ECONOMETRICS II

Cross-sectional and Panel Data

Haimo Fang

Notes on ECON130277H by Prof. Ruochen Wu

Contents

1 Introduction

Welcome to the world of Econometrics II!

This course is about learning how to make sense of data when we want to answer economic questions. In your first econometrics class you saw the basics of regression and estimation; here we take those tools much further, focusing on two broad settings that economists frequently encounter:

- **Cross-sectional data:** many units observed at one point in time (e.g., household surveys).
- **Panel data:** many units tracked over multiple time periods (e.g., firm-level financials).

These data structures are powerful because they let us control for hidden differences across individuals and study dynamics over time. But with that power come new statistical headaches: *dependence*, *heteroskedasticity*, *correlation across equations*, and (most importantly) the need for *robust inference*.

Along the way we will build a foundation in the statistical machinery that underlies modern econometrics. Roughly speaking, the storyline is:

1. **Asymptotics:** how estimators behave in large samples (LLN, CLT, continuous mapping, Slutsky, etc.).
2. **OLS:** small-sample properties (BLUE) and large-sample properties (consistency and asymptotic normality).
3. **Beyond OLS when assumptions fail:** GLS, SURE, and robust variance ideas.
4. **Endogeneity and identification:** what goes wrong, and how we fix it (IV).
5. **Moment-based inference:** GMM and its logic (identification, estimation, testing).
6. **Likelihood-based inference:** MLE as the "full model" counterpart to moment methods.
7. **Binary choice models:** (logit/probit) as a canonical nonlinear likelihood setting.
8. **Panel data:** fixed effects, random effects, and the logic of within vs quasi-demeaning (and friends).

Sometimes our path will be rigorous and mathematical, other times intuitive and story-driven—and occasionally we will slip in a fun fact or two (because even econometricians deserve coffee-break trivia). My hope is that these notes will serve not only as a technical reference, but also as a friendly companion on your path through cross-sectional and panel data analysis. `=D`

1.1 What are we actually doing when we "do inference"?

Before we set off, let us pause for a moment on a deeper question: what does it even mean to "make inference" from data? The answer depends on your view of probability itself.

If you flip a fair coin, is the probability of a head exactly 1/2? Or could it itself be uncertain, varying across trials as if drawn from its own distribution? Your answer reveals which statistical "philosophy" you lean toward:

- **Frequentists** treat parameters as fixed but unknown numbers. Inference is about learning these constants from repeated samples; the randomness lies in the estimator.
- **Bayesians** treat parameters themselves as random variables. Inference is about describing uncertainty about parameters with probability distributions; the randomness lies in the parameter.

Although this philosophical distinction matters, in practice both camps share the same goal: learning about unknown features of the data-generating process. What differs is the machinery they use.

Frequentist inference. We report point estimates (e.g., $\hat{\theta}$) and quantify uncertainty via the *sampling distribution* of $\hat{\theta}$: confidence intervals and hypothesis tests are the workhorses.

Bayesian inference. We combine a prior and a likelihood to obtain a *posterior* $f(\theta \mid y)$ and report credible intervals, posterior probabilities (e.g., $\mathbb{P}(\beta > 0 \mid y)$), or decision rules under a loss function.

Fun Facts 1.1. A random fact and I should probably keep it as a definition: A collection of observations $\{(x_i, y_i)\}_{i=1}^n$ is called a *sample*.

1.2 A minimalist Bayesian skeleton

To formalize the Bayesian perspective, suppose our data are generated from a parametric family. Write the sampling model as

$$y_i \sim f(y_i; \theta).$$

A Bayesian introduces an additional layer:

$$y_i \mid \theta \sim f_Y(y_i; \theta) \quad \text{(likelihood)}, \qquad \theta \sim f_H(\theta) \quad \text{(prior)},$$

and updates beliefs via the posterior

$$f(\theta \mid y) \propto f_H(\theta) \prod_{i=1}^{n} f_Y(y_i; \theta).$$

Fun Facts 1.2 (Hyperparameters)**.** Parameters we must specify to fit the model, but which are not themselves the main object of interest.

Conjugate priors. If the prior and posterior belong to the same distributional family, the prior is said to be *conjugate* to the likelihood. Conjugacy makes Bayesian updating algebraically simple.

Example 1.3. *If the prior is normal and the likelihood is normal with known variance, then the posterior is also normal.*

Example 1.4 (Exponential Family)**.** *Many exponential family likelihoods (normal, binomial, Poisson, etc.) admit conjugate priors.*

When priors are not conjugate, deriving the posterior analytically may be impossible. In such cases, modern methods such as MCMC (e.g., Gibbs sampling, Metropolis–Hastings) allow us to generate draws that approximate the true posterior distribution.

Fun Facts 1.5. If $X \sim N(\mu, \sigma^2)$ and $Y = e^X$, then Y is log-normally distributed, written $Y \sim \ln N(\mu, \sigma^2)$.

1.3 Why inference?

At this point, you might be wondering: why care about inference at all, if prediction accuracy is what matters? After all, much of modern machine learning celebrates metrics like R^2, mean squared error, or cross-validation accuracy. And indeed, these are valuable: they tell us how well a model predicts unseen data.

But econometrics cares about more than prediction. Our central questions are often causal or structural: does education raise wages? does monetary policy affect inflation? Here, we are not satisfied with saying the model predicts well. We need to measure uncertainty, test hypotheses, and construct confidence intervals. R^2 is not enough when your job is to guide policy or test theory.

Interestingly, even the machine learning community is moving in this direction. Researchers now increasingly ask: how can we put error bars around complex black-box predictors? This is precisely an inference problem. Two examples (which you do *not* need for this course) come from ensemble methods:

- **Random Forests (RF).** Though designed as a prediction algorithm, under certain conditions Random Forest predictions satisfy a Central Limit Theorem (CLT). This means that as the number of trees grows, the prediction at a fixed point becomes approximately normal, allowing us to attach standard errors to forest predictions.
- **Bayesian Additive Regression Trees (BART).** BART takes an explicitly Bayesian route, putting priors on trees and learning posteriors through MCMC. The resulting

posterior predictive distribution provides a natural way to quantify uncertainty. Here too, asymptotic normality results exist, letting us treat BART as not just a prediction engine but an inference tool.

1.4 What's next?

Whether you lean Frequentist, Bayesian, or "I just want the standard errors to work," you eventually run into the same bottleneck: we need a language for uncertainty that behaves well as sample size grows. That language is asymptotics. This is why we begin in Chapter 2 with convergence concepts, LLN, and CLT—the infrastructure that will quietly power nearly everything later.

2 Asymptotics

2.1 Convergence in Probability

Think about how we can give a definition for random convergence: we say we can deterministically bound the series within a small enough interval given that N is large enough. But how can we define such behavior if the sequence is random? A deterministic convergence statement cannot be obtained anymore. Still, the idea of convergence can be accommodated in a probability perspective. One can make an attempt below:

Conjecture 2.1 (Asymptotics of deterministic series)**.** High level idea: when defining $\lim_{n\to\infty} a_n = a^*$, we are really just using the ε–N language.

Here is the conceptual swap: in deterministic analysis, we control the tail of the sequence *for sure*. in probability, we typically can't promise "for sure"—we can only promise "with probability close to one". So asymptotics becomes a language for saying: *for large n, bad events happen rarely.*

Asymptotics is the terminology you are looking for to define such analogs in probability arguments. And the first type of asymptotic we can give is **convergence in probability**.

Definition 2.1 (Convergence in probability)**.** Given a random sequence $\{X_n\}_{n=1}^{\infty}$, we say X_n converges to c **in probability** if for every $\varepsilon > 0$,

$$\mathbb{P}\big(|X_n - c| > \varepsilon\big) \to 0 \qquad (n \to \infty).$$

We denote this by $X_n \xrightarrow{p} c$.

Equivalently $\lim_{n\to\infty} \mathbb{P}(|X_n - c| < \varepsilon) = 1$ for every $\varepsilon > 0$.

A quick sanity check: this is really just the ε–N idea with one word changed. Instead of "for all $n \geq N$ the inequality holds," we ask "for all $n \geq N$ the inequality holds *with probability close to one*."

