGLS recipe

1. Run OLS of y on $x_1, \ldots, x_k$ and save the residuals $\hat{u}_i$.
2. Run OLS of $\log(\hat{u}_i^2)$ on $x_1, \ldots, x_k$ and save the fitted values $\hat{g}_i$.
3. Form the estimated weights $\hat{h}_i = \exp(\hat{g}_i)$.
4. Run WLS of y on $x_1, \ldots, x_k$ with weights $1/\hat{h}_i$.

The log in step 2 and the exp in step 3 are the two halves of a single trick: regressing on the log keeps the model linear, and exponentiating back guarantees the weights are positive.

7.5.1 Misspecified Variance Functions

The variance model $h(\vec{x})$ is, like any model, possibly wrong. What happens then? Remarkably little, and the reason is the central theme of this chapter once more.

WLS is robust to a misspecified variance function

If the variance function $h(\vec{x})$ is misspecified but MLR.1–MLR.4 hold, WLS is still *consistent*. As we have stressed, heteroskedasticity — and hence any modeling of it — has nothing to do with unbiasedness or consistency, which depend only on the conditional mean MLR.4, not on the conditional variance. A wrong $h(\vec{x})$ costs efficiency, not consistency.

Two practical corollaries follow. First, if OLS and WLS produce *very different* coefficient estimates, the culprit is usually not the variance model but a failure of MLR.4 (e.g. omitted variables or a wrong functional form): both estimators should be consistent for the same β_j when MLR.4 holds, so a large gap signals that one of them is inconsistent. Second, when heteroskedasticity is strong, it is often still worthwhile to use an admittedly imperfect form of $h(\vec{x})$ in WLS, because even an approximate reweighting buys an efficiency gain over plain OLS. (If you remain uneasy about the variance model, compute robust standard errors after WLS, which protect inference even when the weights are wrong.)

7.5.2 WLS for the Linear Probability Model

A textbook case where the variance function is *exactly* known — not merely assumed — is the linear probability model (LPM) of Chapter 6, where the dependent variable is binary. There the conditional mean is a probability,

$$\mathrm{P}(y = 1 \mid \vec{x}) = p(\vec{x}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k,$$

and because a Bernoulli variable's variance is determined by its mean, the conditional variance is pinned down with no extra assumption:

$$\mathrm{Var}(y \mid \vec{x}) = p(\vec{x})\,[1 - p(\vec{x})]\,.$$

The LPM is therefore *inherently* heteroskedastic, and its variance function is known. The natural WLS weight uses the fitted probability $\hat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \cdots + \widehat{\beta}_k x_{ik}$ in place of $p(\vec{x}_i)$:

$$\hat{h}_i = \hat{y}_i\,(1 - \hat{y}_i).$$

Example (WLS for the LPM).

Estimate an LPM for whether a household owns its home, $\mathrm{P}(\texttt{own} = 1 \mid \vec{x}) = \beta_0 + \beta_1\,\texttt{income} + \beta_2\,\texttt{age}$, on n households. Describe a WLS estimator that accounts for the model's built-in heteroskedasticity.

Solution.

First run OLS of `own` on `income` and `age` to obtain fitted probabilities $\hat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1\,\texttt{income}_i + \widehat{\beta}_2\,\texttt{age}_i$. Form the known variance estimates $\hat{h}_i = \hat{y}_i(1 - \hat{y}_i)$. Then run WLS of `own` on `income` and `age` with weights $1/\hat{h}_i = 1/[\hat{y}_i(1 - \hat{y}_i)]$. The reweighting downweights households whose fitted probability is near 0 or 1 (where the binary outcome is nearly deterministic and $\hat{h}_i$ is small) relative to those near $\hat{y}_i = 0.5$ (where $\hat{h}_i$ is

largest). No separate variance model needs to be assumed, because the LPM's variance function is exact.

Remark (When the LPM weights misbehave).

The procedure is infeasible whenever a fitted probability falls outside $(0, 1)$, since then $\hat{h}_i = \hat{y}_i(1 - \hat{y}_i) \leq 0$ and the weight $1/\hat{h}_i$ is undefined or negative. Two remedies:

- if such cases are rare, truncate the offending fitted values to, say, 0.01 or 0.99 and proceed with WLS;
- if they are common, abandon WLS and report OLS with heteroskedasticity-robust standard errors instead.

The out-of-bounds fitted values are themselves a known limitation of the LPM, which the robust-standard-error route sidesteps entirely.

Remark (Where this leaves us).

Heteroskedasticity is the gentlest of the classical-assumption failures: it leaves OLS unbiased and consistent and leaves the meaning of R^2 intact, demanding only that we repair our standard errors. The robust standard errors of Section 7.2 do this with no extra assumptions; WLS and FGLS go further and recover efficiency when the variance function is known or estimable. Chapter 8 turns to a graver class of problems — misspecified functional form, measurement error, and missing or non-random data — where, unlike here, the conditional-mean assumption MLR.4 itself is at stake and consistency is no longer guaranteed.

Chapter 8

Specification and Data Issues

By now we have a thorough understanding of what makes ordinary least squares work. Under the Gauss–Markov assumptions the slope estimators are unbiased and consistent; under homoskedasticity their variances take a simple form; and Chapter 7 showed how to repair inference when that last convenience fails. In every case the load-bearing assumption was zero conditional mean, $\mathbb{E}(u \mid \vec{x}) = 0$: the error must be uncorrelated with the regressors. We have repeatedly warned that this assumption is the fragile one, and that omitting a relevant variable correlated with $\vec{x}$ breaks it. This chapter is about the other, subtler ways the same assumption can fail — failures that are not about *which* variables we include, but about the *form* the model takes and the *quality of the data* we feed it.

Three threads run through the chapter. The first is *specification*: have we written down the right functional form, and how would we know? We develop the RESET test for functional-form misspecification, and a way to choose between two rival models that are not nested inside one another. The second thread is *unobservables*: a variable we know matters cannot be measured, so we substitute something we can measure — a proxy — and ask when that substitution leaves our coefficients of interest intact. We also study the random coefficient model, where the very parameters differ across individuals. The third thread is *imperfect data*: variables measured with error, observations missing entirely, and a handful of extreme observations that distort the fit. Each problem is diagnosed the same way — write down what the contaminated regression secretly estimates, and ask whether its error term is still uncorrelated with the regressors. When it is not, we either find a repair or characterize the bias precisely. The single most important quantitative result of the chapter, *attenuation bias*, falls out of exactly this bookkeeping.

8.1 Testing for Functional-Form Misspecification

Recall the assumption that does all the work, here written for the multiple regression model with regressor vector $\vec{x} = (x_1, \ldots, x_k)$.

Assumption 8.1: MLR.4 (Zero Conditional Mean)

The error has zero conditional mean,

$$\mathbb{E}(u \mid \vec{x}) = \mathbb{E}(u \mid x_1, \ldots, x_k) = 0.$$

We usually think of this assumption being broken by an *omitted variable* — some determinant of y left inside u and correlated with the included regressors. But it can also be broken even when every relevant variable is present, if those variables enter in the wrong *form*. Suppose the true conditional mean depends on x_1^2 or on an interaction $x_1 x_2$, but we fit a model that is linear in x_1 and x_2 alone. Then the omitted nonlinear terms are swept into u, they are functions of the included regressors, and so they are mechanically correlated with $\vec{x}$: zero conditional mean fails. We call this *functional-form misspecification*. Its consequence is the usual one — the slope estimators are biased and inconsistent. And the damage is not confined to one coefficient: because the regressors are correlated with one another, a single contaminated term typically biases all of the slopes at once.

Functional form is part of the specification

Getting the list of variables right is necessary but not sufficient. If the right variables enter in the wrong form — a missing square, a missing interaction, a level where a log belongs — the omitted nonlinear terms are functions of the included regressors, so they correlate with $\vec{x}$ and MLR.4 fails. The symptom is the same as a classic omitted variable: biased, inconsistent slopes.

8.1.1 The RESET Test

We would like a general test that asks, "is there evidence of *any* omitted nonlinearity?" without our having to guess in advance which squares or interactions are missing. Ramsey's *Regression Specification Error Test* (RESET) supplies one. The idea is economical: instead of adding back every conceivable squared and interaction term — of which there are many once k is large — we add back low-order powers of the *fitted values*. The fitted value

$$\hat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \cdots + \widehat{\beta}_k x_k$$

is a single linear combination of all the regressors, so $\hat{y}^2$ is a particular combination of all the squares and cross-products x_j^2 and $x_j x_\ell$, and $\hat{y}^3$ brings in cubic terms. Thus two extra regressors stand in for a whole family of nonlinear terms at once.

Definition 8.2: RESET (Ramsey's Regression Specification Error Test)

Let $\hat{y}_i$ be the fitted values from the original OLS regression of y on $x_1, \ldots, x_k$. The RESET regression is the expanded model

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \delta_1 \hat{y}^2 + \delta_2 \hat{y}^3 + \text{error},$$

and the test of correct functional form is the joint test

$$H_0 : \ \delta_1 = \delta_2 = 0.$$

If the original model has the correct functional form — if $\mathbb{E}(y \mid \vec{x})$ really is linear in $x_1, \ldots, x_k$ — then no power of $\hat{y}$ should help explain y, so under H_0 the population coefficients δ_1, δ_2 are zero. Rejecting H_0 is evidence that omitted higher-order terms (squares, interactions) belong in the model: functional-form misspecification. The test is run as an ordinary F test of the two exclusion restrictions.

Theorem 8.3: Distribution of the RESET Statistic

Under MLR.1–MLR.5 and the null hypothesis $H_0 : \delta_1 = \delta_2 = 0$, the F statistic for the two restrictions has, in large samples, an F distribution with two numerator degrees of freedom:

$$F \sim F_{2,\, n-k-3}.$$

The numerator carries $q = 2$ degrees of freedom (one for each excluded term, $\hat{y}^2$ and $\hat{y}^3$); the denominator carries $n - k - 3$, because the unrestricted RESET regression estimates $k + 3$ parameters in all — the intercept, the k original slopes, and the two coefficients δ_1, δ_2.

Two cautions complete the picture. First, one must never add $\hat{y}$ *itself* (its first power) to the regression: since $\hat{y}$ is by construction an exact linear combination of $1, x_1, \ldots, x_k$, including it produces perfect multicollinearity and the regression cannot be estimated. We start the powers at the square. Second, and more important in practice, RESET is a *detector*, not a *diagnosis*. A rejection tells us that *some* nonlinearity is missing, but it gives no guidance about *which* variable enters wrongly or *what* form would fix it. It is a smoke alarm: useful for telling you there is a fire, useless for telling you which room.

Remark (Why fitted powers and not the regressors' own powers).

RESET is deliberately frugal. Adding $\hat{y}^2$ and $\hat{y}^3$ costs only two degrees of freedom regardless of how many regressors the model has, whereas adding every x_j^2 and every interaction $x_j x_\ell$ would cost $k + \binom{k}{2}$ degrees of freedom and quickly exhaust a small sample. The price of this frugality is the loss of specificity just noted: by collapsing all nonlinear terms into powers of a single index $\hat{y}$, we can detect misspecification cheaply but cannot localize it.

8.1.2 Testing Against Nonnested Alternatives

The exclusion tests we have used so far — and RESET itself — all compare a restricted model with an unrestricted one that *contains* it. The restricted model is a special case of the larger model, obtained by setting some coefficients to zero. Such models are called *nested*. But applied work often confronts two genuinely different specifications, neither of which is a special case of the other. The classic example is a choice between levels and logs:

$$\begin{aligned} \text{Model 1:} \quad & y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u, \\ \text{Model 2:} \quad & y = \beta_0 + \beta_1 \log x_1 + \beta_2 \log x_2 + u. \end{aligned}$$

There is no restriction on Model 2 that collapses it to Model 1, nor vice versa. They are *nonnested* alternatives, and the ordinary F test does not apply.

The encompassing (comprehensive) model. One natural approach is to build a single *comprehensive* model that contains both rivals as special cases, and then test each rival as a set of exclusion restrictions inside it. For the levels-versus-logs example the comprehensive model is

$$y = \gamma_0 + \gamma_1 x_1 + \gamma_2 x_2 + \gamma_3 \log x_1 + \gamma_4 \log x_2 + u.$$

Model 1 is the special case $\gamma_3 = \gamma_4 = 0$; Model 2 is the special case $\gamma_1 = \gamma_2 = 0$. We can now run two ordinary F tests:

$$H_0^{(1)} : \ \gamma_3 = \gamma_4 = 0 \quad \text{(test of Model 1)}, \qquad H_0^{(2)} : \ \gamma_1 = \gamma_2 = 0 \quad \text{(test of Model 2)}.$$

If $H_0^{(1)}$ cannot be rejected we say Model 1 is *preferred*; if $H_0^{(2)}$ cannot be rejected we say Model 2 is preferred. The word "preferred" is chosen carefully. Preference is not truth: failing to reject Model 1 does not prove it is correctly specified, and it is entirely possible that *neither* rival is the true model, since the truth requires zero conditional mean to hold and neither levels nor logs need satisfy it. The four outcomes (reject both, reject neither, reject exactly one of each) are all possible, and a clear winner need not emerge.

Remark (A subtlety of the comprehensive approach).

> There is a logical wrinkle worth stating. Suppose Model 1 really is the true model, so $\mathbb{E}(u \mid x_1, x_2) = 0$. Then *any* transformation of the regressors is also uncorrelated with the error — in particular $\log x_1$ and $\log x_2$ carry no additional explanatory power, so $\gamma_3 = \gamma_4 = 0$ in the population. The comprehensive test of Model 1 is then valid. But the comprehensive framework treats the two models symmetrically, and that symmetry can blur which rival actually fails. A more pointed procedure tests each model directly against its rival's *fit*.

The Davidson–MacKinnon regression. A sharper test, due to Davidson and MacKinnon, takes the null model seriously and confronts it with the *fitted values* of the rival. The recipe, with Model 1 as the null, is:

1. Estimate the alternative, Model 2, by OLS and save its fitted values $\tilde{y}_i$ (the fitted values of the rival).

2. Estimate the augmented null regression

$$y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 \tilde{y} + \text{error},$$

and examine the t statistic on α_3. The test is $H_0 : \alpha_3 = 0$.