Here's an example of convergence in probability (if you are interested):

Example 2.2. *Consider a random variable X_n with $\mathbb{P}(X_n = 0) = 1 - \frac{1}{n}$, and $\mathbb{P}(X_n = n) = \frac{1}{n}$. It converges to 0 in probability since for any $\varepsilon > 0$,*

$$\mathbb{P}(|X_n - 0| > \varepsilon) = \mathbb{P}(X_n = n) = \frac{1}{n} \to 0.$$

This example is worth remembering because it shows what convergence in probability *does and does not* care about: X_n occasionally takes a huge value ($n!$), but the probability of that "jump" goes to 0. So probability convergence is about *frequency of failure*, not *size of failure*.

There's another type of convergence called *almost sure* convergence, which will be less of an interest in this course, though. Almost sure convergence is the "stronger, pathwise"

notion: it says that for *almost every* outcome ω, the sequence eventually behaves. Convergence in probability is weaker: it allows different ω's to misbehave at different n's, as long as the total probability of misbehavior goes to zero.

Fun Facts 2.3 (Almost sure convergence)**.** Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and random variables X_n, X, we say X_n converges to X **almost surely** if

$$\mathbb{P}\Big(\big\{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\big\}\Big) = 1.$$

2.2 Markov Inequality and Chebyshev Inequality

Often, convergence in probability does not come for free. It requires us to leverage information we know about the random variable at hand.

And here comes the practical problem: the definition of $X_n \xrightarrow{p} c$ asks us to control probabilities like $\mathbb{P}(|X_n - c| > \varepsilon)$. But in real life, we often don't know that probability exactly. What we *do* know (or can compute) are things like $\mathbb{E}[X_n]$ or $\mathrm{Var}(X_n)$. So we need inequalities that convert *moments* into *tail probabilities*. This is why Markov and Chebyshev show up everywhere: they are the duct tape of probability.

In proving convergence in probability (e.g. WLLN), lots of techniques have been developed. Some call them concentration inequalities. But among the earliest and most useful is the **Markov inequality**.

Lemma 2.1 (Markov inequality)**.** Let $X \geq 0$ almost surely and let $\delta > 0$. Then

$$\mathbb{P}(X > \delta) \leq \frac{\mathbb{E}[X]}{\delta}.$$

Proof.

$$\mathbb{E}[X] \geq \mathbb{E}\big[X\mathbf{1}(X \geq \delta)\big] \geq \delta\mathbb{E}\big[\mathbf{1}(X \geq \delta)\big] = \delta\,\mathbb{P}(X \geq \delta).$$

Divide both sides by δ. □

Markov is extremely general (it only needs nonnegativity), which is code for: the bound can be extremely loose. But loose is still better than nothing when you have nothing.

Another important inequality is the **Chebyshev inequality**, historically connected to Markov.

Lemma 2.2 (Chebyshev inequality)**.** Let X be a random variable with $\mathbb{E}[(X-c)^2] < \infty$, and let $\varepsilon > 0$. Then

$$\mathbb{P}(|X - c| > \varepsilon) \leq \frac{\mathbb{E}[(X-c)^2]}{\varepsilon^2}.$$

In particular, taking $c = \mathbb{E}[X]$ yields

$$\mathbb{P}\big(|X - \mathbb{E}[X]| > \varepsilon\big) \leq \frac{\mathrm{Var}(X)}{\varepsilon^2}.$$

Proof. Apply Markov to the nonnegative r.v. $(X-c)^2$:

$$\begin{aligned}\mathbb{E}[(X-c)^2] &\geq \mathbb{E}\big[(X-c)^2\mathbf{1}((X-c)^2 \geq \varepsilon^2)\big] \\ &\geq \varepsilon^2\,\mathbb{E}\big[\mathbf{1}((X-c)^2 \geq \varepsilon^2)\big] \\ &= \varepsilon^2\,\mathbb{P}((X-c)^2 \geq \varepsilon^2).\end{aligned}$$

Divide both sides by ε^2. □

Chebyshev is the first time variance becomes a protagonist: it tells you that if variance goes to zero, large deviations become rare. Which is exactly the vibe we need for proving things like "sample averages stabilize."

Both inequalities bound the probability of a deviation of our sequence from a proposed "limit," which matches the definition of convergence in probability. However, convergence in probability is sometimes less handy than a stronger notion of convergence, which we will call **convergence in higher order means**, also known as **convergence in L^p**.

The big idea: instead of controlling tail probabilities directly, we control expected losses like $\mathbb{E}\left[|X_n - c|^2\right]$, which can be easier to compute.

2.3 Convergence in higher order mean (L^p)

Definition 2.2 (Convergence in second order mean)**.** If $\lim_{n\to\infty} \mathbb{E}[|X_n - c|^2] = 0$, then we say X_n converges to c in **quadratic mean** (or L^2), denoted $X_n \xrightarrow{L^2} c$.

This type of convergence is stronger than convergence in probability. Intuition: large deviations get squared, so "rare but huge" events matter much more. So L^2 convergence is basically saying: not only are bad events rare, even their squared damage becomes negligible on average.

Now comes the key bridge: we introduced Markov/Chebyshev precisely so we can translate moment statements into probability statements. So it should not be surprising that L^2 convergence implies convergence in probability.

Theorem 2.4 (L^2 convergence implies convergence in probability)**.** If $X_n \xrightarrow{L^2} c$, then $X_n \xrightarrow{p} c$.

Proof. For any $\varepsilon > 0$, apply Markov to the nonnegative r.v. $(X_n - c)^2$:

$$\mathbb{P}(|X_n - c| > \varepsilon) = \mathbb{P}((X_n - c)^2 > \varepsilon^2) \leq \frac{\mathbb{E}[(X_n - c)^2]}{\varepsilon^2} \to 0.$$

□

So, if you can prove an L^2 statement, you automatically get a convergence-in-probability statement for free. (This is one of the rare times in life that "stronger implies weaker" works in your favor.)

There are lots of estimators (who themselves are random variables) we wish to prove convergence for. The sample mean is the first celebrity in this story. Before proving its L^2 convergence, we isolate a small reusable lemma below. In English, Theorem 2.5 is saying if the mean goes to c and the variance collapses to zero, then the whole random variable collapses to c in L^2.

Theorem 2.5. Given a sequence $\{X_n\}_{\mathbb{N}^+}$ with mean $\mathbb{E}[X_n] = \mu_n < \infty$ and variance $\sigma_n^2 < \infty$, if $\mu_n \to c$ and $\sigma_n^2 \to 0$, then $X_n \xrightarrow{L^2} c$.

Proof. Use the identity

$$\mathbb{E}\big[(X_n - c)^2\big] = \mathrm{Var}\,(X_n) + (\mathbb{E}[X_n] - c)^2 = \sigma_n^2 + (\mu_n - c)^2 \to 0.$$

□

Theorem 2.5 is mainly a tool for proving L^2 convergence of sample means later. This is intuitive: as n progresses, we have a pile of distributions that become more and more concentrated, and eventually collapse to a delta function at c.

Fun Facts 2.6. Here are some specification on what do we mean by each terminology: **mean**: sample mean. **Expectation**: population mean.

Fun Facts 2.7. You never say "I ate my breakfast" before, but rather "I had my breakfast." "Maketh" is just a fancy word for "Makes".

Enough fun facts. Utilizing 2.5, now we can state and prove a quadratic mean convergence of sample means.

Theorem 2.8 (Convergence of Sample Mean in Quadratic Mean)**.** Define $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. If $\{X_i\}_{i\geq 1}$ are i.i.d. with mean μ and variance $\sigma^2 < \infty$, then $\bar{X}_n \xrightarrow{L^2} \mu$.

Proof. First,

$$\mathbb{E}[\bar{X}_n] = \mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^n X_i\Big] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[X_i] = \mu.$$

Next, using independence,

$$\mathrm{Var}\left(\bar{X}_n\right) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^n X_i\right) = \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}\,(X_i) = \frac{\sigma^2}{n} \to 0.$$

Apply Theorem 2.5. □

Since L^2 convergence implies convergence in probability (Theorem 2.4), we can safely conclude that **the sample mean converges to the population expectation (mean) in probability**.

At this point we have enough machinery to talk about estimators in econometrics: we want them to hit the right target (eventually), and ideally not be systematically off-target even in finite samples. That leads to unbiasedness and consistency.

2.4 Unbiasedness & Consistency

Definition 2.3 (Unbiasedness)**.** Consider a parameter $\theta \in \mathbb{R}$ and an estimator $\hat{\theta}_n$ (a random variable). If $\mathbb{E}[\hat{\theta}_n] = \theta$, then we say $\hat{\theta}_n$ is unbiased for θ.

Definition 2.4 (Consistency)**.** Consider a parameter $\theta \in \mathbb{R}$ and an estimator $\hat{\theta}_n$ (a random variable). If $\hat{\theta}_n \xrightarrow{p} \theta$, then we say $\hat{\theta}_n$ is a consistent estimator for θ.

Unbiasedness is a concept under finite sample size. Consistency is a metric when you let the sample size grow. You can think of it this way: unbiasedness is "on average I'm right, even for fixed n" (a finite-sample promise), while consistency is "I might be wrong now, but give me enough data and I'll stop embarrassing myself" (an asymptotic promise). These are different virtues. Econometrics likes both, but settles for consistency a lot.

Notice that consistency does not imply unbiasedness, nor does unbiasedness imply consistency. The latter direction is easy to deal with. Here's a counterexample of the "consistent but biased" direction:

Example 2.9. *Let $\hat{\mu}_n = \bar{X}_n + \frac{1}{n}$. It is consistent for μ (because $\bar{X}_n \xrightarrow{p} \mu$ and $\frac{1}{n} \to 0$), but it is not unbiased:*

$$\mathbb{E}[\hat{\mu}_n] = \mathbb{E}[\bar{X}_n] + \frac{1}{n} = \mu + \frac{1}{n} \neq \mu.$$

This example also gives us a spoiler on a common pattern: "small deterministic biases" often vanish asymptotically, which means asymptotics can be forgiving—sometimes too forgiving, if you're not careful.

As promised before, L^p convergence sometimes comes off more handy as convergence in probability. If we take a closer look at Definition 2.4, a theorem on the consistency of the sample mean can be given by referring to the convergence of $\bar{X}_n$ (Theorem 2.8).

Theorem 2.10 (Consistency of sample means)**.** A sample mean $\bar{X}_n$ is a consistent estimator for the population mean μ.

At this point, you might reasonably ask: "why did we work so hard to prove consistency of the sample mean, when everyone already believes it in their bones?" Because this is the template: compute mean/variance ⇒ prove a mode of convergence ⇒ conclude consistency. The weak law of large numbers packages this template into a named theorem.

2.5 Univariate Weak Law of Large Numbers

Phew... that's quite a lot going on in previous sections. We've been discussing different types of convergence and something on unbiasedness and consistency as well. But what is really governing those properties? How do we know that a sample mean is (usually) consistent?

The short answer is: averaging kills noise. The longer answer is: averaging kills noise *under conditions*, and probability theory is where we list those conditions in public.

When you place your cursor on the `select` button on `jwfw.com` to sign up for classes, why you don't see yourself enrolling in none of them?

Well, if you haven't heard of the weak law of large numbers before, here, I quote from Dr. Sheldon Cooper:

> Oh, well, this would be one of those circumstances that people unfamiliar with the law of large numbers would call a coincidence.

Take this!

Theorem 2.11 (Khintchine's WLLN)**.** If $\{X_i\}_{i\geq 1}$ are i.i.d. and $\mathbb{E}[|X_1|] < \infty$ with $\mathbb{E}[X_1] = \mu$, then

$$\bar{X}_n \xrightarrow{p} \mu.$$

This is the first WLLN we see in this class. And if you are interested in probability theory, you know there are many versions of LLN. A helpful way to distinguish between them is to look at assumptions: here, we only assume i.i.d. and finite mean. We make *no* assumptions on the variance.

Why does that matter, though? Because in real data you can absolutely have heavy tails: finite mean might be plausible, finite variance might not. So Khintchine is a nice reminder that "variance exists" is not a law of nature.

Fun Facts 2.12 (A little digress on the Student-t Distribution. You can skip it if you are annoyed.)**.** A random variable following a t-distribution has density

$$f(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\pi\nu}\Gamma(\frac{\nu}{2})}\Big(1+\frac{t^2}{\nu}\Big)^{-\frac{\nu+1}{2}}, \qquad \nu > 0,\ t \in \mathbb{R}.$$

ν is called the **degree of freedom** (roughly: number of data points minus number of estimated "free" parameters).
A t-statistic for an estimator is given by

$$t = \frac{\hat{\theta}_n - \theta}{s.e.(\hat{\theta}_n)}.$$

A t distribution approaches $\mathcal{N}(0,1)$ as $\nu \to \infty$. A t distribution with df 1 is a Cauchy distribution.
For a $t(\nu)$ distribution: the mean exists iff $\nu > 1$ and the variance exists iff $\nu > 2$. So for $1 < \nu \leq 2$ you get finite mean but infinite variance.

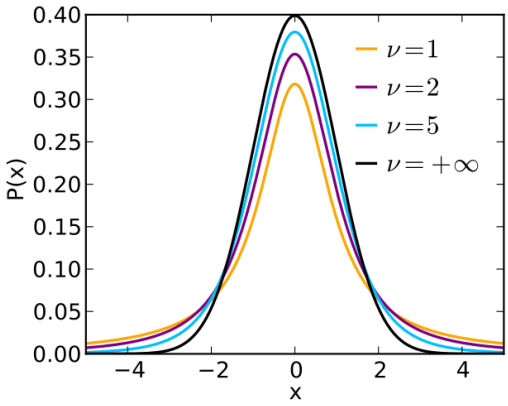

Figure 1: t-distributions

Fun Facts 2.13 (Student-t Distribution (cont.))**.** Since we are talking about t-statistics, consider the linear model

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

The degree of freedom of the usual t-statistic in this model would be $n-2$. If we only have two data points, can we test the null hypothesis $H_0 : \beta_1 = 1$? Of course not! We are not estimating (all information from the data is exhausted), i.e. we are **extrapolating**. Inference is impossible. If we have 3 data points, we can actually estimate $\hat{\beta}_1$ and H_0 becomes testable.
Rule of thumb: having a t-statistic with $|t| > 2$ often corresponds to something around 95% significance (in large-ish samples), but yes, this is a rule of thumb, not a law of physics.

Fun Facts 2.14 (Bezier curve (Splines))**.** A way to fit data points without interpolating them. For example, with 15 data points, you don't want to use 14-th order polynomials to fit it.

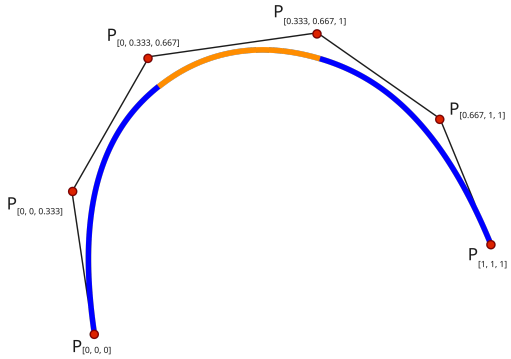

Figure 2: Splines

Now, there is another WLLN you will see constantly because it is "one line" once Chebyshev exists. Philosophically, it says: if the variance of the average goes to zero, then the average concentrates around its mean. (Yes, this is basically the same movie as before, with a slightly different cast.)

Theorem 2.15 (Chebyshev's WLLN)**.** Let $\{X_i\}_{i\geq 1}$ be independent with $\mathbb{E}[X_i] = \mu_i < \infty$ and $\text{Var}(X_i) = \sigma_i^2 < \infty$. Define

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad \bar{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} \mu_i.$$

If

$$\frac{1}{n^2}\sum_{i=1}^{n} \sigma_i^2 \to 0,$$

then

$$\bar{X}_n - \bar{\mu}_n \xrightarrow{p} 0.$$

Fun Facts 2.16. Note: writing "$\bar{X}_n \xrightarrow{p} \mu_n$" is not a standard convergence-in-probability statement if μ_n changes with n. The clean object is $\bar{X}_n - \bar{\mu}_n \xrightarrow{p} 0$ (as above). If additionally $\bar{\mu}_n \to \mu$, then $\bar{X}_n \xrightarrow{p} \mu$ by Slutsky/continuous mapping.