The logic is that if Model 1 is adequate, the rival's fitted value $\tilde{y}$ — which encodes the log specification — should add nothing once the level regressors are present, so $\alpha_3 = 0$. A significant α_3 is evidence against Model 1 in favor of Model 2. By the same construction with the roles reversed, one obtains a t test of Model 2 against Model 1: estimate Model 1, save its fitted values $\hat{y}_i$, and test the coefficient on $\hat{y}$ in a regression of y on $\log x_1, \log x_2$, and $\hat{y}$.

Remark (Three things to remember about nonnested testing).

1. It matters which model you name the null — the two roles are not symmetric, and reversing them can reverse the verdict.
2. No clear winner need emerge: both models can be rejected (both misspecified), or neither can be rejected (the data cannot tell them apart). Even when one is preferred, it may still differ from the true model.
3. These tests require the rival models to share the *same* dependent variable. They cannot compare, say, a model for y against a model for $\log y$, because the sums of squares are then not on the same scale.

8.2 Using Proxy Variables for Unobserved Regressors

Some determinants of y are real, important, and impossible to measure directly. Individual "ability" in a wage equation, "management quality" in a firm-performance equation, "health" in a mortality equation — each clearly belongs in the model, yet none has a thermometer. Leaving such a variable out is a textbook omitted-variable problem: if it is correlated with the included regressors, every slope is biased. When we cannot measure the variable itself, the next best thing is to measure a *proxy* — an observable variable that stands in for the unobservable — and include the proxy instead. The question is exactly when this substitution rescues the coefficients we care about.

8.2.1 The Plug-In Solution and When It Works

Suppose the population model is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3^* + u,$$

where x_3^* (think: ability) is unobserved while x_1, x_2 (think: education, experience) are observed. We have available a proxy x_3 (think: an IQ score) related to the unobservable by

$$x_3^* = \delta_0 + \delta_3 x_3 + v_3,$$

where v_3 is the part of x_3^* that the proxy fails to capture. The *plug-in solution* is simply to run OLS of y on x_1, x_2, x_3 — substituting the proxy for the unobservable. Substituting the proxy relation into the population model gives

$$y = (\beta_0 + \beta_3\delta_0) + \beta_1 x_1 + \beta_2 x_2 + (\beta_3\delta_3)\, x_3 + (u + \beta_3 v_3).$$

This is a regression of y on x_1, x_2, x_3 with a composite error $u + \beta_3 v_3$. For OLS to deliver unbiased, consistent estimates of β_1 and β_2 — our coefficients of interest — the composite error must be uncorrelated with all three included regressors. Two assumptions secure exactly this.

Assumption 8.4: Conditions for a Good Proxy Variable

(i) The original error is exogenous. u is uncorrelated with x_1, x_2, and x_3^*; in particular $\mathbb{E}(u \mid x_1, x_2, x_3^*) = 0$. The deeper requirement embedded here is that the proxy be irrelevant once the true variable is held fixed — u is also uncorrelated with x_3. The proxy is "just a proxy": it does not belong in the structural model in its own right, and would drop out if x_3^* could be observed.

(ii) The proxy is a good stand-in. The error v_3 in the proxy relation is uncorrelated with x_1, x_2, and x_3. Equivalently, $\mathbb{E}(x_3^* \mid x_1, x_2, x_3) = \mathbb{E}(x_3^* \mid x_3) = \delta_0 + \delta_3 x_3$: once we know the proxy, the other regressors carry no further information about the unobservable.

Theorem 8.5: Validity of the Plug-In Estimator

Under conditions (i) and (ii), the composite error $u + \beta_3 v_3$ is uncorrelated with x_1, x_2, x_3. Consequently the OLS regression of y on x_1, x_2, x_3 consistently estimates β_1 and β_2, and the coefficient on x_3 consistently estimates $\beta_3\delta_3$ — a (scaled) version of the unobservable's effect.

Proof. Condition (i) makes u uncorrelated with x_1, x_2, x_3. Condition (ii) makes v_3 uncorrelated with x_1, x_2, x_3. Hence the linear combination $u + \beta_3 v_3$ is uncorrelated with each of x_1, x_2, x_3, which is precisely the exogeneity requirement for OLS on the rewritten equation. By the consistency of OLS under zero correlation between regressors and error (Chapter 4), all slope coefficients in that equation are consistently estimated. The slope on x_3 is $\beta_3\delta_3$ by construction. □

It is worth dwelling on why both assumptions are needed and what goes wrong if either fails. If condition (ii) is violated — say x_3 is correlated with the omitted part v_3 through x_1 — then x_1 is correlated with the composite error and $\widehat{\beta}_1$ is biased: the proxy is not good enough, and including it does not purge the omitted-variable bias. If condition (i) is violated — if the proxy belongs in the model in its own right, so u is correlated with x_3 — then the composite error is correlated with x_3 and the proxy itself is endogenous. The coefficient on x_3 is then a tangle of $\beta_3\delta_3$ and the spurious direct effect, and it is no longer a clean multiple of the unobservable's coefficient. Note throughout that a proxy is *not* the

same object as the variable it proxies; we never claim to have measured ability, only to have controlled for it well enough that the other slopes come out right.

8.2.2 A Lagged Dependent Variable as a Proxy

A particularly useful proxy for a stew of unmeasured factors is the *past value of the dependent variable itself*. Many outcomes carry inertia: a city with a high crime rate this year almost certainly had a high crime rate last year, and the gap between high-crime and low-crime cities reflects a long list of unobserved factors — enforcement intensity, social cohesion, economic structure — that persist over time. Conditioning on last year's outcome controls, at least partly, for all of those persistent unobservables at once. Consider modeling this year's crime rate as a function of current unemployment, with last year's crime rate included as a proxy:

$$crime = \beta_0 + \beta_1\, crime_{-1} + \beta_2\, unem + u.$$

The estimate $\widehat{\beta}_2$ now answers a sharper question. Rather than comparing cities at large — which differ in countless unobserved ways — it compares cities that *started from the same crime level last year*, and asks how their crime rates this year differ with unemployment. By holding fixed last year's outcome, we hold fixed the bulk of the persistent unobserved heterogeneity, and the unemployment coefficient is far more credible than in a cross-sectional regression without the lag.

8.3 The Random Coefficient Model

Every model so far has assumed that a single slope β_1 describes the entire population: the effect of x on y is the same for everyone. That is often too rigid. The return to a year of schooling may genuinely differ from person to person; the effect of a price cut may differ across stores. The *random coefficient model* (also called the random slope model) takes this seriously by letting each individual have her own intercept and slope.

For the simple regression case, write

$$y_i = a_i + b_i x_i,$$

where the intercept a_i and slope b_i vary across individuals i. With only one observation (y_i, x_i) per person we obviously cannot estimate a separate (a_i, b_i) for each one. What we *can* hope to estimate is the population *average* intercept and slope. Decompose each random coefficient into a fixed population mean plus a mean-zero individual deviation:

$$a_i = \alpha + c_i, \qquad b_i = \beta + d_i, \qquad \mathbb{E}(c_i) = \mathbb{E}(d_i) = 0,$$

where $\alpha = \mathbb{E}(a_i)$ is the average intercept and $\beta = \mathbb{E}(b_i)$ is the average slope — the parameter of central interest. Substituting,

$$y_i = (\alpha + c_i) + (\beta + d_i)x_i = \underbrace{(\alpha + \beta x_i)}_{\text{systematic part}} + \underbrace{(c_i + d_i x_i)}_{\text{composite error}}.$$

This is just a simple regression of y_i on x_i with a particular composite error $u_i := c_i + d_i x_i$.

Whether OLS recovers the average slope β depends, as always, on whether that error is mean-independent of x_i.

Assumption 8.6: Exogeneity of the Random Components

The individual deviations are mean-independent of the regressor:

$$\mathbb{E}(c_i \mid x_i) = \mathbb{E}(d_i \mid x_i) = 0,$$

equivalently $\mathbb{E}(a_i \mid x_i) = \alpha$ and $\mathbb{E}(b_i \mid x_i) = \beta$. The random parts of the intercept and slope are unrelated to the value of x.

Theorem 8.7: OLS Estimates the Average Slope

Under the exogeneity assumption, the composite error has zero conditional mean,

$$\mathbb{E}(c_i + d_i x_i \mid x_i) = \mathbb{E}(c_i \mid x_i) + x_i \, \mathbb{E}(d_i \mid x_i) = 0,$$

so OLS of y_i on x_i consistently and unbiasedly estimates the *average* intercept α and the *average* slope β.

So far this looks like ordinary OLS, and indeed the point estimates are fine. The catch lies in the variance. Even if the underlying components c_i and d_i are each homoskedastic, the composite error is *not*: its conditional variance depends on x_i.

Theorem 8.8: Induced Heteroskedasticity

Suppose $\operatorname{Var}(c_i \mid x_i) = \sigma_c^2$, $\operatorname{Var}(d_i \mid x_i) = \sigma_d^2$, and $\operatorname{Cov}(c_i, d_i) = 0$ given x_i. Then the composite error in the random coefficient model is heteroskedastic, with conditional variance

$$\operatorname{Var}(c_i + d_i x_i \mid x_i) = \sigma_c^2 + \sigma_d^2 \, x_i^2.$$

Proof. Conditional on x_i, x_i is a constant, so

$$\operatorname{Var}(c_i + d_i x_i \mid x_i) = \operatorname{Var}(c_i \mid x_i) + x_i^2 \operatorname{Var}(d_i \mid x_i) + 2 x_i \operatorname{Cov}(c_i, d_i) = \sigma_c^2 + \sigma_d^2 \, x_i^2,$$

using $\operatorname{Cov}(c_i, d_i) = 0$. The variance rises with x_i^2 because the slope deviation d_i is scaled by x_i. □

This is heteroskedasticity, but of an unusually friendly kind: its *functional form is known*, $h(x_i) = \sigma_c^2 + \sigma_d^2 x_i^2$. We may therefore proceed in either of the two ways developed in Chapter 7. We can run plain OLS and report heteroskedasticity-robust standard errors — valid because OLS still consistently estimates α and β. Or, for efficiency, we can use weighted least squares with weights based on the known variance form. Either route consistently estimates the population-average intercept and slope.

What the random coefficient model cannot do

The random coefficient model recovers *population averages*, α and β — and nothing more. Because each individual carries her own unobserved deviations c_i and d_i, the model cannot predict the outcome for any particular individual, nor identify any individual's personal slope. It tells us the average effect in the population, while remaining silent about who is above or below that average.

8.4 Measurement Error

Often a variable in our model is recorded with error: survey respondents misreport income, accounting figures are revised, instruments are imprecise. The consequences of measurement error depend sharply on *which* variable is mismeasured — the dependent variable or an explanatory variable — and the contrast between the two cases is one of the most important lessons in this chapter. The method, as always, is to write the observed-data regression explicitly and inspect the correlation between its error and its regressors.

8.4.1 Measurement Error in the Dependent Variable

Let the true model be

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

where y^* is the correctly measured outcome, which we cannot observe. Instead we observe y, which differs from y^* by a measurement error e_0:

$$e_0 = y - y^*, \qquad \text{so} \qquad y = y^* + e_0.$$

Substituting $y^* = y - e_0$ into the true model and rearranging,

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + (u + e_0).$$

We can only regress the observed y on the regressors, which means living with the composite error $u + e_0$. The decisive question is whether e_0 is correlated with the regressors. If the measurement error is mean-independent of the regressors — $\mathbb{E}(u + e_0 \,|\, x_1, \ldots, x_k) = 0$, which holds when e_0 has zero mean and is uncorrelated with the x_j — then OLS of y on $x_1, \ldots, x_k$ remains unbiased and consistent for every slope.

Theorem 8.9: Measurement Error in y Is Harmless (in Expectation)

If the measurement error e_0 in the dependent variable is uncorrelated with the explanatory variables and with u, then OLS of the observed y on $x_1, \dots, x_k$ is unbiased and consistent for $\beta_0, \beta_1, \dots, \beta_k$. The only cost is a loss of precision: the error variance is inflated from σ_u^2 to

$$\operatorname{Var}(u + e_0) = \sigma_u^2 + \sigma_{e_0}^2 \;>\; \sigma_u^2,$$

so the sampling variances of the OLS estimators are larger than they would be with the true y^*.

The intuition is reassuring: noise in the outcome simply pools with the existing disturbance. As long as that noise does not systematically move with the regressors, it cannot bias the slopes; it only adds to the unexplained variation, widening standard errors. (If e_0 *were* correlated with some x_j — say people with more education systematically over-report wages — this benign verdict would fail, and the affected slopes would be biased. The classical case assumes this away.)

8.4.2 Measurement Error in an Explanatory Variable

Mismeasuring a regressor is a different and more serious matter. Take the simple regression with a single mismeasured regressor,

$$y = \beta_0 + \beta_1 x^* + u,$$

where the true x^* is unobserved and we observe instead

$$x = x^* + e,$$

with e the measurement error. Substituting $x^* = x - e$,

$$y = \beta_0 + \beta_1 x + (u - \beta_1 e).$$

We regress y on the observed x, carrying the composite error $u - \beta_1 e$. Whether $\widehat{\beta}_1$ is consistent turns entirely on whether x is correlated with this composite error — and the answer depends on which of two assumptions we make about e.

Case 1: error uncorrelated with the observed regressor. If we assume $\operatorname{Cov}(x, e) = 0$ — the measurement error is uncorrelated with the *observed* value — then $\mathbb{E}(x^* \mid x) = x$, the observed x is the best predictor of the truth, and x plays exactly the role of a good proxy for x^*. In this case the composite error is uncorrelated with x, and $\widehat{\beta}_1$ is unbiased and consistent. This assumption is convenient but often implausible: it requires the noise to be larger precisely when the recorded value is larger.