At this stage, a natural question is: what if we apply some function to our convergent sequence? In econometrics we do this all the time: we take square roots, inverses, ratios, plug-in variance estimates, logs, etc. So we need a theorem that says "convergence survives reasonable transformations." The following theorem shows that convergence in probability is preserved under continuous transformations.

Theorem 2.17 (Continuous Mapping Theorem (for convergence in probability))**.** If $g(\cdot)$ is continuous and $X_n \xrightarrow{p} c$, then

$$g(X_n) \xrightarrow{p} g(c).$$

2.6 Multivariate WLLN

Next we extend the discussion from univariate to multivariate. This is not optional: most econometrics objects are vectors and matrices, and pretending everything is scalar is a lifestyle choice, not a methodology.

Fun Facts 2.18. Antonio Lucio Vivaldi's works are labeled beginning with *R.V.* followed by a bunch of numbers.

Definition 2.5 (Convergence of random vectors/matrices in probability)**.** Let $\mathbf{X}_n \in \mathbb{R}^d$ and $\mathbf{c} \in \mathbb{R}^d$. We say $\mathbf{X}_n \xrightarrow{p} \mathbf{c}$ if for every $\varepsilon > 0$,

$$\mathbb{P}\big(\|\mathbf{X}_n - \mathbf{c}\| > \varepsilon\big) \to 0.$$

For random matrices of fixed dimension, the same definition applies using any matrix norm (equivalent in finite dimension).

You can think entrywise convergence as an equivalent way in fixed dimension (norm equivalence), but note the entries are not "independent objects"—they can be tightly linked. Example: $\mathbf{v}_n = (X_n, -X_n)^\top$ with $X_n \sim \mathcal{N}(0, 1/n)$. Then $X_n \xrightarrow{p} 0$ and hence $\mathbf{v}_n \xrightarrow{p} \mathbf{0}$, even though the coordinates are perfectly (negatively) dependent.

Finally, we want algebra rules. Because in practice we build estimators by adding, multiplying, dividing, and inverting things. If convergence is not stable under these operations, asymptotics would be a very short subject.

Theorem 2.19 (Algebra under $\xrightarrow{p}$ (a.k.a. "Slutsky for probability"))**.** If $X_n \xrightarrow{p} c$ and $Y_n \xrightarrow{p} d$, then

$$\begin{aligned} X_n + Y_n &\xrightarrow{p} c + d, \\ X_n Y_n &\xrightarrow{p} cd, \\ X_n / Y_n &\xrightarrow{p} c/d \qquad \text{provided } d \neq 0. \end{aligned}$$

Moreover, if $\mathbf{W}_n \xrightarrow{p} \Omega$ where Ω is nonsingular and $\mathbf{W}_n$ is nonsingular w.p. $\to 1$, then

$$\mathbf{W}_n^{-1} \xrightarrow{p} \Omega^{-1}, \qquad \mathbf{W}_n \mathbf{V}_n \xrightarrow{p} \Omega \mathbf{v} \;\text{ if }\; \mathbf{V}_n \xrightarrow{p} \mathbf{v}.$$

A sketch of proof: the mappings above are continuous (on the domain where they are defined), so Theorem 2.17 applies.

Fun Facts 2.20. About the t-statistic in the linear model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$: *exact* finite-sample t-distribution statements usually require normality of ϵ_i (plus the usual linear model conditions). But *asymptotically*, t-statistics often converge to $\mathcal{N}(0, 1)$ under much weaker conditions (CLT + Slutsky).

So much for convergence in probability. However, is convergence in probability alone sufficient for econometric practice? Here's the thing: convergence in probability is great for estimation (location), but useless for inference unless you also know the shape of fluctuations. To reject a null hypothesis, you need a limiting distribution, not just a limiting point.

WLLN tells you that with infinite sample size you end up somewhere, but it does not tell you how fast you get there, or what the error looks like for large but finite n. Hence we need **convergence in distribution**, and that's where the famous CLT(s?) comes in.

2.7 Convergence in Distribution (Weak Convergence)

What is convergence in distribution? Here's the definition:

Definition 2.6 (Convergence in distribution)**.** Given $X_n \sim F_n$ and $X \sim F$, we say X_n converges to X **in distribution** if

$$F_n(x) \to F(x) \qquad \text{for every continuity point } x \text{ of } F.$$

We denote this by $X_n \xrightarrow{d} X$.

Before going further, it helps to interpret what weak convergence is saying: it is convergence of probability laws, not convergence of outcomes. So it is weaker than almost sure or in-probability convergence, but it is exactly what we need for asymptotic inference.

A limiting distribution is described below.

Definition 2.7 (Limiting distribution & moments)**.** If $X_n \xrightarrow{d} X \sim F$, then F is the limiting distribution of X_n.
If F has moments, then the moments generated by F are called the limiting moments of X_n.

Fun Facts 2.21. Warning: $X_n \xrightarrow{d} X$ does *not* automatically imply $\mathbb{E}[X_n^k] \to \mathbb{E}[X^k]$. To pass moments through limits you typically need extra conditions (e.g. uniform integrability / dominated convergence style assumptions).

Here's an example:

Example 2.22. *If $X_n \sim t(n)$, then $X_n \xrightarrow{d} \mathcal{N}(0,1)$. The limiting expectation is 0 (since $\mathbb{E}[X_n] = 0$ for $n > 1$) and the variance is $\frac{n}{n-2}$ for $n > 2$, whose limit is 1.*

Boom! Slutsky again (now in the weak convergence setting).

Theorem 2.23 (Continuous mapping + Slutsky (weak convergence))**.** If $X_n \xrightarrow{d} X$ and $g(\cdot)$ is continuous, then $g(X_n) \xrightarrow{d} g(X)$.
Moreover, if $Y_n \xrightarrow{p} c$ where c is constant, then

$$X_n Y_n \xrightarrow{d} cX, \qquad X_n + Y_n \xrightarrow{d} X + c, \qquad \frac{X_n}{Y_n} \xrightarrow{d} \frac{X}{c} \ \text{ if } c \neq 0.$$

What you can read off Theorem 2.23 is as follow: you can replace unknown constants by consistent estimators inside limiting distributions. This is the legal justification for half of econometrics.

Notice: for $\frac{1}{X_n}$, the map $f(x) = 1/x$ is continuous on $\mathbb{R} \setminus \{0\}$. So the issue is not "$1/x$ is never continuous"; the issue is whether your sequence puts mass near 0 (or the limit puts

positive probability at 0), in which case continuity at the relevant points fails.

Another way to check weak convergence is via the Cramér–Wold device: if after any linear transformation the random vector still converges to the linearly transformed limit, then the whole vector converges. This is the multivariate version of "if all shadows match, the object matches."

Theorem 2.24 (Cramér–Wold device)**.** Let $\mathbf{X}_n$ be a random vector in $\mathbb{R}^d$ and $\mathbf{X}$ another random vector in $\mathbb{R}^d$. Then $\mathbf{X}_n \xrightarrow{d} \mathbf{X}$ if and only if

$$\theta^\top \mathbf{X}_n \xrightarrow{d} \theta^\top \mathbf{X} \qquad \text{for all } \theta \in \mathbb{R}^d.$$

The main reason we are introducing Cramé-Wold is because that marginal limits do not determine joint limits: knowing what happens to each coordinate separately does not tell you how they depend on each other. And dependence is basically the entire plot in multivariate probability.

Fun Facts 2.25 (Survivor's bias)**.** During WWII, US AAF (Army Air Force) conducted a survey on their planes. They found the planes were more likely to be hit on the wings and the back. It seems those places are more likely to be hit, so why don't we just enhance this part? It turned out the really weak part was the *unhit* part: planes hit there did not come back, so you never saw those bullet holes.

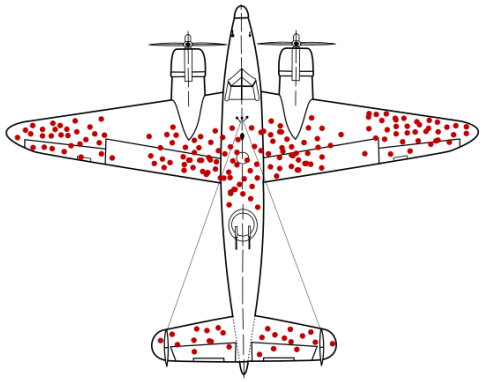

Figure 3: Diagram in which red dots stand for places where surviving planes were shot. This only tells you where planes can get shot and still come back to base. Survivorship bias: your only information is what has survived.

Nice! Now we have the tools we need and we are ready for CLT.

2.8 Central Limit Theorems

The CLT is the punchline for inference: after the right scaling, averages behave approximately normal. A mental picture is that the sums of many small-ish independent-ish things tend to forget their original shape.

Theorem 2.26 (Lindeberg–Lévy CLT (classical CLT))**.** If $\{X_i\}_{i\geq 1}$ are i.i.d. with $\mathbb{E}[X_1] = \mu$ and $\mathrm{Var}\,(X_1) = \sigma^2 < \infty$, and $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$, then

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2).$$

Classical CLT is the cleanest case: i.i.d. + finite variance. But econometrics rarely hands you i.i.d. on a silver platter. So the natural next question is: what if observations are not identically distributed? Or what if one observation has a giant variance and tries to dominate the sum like a main character?

Another CLT is named after Lindeberg and Feller. Here we drop i.i.d. and keep (roughly) independence + a condition preventing any single term from dominating the sum.

Theorem 2.27 (Lindeberg–Feller CLT (one standard version))**.** Let $\{X_i\}_{i\geq 1}$ be independent with $\mathbb{E}[X_i] = \mu_i$ and $\mathrm{Var}\,(X_i) = \sigma_i^2 < \infty$. Let

$$s_n^2 = \sum_{i=1}^n \sigma_i^2, \qquad s_n^2 \to \infty, \qquad S_n = \sum_{i=1}^n (X_i - \mu_i).$$

Assume the **Lindeberg condition**: for every $\varepsilon > 0$,

$$\frac{1}{s_n^2}\sum_{i=1}^n \mathbb{E}\Big[(X_i - \mu_i)^2 \mathbf{1}\big(|X_i - \mu_i| > \varepsilon s_n\big)\Big] \to 0.$$

Then

$$\frac{S_n}{s_n} \xrightarrow{d} \mathcal{N}(0, 1).$$

Equivalently,

$$\frac{\sum_{i=1}^n (X_i - \mu_i)}{\sqrt{\sum_{i=1}^n \sigma_i^2}} \xrightarrow{d} \mathcal{N}(0, 1).$$

It is mainly a "no freak accidents" rule. It forces the contribution of large deviations (relative to s_n) to become negligible. So the normalized sum behaves like a sum of many small contributions, which is exactly when the Gaussian approximation becomes believable.

Finally, the multivariate case. And again because in econometrics your estimator is usually a vector, and your asymptotic distribution is usually a multivariate normal with some covariance matrix you then spend weeks estimating. The last CLT we will discuss here is the multivariate case. The proof is sketched via Cramér–Wold.

Theorem 2.28 (Multivariate CLT (i.i.d. case))**.** Let $\mathbf{X}_1, \ldots, \mathbf{X}_n \in \mathbb{R}^d$ be i.i.d. with

$$\mathbb{E}[\mathbf{X}_1] = \mathbf{M}, \qquad \mathrm{Var}\,(\mathbf{X}_1) = \Sigma \ \text{ (finite covariance matrix).}$$

Let $\bar{\mathbf{X}}_n = \frac{1}{n}\sum_{i=1}^n \mathbf{X}_i$. Then

$$\sqrt{n}\big(\bar{\mathbf{X}}_n - \mathbf{M}\big) \xrightarrow{d} \mathcal{N}\big(\mathbf{0}, \Sigma\big).$$

Proof. Let

$$\mathbf{Z}_n := \sqrt{n}\big(\bar{\mathbf{X}}_n - \mathbf{M}\big) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\mathbf{X}_i - \mathbf{M}) \in \mathbb{R}^d.$$

By the Cramér–Wold device (Theorem 2.24), it suffices to show that for every fixed $\theta \in \mathbb{R}^d$,

$$\theta^\top \mathbf{Z}_n \xrightarrow{d} \theta^\top \mathbf{Z},$$

where $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \Sigma)$.

Fix θ. Define the scalar i.i.d. variables

$$Y_i := \theta^\top (\mathbf{X}_i - \mathbf{M}).$$

Then $\mathbb{E}[Y_i] = \theta^\top \mathbb{E}[\mathbf{X}_i - \mathbf{M}] = 0$, and

$$\mathrm{Var}(Y_i) = \mathrm{Var}\left(\theta^\top (\mathbf{X}_i - \mathbf{M})\right) = \theta^\top \Sigma \theta,$$

which is finite because Σ is finite.

Now note

$$\theta^\top \mathbf{Z}_n = \theta^\top \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\mathbf{X}_i - \mathbf{M}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \theta^\top (\mathbf{X}_i - \mathbf{M}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i.$$

By the univariate CLT (Theorem 2.26),

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i \xrightarrow{d} \mathcal{N}\big(0, \theta^\top \Sigma \theta\big).$$

On the other hand, if $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \Sigma)$, then $\theta^\top \mathbf{Z} \sim \mathcal{N}(0, \theta^\top \Sigma \theta)$. Therefore, for every θ,

$$\theta^\top \mathbf{Z}_n \xrightarrow{d} \theta^\top \mathbf{Z}.$$

By Cramér–Wold, this implies

$$\mathbf{Z}_n \xrightarrow{d} \mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \Sigma),$$

i.e. $\sqrt{n}(\bar{\mathbf{X}}_n - \mathbf{M}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Sigma)$. □

Fun Facts 2.29 (Triangular distribution)**.** Just so you know: $U := U_1 + U_2$ with $U_i \overset{i.i.d.}{\sim} \mathrm{Unif}[0, 1]$ follows a so-called triangular distribution because the pdf plot looks like a triangle. Try it yourself!

2.9 Summary

Let's take stock before we sprint into regression. Up to this point, we built a little asymptotic toolkit:

- **Convergence in probability** ($\xrightarrow{p}$): the "location" story. If your estimator converges in probability to the truth, you get consistency.

- **Convergence in distribution** ($\xrightarrow{d}$): the "shape" story. This is where inference lives: test statistics, confidence intervals, and all the things you pretend are exact in finite samples.
- **Continuous mapping + Slutsky**: the legal loopholes that let you transform convergences without re-proving everything from scratch. (Yes, this is basically why asymptotic theory is scalable.)
- **LLN and CLT**: the two workhorses. LLN explains why averages stabilize; CLT explains why we get normal approximations and can stop crying.

In econometrics, most estimators are not random variables but *vectors* (a whole stack of parameters), and their asymptotic distributions are typically multivariate normal with some covariance matrix that you then spend weeks estimating (and another week debugging why it's not PSD). So the workflow you should keep in mind is:

$$\textbf{estimator} \overset{\text{LLN}}{\Rightarrow} \textbf{consistency}, \textbf{estimator} \overset{\text{CLT + Slutsky}}{\Rightarrow} \textbf{asymptotic normality}.$$

Once you have those two, inference is basically: plug in an estimate of the asymptotic variance and hope the sample size is "large enough" (whatever that means this week).

OLS is the first place where all of this machinery becomes more than philosophical. As you might recall from your Econometric I, the OLS estimator is a clean algebraic object:

$$\hat{\beta} = (X^\top X)^{-1} X^\top y,$$

so it's a perfect testbed for continuous mapping + Slutsky. Moreover, Consistency becomes a question about whether sample moments converge:

$$\frac{1}{n} X^\top X \xrightarrow{p} Q \qquad \text{and} \qquad \frac{1}{n} X^\top \varepsilon \xrightarrow{p} 0,$$

which is literally LLN in matrix form. What's even better is that asymptotic normality becomes a CLT for sums like

$$\frac{1}{\sqrt{n}} X^\top \varepsilon = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} x_i \varepsilon_i,$$

and then Slutsky turns it into a distribution for $\sqrt{n}(\hat{\beta} - \beta)$. So yes, OLS will be the next victim.

As a TLDR, Asymptotics is our way of saying: *if n is large, the messy random things behave like deterministic limits plus a normal fluctuation.* OLS is where we can cash that check for the first time. So next up: we introduce the linear model, define OLS properly, and then redo the two-step asymptotic routine mentioned above. And yes, we will eventually talk about standard errors. No, they will not be as innocent as you want them to be.