Case 2: the classical errors-in-variables assumption. The more standard and realistic assumption is that the measurement error is uncorrelated with the *true, unobserved* value:

$$\boxed{\text{Cov}(x^*, e) = 0} \qquad \text{(classical errors-in-variables, CEV).}$$

Definition 8.10: Classical Errors-in-Variables (CEV)

The measurement error $e = x - x^*$ in an explanatory variable satisfies the *classical errors-in-variables assumption* if it is uncorrelated with the true value of the variable,

$$\text{Cov}(x^*, e) = 0,$$

and (as throughout) uncorrelated with the structural error u.

Under CEV, the error contaminates the observed regressor in a way that cannot be wished away. Compute the covariance between the observed regressor x and the composite error $u - \beta_1 e$:

$$\begin{aligned}\text{Cov}(x, u - \beta_1 e) &= \text{Cov}(x^* + e, u - \beta_1 e) \\ &= \underbrace{\text{Cov}(x^*, u)}_{=0} - \beta_1 \underbrace{\text{Cov}(x^*, e)}_{=0 \text{ (CEV)}} + \underbrace{\text{Cov}(e, u)}_{=0} - \beta_1 \text{ Cov}(e, e) \\ &= -\beta_1 \sigma_e^2.\end{aligned}$$

The observed regressor is correlated with the error whenever $\beta_1 \neq 0$ and there is any measurement error at all ($\sigma_e^2 > 0$). Endogeneity is unavoidable, and OLS is inconsistent. We can pin down the inconsistency exactly. Writing the estimated equation as $y = \beta_0 + \beta_1 x + v$ with $v = u - \beta_1 e$,

$$\text{plim}\, \widehat{\beta}_1 = \beta_1 + \frac{\text{Cov}(x, v)}{\text{Var}(x)} = \beta_1 + \frac{-\beta_1 \sigma_e^2}{\sigma_{x^*}^2 + \sigma_e^2}.$$

The last step uses $\text{Var}(x) = \text{Var}(x^* + e) = \sigma_{x^*}^2 + \sigma_e^2$ under CEV (because x^* and e are uncorrelated). Simplifying,

Theorem 8.11: Attenuation Bias Under CEV

Under the classical errors-in-variables assumption, OLS of y on the mismeasured regressor x is inconsistent, with probability limit

$$\text{plim}\, \widehat{\beta}_1 = \beta_1 \left(\frac{\sigma_{x^*}^2}{\sigma_{x^*}^2 + \sigma_e^2} \right) = \beta_1 \cdot \frac{\text{Var}(x^*)}{\text{Var}(x)}.$$

The multiplier

$$\frac{\text{Var}(x^*)}{\text{Var}(x)} = \frac{\sigma_{x^*}^2}{\sigma_{x^*}^2 + \sigma_e^2} \in (0, 1)$$

is always strictly between zero and one. Hence $\text{plim}\, \widehat{\beta}_1$ has the same sign as β_1 but is smaller in magnitude: the estimated effect is biased *toward zero*. This is called *attenuation bias*.

Proof. From the covariance computation, $\operatorname{plim}\widehat{\beta}_1 = \beta_1 - \beta_1\sigma_e^2/(\sigma_{x^*}^2 + \sigma_e^2)$. Factoring out β_1,

$$\operatorname{plim}\widehat{\beta}_1 = \beta_1\left(1 - \frac{\sigma_e^2}{\sigma_{x^*}^2 + \sigma_e^2}\right) = \beta_1 \cdot \frac{\sigma_{x^*}^2}{\sigma_{x^*}^2 + \sigma_e^2}.$$

Since $\sigma_{x^*}^2 > 0$ and $\sigma_e^2 > 0$, the ratio lies strictly in $(0,1)$, so $\left|\operatorname{plim}\widehat{\beta}_1\right| < |\beta_1|$ while $\operatorname{sgn}(\operatorname{plim}\widehat{\beta}_1) = \operatorname{sgn}(\beta_1)$. □

The two measurement-error verdicts

Error in y: harmless to the slopes if uncorrelated with the regressors — it only inflates the error variance and widens standard errors.

Error in a regressor (CEV): not harmless — it makes the mismeasured regressor endogenous and biases its slope *toward zero* by the factor $\operatorname{Var}(x^*)\,/\,\operatorname{Var}(x) \in (0,1)$. The noisier the measurement (larger σ_e^2), the worse the attenuation.

Example (Attenuation in a wage regression).

Suppose the true return to actual labor-market experience is $\beta_1 = 0.06$ (six percent per year), but survey-reported experience x equals true experience x^* plus a reporting error e satisfying CEV. If the variance of true experience is $\sigma_{x^*}^2 = 64$ and the measurement-error variance is $\sigma_e^2 = 16$, what does OLS estimate in a large sample?

Solution.

The attenuation factor is

$$\frac{\sigma_{x^*}^2}{\sigma_{x^*}^2 + \sigma_e^2} = \frac{64}{64 + 16} = \frac{64}{80} = 0.8.$$

Hence $\operatorname{plim}\widehat{\beta}_1 = 0.06 \times 0.8 = 0.048$. OLS will, on average in large samples, report a return of about 4.8% rather than the true 6% — the effect is pulled 20% of the way toward zero, the proportion of the observed variance that is pure noise.

8.4.3 Measurement Error in the Multiple Regression Case

When the mismeasured regressor sits inside a multiple regression, the situation is messier. Let

$$y = \beta_0 + \beta_1 x_1^* + \beta_2 x_2 + \cdots + \beta_k x_k + u, \qquad x_1 = x_1^* + e,$$

with $x_2, \ldots, x_k$ measured without error. Substituting gives

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + (u - \beta_1 e).$$

As before, x_1 is correlated with the composite error $u - \beta_1 e$ under CEV. But now the contamination spreads: even though $u - \beta_1 e$ is uncorrelated with the correctly measured regressors $x_2, \ldots, x_k$, those regressors are in general correlated with x_1, and through that channel the bias in $\widehat{\beta}_1$ leaks into the other coefficients. The upshot is that *all* of the slopes can be biased and inconsistent, not just the one on the mismeasured variable. The

exact expressions for the various plim $\widehat{\beta}_j$ are algebraically involved and depend on the full correlation structure among the regressors; we do not reproduce them here. The qualitative lesson is what matters: mismeasuring even a single regressor can corrupt every coefficient in the model, and the only fully general repair is instrumental variables (Chapter 10).

8.5 Missing Data and Nonrandom Samples

Real datasets have holes. A respondent skips the income question; a firm does not report R&D spending; a record is simply lost. Missing data is best understood as a special case of *nonrandom sampling* (sample selection): the observations we actually use are not a random draw from the population, because the ones with missing values are dropped. Whether this matters follows the same principle that has organized the whole chapter.

The selection principle

Dropping observations is harmless to OLS *if the rule that decides which observations are kept is uncorrelated with the error term.* Selection on the explanatory variables (*exogenous* sample selection) is fine, because the regression already conditions on those variables. Selection on the dependent variable or on the error itself (*endogenous* sample selection) is a genuine problem and biases the estimates.

To make the missingness mechanism precise, let $m_{ik} = 1$ if observation i has a valid value of variable x_k and $m_{ik} = 0$ if it is missing. Two benchmark assumptions are standard.

Definition 8.12: MCAR and MAR

The data are *Missing Completely at Random* (MCAR) if missingness is unrelated to everything — to the error u and to all regressors:

$$\mathrm{P}(m_k = 1 \,|\, u, x_1, \ldots, x_k) = \mathrm{P}(m_k = 1)\,.$$

The data are *Missing at Random* (MAR) if, once we condition on the *observed* regressors, missingness is unrelated to the error:

$$\mathrm{P}(m_k = 1 \,|\, u, x_1, \ldots, x_k) = \mathrm{P}(m_k = 1 \,|\, x_1, \ldots, x_k)\,.$$

Under MAR the missingness mechanism is allowed to depend on $x_1, \ldots, x_k$ but must be conditionally independent of u given the regressors.

MCAR is the strongest and cleanest assumption: missingness is pure coin-flipping, so the complete cases are themselves a random sample and ordinary OLS on them is unbiased — just based on fewer observations. MAR is weaker and more realistic, allowing, for example, that high earners are more likely to skip the income question, as long as nothing in the unexplained part of y drives the missingness.

8.5.1 The Missing-Indicator Method

Suppose we are missing values of a single regressor x_k, while y and the other regressors $x_1, \dots, x_{k-1}$ are fully observed. The simplest response is to use only the *complete cases* — discard any observation with a missing x_k — yielding the complete-case estimator. This is robust but wasteful: it throws away the fully observed information on y and $x_1, \dots, x_{k-1}$ for every dropped row. The *missing-indicator method* (MIM) tries to keep all the rows. It builds two new variables:

$$z_{ik} = \begin{cases} x_{ik}, & \text{if } x_{ik} \text{ is observed,} \\ 0, & \text{if } x_{ik} \text{ is missing,} \end{cases} \qquad m_{ik} = \begin{cases} 1, & \text{if } x_{ik} \text{ is observed,} \\ 0, & \text{if } x_{ik} \text{ is missing,} \end{cases}$$

where z_{ik} is the regressor zeroed-out where missing and m_{ik} is the missing-data indicator. One then regresses y_i on $x_{i1}, \dots, x_{i,k-1}, z_{ik}, m_{ik}$ using *all* observations. Including the indicator m_{ik} is essential: it allows the regression to assign a separate intercept to the observations whose x_k was set to zero, so that they are not forced to behave as though x_k truly equalled zero.

Remark (The strong assumption behind MIM).

> The missing-indicator method buys back the discarded rows only at the price of a stringent assumption. It delivers consistent estimates of the other coefficients essentially only when the partially-missing regressor x_k is *uncorrelated with the remaining regressors*,
>
> $$\operatorname{Cov}(x_k, x_j) = 0 \quad \text{for all } j \neq k.$$
>
> This is rarely true in practice — regressors in economic models are routinely correlated — so MIM, despite its convenience, can introduce its own bias. Note also that simply *omitting* the indicator m_{ik} and using z_{ik} alone is equivalent to pretending $x_{ik} = 0$ wherever it is missing, which is clearly worse.

8.6 Outliers and Influential Observations

A handful of extreme observations can exert outsized leverage on an OLS fit, because least squares minimizes the *sum of squared* residuals: a single point far from the cloud contributes its squared deviation, and the line bends to accommodate it. Such an *outlier* is *influential* when its presence or absence noticeably changes the estimated coefficients. The first question is always the source. If an outlier is a data-entry mistake — a misplaced decimal, a coding error — the right action is simply to correct or discard it. But if the outlier is a genuine product of the data-generating process — a real, if unusual, observation — the decision is delicate: discarding it discards real information, while keeping it lets one point dominate the analysis.

8.6.1 Least Absolute Deviations

When outliers are real and we do not wish to drop them, we can change the *estimator* to one that is less sensitive to extreme observations. *Least absolute deviations* (LAD) replaces

the squared residuals of OLS with absolute residuals:

$$\min_{b_0,b_1,\ldots,b_k} \sum_{i=1}^{n} | \, y_i - b_0 - b_1 x_{i1} - \cdots - b_k x_{ik} \, | \, .$$

Because deviations are not squared, a far-flung point contributes its distance only linearly rather than quadratically, so LAD pulls toward the bulk of the data far less than OLS does. This robustness has a precise statistical counterpart.

Theorem 8.13: LAD Estimates the Conditional Median

The least absolute deviations estimator estimates the parameters of the *conditional median* of y given $\vec{x}$, whereas OLS estimates the parameters of the *conditional mean*. The two coincide when the error distribution is *symmetric* about zero — then the conditional mean and conditional median are equal — but they differ when the error distribution is skewed.

The median is the natural target for a robust procedure: it is the value that minimizes expected absolute deviation, just as the mean minimizes expected squared deviation, and the median is famously insensitive to extreme observations. Finally, LAD is not an isolated trick but the simplest member of a broad family: it is the special case of *quantile regression* (Chapter-level developments aside) corresponding to the 0.5 quantile. Quantile regression generalizes the idea by minimizing an asymmetrically weighted sum of absolute residuals to estimate any conditional quantile — the conditional first quartile, the ninth decile, and so on — giving a far richer picture of how $\vec{x}$ shifts the entire conditional distribution of y, not merely its center.

Remark (Where this leaves us).

Every problem in this chapter was diagnosed by the same maneuver: write down the regression we can actually run, identify its error term, and ask whether that error is correlated with the regressors. Functional-form errors, omitted unobservables, and mismeasured regressors all break that correlation condition and bias OLS; measurement error in y, exogenous selection, and symmetric outliers leave the slopes intact (costing only precision or efficiency). When a regressor is genuinely endogenous — as under classical errors-in-variables — no reweighting or proxy fully repairs the damage, and we need a new tool that manufactures exogenous variation from outside the model. That tool is instrumental variables, the subject of Chapter 10.

Chapter 9

Panel Data Methods

So far every dataset has been a single cross section: one observation per unit, taken at one moment. The recurring danger has been omitted variables — some unobserved factor, correlated with our regressor, that biases the slope (Chapter 2). If that factor were observed we would simply control for it; the trouble is precisely that it is not. *Panel data* offer a way out that needs no proxy and no instrument. A panel follows the *same* units — people, firms, cities, countries — over two or more time periods. The extra dimension lets each unit serve as its own control: by comparing a unit to *itself* across time, any characteristic of the unit that does not change over the window of observation cancels out, whether we ever observe it or not.

This is a powerful idea, and most of this chapter is one idea worn four ways. We split the error into a time-constant unobserved *effect* a_i (ability, location, management culture, soil quality) and a time-varying *idiosyncratic* error u_{it}, and then we eliminate a_i by differencing or demeaning. Difference-in-differences compares the before/after change in a treated group to the before/after change in a control group. First differencing subtracts adjacent periods. The fixed-effects (within) estimator subtracts each unit's time average. Random effects subtracts only a *fraction* of that average, and correlated random effects ties the whole family together and hands us a clean test for which method to use. Throughout, the prize is a causal slope in the presence of unobserved, time-constant endogeneity — exactly the situation a single cross section cannot handle.

We index a unit by $i = 1, \ldots, N$ and a time period by $t = 1, \ldots, T$. A central modeling premise is that units are independent draws — each i is i.i.d. — while observations *within* a unit are allowed to be correlated across time. Cross-unit correlation is ruled out; within-unit serial correlation is not, and managing it will be a recurring theme.

9.1 Difference-in-Differences

Most policy questions concern an event located in time — a law passed, a plant opened, a tax changed — and ideally one we can treat as exogenous to the units it affects. We sort units into a *treatment* group (those exposed to the event) and a *control* group (those not), and we ask how the event moved the outcome. The obstacle is that the two groups usually differ for reasons unrelated to the event: treated units may have been on a different level all along. Difference-in-differences (DiD) handles this by focusing not on the levels, nor even

on the trends, but on the *difference between the trends*.

9.1.1 The parallel-trends idea

Consider the effect of building a garbage incinerator on nearby house prices. Let *rprice* be the real house price and *nearinc* a dummy equal to one for houses near the incinerator site. A naive analyst takes data from the year construction began and runs

$$rprice = \gamma_0 + \gamma_1\, nearinc + u,$$

reading γ_1 as "the effect of the incinerator." But houses near the future site were farther from the city center and cheaper *regardless* of the incinerator. The coefficient γ_1 confounds the incinerator's effect with a pre-existing level difference between the two neighborhoods.

The fix is to look at how the gap between the groups *changed*. We need data from at least two years, one before the event and one after, and we maintain the *parallel-trends assumption*: in the absence of the event, the average outcome in the two groups would have moved by the same amount in the same direction. Under parallel trends, any divergence in the gap between groups, before versus after, must be attributable to the event.

Assumption 9.1: Parallel Trends

Let y^0_{it} denote the outcome unit i would have at time t if untreated (its potential outcome under no treatment). The *parallel-trends* assumption states that, in the absence of treatment, the average change in y^0 is the same for the treatment (T) and control (C) groups:

$$\mathbb{E}\left(y^0_{i,\text{after}} - y^0_{i,\text{before}} \,|\, T\right) = \mathbb{E}\left(y^0_{i,\text{after}} - y^0_{i,\text{before}} \,|\, C\right).$$

Equivalently, the two groups share a common time trend; they may sit at different *levels* but they drift in parallel. This is what makes the untreated control group a valid stand-in for what would have happened to the treated group.

9.1.2 The two-period regression

The cleanest way to compute the DiD estimator is by a single regression with an interaction term. Define a time dummy $y81 = 1$ for the post-event year (1981, when incinerator construction started) and 0 otherwise. Estimate

$$rprice = \beta_0 + \delta_0\, y81 + \beta_1\, nearinc + \delta_1\,(y81 \cdot nearinc) + u.$$

The coefficient δ_1 on the interaction is the DiD estimator. In the general notation, with $dT = 1$ for the treatment group, $d2 = 1$ for the second (post) period, and possibly other controls,

$$y = \beta_0 + \delta_0\, d2 + \beta_1\, dT + \delta_1\,(d2 \cdot dT) + (\text{other factors}) + u.$$

Each coefficient has a job, and reading them off is the key to understanding the method:

- $\beta_1\, dT$ controls for the *fixed level difference* between the treatment and control groups (the pre-existing gap).
- $\delta_0\, d2$ controls for the *common time trend* affecting both groups.
- $\delta_1\, (d2 \cdot dT)$ captures the *extra* change experienced only by the treated group after the event — the difference in trends, which is the treatment effect.

Note that to detect a slope difference at all, one must first control for the parallel trend; the group dummy dT and the time dummy $d2$ are therefore both indispensable, never to be dropped.

9.1.3 The 2×2 table

The four conditional means produced by the model lay out as a small table, and reading the margins two different ways reveals the two equivalent meanings of δ_1.

	Before	After	After − Before
Control	β_0	$\beta_0 + \delta_0$	δ_0
Treatment	$\beta_0 + \beta_1$	$\beta_0 + \beta_1 + \delta_0 + \delta_1$	$\delta_0 + \delta_1$
Treatment − Control	β_1	$\beta_1 + \delta_1$	δ_1

The lower-right cell is δ_1 no matter which way we reach it, and that is the whole point.

Two readings of the same number

Let $\bar{y}_{p,g}$ be the sample mean for period $p \in \{0,1\}$ (before/after) and group $g \in \{C,T\}$. The DiD estimator $\hat{\delta}_1$ admits two algebraically identical expressions:

$$\hat{\delta}_1 = \underbrace{(\bar{y}_{1,T} - \bar{y}_{1,C})}_{\text{after-gap}} - \underbrace{(\bar{y}_{0,T} - \bar{y}_{0,C})}_{\text{before-gap}} \quad \text{(difference of cross-group differences)},$$

$$\hat{\delta}_1 = \underbrace{(\bar{y}_{1,T} - \bar{y}_{0,T})}_{\text{treated change}} - \underbrace{(\bar{y}_{1,C} - \bar{y}_{0,C})}_{\text{control change}} \quad \text{(difference of over-time differences)}.$$

First reading: how much did the treated-minus-control gap widen from before to after. Second reading: how much more did the treated group change over time than the control group did. They are equal because both compute the same corner of the 2×2 table.

9.1.4 Why δ_1 is the ATE

The subtraction is what does the work. Comparing the change in the treated group to the change in the control group nets out anything common to both groups over the period — macroeconomic shocks, seasonality, secular drift — because those common factors enter the control change too and cancel. What remains is the part of the treated group's movement

that the control group did *not* share, which, under parallel trends, is the causal effect of the event. When assignment to treatment is effectively random (a natural or quasi-natural experiment), δ_1 identifies the *average treatment effect* (ATE).

Remark (Natural experiments).

> DiD is the workhorse for (quasi-)natural experiments. In a *true* experiment, subjects are randomly assigned to treatment or control, so the groups are comparable by construction. In a *natural* experiment, the two groups arise from some change in policy or circumstance and may differ systematically; the before/after differencing is precisely how we control for that systematic difference. DiD therefore evaluates policy changes and other exogenous events when randomization was not under our control.

9.1.5 When parallel trends fails: triple differences

DiD is only as good as its identifying assumption. If the two groups would have followed *different* trends even without the event — a differential trend — then δ_1 mixes the treatment effect with that pre-existing divergence, and we can no longer say it isolates the policy. We cannot be sure whether δ_1 reflects the policy or merely some unaccounted factor pushing the groups apart.

One remedy is to add flexibility through a second control dimension. Suppose we worry that the treatment and control neighborhoods are trending differently because of, say, their differing age composition. We can find a second pair of groups — treated and untreated — that shares the suspect trend but is unaffected by the policy, and difference *again*. The result is the *difference-in-difference-in-differences* (DDD, or triple-differences) estimator: it adds further interaction terms so that any common differential trend across the extra dimension is itself differenced away, leaving a cleaner estimate of the policy effect.

9.2 First Differencing

We now move from the two-group, before/after setup to a panel of many units each observed over time, and to a continuous (rather than binary) regressor. The mechanics differ but the spirit is identical: subtract to eliminate the unobserved, time-constant nuisance.

9.2.1 The unobserved-effects model

Suppose we want the causal effect of x_{it} on y_{it} and have no other regressors at hand. We can still hope to recover it if each unit is observed for at least two periods and the *other* factors affecting y_{it} stay roughly constant over the window. Write the model as

$$y_{it} = \beta_0 + \delta_0\, d_t + \beta_1\, x_{it} + v_{it}, \qquad v_{it} = a_i + u_{it},$$

where:

- d_t is a time dummy for the second period (allowing a different intercept across periods);
- a_i is the *unobserved effect* (also called the fixed effect, individual heterogeneity, or unobserved heterogeneity): time-constant factors specific to unit i;

- u_{it} is the *idiosyncratic error*: time-varying unobserved factors.

The composite error $v_{it} = a_i + u_{it}$ packages the two together.

9.2.2 The heterogeneity bias of pooled OLS

The crudest approach ignores the panel structure: stack all NT observations and run OLS on the original equation. This *pooled OLS* estimator is consistent only if the entire composite error v_{it} is uncorrelated with x_{it}. Even granting that the idiosyncratic part is clean, $\text{Cov}(x_{it}, u_{it}) = 0$, consistency still requires $\text{Cov}(x_{it}, a_i) = 0$. But the unobserved effect is exactly the sort of thing — ability, location quality, firm culture — that we expect to be correlated with the regressor.

Heterogeneity bias

If the time-constant unobserved effect a_i is correlated with x_{it}, then pooled OLS on $y_{it} = \beta_0 + \delta_0 d_t + \beta_1 x_{it} + v_{it}$ is biased and inconsistent for β_1. This is called *heterogeneity bias* — it is simply omitted-variable bias (Chapter 2) with the omitted variable being a_i.

This is the whole reason we collected a panel: to allow a_i to be *arbitrarily* correlated with the regressors and still estimate β_1 consistently.

9.2.3 Differencing out the effect

Because a_i does not change over time, it disappears when we subtract one period from another. Take the model in two periods t with $d_t = 0$ (period 1) and $d_t = 1$ (period 2) and difference:

$$\Delta y_i = \delta_0 + \beta_1 \, \Delta x_i + \Delta u_i,$$

where $\Delta y_i = y_{i2} - y_{i1}$, and likewise for Δx_i and Δu_i. The term a_i is gone: $a_i - a_i = 0$. This is the *first-differenced (FD) equation.* The time dummy's coefficient δ_0 survives as the new intercept (since $\Delta d_t = 1 - 0 = 1$).

The FD equation is just a cross-sectional regression of Δy_i on Δx_i, so all the OLS machinery returns, provided one new condition holds.

Assumption 9.2: Strict Exogeneity

The idiosyncratic error is *strictly exogenous* if, in every period, it is uncorrelated with the explanatory variables in *all* periods:

$$\mathbb{E}(u_{it} \,|\, x_{i1}, x_{i2}, \ldots, x_{iT}, \, a_i) = 0 \qquad \text{for all } t.$$

This is stronger than contemporaneous exogeneity $\mathbb{E}(u_{it} \,|\, x_{it}) = 0$: it forbids feedback from past shocks to future regressors and from future shocks to current regressors. Strict exogeneity is exactly what guarantees $\text{Cov}(\Delta x_i, \Delta u_i) = 0$, so that FD-OLS is unbiased and consistent for β_1.

Theorem 9.3: Consistency of First Differencing

In the unobserved-effects model $y_{it} = \beta_0 + \delta_0 d_t + \beta_1 x_{it} + a_i + u_{it}$ with $T = 2$, if u_{it} is strictly exogenous and Δx_i has variation across i, then OLS on the first-differenced equation is consistent for β_1 regardless of how a_i correlates with the regressors.

The crucial freedom is that we placed *no* restriction on $\text{Cov}(x_{it}, a_i)$: the differencing annihilates a_i, so any form of correlation between a_i and the regressors is permitted. The effect a_i is left completely general. "General," though, is not the same as "random": here a_i is treated as a fixed, unit-specific quantity, not as a draw from a distribution we model.

Remark (FD equals DiD).

> When x_{it} is itself a treatment dummy, first differencing reproduces DiD exactly. In DiD, the group dummy dT absorbs the fixed level difference, the time dummy $d2$ absorbs the common trend, and the interaction $\delta_1(d2 \cdot dT)$ is the effect. In FD, differencing wipes out the fixed effect at the outset, the common trend lands in the intercept δ_0, and the term $\beta_1 \Delta x_i$ *is* the interaction by construction. Same estimate, two routes.

9.2.4 Costs of differencing

Eliminating a_i is not free. Two limitations recur for every differencing/demeaning method in this chapter.

Lost variation, larger standard errors. Differencing can sharply reduce the variation in the regressor. Even if x_{it} varies a lot across units in levels, the within-unit change Δx_i may move very little. Recall from Chapter 1 that $\text{Var}\left(\widehat{\beta}_1\right) = \sigma^2/\text{SST}_x$: little variation in Δx_i inflates the standard error and costs precision.

Time-invariant regressors vanish. Any regressor that is constant over time — gender, race, a person's education if fixed over the window — has $\Delta x_i = 0$ and is differenced away. Its effect simply cannot be estimated by FD (or, as we will see, by fixed effects). This is the price of letting a_i be arbitrary: the method cannot distinguish a time-constant regressor from the time-constant effect it eliminates.

Strict exogeneity is essential. If strict exogeneity fails — for instance, if a shock u_{it} today feeds into next period's $x_{i,t+1}$ — then Δx_i is correlated with Δu_i and the FD estimator loses consistency.

9.2.5 More than two periods, and the serial-correlation check

First differencing extends to $T > 2$. With three periods, allow a separate intercept in each:

$$y_{it} = \delta_1 + \delta_2\, d2_t + \delta_3\, d3_t + \beta_1\, x_{it} + u_{it},$$

where $d2_t, d3_t$ are dummies for periods 2 and 3 (so a_i is folded into δ_1 for exposition, or carried along and differenced out). Differencing adjacent periods gives

$$\Delta y_{it} = \delta_2\, \Delta d2_t + \delta_3\, \Delta d3_t + \beta_1\, \Delta x_{it} + \Delta u_{it}.$$

Here $\Delta d2_t = 1,\ \Delta d3_t = 0$ at $t = 2$, while $\Delta d2_t = -1,\ \Delta d3_t = 1$ at $t = 3$. This differenced equation has no intercept, which is inconvenient for computing R^2. It is cleaner to estimate the algebraically equivalent form with an intercept,

$$\Delta y_{it} = \alpha_0 + \alpha_3\, d3_t + \beta_1\, \Delta x_{it} + \Delta u_{it},$$

which reports the same $\widehat{\beta}_1$ but a usable R^2.