3 Ordinary Least Squares

We now cash in the CLT we just earned and spend it on the most overused estimator in social sciences: *ordinary least squares*. It is the estimator you run first, defend later, and then spend the rest of the paper fixing with robust standard errors.

3.1 Setup and notation

Consider the linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon,$$

where $\mathbf{y} \in \mathbb{R}^n$, $\mathbf{X} \in \mathbb{R}^{n\times k}$, $\boldsymbol{\beta} \in \mathbb{R}^k$, and $\varepsilon \in \mathbb{R}^n$. Write the i-th row of $\mathbf{X}$ as $\mathbf{x}_i^\top \in \mathbb{R}^{1\times k}$ and the i-th component of $\mathbf{y}$ as y_i. Then the stacked model is equivalent to the observation-level model

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_i, \qquad i = 1, \dots, n.$$

Definition 3.1 (Data arrangement)**.** For n observations and k regressors,

- $\mathbf{y} \in \mathbb{R}^n$ is the response vector;
- $\mathbf{X} \in \mathbb{R}^{n\times k}$ is the regressor matrix;
- the i-th row of $\mathbf{X}$ is $\mathbf{x}_i^\top$ (a $1 \times k$ vector);
- the j-th column of $\mathbf{X}$ is the $n \times 1$ vector of regressor j across all observations.

Example 3.1 (A simple wage regression)**.** *Let $y_i = \ln(\mathit{Wage}_i)$ and regressors include an intercept, education, and experience:*

$$\mathbf{y} = \begin{bmatrix} \ln(\mathit{Wage}_1) \\ \vdots \\ \ln(\mathit{Wage}_n) \end{bmatrix}, \qquad \mathbf{X} = \begin{bmatrix} 1 & \mathit{Edu}_1 & \mathit{Exp}_1 \\ \vdots & \vdots & \vdots \\ 1 & \mathit{Edu}_n & \mathit{Exp}_n \end{bmatrix}.$$

Each row corresponds to one observation.

3.2 Design

The OLS idea is simple: choose $\boldsymbol{\beta}$ to make fitted values $\mathbf{X}\boldsymbol{\beta}$ as close to $\mathbf{y}$ as possible. We define the degree of 'closedness' using the sum of squared errors (SSE)

$$SSE(\boldsymbol{\beta}) := \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

The OLS estimator is

$$\hat{\boldsymbol{\beta}} := \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^k} SSE(\boldsymbol{\beta}).$$

Expand the quadratic:

$$SSE(\boldsymbol{\beta}) = \mathbf{y}^\top \mathbf{y} - 2\boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{y} + \boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{X}\boldsymbol{\beta}.$$

Differentiate w.r.t. $\boldsymbol{\beta}$:

$$\nabla_{\boldsymbol{\beta}} SSE(\boldsymbol{\beta}) = -2\mathbf{X}^\top \mathbf{y} + 2\mathbf{X}^\top \mathbf{X}\boldsymbol{\beta}.$$

The first-order condition gives

$$\mathbf{X}^\top \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^\top \mathbf{y}.$$

If $\mathbf{X}$ has full column rank (so $\mathbf{X}^\top \mathbf{X}$ is invertible), then

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}.$$

Fun Facts 3.2 (Useful matrix derivatives)**.** For conformable vectors/matrices (and symmetric A where needed),

- $\nabla_x (a^\top x) = a$;
- $\nabla_x (x^\top A x) = (A + A^\top)x$ (in particular, $= 2Ax$ if A is symmetric);
- $\nabla_x \|x\|^2 = 2x$.

3.3 Gauss–Markov assumptions

To talk about statistical properties of $\hat{\boldsymbol{\beta}}$, we need assumptions. Some assumptions buy *unbiasedness*. Others buy the *BLUE* badge. And a different set buys asymptotic normality.

(GM1) **Linearity.** $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon$.

(GM2) **Full column rank.** $\operatorname{rank}(\mathbf{X}) = k$ (so $\mathbf{X}^\top \mathbf{X}$ is invertible). This means the data has enough variation across rows.

(GM3) **Zero conditional mean.** $\mathbb{E}[\varepsilon \mid \mathbf{X}] = 0$.

(GM4) **Homoskedasticity.** $\operatorname{Var}(\varepsilon \mid \mathbf{X}) = \sigma^2 I_n$ for some $\sigma^2 \in (0, \infty)$.

A useful way to remember this:

- (GM1–GM3) $\Rightarrow$ OLS is **unbiased** (and typically consistent under LLN-type conditions).
- Add (GM4) $\Rightarrow$ OLS becomes **BLUE**.

3.4 Small-sample properties: unbiasedness and BLUE

3.4.1 Unbiasedness

Proposition 3.3 (Unbiasedness of OLS)**.** Assume (GM1)–(GM3). Then

$$\mathbb{E}\left[\hat{\boldsymbol{\beta}} \mid \mathbf{X}\right] = \boldsymbol{\beta}.$$

Proof. Using $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon$,

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{y} = \boldsymbol{\beta} + (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \varepsilon.$$

Taking conditional expectation and using $\mathbb{E}[\varepsilon \mid \mathbf{X}] = 0$ yields $\mathbb{E}\left[\hat{\boldsymbol{\beta}} \mid \mathbf{X}\right] = \boldsymbol{\beta}$. □

3.4.2 Variance and the classical formula

Definition 3.2 (Variance–covariance matrix)**.** For a random vector $\mathbf{z}$, define

$$\mathrm{Var}(\mathbf{z}) := \mathbb{E}\left[(\mathbf{z} - \mathbb{E}[\mathbf{z}])(\mathbf{z} - \mathbb{E}[\mathbf{z}])^\top\right].$$

Proposition 3.4 (Conditional variance of OLS)**.** Assume (GM1)–(GM3). Then

$$\mathrm{Var}\left(\hat{\boldsymbol{\beta}} \mid \mathbf{X}\right) = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathrm{Var}(\varepsilon \mid \mathbf{X})\ \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}.$$

If in addition (GM4) holds, $\mathrm{Var}(\varepsilon \mid \mathbf{X}) = \sigma^2 I_n$, then

$$\mathrm{Var}\left(\hat{\boldsymbol{\beta}} \mid \mathbf{X}\right) = \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}.$$

Proof. From $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \varepsilon$, apply $\mathrm{Var}(A\mathbf{z} \mid \mathbf{X}) = \mathbf{A}\mathrm{Var}(\mathbf{z} \mid \mathbf{X})\,\mathbf{A}^\top$ with $A = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top$. □

3.4.3 Gauss–Markov theorem (BLUE)

Theorem 3.5 (Gauss–Markov)**.** Assume (GM1)–(GM4). Among all *linear unbiased* estimators of $\boldsymbol{\beta}$ of the form $\tilde{\boldsymbol{\beta}} = C\mathbf{y}$ (where C may depend on $\mathbf{X}$ but not on $\mathbf{y}$), the OLS estimator $\hat{\boldsymbol{\beta}}$ has the smallest conditional variance in the Loewner order:

$$\mathrm{Var}\left(\tilde{\boldsymbol{\beta}} \mid \mathbf{X}\right) - \mathrm{Var}\left(\hat{\boldsymbol{\beta}} \mid \mathbf{X}\right) \succeq 0.$$

Equivalently, $\hat{\boldsymbol{\beta}}$ is **BLUE**.