A standing requirement of FD is that the *differenced* errors Δu_{it} be serially uncorrelated, and this is easy to test. Let $r_{it} = \Delta u_{it}$ (estimated by the FD residuals) and posit an AR(1) structure $r_{it} = \rho\, r_{i,t-1} + e_{it}$. No serial correlation means $\rho = 0$. Regress the residuals on their own lag, and test $H_0 : \rho = 0$ with the usual t statistic. Rejection signals serial correlation in Δu_{it}, calling for robust standard errors or for the fixed-effects estimator instead.

9.3 Fixed Effects Estimation

First differencing subtracts *adjacent* periods. Fixed effects subtracts each unit's *time average*. For $T = 2$ the two coincide, but for $T > 2$ they are genuinely different estimators with different efficiency properties. We work with the model

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

where a_i is the fixed effect for unit i and u_{it} the idiosyncratic error. (An overall intercept is omitted because it is perfectly absorbed into a_i.)

9.3.1 The within (time-demeaning) estimator

The idea is to exploit the variation of each unit *around its own mean over time*. Average the model over the T periods for a fixed unit i:

$$\bar{y}_i = \beta_1\, \bar{x}_i + a_i + \bar{u}_i, \qquad \bar{y}_i := \frac{1}{T}\sum_{t=1}^{T} y_{it},$$

and similarly $\bar{x}_i := \frac{1}{T}\sum_{t=1}^{T} x_{it}$, $\bar{u}_i := \frac{1}{T}\sum_{t=1}^{T} u_{it}$. Note that a_i is constant in t, so its average over time is just a_i itself. Subtract this time-average equation from the original:

$$y_{it} - \bar{y}_i = \beta_1\,(x_{it} - \bar{x}_i) + (u_{it} - \bar{u}_i).$$

Writing $\ddot{y}_{it} := y_{it} - \bar{y}_i$ (and likewise $\ddot{x}_{it}, \ddot{u}_{it}$) gives the *within* or *time-demeaned* equation

$$\ddot{y}_{it} = \beta_1\, \ddot{x}_{it} + \ddot{u}_{it}.$$

Definition 9.4: Fixed-Effects (Within) Estimator

The *fixed-effects estimator* $\widehat{\beta}_{\text{FE}}$ is the pooled OLS estimator of the time-demeaned equation $\ddot{y}_{it} = \beta_1 \ddot{x}_{it} + \ddot{u}_{it}$. The transformation $x_{it} \mapsto \ddot{x}_{it} = x_{it} - \bar{x}_i$ is the *within transformation*; it removes a_i because the effect is constant over time. The within equation has no intercept — the intercept is absorbed by a_i.

As with FD, strict exogeneity of u_{it} (conditional on all x_{it} and a_i) is what makes $\ddot{x}_{it}$ uncorrelated with $\ddot{u}_{it}$, so that $\widehat{\beta}_{\text{FE}}$ is consistent for any correlation pattern between a_i and the regressors.

Once $\widehat{\beta}_{\text{FE}}$ is in hand, the individual effects can be recovered. Since $\bar{u}_i \xrightarrow{p} 0$ as T grows, an estimate of the fixed effect is

$$\hat{a}_i = \bar{y}_i - \widehat{\beta}_{\text{FE}}\, \bar{x}_i.$$

The within transformation has the same Achilles' heel as FD: time-invariant regressors are demeaned to zero and their effects cannot be estimated. A constant-over-time variable $x_{it} = x_i$ has $\ddot{x}_{it} = 0$.

9.3.2 Degrees of freedom

A subtle but important accounting point: forming the within-transformed data uses up degrees of freedom, because we estimated N time-averages (one $\bar{y}_i$ per unit) in addition to the k slope parameters. With NT total observations, N averages, and k regressors, the residual degrees of freedom are

$$df = NT - N - k = N(T-1) - k.$$

Software that runs OLS on the demeaned data without this correction will report degrees of freedom that are too large and standard errors that are too small; the correction is essential.

9.3.3 The dummy-variable interpretation

There is an equivalent and illuminating way to see fixed effects: put a separate intercept dummy for each unit directly into the level equation,

$$y_{it} = a_1\, ind1_{it} + a_2\, ind2_{it} + \cdots + a_N\, indN_{it} + \beta_1\, x_{it} + u_{it},$$

where $indK_{it} = 1$ if the observation belongs to unit K and 0 otherwise. Running OLS on this is the *least-squares dummy-variable* (LSDV) regression, and it returns exactly $\widehat{\beta}_{\text{FE}}$ together with $\hat{a}_i$ as the dummy coefficients. Because the overall intercept is excluded, the full set of N dummies does not create a dummy-variable trap; including both the N dummies and a separate intercept *would* cause perfect multicollinearity. The LSDV view also makes the $NT - N - k$ degree-of-freedom count transparent: we literally estimate $N + k$ parameters. When N is large the LSDV regression becomes impractical (too many dummies), which is exactly why the algebraically identical within transformation is used in practice.

9.3.4 The between estimator

The within estimator throws away the cross-sectional (level) information and keeps only within-unit variation. The *between estimator* does the opposite: it discards time variation and uses only the cross-sectional means. Take the time-averaged equation

$$\bar{y}_i = \beta_1 \bar{x}_i + a_i + \bar{u}_i, \qquad i = 1, \ldots, N,$$

and run OLS of $\bar{y}_i$ on $\bar{x}_i$ across the N units. The resulting slope is the *between estimator*. Because a_i remains in this equation and is allowed to correlate with $\bar{x}_i$, the between estimator is generally biased and inconsistent for β_1; it ignores the within-unit information that fixed effects exploits. (If one is willing to assume a_i is uncorrelated with x_{it}, the random-effects estimator of the next section uses the data more efficiently.) Even for unbiasedness of the between estimator one needs strict exogeneity, u_{it} uncorrelated with x_{it} at every lead and lag.

9.3.5 Some practical points on within estimation

- The R^2 from the demeaned regression is misleading as a measure of overall fit, because it ignores the variation explained by the a_i.
- Effects of *time-invariant* regressors cannot be estimated. However, the effect of an *interaction* between a time-varying variable and a time-invariant one *can* be estimated, because the interaction itself varies over time.
- If a full set of time dummies is included, the effect of any variable whose change over time is constant across units (e.g. work experience measured in years) cannot be separately identified — it is perfectly collinear with the time dummies.
- Remember to adjust degrees of freedom to $NT - N - k$.

9.3.6 FD versus FE

Both estimators kill a_i and require strict exogeneity, and they suit slightly different circumstances.

When to use which

- $T = 2$**:** FD and FE are *numerically identical*. (For exact equivalence the FE model must include a dummy for the second period, since FD carries a time dummy automatically.)
- $T > 2$**, classical assumptions:** If the idiosyncratic errors u_{it} are serially uncorrelated and homoskedastic, FE is *more efficient* than FD.
- **Strong serial correlation:** FD can be better when the u_{it} are highly serially correlated. In the extreme where the u_{it} follow a random walk ($u_{it} = \rho u_{i,t-1} + e_{it}$ with $\rho = 1$), differencing yields $\Delta y_{it} = \beta_1 \Delta x_{it} + e_{it}$ with a well-behaved error e_{it}. When T is large relative to N the panel takes on a time-series character and serial correlation matters more.

- **Simplicity:** FD is more transparent, and heteroskedasticity-robust standard errors are easier to compute for it.
- **Unbalanced panels:** FE tolerates units with some missing periods (it just averages over the periods that are present), whereas FD needs both t and $t-1$ present to form Δ. FE generally preserves more data.
- **In practice:** compute both and check that the conclusions are robust.

Example (FD and FE coincide at $T = 2$).

With two periods, show that the FE within transformation and first differencing give the same slope.

Solution.

With $T = 2$ the time average for unit i is $\bar{y}_i = \frac{1}{2}(y_{i1} + y_{i2})$, so the demeaned values are

$$\ddot{y}_{i1} = y_{i1} - \tfrac{1}{2}(y_{i1} + y_{i2}) = -\tfrac{1}{2}\,\Delta y_i, \qquad \ddot{y}_{i2} = y_{i2} - \tfrac{1}{2}(y_{i1} + y_{i2}) = +\tfrac{1}{2}\,\Delta y_i,$$

and identically $\ddot{x}_{i1} = -\frac{1}{2}\Delta x_i$, $\ddot{x}_{i2} = +\frac{1}{2}\Delta x_i$. The within estimator is

$$\widehat{\beta}_{\text{FE}} = \frac{\sum_i \sum_{t=1}^{2} \ddot{x}_{it}\,\ddot{y}_{it}}{\sum_i \sum_{t=1}^{2} \ddot{x}_{it}^{\,2}} = \frac{\sum_i \left[\frac{1}{4}\Delta x_i \Delta y_i + \frac{1}{4}\Delta x_i \Delta y_i\right]}{\sum_i \left[\frac{1}{4}\Delta x_i^2 + \frac{1}{4}\Delta x_i^2\right]} = \frac{\sum_i \Delta x_i\,\Delta y_i}{\sum_i \Delta x_i^2},$$

which is exactly the OLS slope from regressing Δy_i on Δx_i — the FD estimator. The two are identical.

9.4 Random Effects Estimation

Fixed effects buys robustness — it allows a_i to correlate freely with the regressors — at the cost of efficiency and of any ability to estimate time-invariant effects. If we are willing to assume that a_i is *uncorrelated* with the regressors, we can do better on both counts. This is the random-effects (RE) model:

$$y_{it} = \beta_0 + \beta_1\, x_{it} + a_i + u_{it}, \qquad \operatorname{Cov}(x_{it}, a_i) = 0.$$

Now a_i is treated as a random draw, unrelated to the explanatory variables; the composite error $v_{it} = a_i + u_{it}$ is then uncorrelated with x_{it}, so pooled OLS is at least *consistent*. The problem is efficiency.

9.4.1 The serial correlation in the composite error

Even though a_i is unrelated to the regressors, it is shared by all of unit i's observations, which induces serial correlation in v_{it}. Assuming the idiosyncratic errors are serially uncorrelated

($\mathrm{Cov}(u_{it}, u_{is}) = 0$ for $t \neq s$) and writing $\sigma_a^2 = \mathrm{Var}(a_i)$, $\sigma_u^2 = \mathrm{Var}(u_{it})$,

$$\begin{aligned}\mathrm{Cov}(v_{it}, v_{is}) &= \mathrm{Cov}(a_i + u_{it}, a_i + u_{is}) \\ &= \mathrm{Var}(a_i) + \mathrm{Cov}(a_i, u_{is}) + \mathrm{Cov}(a_i, u_{it}) + \mathrm{Cov}(u_{it}, u_{is}) \\ &= \sigma_a^2 + 0 + 0 + 0 = \sigma_a^2 \neq 0, \qquad t \neq s.\end{aligned}$$

Since $\mathrm{Var}(v_{it}) = \sigma_a^2 + \sigma_u^2$, the within-unit (intraclass) correlation is

$$\mathrm{Corr}(v_{it}, v_{is}) = \frac{\mathrm{Cov}(v_{it}, v_{is})}{\sqrt{\mathrm{Var}(v_{it})\,\mathrm{Var}(v_{is})}} = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_u^2} > 0, \qquad t \neq s.$$

This positive serial correlation means pooled OLS, while consistent, is *inefficient*, and its usual standard errors are wrong unless adjusted for the within-unit correlation. We would like a transformation that restores the Gauss–Markov structure (a homoskedastic, serially uncorrelated error). Full time-demeaning, as in FE, over-corrects: $v_{it} - \bar{v}_i$ is still serially correlated.

9.4.2 Quasi-demeaning

The right amount of demeaning is partial. Subtract only a *fraction* λ of the time average:

$$y_{it} - \lambda\,\bar{y}_i = \beta_0(1 - \lambda) + \beta_1\,(x_{it} - \lambda\,\bar{x}_i) + (v_{it} - \lambda\,\bar{v}_i).$$

This is *quasi-demeaning*, and λ is the quasi-demeaning parameter. The hope is to choose λ so that the transformed error $e_{it} = v_{it} - \lambda\bar{v}_i$ is serially uncorrelated, $\mathrm{Cov}(e_{it}, e_{is}) = 0$ for all $t \neq s$. The value that achieves this is

$$\lambda = 1 - \sqrt{\frac{\sigma_u^2}{\sigma_u^2 + T\,\sigma_a^2}}, \qquad \lambda \in [0, 1].$$

The parameters σ_u^2 and σ_a^2 are unknown but can be estimated from residuals of a preliminary regression (e.g. pooled OLS or FE), making the procedure a *feasible generalized least squares* (FGLS) estimator. With $\hat{\lambda}$ in hand, run OLS on the quasi-demeaned data; that is the random-effects estimator.

Theorem 9.5: Random Effects as FGLS Between Pooled OLS and FE

The RE estimator is the FGLS estimator of the unobserved-effects model under $\mathrm{Cov}(x_{it}, a_i) = 0$. The quasi-demeaning weight λ interpolates between two familiar special cases:

$$\lambda = 0 \;\Rightarrow\; \text{pooled OLS}, \qquad \lambda = 1 \;\Rightarrow\; \text{fixed effects (full demeaning)}.$$

- If the random effect is unimportant relative to the idiosyncratic error (σ_a^2 small, $\lambda \to 0$), RE approaches pooled OLS.
- If the random effect dominates (σ_a^2 large or T large, $\lambda \to 1$), RE approaches fixed effects.

Remark (Properties of the transformed error).

It is a worthwhile exercise to verify, with λ as above, that the quasi-demeaned error $e_{it} = v_{it} - \lambda \bar{v}_i$ satisfies $\mathbb{E}(e_{it}) = 0$, $\mathrm{Var}(e_{it}) = \sigma_u^2$, and $\mathrm{Cov}(e_{it}, e_{is}) = 0$ for $t \neq s$. These three facts are exactly the Gauss–Markov conditions, which is why $\widehat{\beta}_1$ from the quasi-demeaned regression is BLUE under the RE assumptions.

9.4.3 What RE buys and what it assumes

A genuine advantage of RE over FE is that it *can* estimate the effects of time-invariant regressors: because it does not fully demean, a constant-over-time variable is not annihilated. The catch is the maintained assumption $\mathrm{Cov}(x_{it}, a_i) = 0$. In economics, unobserved individual effects are rarely plausibly uncorrelated with the regressors — ability is correlated with schooling, location quality with firm decisions — so fixed effects is usually the more convincing default. The choice between RE (efficient but fragile) and FE (robust but less efficient) is the central modeling decision, and the next section turns it into a formal test.

9.5 Correlated Random Effects

Correlated random effects (CRE) is a device that contains FE and RE as special cases, makes the difference between them visible, and delivers a clean test for choosing between them. We start from the same model,

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

but now we neither force a_i to be uncorrelated with the regressors (as RE does) nor leave its correlation entirely unmodeled (as FE does). Instead we model that correlation through Mundlak's device.

9.5.1 Mundlak's device

Project the unobserved effect onto the unit's time-average regressors:

$$a_i = \gamma_0 + \gamma_1 \bar{x}_i + \gamma_i,$$

where $\bar{x}_i$ is the time average of x_{it} and the new error γ_i is assumed random and uncorrelated with each x_{it} (so $\text{Cov}(\gamma_i, \bar{x}_i) = 0$). This says: whatever correlation a_i has with the regressors runs through their time average $\bar{x}_i$. Substituting into the model gives the *CRE equation*

$$y_{it} = \gamma_0 + \beta_1 x_{it} + \gamma_1 \bar{x}_i + v_{it}, \qquad v_{it} = u_{it} + \gamma_i,$$

with $\mathbb{E}(v_{it}) = 0$ and $\text{Cov}(v_{it}, x_{it}) = 0$. The only thing distinguishing the CRE equation from a plain RE (or pooled OLS) equation is the extra regressor $\gamma_1 \bar{x}_i$ — the unit's time-average of x.

9.5.2 CRE reproduces fixed effects

Estimating the CRE equation by RE (or even by pooled OLS) yields a remarkable identity: the coefficient on the time-varying regressor equals the fixed-effects estimate,

$$\widehat{\beta}_{1,\text{CRE}} = \widehat{\beta}_{1,\text{FE}}.$$

In words: adding the time average $\bar{x}_i$ as a regressor and then running RE/OLS is *equivalent* to demeaning out the time average and running pooled OLS. This gives a new reading of fixed effects — when estimating the partial effect of x_{it}, FE is implicitly *controlling for the time average* $\bar{x}_i$.

CRE as a unifying device

The single coefficient γ_1 organizes the whole family:

- CRE estimated as above reproduces the *FE* slope $\widehat{\beta}_{1,\text{FE}}$ exactly.
- Setting $\gamma_1 = 0$ collapses the CRE equation to the *RE* equation.
- So γ_1 measures precisely the gap between RE and FE.

When $\gamma_1 \neq 0$, the regressors $\bar{x}_i$ and x_{it} appear together and are collinear (especially when x_{it} varies little over time), which inflates the standard error of $\widehat{\beta}_{1,\text{FE}}$. This is the formal reason FE is generally less precise than RE. A bonus of the CRE formulation is that, unlike pure FE, it *can* also estimate the effects of time-constant regressors, by including them alongside the time averages.

9.5.3 A formal FE-versus-RE test

Because γ_1 is exactly the wedge between RE and FE, testing whether it is zero *is* the test of RE against FE. RE imposes $\gamma_1 = 0$; FE leaves γ_1 free. The null hypothesis favors the

simpler, more efficient model:

$$H_0 : \gamma_1 = 0 \quad \text{(RE is valid)}.$$

Estimate the CRE equation and test H_0 with a (cluster-robust) t or F statistic. If we fail to reject, RE is adequate and we enjoy its efficiency and its ability to estimate time-invariant effects. If we reject, the unobserved effect is correlated with the regressors and FE is preferred. This Mundlak/CRE test is a robust, regression-based alternative to the classical Hausman test, and it makes precise what "choosing between FE and RE" really means.

9.6 General Policy Analysis with Panel Data

The two-period before/after design of Section 9.1 is the simplest case of a much more general panel policy framework, available whenever $T \geq 2$. Write

$$y_{it} = \delta_1 + \delta_2\, d2_t + \cdots + \delta_T\, dT_t + \beta\, w_{it} + x_{it}\varphi + a_i + u_{it},$$

where $d2_t, \ldots, dT_t$ are time dummies (aggregate shocks common to all units), w_{it} is the binary policy variable, x_{it} are other controls with coefficient vector φ, a_i is the unit fixed effect, and u_{it} the idiosyncratic error. The coefficient β is the ATE of the policy.

The whole point of the panel is to let the policy be *systematically related* to the unobserved effect a_i — as happens under self-selection, when units that adopt a policy differ in unobserved, time-constant ways from those that do not. We accommodate this by estimating with FD or FE (both of which eliminate a_i), using *cluster-robust* standard errors to handle within-unit serial correlation.

Two refinements matter for credible policy work.

Remark (Testing for feedback).

> Strict exogeneity rules out feedback from the error to future policy. If we worry that the policy variable *reacts* to past shocks — the outcome's error feeds back into next period's policy — we can test for it. Add a *lead* of the policy, $\delta\, w_{i,t+1}$, to the model. Under strict exogeneity the future policy should not predict the current outcome, so we test $H_0 : \delta = 0$. Rejection is evidence of feedback (and a violation of strict exogeneity).

Remark (Unit-specific trends).

> If different units are on different time *trends* (not just different levels), and $T \geq 3$, we can add a unit-specific linear trend $g_i\, t$ to the model. This lets the policy be correlated not only with level differences across units (captured by a_i) but also with trend differences. The augmented model is still estimated by FE; under first differencing the term $g_i\, t$ differences to the unit-specific constant g_i, which is then swept out alongside the usual intercept.

Remark (Where this is heading).

Panel methods defeat endogeneity of a very specific kind: *time-constant* unobserved heterogeneity, removable by differencing or demeaning, provided the regressors vary over time and the idiosyncratic error is strictly exogenous. They cannot rescue us from endogeneity that is itself time-varying, nor from time-invariant regressors of direct interest, nor from violations of strict exogeneity such as simultaneity or measurement error. For those problems we need a genuinely external source of variation — an instrument — which is the subject of Chapter 10.

Chapter 10

Instrumental Variables and Two-Stage Least Squares

Every estimator in this book has rested on one fragile assumption: that the explanatory variables are uncorrelated with the error term. When wages depend on education, we assumed that the unobserved factors lumped into the error — innate ability, family background, drive — were unrelated to how much schooling a person got. That assumption is almost never literally true. People who would earn more anyway also tend to acquire more education, so the very thing we cannot measure is tangled up with the thing we can. When this happens we say the regressor is *endogenous*, and ordinary least squares no longer recovers the parameter we care about: it is biased and, worse, inconsistent — more data does not save us.

The previous chapters offered two partial cures. A good proxy variable (Chapter 8) can stand in for the missing factor; fixed effects (Chapter 9) can sweep out an omitted variable when it is constant over time and we have panel data. Both are special-purpose tools that work only under favorable circumstances. This chapter introduces the most general and most widely used remedy for endogeneity: the method of *instrumental variables* (IV), and its operational cousin, *two-stage least squares* (2SLS). The idea is to find a new variable — an *instrument* — that is correlated with the troublesome regressor but otherwise plays no role in the equation and is unrelated to the error. Such a variable provides a clean source of variation in the regressor that we can use to identify its effect, much as a randomized nudge would.

The price of this generality is a new burden of belief. Two of an instrument's three defining properties can be checked in the data; the crucial third one — that the instrument is unrelated to the error — cannot be tested with a single instrument and must be argued from economic reasoning. The IV estimator is also biased in finite samples, even when it is consistent, and it behaves badly when the instrument is only weakly related to the regressor. The bulk of this chapter is about making these trade-offs precise.

10.1 Endogeneity

Recall the central assumption that makes OLS work. In the simple regression model

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

the slope β_1 is consistently estimated by OLS provided the *weak* condition

$$\mathrm{Cov}(x, u) = 0$$

holds. (We call it weak because it is implied by, but weaker than, the zero conditional mean assumption $\mathbb{E}(u \mid x) = 0$ of Chapter 1; for *consistency* of the slope, zero covariance is all we need.) When this condition holds, OLS is the best tool available and there is no reason to look further.

Definition 10.1: Endogenous and Exogenous Regressors

A regressor x is *endogenous* if it is correlated with the error term,

$$\mathrm{Cov}(x, u) \neq 0,$$

and *exogenous* otherwise. In the multiple regression model $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$, the regressor x_j is endogenous whenever $\mathrm{Cov}(x_j, u) \neq 0$ for that particular j.

When a regressor is endogenous, OLS is inconsistent: the part of u that moves with x gets misattributed to x, and the estimated slope absorbs a contaminating term that does not vanish as the sample grows. Endogeneity is, in the words of applied economists, *endemic* in the social sciences. It arises through several distinct channels, but for the purposes of this chapter you need not agonize over which one is operating — it is enough to picture an important variable that has been left out of the equation and folded into u. The main sources are:

- *Omitted variables.* Important personal characteristics — ability, motivation, health — often cannot be observed, and they are typically correlated with the included regressors. Their absence pushes their influence into u, which then moves with x.
- *Measurement error.* If we observe a noisy proxy for the true regressor instead of the regressor itself, the measurement noise enters the error term and is mechanically correlated with the mismeasured variable. We treat this case in detail in Section 10.7.
- *Simultaneity.* When y and x are determined jointly — price and quantity, for instance — feedback from y to x makes x correlated with u.

The tools we already have address endogeneity only under restrictive conditions: the proxy-variable method handles certain omitted regressors, and fixed-effects methods work when panel data are available, the source of endogeneity is time-constant, and the regressors of interest *do* vary over time. Instrumental variables estimation is the general-purpose alternative, and it is today the most well-known and favored approach to the problem.

10.2 The Instrumental Variable

An instrument is a variable z that lets us extract clean variation in an endogenous regressor. To do its job it must satisfy three requirements. The first is a modeling restriction (an exclusion), the second can be verified in the data, and the third must be taken largely on faith.

Definition 10.2: Instrument: the Three Requirements

A variable z is a valid *instrument* for the endogenous regressor x in $y = \beta_0 + \beta_1 x + u$ if it satisfies:

1. **Exclusion.** z does not appear in the structural equation; it has no direct effect on y once x is accounted for.
2. **Relevance.** z is correlated with the endogenous regressor:
$$\text{Cov}(z, x) \neq 0.$$
3. **Exogeneity.** z is uncorrelated with the error term:
$$\text{Cov}(z, u) = 0.$$

It is worth dwelling on each requirement, because everything in this chapter follows from them.

Exclusion. The instrument must not be one of the explanatory variables in the equation of interest — it must not help explain y directly. There is a purely mechanical reason for this, beyond the conceptual one. If z already appeared in the structural equation, then using it as an instrument would leave us with fewer moment conditions than unknown parameters, and the slope coefficients could not be solved for by the method of moments (Section 10.4); in a two-stage least squares implementation the same redundancy shows up as severe multicollinearity. The instrument earns its keep precisely by being *outside* the equation.

Relevance. The instrument must be correlated with the endogenous regressor; otherwise it is not a "representative" of x and carries no usable information about it. This requirement *can* be checked: simply regress x on z (and on any other exogenous variables in the model) and test whether the coefficient on z is statistically significant. When the correlation is present but faint — $\text{Cov}(z, x)$ close to zero — we call z a *weak instrument*, and Section 10.6 shows that weak instruments do real damage.

Exogeneity. The instrument must be uncorrelated with the structural error. This is the requirement that gives IV its power — and its peril — because in general it *cannot be tested* when there is only one instrument per endogenous regressor. The reason is circular: to test whether z is correlated with u we would need the errors u_i, but the errors are unobserved and can only be estimated from residuals, which themselves require consistent coefficient estimates — exactly what we are trying to obtain. Because x is endogenous, OLS gives us

the wrong $\widehat{\beta}_0, \widehat{\beta}_1$ and hence the wrong residuals, so there is nothing reliable to test against. Exogeneity therefore rests on an economic argument, a *belief* about how the world works, rather than on a statistic. (When instruments outnumber endogenous regressors, a partial test becomes available; see Section 10.9.)

10.3 IV Estimation in Simple Regression

We build intuition in the one-regressor case, contrasting the exogenous and endogenous situations and reading off the IV estimator as a natural generalization of OLS.

10.3.1 The Exogenous Benchmark

Start from $y_i = \beta_0 + \beta_1 x_i + u_i$ and suppose x is exogenous. Identification of β_1 flows directly from the zero-covariance condition. Imposing $\mathrm{Cov}(x, u) = 0$ and substituting for $u = y - \beta_0 - \beta_1 x$,

$$\begin{aligned} & \mathrm{Cov}(x_i, u_i) = 0 \\ \Longrightarrow\ & \mathrm{Cov}(x_i, y_i - \beta_0 - \beta_1 x_i) = 0 \\ \Longrightarrow\ & \mathrm{Cov}(x_i, y_i) - \beta_1\, \mathrm{Var}(x_i) = 0 \\ \Longrightarrow\ & \beta_1 = \frac{\mathrm{Cov}(x_i, y_i)}{\mathrm{Var}(x_i)}. \end{aligned}$$

Replacing each population moment by its sample analogue recovers the familiar OLS slope,

$$\widehat{\beta}_1 = \frac{\widehat{\mathrm{Cov}}(x_i, y_i)}{\widehat{\mathrm{Var}}(x_i)} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \qquad \widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x},$$

which is consistent, $\widehat{\beta}_1 \xrightarrow{p} \beta_1$, exactly as long as exogeneity holds.