Proof. Let $\tilde{\boldsymbol{\beta}} = \mathbf{C}\mathbf{y}$ be linear and unbiased. Unbiasedness means

$$\mathbb{E}\left[\tilde{\boldsymbol{\beta}} \mid \mathbf{X}\right] = \mathbf{C}\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta} \quad \Rightarrow \quad \mathbf{C}\mathbf{X} = \mathbf{I}_k.$$

Write

$$\mathbf{C} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top + \mathbf{D}, \qquad \text{where } \mathbf{D}\mathbf{X} = 0$$

which is always possible because $\mathbf{C}\mathbf{X} = \mathbf{I}_k$ and $(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{X} = \mathbf{I}_k$. Then

$$\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta} = \mathbf{C}\varepsilon = \big((\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top + \mathbf{D}\big)\varepsilon = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + \mathbf{D}\varepsilon.$$

Under (GM4), $\mathrm{Var}(\varepsilon \mid \mathbf{X}) = \sigma^2 \mathbf{I}_n$. Hence

$$\mathrm{Var}\left(\tilde{\boldsymbol{\beta}} \mid \mathbf{X}\right) = \sigma^2 \mathbf{C}\mathbf{C}^\top = \sigma^2 \Big((\mathbf{X}^\top \mathbf{X})^{-1} + \mathbf{D}\mathbf{D}^\top\Big),$$

because the cross terms vanish:

$$(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{D}^\top = (\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^\top\mathbf{D}^\top) = (\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{D}\mathbf{X})^\top = 0.$$

Since $\mathbf{D}\mathbf{D}^\top \succeq 0$, we conclude

$$\mathrm{Var}\left(\tilde{\boldsymbol{\beta}} \mid \mathbf{X}\right) - \mathrm{Var}\left(\hat{\boldsymbol{\beta}} \mid \mathbf{X}\right) = \sigma^2\mathbf{D}\mathbf{D}^\top \succeq 0.$$

□

3.5 Large-sample properties

Finite-sample results are comforting, but inference in econometrics lives on asymptotics. The two core questions:

- **Consistency:** does $\hat{\boldsymbol{\beta}} \to \boldsymbol{\beta}$ as $n \to \infty$?
- **Asymptotic normality:** does $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ converge in distribution to a normal law?

3.5.1 Consistency

Assume we have i.i.d. data $\{(\mathbf{x}_i, \varepsilon_i)\}_{i=1}^n$ with $\mathbb{E}[\varepsilon_i \mid \mathbf{x}_i] = 0$ and finite second moments. Rewrite

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\varepsilon = \Big(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\Big)^{-1}\Big(\frac{1}{n}\mathbf{X}^\top\varepsilon\Big) = \Big(\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top\Big)^{-1}\Big(\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\varepsilon_i\Big).$$

Theorem 3.6 (Consistency of OLS)**.** Suppose $\{(\mathbf{x}_i, \varepsilon_i)\}_{i=1}^n$ are i.i.d. and:

(C1) $\mathbb{E}\left[\|\mathbf{x}_i\|^2\right] < \infty$ and $Q := \mathbb{E}\left[\mathbf{x}_i\mathbf{x}_i^\top\right]$ is positive definite;

(C2) $\mathbb{E}[\varepsilon_i \mid \mathbf{x}_i] = 0$ and $\mathbb{E}\left[\varepsilon_i^2\right] < \infty$.

Then $\hat{\boldsymbol{\beta}} \xrightarrow{\mathbb{P}} \boldsymbol{\beta}$.

One-line proof. By LLN, $\frac{1}{n}\sum \mathbf{x}_i\mathbf{x}_i^\top \xrightarrow{\mathbb{P}} Q$ and $\frac{1}{n}\sum \mathbf{x}_i\varepsilon_i \xrightarrow{\mathbb{P}} \mathbb{E}[\mathbf{x}_i\varepsilon_i] = 0$. By the continuous mapping theorem, the inverse converges; multiplying gives $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \xrightarrow{\mathbb{P}} 0$. □

3.5.2 Asymptotic normality

Now we ask: how does $\hat{\boldsymbol{\beta}}$ fluctuate around $\boldsymbol{\beta}$ when n is large?

Theorem 3.7 (Asymptotic normality of OLS (i.i.d. case))**.** Suppose $\{(\mathbf{x}_i, \varepsilon_i)\}_{i=1}^n$ are i.i.d. and:

(11) $\mathbb{E}[\varepsilon_i \mid \mathbf{x}_i] = 0$;

(22) $\mathbb{E}\left[\varepsilon_i^2 \mid \mathbf{x}_i\right] < \infty$ and $\mathbb{E}\left[\|\mathbf{x}_i\|^2 \varepsilon_i^2\right] < \infty$;

(33) $\mathbf{Q} := \mathbb{E}\left[\mathbf{x}_i \mathbf{x}_i^\top\right]$ is positive definite;

(44) define $\Omega := \mathbb{E}\left[\varepsilon_i^2 \mathbf{x}_i \mathbf{x}_i^\top\right]$ (finite).

Then

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}(0, \mathbf{Q}^{-1}\Omega\mathbf{Q}^{-1}).$$

If additionally $\mathrm{Var}(\varepsilon_i \mid \mathbf{x}_i) = \sigma^2$ (homoskedasticity), then $\Omega = \sigma^2\mathbf{Q}$ and the limit simplifies to

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}(0, \sigma^2\mathbf{Q}^{-1}).$$

Proof. Start from

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \Big(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\Big)^{-1}\Big(\frac{1}{\sqrt{n}}\mathbf{X}^\top\varepsilon\Big) = \Big(\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top\Big)^{-1}\Big(\frac{1}{\sqrt{n}}\sum_{i=1}^n \mathbf{x}_i\varepsilon_i\Big).$$

By LLN, $\frac{1}{n}\sum \mathbf{x}_i\mathbf{x}_i^\top \xrightarrow{\mathbb{P}} Q$. By the multivariate CLT applied to $\{\mathbf{x}_i\varepsilon_i\}$ (mean zero by $\mathbb{E}[\varepsilon_i \mid \mathbf{x}_i] = 0$),

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n \mathbf{x}_i\varepsilon_i \xrightarrow{d} \mathcal{N}(0, \boldsymbol{\Omega}), \qquad \boldsymbol{\Omega} = \mathbb{E}\left[\varepsilon_i^2\mathbf{x}_i\mathbf{x}_i^\top\right].$$

Slutsky then yields

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}(0, \mathbf{Q}^{-1}\Omega\mathbf{Q}^{-1}).$$

Under homoskedasticity, $\boldsymbol{\Omega} = \mathbb{E}\left[(\mathbb{E}\left[\varepsilon_i^2 \mid \mathbf{x}_i\right])\mathbf{x}_i\mathbf{x}_i^\top\right] = \sigma^2\mathbf{Q}$, giving the simplified form. □

3.7 actually allows you to do lots of things. A point estimate $\hat{\boldsymbol{\beta}}$ is a number. Inference is the part where we admit we might be wrong *in a quantifiable way.*

Once we have

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \approx \mathcal{N}(0, \text{asymptotic VCV}),$$

we unlock:

- **Standard errors** (square roots of diagonal VCV entries),
- **Confidence intervals** (estimate ± critical value × SE),
- t**-tests** for single restrictions,
- F **/ Wald tests** for joint restrictions.

In other words: most empirical papers.

3.6 Summary

In the classical world, $\text{Var}(\varepsilon \mid \mathbf{X}) = \sigma^2 \mathbf{I}_n$ and everything is clean: OLS is unbiased, BLUE, and its variance formula is pretty enough to put on a mug.

In the real world, $\text{Var}(\varepsilon \mid \mathbf{X})$ is rarely a multiple of the identity. Heteroskedasticity, serial correlation, and clustering show up like uninvited guests: they do not ruin $\hat{\boldsymbol{\beta}}$ as an estimator of $\boldsymbol{\beta}$ (under exogeneity), but they *do* ruin the naive standard errors. So next, we build the toolbox for that: robust covariance estimators, GLS/FGLS ideas, and (eventually) the moment-based perspective that leads to GMM.

4 Generalized Least Squares

Up to now, we have been living in the comfortable world of the Gauss–Markov assumptions: errors are homoskedastic, uncorrelated, and everything falls neatly into place. But life is rarely that kind.

The basic OLS framework assumes spherical errors (homoskedasticity):

$$\operatorname{Var}(\varepsilon \mid \mathbf{X}) = \sigma^2 \mathbf{I}_n.$$

This is the magical identity matrix that guarantees OLS is BLUE. But what if reality disagrees? What if error variances grow with $\mathbf{x}_i$, or if today's error is correlated with yesterday's? In financial returns, medical data, and time series, such "violations" are the rule, not the exception.

That's where **GLS** (as a friend) comes in. Generalized least squares keeps the spirit of OLS (linear in $\boldsymbol{\beta}$) but relaxes the spherical-error assumption: we give up the simplicity of $\sigma^2 \mathbf{I}_n$ and replace it with a more general covariance matrix $\boldsymbol{\Omega}$:

$$\operatorname{Var}(\varepsilon \mid \mathbf{X}) = \boldsymbol{\Omega}.$$

This certainly complicates life. But it also makes this course far less boring. (And if you ever feel tempted to look down on yourself as a mere statistician or econometrician, remember: even physicists with their grand thermo-statistics occasionally rely on the same math.)