10.3.2 The Endogenous Case and the IV Estimator

Now suppose $\mathrm{Cov}(x_i, u_i) \neq 0$, so x is endogenous and the OLS argument above collapses — we can no longer set $\mathrm{Cov}(x, u)$ to zero. Suppose, however, that we have found an instrument z_i satisfying the three requirements, in particular exogeneity $\mathrm{Cov}(z_i, u_i) = 0$ and relevance $\mathrm{Cov}(z_i, x_i) \neq 0$. Repeating the identification argument but anchoring it on z rather than x,

$$\begin{aligned} & \mathrm{Cov}(z_i, u_i) = 0 \\ \Longrightarrow\ & \mathrm{Cov}(z_i, y_i - \beta_0 - \beta_1 x_i) = 0 \\ \Longrightarrow\ & \mathrm{Cov}(z_i, y_i) - \beta_1\, \mathrm{Cov}(z_i, x_i) = 0 \\ \Longrightarrow\ & \beta_1 = \frac{\mathrm{Cov}(z_i, y_i)}{\mathrm{Cov}(z_i, x_i)}. \end{aligned}$$

The relevance condition $\mathrm{Cov}(z_i, x_i) \neq 0$ is what makes this division legitimate; it is the reason relevance is indispensable. Replacing population moments by sample moments gives the *instrumental variables estimator*.

Definition 10.3: IV Estimator (Simple Regression)

Given an instrument z for the endogenous regressor x, the IV estimator of the slope is the ratio of sample covariances

$$\widehat{\beta}_{1,\mathrm{IV}} = \frac{\widehat{\mathrm{Cov}}(z_i, y_i)}{\widehat{\mathrm{Cov}}(z_i, x_i)} = \frac{\sum_{i=1}^{n}(z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^{n}(z_i - \bar{z})(x_i - \bar{x})},$$

and the intercept is $\widehat{\beta}_{0,\mathrm{IV}} = \bar{y} - \widehat{\beta}_{1,\mathrm{IV}}\,\bar{x}$.

The denominator is the sample covariance *between the instrument and the regressor*, $\widehat{\mathrm{Cov}}(z_i, x_i)$ — not a variance. The intercept formula is valid because $\bar{y} = \beta_0 + \beta_1 \bar{x}$ continues to hold whenever $\mathbb{E}(u) = 0$, regardless of endogeneity.

Two special cases tie this back to what we know. First, if x happens to be exogenous, we may simply take $z \equiv x$: the IV formula then reduces to the OLS formula, because $\widehat{\mathrm{Cov}}(x, x) = \widehat{\mathrm{Var}}(x)$. In this sense *every exogenous variable is its own instrument.* Second, the IV estimator has a finite-sample property worth flagging.

IV is biased but consistent

Unlike OLS under exogeneity, the IV estimator is *biased* in finite samples:

$$\mathbb{E}\left(\widehat{\beta}_{1,\mathrm{IV}}\right) \neq \beta_1, \qquad \text{yet} \qquad \widehat{\beta}_{1,\mathrm{IV}} \xrightarrow{p} \beta_1.$$

The bias arises because the estimator is a ratio of random quantities, and the expectation of a ratio is not the ratio of expectations. Consistency holds as the sample grows because sample covariances converge to their population counterparts and the ratio converges to $\mathrm{Cov}(z, y) \,/\, \mathrm{Cov}(z, x) = \beta_1$.

10.3.3 What a Poor Instrument Costs: IV versus OLS

When can IV actually be *worse* than the OLS estimator it is meant to replace? The danger is an instrument that is not perfectly exogenous (a small $\mathrm{Cov}(z, u)$) and only weakly relevant (a small $\mathrm{Cov}(z, x)$). To compare the two estimators we examine their probability limits — the values they converge to, which capture their inconsistency.

Write the IV estimator in terms of the errors. Substituting $y_i = \beta_0 + \beta_1 x_i + u_i$ into the numerator and using $\sum_{i=1}^{n}(z_i - \bar{z}) = 0$,

$$\begin{aligned}\widehat{\beta}_{1,\mathrm{IV}} &= \frac{\sum_{i=1}^{n}(z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^{n}(z_i - \bar{z})(x_i - \bar{x})} = \frac{\sum_{i=1}^{n}(z_i - \bar{z})(\beta_0 + \beta_1 x_i + u_i)}{\sum_{i=1}^{n}(z_i - \bar{z})(x_i - \bar{x})} \\ &= \beta_1 + \frac{\sum_{i=1}^{n}(z_i - \bar{z})\,u_i}{\sum_{i=1}^{n}(z_i - \bar{z})(x_i - \bar{x})}.\end{aligned}$$

Dividing numerator and denominator of the second term by n and taking probability limits (sample covariances converge to population covariances, which we may write in correlation

form),

$$\operatorname{plim} \widehat{\beta}_{1,\mathrm{OLS}} = \beta_1 + \operatorname{Corr}(x,u) \cdot \frac{\sigma_u}{\sigma_x},$$

$$\operatorname{plim} \widehat{\beta}_{1,\mathrm{IV}} = \beta_1 + \frac{\operatorname{Corr}(z,u)}{\operatorname{Corr}(z,x)} \cdot \frac{\sigma_u}{\sigma_x}.$$

The OLS expression is the familiar inconsistency from endogeneity: its size is governed by $\operatorname{Corr}(x,u)$. The IV expression replaces this with the ratio $\operatorname{Corr}(z,u)/\operatorname{Corr}(z,x)$. Comparing the magnitudes of the two contaminating terms gives a sharp criterion.

When IV is worse than OLS

The IV estimator is *more* inconsistent than OLS if and only if

$$\frac{|\operatorname{Corr}(z,u)|}{|\operatorname{Corr}(z,x)|} > |\operatorname{Corr}(x,u)|.$$

Two readings follow. A small violation of exogeneity ($\operatorname{Corr}(z,u)$ slightly nonzero) can be amplified into a large inconsistency when relevance is weak ($\operatorname{Corr}(z,x)$ small), because the violation is divided by a small number. Conversely, a strong, clean instrument — $\operatorname{Corr}(z,u) \approx 0$ and $\operatorname{Corr}(z,x)$ large — makes the IV term negligible and IV strictly preferable.

This is the formal warning behind the slogan that a slightly invalid, weak instrument can be worse than no instrument at all.

10.4 IV Estimation in Multiple Regression

Real applications have several regressors, some exogenous and some not. Write the equation of interest — the *structural equation* — as

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \cdots + \beta_k z_{k-1} + u_1,$$

where y_2 is the endogenous explanatory variable (the change of name from x to y_2 signals that it, like y_1, is correlated with the error), and $z_1, \ldots, z_{k-1}$ are exogenous explanatory variables already in the equation. We call this a structural equation because the β_j are the parameters of economic interest; the label does *not* assert that the equation represents a causal mechanism — it is simply the relationship we wish to estimate.

To handle the endogenous y_2 we need an instrument z_k satisfying the three requirements, restated for the multiple-regression setting:

1. z_k does not appear in the structural equation (exclusion);
2. z_k is uncorrelated with the error term u_1 (exogeneity);
3. z_k is *partially* correlated with y_2 after the other exogenous variables are controlled for (relevance).

The word "partially" matters. With other regressors present, the relevant notion of relevance is the correlation of z_k with y_2 *net of* $z_1, \ldots, z_{k-1}$. To make this precise, project y_2 onto *all* the exogenous variables:

Definition 10.4: Reduced Form (First-Stage) Equation

The *reduced form* for the endogenous regressor y_2 regresses it on every exogenous variable in the system — the exogenous regressors already in the structural equation *and* the excluded instrument(s):

$$y_2 = \pi_0 + \pi_1 z_1 + \cdots + \pi_{k-1} z_{k-1} + \pi_k z_k + v_2,$$

with $\mathbb{E}(v_2) = 0$ and v_2 uncorrelated with every exogenous variable. The included exogenous regressors $z_1, \ldots, z_{k-1}$ are present to prevent omitted-variable bias in the reduced form.

The partial-relevance condition is now an honest statistical hypothesis: z_k is relevant precisely when $\pi_k \neq 0$. The coefficient π_k measures the strength of the (partial) correlation between y_2 and z_k, and we want the null hypothesis $\pi_k = 0$ to be firmly rejected. When there is more than one excluded instrument, relevance becomes the joint hypothesis that *all* their reduced-form coefficients are zero, which we again want to reject.

IV estimation in this setting can be carried out in two equivalent ways: the *method of moments* and *two-stage least squares*. We take them in turn.

10.4.1 The Method of Moments

Consider the compact case

$$y_{1i} = \beta_0 + \beta_1 y_{2i} + \beta_2 z_{1i} + u_i,$$

with y_2 endogenous, z_1 exogenous, and an instrument z_2 found for y_2. We have three parameters $(\beta_0, \beta_1, \beta_2)$ and exactly three population moment conditions: $\mathbb{E}(u) = 0$ (from the intercept), and the exogeneity of the two exogenous variables z_1 and z_2. Writing each condition out and replacing it by its sample analogue:

$$\begin{cases} \mathbb{E}(u_i) = 0 & \Longrightarrow \ \dfrac{1}{n}\sum_{i=1}^{n} (y_{1i} - \beta_0 - \beta_1 y_{2i} - \beta_2 z_{1i}) = 0, \\ \mathrm{Cov}(z_{1i}, u_i) = 0 & \Longrightarrow \ \dfrac{1}{n}\sum_{i=1}^{n} z_{1i}\,(y_{1i} - \beta_0 - \beta_1 y_{2i} - \beta_2 z_{1i}) = 0, \\ \mathrm{Cov}(z_{2i}, u_i) = 0 & \Longrightarrow \ \dfrac{1}{n}\sum_{i=1}^{n} z_{2i}\,(y_{1i} - \beta_0 - \beta_1 y_{2i} - \beta_2 z_{1i}) = 0. \end{cases}$$

(For the latter two, $\mathrm{Cov}(z_j, u) = 0$ together with $\mathbb{E}(u) = 0$ gives $\mathbb{E}(z_j u) = 0$, whose sample analogue is the displayed equation.) These are three linear equations in three unknowns, and solving them yields the method-of-moments IV estimators $\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2$.

The counting of equations against unknowns gives the language of identification.

Definition 10.5: Identification by Counting Instruments

Let the structural equation have g endogenous regressors and suppose we have m excluded instruments available. Then:

- $m = g$: *just-identified*. The number of moment conditions equals the number of unknowns; the method of moments has a unique solution.
- $m > g$: *over-identified*. There are more moment conditions than unknowns.
- $m < g$: *under-identified*. There are too few; the parameters cannot be estimated.

The method of moments handles only the just-identified case cleanly: with one endogenous variable it needs exactly one instrument; with two endogenous variables it needs exactly two; in general the number of instruments must equal the number of endogenous regressors so that the system has exactly one solution. If we are fortunate enough to have *more* instruments than endogenous variables — say three instruments for one endogenous regressor — the method of moments breaks down, because the system of equations would be overdetermined (one solution per equation, but no single set of coefficients satisfies all of them). The valuable extra instruments simply cannot be used. This limitation is exactly what two-stage least squares overcomes.

10.5 Two-Stage Least Squares

Two-stage least squares (2SLS) is the workhorse implementation of IV. It accommodates over-identification gracefully, extends to several endogenous regressors, and reduces to plain IV in the just-identified case. As its name promises, it runs two ordinary regressions in sequence.

10.5.1 The Two Stages

Consider the general model

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \cdots + \beta_k z_{k-1} + u,$$

with y_2 endogenous, the remaining explanatory variables exogenous, and an instrument z_k found for y_2.

The 2SLS Algorithm

First stage (reduced form). Regress the endogenous variable y_2 on *all* exogenous variables — the exogenous regressors $z_1, \ldots, z_{k-1}$ already in the equation together with the excluded instrument(s) z_k — by OLS, and form the fitted values:

$$\widehat{y}_2 = \widehat{\pi}_0 + \widehat{\pi}_1 z_1 + \cdots + \widehat{\pi}_{k-1} z_{k-1} + \widehat{\pi}_k z_k.$$

This strips y_2 of its endogenous part, keeping only the component explained by exogenous information.

Second stage. Run OLS on the structural equation with y_2 replaced by its first-stage fitted value $\widehat{y}_2$:

$$y_1 = \beta_0 + \beta_1 \widehat{y}_2 + \beta_2 z_1 + \cdots + \beta_k z_{k-1} + \text{error}.$$

The resulting coefficient on $\widehat{y}_2$ is the 2SLS estimate of β_1.

Note the asymmetry between stages: the first stage uses the actual instrument z_k to build $\widehat{y}_2$, but the instrument z_k itself never appears in the second stage — it has done its work by purifying y_2.

10.5.2 Why 2SLS Works

The intuition. Every regressor in the second-stage equation is exogenous. The original exogenous variables $z_1, \ldots, z_{k-1}$ were exogenous to begin with, and the constructed regressor $\widehat{y}_2$ is a linear combination of exogenous variables only, so it too is uncorrelated with u. By replacing y_2 with a prediction built from exogenous information, we have purged it of the endogenous component that was correlated with the error. With every regressor now exogenous, OLS in the second stage is consistent.

The reduced-form derivation. To see why the fitted value is the *best* possible instrument, take the simplest over-identified case: the structural equation $y_1 = \beta_0+\beta_1 y_2+\beta_2 z_1+u$, with two excluded exogenous variables z_2 and z_3 available. By assumption z_2 and z_3 do not appear in the structural equation and each is uncorrelated with u — conditions known as the *exclusion restrictions* (these are the IV requirements minus relevance).

Now observe: since the included exogenous z_1 is uncorrelated with u, and the excluded z_2, z_3 are each uncorrelated with u, *any* linear combination of z_1, z_2, z_3 is uncorrelated with u and hence a valid instrument. There are infinitely many candidates. The best one — the combination most strongly correlated with y_2, and therefore the strongest instrument — is the linear combination delivered by the reduced form of y_2:

$$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3 + v, \qquad \mathbb{E}(v) = 0, \;\; \text{Cov}(z_1, v) = \text{Cov}(z_2, v) = \text{Cov}(z_3, v) = 0.$$

Call the systematic part $y_2^* = \pi_0+\pi_1 z_1+\pi_2 z_2+\pi_3 z_3$. The reduced form cleanly splits y_2 into the exogenous piece y_2^*, which is uncorrelated with u, and the disturbance v, which carries the endogeneity: v is the reason y_2 may be correlated with u. For this best instrument to add information beyond z_1 — that is, for the relevance / identification condition to hold — y_2^* must not be perfectly explained by z_1 alone. We check this by testing the joint hypothesis $\pi_2 = \pi_3 = 0$ in the reduced form. If we *cannot* reject it, the excluded instruments add nothing and the structural equation is not identified.

We do not know the true π_j, but we can estimate the reduced form by OLS to obtain

$$\widehat{y}_2 = \widehat{\pi}_0 + \widehat{\pi}_1 z_1 + \widehat{\pi}_2 z_2 + \widehat{\pi}_3 z_3,$$

and use $\widehat{y}_2$ as the single instrument for y_2. This is the key economy of 2SLS: no matter how many instruments we started with, after estimating the reduced form there is exactly *one* constructed instrument, $\widehat{y}_2$, the best linear combination of all the exogenous variables.

Feeding $\widehat{y}_2$ into the just-identified method-of-moments system makes the equivalence with two-stage least squares transparent:

$$\begin{cases} \sum_{i=1}^{n} \left(y_{1i} - \widehat{\beta}_0 - \widehat{\beta}_1 y_{2i} - \widehat{\beta}_2 z_{1i} \right) = 0, \\ \sum_{i=1}^{n} z_{1i} \left(y_{1i} - \widehat{\beta}_0 - \widehat{\beta}_1 y_{2i} - \widehat{\beta}_2 z_{1i} \right) = 0, \\ \sum_{i=1}^{n} \widehat{y}_{2i} \left(y_{1i} - \widehat{\beta}_0 - \widehat{\beta}_1 y_{2i} - \widehat{\beta}_2 z_{1i} \right) = 0. \end{cases}$$

Solving this system gives the same estimates as the two-stage procedure. We can also read the consistency off the second-stage equation directly. Writing $y_2 = y_2^* + v$ and substituting,

$$y_1 = \beta_0 + \beta_1 y_2^* + \beta_2 z_1 + (u + \beta_1 v),$$

where the composite error $u + \beta_1 v$ has mean zero and is uncorrelated with both y_2^* and z_1 (because u is uncorrelated with the exogenous variables by exogeneity, and v is uncorrelated with them by construction of the reduced form). That is exactly why regressing y_1 on $\widehat{y}_2$ and z_1 delivers consistent estimates.

Remark (2SLS contains IV as a special case).

> If there is one endogenous variable and exactly one instrument, 2SLS is numerically identical to the simple IV estimator of Section 10.3. When instruments outnumber endogenous regressors, 2SLS uses all of them at once through the reduced form, whereas the basic method of moments cannot. This is why we say 2SLS is more general than the method of moments.

10.5.3 Several Endogenous Variables

The procedure scales directly. Suppose the structural equation has two endogenous regressors,

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 y_3 + \beta_3 z_1 + u,$$

and we have located two excluded instruments z_2, z_3 (at least as many instruments as endogenous variables). Then:

- **First stage:** regress y_2 on z_1, z_2, z_3 to get $\widehat{y}_2$, and *separately* regress y_3 on z_1, z_2, z_3 to get $\widehat{y}_3$. Each endogenous regressor gets its own reduced form, using the same full set of exogenous variables.
- **Second stage:** regress y_1 on $\widehat{y}_2, \widehat{y}_3, z_1$.

In general, identification by 2SLS requires at least as many excluded instruments as endogenous regressors.

10.5.4 Two Cautions in Practice

Remark (Multicollinearity and wrong standard errors).

Two practical points govern any 2SLS application.

(1) Multicollinearity. With more than one endogenous regressor, the fitted values $\widehat{y}_2$ and $\widehat{y}_3$ are both linear combinations of the *same* exogenous variables z_1, z_2, z_3, so they tend to be highly correlated with each other. As long as the collinearity is not perfect, 2SLS proceeds without difficulty, but the estimates may be imprecise.

(2) The naive second-stage standard errors are wrong. If one literally runs the two OLS regressions by hand, the standard errors reported by the second stage are incorrect. The reason is that the second stage treats $\widehat{y}_2$ as if it were data, ignoring that $\widehat{y}_2$ is itself an *estimate* from the first stage. Correct standard errors must propagate the first-stage estimation uncertainty into the second stage. Crucially, the *coefficient estimates* (the magnitudes of the slopes) are correct — it is only their reported standard errors that need fixing. In practice, dedicated 2SLS software computes the right standard errors automatically; one should never report the hand-run second-stage OLS standard errors.

10.6 Weak Instruments

Return to the simple IV estimator and look at its denominator,

$$\widehat{\beta}_{1,\text{IV}} = \frac{\widehat{\text{Cov}}(z_i, y_i)}{\widehat{\text{Cov}}(z_i, x_i)} = \frac{\sum_{i=1}^{n}(z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^{n}(z_i - \bar{z})(x_i - \bar{x})}.$$

If z is a weak instrument — only faintly correlated with x — the denominator $\sum_{i=1}^{n}(z_i - \bar{z})(x_i - \bar{x})$ is close to zero. Dividing by a number near zero makes the estimator volatile: its sampling distribution becomes *fat-tailed*, and the associated t statistic departs sharply from the standard normal that the usual asymptotics promise. The practical consequence is an inflated Type-I error rate. A test conducted at a nominal 5% level may in truth reject a true null far more often, so one risks "finding" an effect that is not there. Because this distortion is invisible if one trusts the textbook formulas blindly, weak instruments must be diagnosed explicitly — typically by checking that the first-stage relevance is strong (a large first-stage F statistic on the excluded instruments), not merely nonzero.

10.7 IV as a Cure for Measurement Error

Instrumental variables address more than omitted variables; they also repair the bias caused by *measurement error* in a regressor. Suppose the true model is

$$y = \beta_0 + \beta_1 x^* + u, \qquad \text{Cov}(x^*, u) = 0,$$

where the true regressor x^* is unobservable. What we observe instead is a noisy measurement

$$x = x^* + e,$$

where e is the measurement error. Under the *classical errors-in-variables* (CEV) assumption, the error is uncorrelated with the true value, $\text{Cov}(x^*, e) = 0$, which implies $\text{Cov}(x, e) = \text{Cov}(x^* + e, e) = \text{Var}(e) = \sigma_e^2 \neq 0$. Substituting $x^* = x - e$ into the model rewrites it in

terms of the observable x:

$$y = \beta_0 + \beta_1 x + (u - \beta_1 e).$$

The composite error $u-\beta_1 e$ is correlated with the regressor x, because both contain e: indeed $\mathrm{Cov}(x, u - \beta_1 e) = -\beta_1 \sigma_e^2 \neq 0$. So x is endogenous, purely as a result of measurement error, and OLS is inconsistent (it suffers from *attenuation bias*, pulling the estimate toward zero).

What we need is an instrument for x — a variable strongly correlated with x but uncorrelated with the measurement error e (and with u). A natural candidate is a *second, independent measurement* of the same underlying quantity. Let

$$z = x^* + \eta,$$

a different noisy reading of x^* with its own measurement error η. Assume the two measurement errors are uncorrelated, $\mathrm{Cov}(e, \eta) = 0$ (and that η is uncorrelated with u). Then z qualifies as an instrument for x:

- *Relevance:* z and x are correlated because both track the common signal x^* — they share the term x^*.
- *Exogeneity:* z is uncorrelated with the composite error $u - \beta_1 e$, since η is uncorrelated with both u and e.

Using the second measurement z as an instrument for the first measurement x therefore restores consistency.

10.8 Testing for Endogeneity

IV/2SLS is the right tool only when a regressor really is endogenous. If y_2 is in fact exogenous, OLS is consistent and *more efficient* than 2SLS, so we would not want to pay the variance cost of IV needlessly. We therefore want a test of whether y_2 is endogenous. Consider the structural equation

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \cdots + \beta_k z_{k-1} + u_1,$$

where y_2 is suspected to be endogenous and $z_1, \ldots, z_{k-1}$ are exogenous.

One could simply compare the OLS and 2SLS coefficient estimates: if they differ significantly, y_2 is likely endogenous (this is the logic of the Hausman test). A more convenient implementation runs entirely through a single auxiliary regression — the *control-function* approach. Begin with the reduced form

$$y_2 = \pi_0 + \pi_1 z_1 + \cdots + \pi_k z_k + v_2,$$

where z_k satisfies the exclusion restrictions. As we saw, $\widehat{y}_2$ is the clean, exogenous part of y_2, and the reduced-form residual v_2 absorbs all the potential endogeneity. It follows that y_2 is exogenous if and only if v_2 is uncorrelated with u_1. Write u_1 as a linear function of v_2,

$$u_1 = \delta_1 v_2 + e_1,$$

so that the endogeneity question reduces to the single hypothesis $\delta_1 = 0$.

The Regression-Based (Wu–Hausman) Endogeneity Test

1. Estimate the reduced form of y_2 by OLS and save the residuals $\widehat{v}_2$.
2. Add $\widehat{v}_2$ to the structural equation and estimate by OLS:

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \cdots + \beta_k z_{k-1} + \delta_1 \widehat{v}_2 + \text{error}.$$

3. Test $H_0 : \delta_1 = 0$ with the usual t statistic (using heteroskedasticity-robust standard errors if needed). Rejecting H_0 is evidence that v_2 and u_1 are correlated, hence that y_2 is endogenous.

When several regressors are suspected of endogeneity, the procedure generalizes naturally: regress each suspect variable on *all* exogenous variables (those in the structural equation and the excluded instruments) to obtain its reduced-form residual; add all the residuals to the structural equation and run OLS; then test the *joint* hypothesis that all the residual coefficients are zero with an F test. A rejection signals that at least one of the suspect regressors is endogenous.

10.9 Testing the Over-Identifying Restrictions

The exogeneity of an instrument cannot be tested when the model is just-identified. But when we have *more* instruments than we strictly need — the over-identified case — a partial test becomes possible. The idea is intuitive: if all the instruments are valid, then each one alone should produce an estimate consistent with the others, so estimates based on different instruments should agree up to sampling error. Equivalently, if every instrument is exogenous, the 2SLS residuals should be uncorrelated with all the instruments. A correlation between the residuals and the instruments is evidence that some instrument is invalid.

The Over-Identification (Sargan) Test

1. Estimate the structural equation by 2SLS and obtain the residuals $\widehat{u}_1$.
2. Regress $\widehat{u}_1$ on *all* the exogenous variables (the included exogenous regressors and every instrument), and record the R^2 from this auxiliary regression.
3. Under the null hypothesis that all instruments are uncorrelated with u_1,

$$nR^2 \overset{a}{\sim} \chi^2_q, \qquad q = (\text{number of instruments}) - (\text{number of endogenous regressors}).$$

If nR^2 exceeds the chi-square critical value with q degrees of freedom, reject H_0: at least one instrument fails exogeneity.

The degrees of freedom q are exactly the number of *surplus* instruments, the over-identifying restrictions being tested. Note the limits of the test. It can only be run when $q \geq 1$, so a just-identified model offers nothing to test. And a rejection tells us that *some* instrument is invalid without saying which; a failure to reject is reassuring but is not proof

that all instruments are exogenous, since the test has no power to detect violations common to all of them.

10.10 A Worked Example

Example (Schooling and wages).

A researcher wants the return to education in the log-wage equation

$$\log(\text{wage}) = \beta_0 + \beta_1 \,\text{educ} + \beta_2 \,\text{exper} + u,$$

but worries that *educ* is endogenous: unobserved ability raises both schooling and wages, so $\text{Cov}(\text{educ}, u) > 0$ and OLS overstates β_1. She proposes the subject's *number of siblings* as an instrument z, on the grounds that children from larger families tend to receive less schooling (relevance) but that family size has no direct effect on the wage once education and experience are controlled for (exclusion and exogeneity). Outline the 2SLS estimation, the endogeneity test, and explain why the over-identification test cannot be run here.

Solution.

Identification. There is one endogenous regressor (*educ*) and one excluded instrument (*siblings*), so the model is just-identified, $m = g = 1$. Here 2SLS coincides with the simple IV estimator.

First stage (relevance check). Regress the endogenous variable on all exogenous variables, including the instrument:

$$\text{educ} = \pi_0 + \pi_1 \,\text{exper} + \pi_2 \,\text{siblings} + v,$$

and confirm that $\widehat{\pi}_2$ is significantly negative (a strong first stage). A weak first stage would warn of the weak-instrument problem of Section 10.6; check the first-stage F on *siblings*.

Second stage. Form $\widehat{\text{educ}}$ from the first stage and regress

$$\log(\text{wage}) = \beta_0 + \beta_1 \,\widehat{\text{educ}} + \beta_2 \,\text{exper} + \text{error}.$$

The coefficient on $\widehat{\text{educ}}$ is the 2SLS estimate of β_1. Report 2SLS (not hand-run OLS) standard errors, since the second stage uses a generated regressor.

Endogeneity test. Save the first-stage residual $\widehat{v}$, add it to the original (uninstrumented) equation,

$$\log(\text{wage}) = \beta_0 + \beta_1 \,\text{educ} + \beta_2 \,\text{exper} + \delta_1 \widehat{v} + \text{error},$$

and test $H_0 : \delta_1 = 0$ with a t test. Rejection confirms that *educ* is endogenous and that IV was warranted.

Why no over-identification test. The model is just-identified, so $q = 1 - 1 = 0$: there are no surplus instruments and the Sargan statistic has zero degrees of freedom. The exogeneity of *siblings* therefore cannot be tested and must be defended by argument.

To enable the over-identification test, she would need a *second* instrument for *educ* (giving $q = 1$).

Remark (Where this leaves us).

Instrumental variables convert the untestable assumption $\text{Cov}(x, u) = 0$ into a different, often more credible one: the existence of a variable that shifts x without otherwise touching y. The estimator is biased but consistent, vulnerable to weak and invalid instruments, and best implemented as two-stage least squares so that over-identifying instruments can all be used and the standard errors come out right. We can test whether IV is needed (the endogeneity test) and, when instruments are plentiful, partially test their validity (the over-identification test) — but the core exogeneity assumption, in the just-identified case, remains a matter of economic judgment.