Once we leave the Gauss–Markov paradise, OLS is no longer efficient, and inference is trickier. But by exploiting special structures of $\boldsymbol{\Omega}$, we can design remedies: robust standard errors, weighted least squares, feasible GLS, clustered inference, and system GLS.

We begin our journey with the most common departure: **heteroskedasticity**.

4.1 Diagnosing Heteroskedasticity

The OLS estimator is BLUE under the classical Gauss–Markov assumptions, which include homoskedasticity and no serial correlation. But life is not always a bed of roses. Take, for instance, the annual rate of return of a financial asset. Intuition tells us that higher expected returns usually come with greater risk. In other words, the error variance is not constant:

$$\operatorname{Var}(\varepsilon_i \mid \mathbf{x}_i) = \sigma_i^2, \qquad \text{with } \sigma_i^2 \text{ varying in } \mathbf{x}_i.$$

This phenomenon is called **heteroskedasticity**. Put differently, the conditional variance is no longer flat—it bends and twists with the data.

Definition 4.1 (Heteroskedasticity)**.** If $\operatorname{Var}(\varepsilon_i \mid \mathbf{x}_i) = \sigma_i^2 = f(\mathbf{x}_i)$ and f is not a constant function of $\mathbf{x}_i$, then we refer to this variation in conditional variance as **heteroskedasticity**.

Under heteroskedasticity, OLS remains unbiased (under exogeneity), but it loses efficiency—a fact we already saw in the proof of Theorem 3.5. Still, before resigning ourselves to this bad news, it makes sense to check whether heteroskedasticity is actually present.

A natural (and slightly cheeky) idea is: run your OLS first, grab the squared residuals $\hat{\varepsilon}_i^2$, and then see if they depend on $\mathbf{x}_i$. In practice, this means regressing $\hat{\varepsilon}_i^2$ on some functions of regressors and asking whether those variables help predict the error variance. There are two classic ways to formalize this idea. We introduce the Breusch–Pagan test first.

Algorithm 1 Breusch–Pagan Test

Input: regressor matrix $\mathbf{X} \in \mathbb{R}^{n \times k}$ and response vector $\mathbf{y} \in \mathbb{R}^n$
Propose the null hypothesis H_0: $\{\varepsilon_i\}$ are homoskedastic.
Run OLS under $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon$. Obtain residuals $\hat{\varepsilon}_i$ for $i = 1, \ldots, n$.
Compute squared residuals $\hat{\varepsilon}_i^2$.
Choose auxiliary regressors $\mathbf{z}_i$ (often $\mathbf{z}_i = \mathbf{x}_i$ or $\mathbf{z}_i = (1, \mathbf{x}_i^\top)^\top$), and stack them into $\mathbf{Z}$.
Run the auxiliary regression: $\hat{\varepsilon}_i^2 = \alpha + \mathbf{z}_i^\top \boldsymbol{\gamma} + u_i$.
Obtain R^2 from the auxiliary regression.
Compute the LM statistic $LM = n \cdot R^2$.
Let q be the number of *slope* regressors in the auxiliary regression (excluding the intercept). Under H_0, $LM \sim \chi_q^2$.
if p-value $< \alpha$ **then**
    Reject H_0 (evidence of heteroskedasticity).
else
    Fail to reject H_0.
end if

The key philosophy of Breusch–Pagan is: if there is no heteroskedasticity, then the auxiliary regressors should have no explanatory power for the squared residuals. Equivalently, under H_0, the slope vector $\boldsymbol{\gamma}$ in

$$\hat{\varepsilon}_i^2 = \alpha + \mathbf{z}_i^\top \boldsymbol{\gamma} + u_i$$

should be "close to zero" in a joint sense. The LM statistic measures how far the score (at the null) wanders away from zero, and the chi-square distribution is the natural limit for such quadratic forms.

I should mention that the justifications below are optional. But the formula $LM = nR^2$ confused me when I was taking the course. Anyway, if anyone does feel the same way as I did, here's the clean (in a sense that you are familiar with Chapter 8) takeaway

Fun Facts 4.1. On why LM follows a χ^2_q
Consider the regression model

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_i, \qquad i = 1, \dots, n,$$

with null hypothesis

$$H_0 : \ \mathrm{Var}\,(\varepsilon_i \mid \mathbf{X}) = \sigma^2.$$

Under the alternative, model the variance as

$$\mathrm{Var}\,(\varepsilon_i \mid \mathbf{X}) = \sigma^2\, h(\mathbf{z}_i; \boldsymbol{\delta}), \qquad h(\mathbf{z}_i; \boldsymbol{\delta}) = 1 + \delta_1 z_{i1} + \cdots + \delta_q z_{iq},$$

so H_0 corresponds to $\boldsymbol{\delta} = \mathbf{0}$. Define the log-likelihood contribution

$$\ell_i(\boldsymbol{\delta}) := \log f(y_i \mid \mathbf{x}_i; \boldsymbol{\delta}), \qquad \ell(\boldsymbol{\delta}) := \sum_{i=1}^{n} \ell_i(\boldsymbol{\delta}),$$

and the score (gradient) w.r.t. $\boldsymbol{\delta}$:

$$\mathbf{S}(\boldsymbol{\delta}) := \frac{\partial \ell(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}}.$$

Under standard regularity conditions, under H_0 we have asymptotic normality:

$$\mathbf{S}(\mathbf{0}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{I}(\mathbf{0})),$$

where $\mathbf{I}(\mathbf{0})$ is the Fisher information at the null. The LM statistic is the quadratic form

$$LM := \mathbf{S}(\mathbf{0})^\top \mathbf{I}(\mathbf{0})^{-1} \mathbf{S}(\mathbf{0}),$$

hence

$$LM \xrightarrow{d} \chi^2_q.$$

Fun Facts 4.2. On why $LM = nR^2$ (in the BP construction)
Let $\hat{\varepsilon}_i$ be OLS residuals. In the BP setup, the score at the null is proportional to

$$\mathbf{S}(\mathbf{0}) \propto \sum_{i=1}^{n} \mathbf{z}_i\big(\hat{\varepsilon}_i^2 - \hat{\sigma}^2\big), \qquad \hat{\sigma}^2 := \frac{1}{n}\sum_{i=1}^{n} \hat{\varepsilon}_i^2.$$

The information at the null is proportional to

$$\mathbf{I}(\mathbf{0}) \propto \sum_{i=1}^{n} \mathbf{z}_i \mathbf{z}_i^\top = \mathbf{Z}^\top \mathbf{Z}.$$

Thus the LM statistic has the generic "score–information–score" shape:

$$LM \propto \Big(\mathbf{Z}^\top(\hat{\boldsymbol{\varepsilon}}^{\,2} - \hat{\sigma}^2\mathbf{1})\Big)^\top (\mathbf{Z}^\top\mathbf{Z})^{-1}\Big(\mathbf{Z}^\top(\hat{\boldsymbol{\varepsilon}}^{\,2} - \hat{\sigma}^2\mathbf{1})\Big),$$

where $\hat{\boldsymbol{\varepsilon}}^{\,2} := (\hat{\varepsilon}_1^2, \ldots, \hat{\varepsilon}_n^2)^\top$. Now consider the auxiliary regression

$$\hat{\varepsilon}_i^2 = \alpha + \mathbf{z}_i^\top \boldsymbol{\gamma} + u_i.$$

The explained sum of squares can be written as

$$ESS = (\hat{\boldsymbol{\varepsilon}}^{\,2} - \overline{\hat{\varepsilon}^2}\mathbf{1})^\top \mathbf{P}_{\mathbf{Z}} (\hat{\boldsymbol{\varepsilon}}^{\,2} - \overline{\hat{\varepsilon}^2}\mathbf{1}), \qquad \mathbf{P}_{\mathbf{Z}} := \mathbf{Z}(\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top.$$

Up to scaling conventions, this is the same quadratic form as the numerator in LM. Since $R^2 = ESS/TSS$ and TSS is proportional to $n\hat{\sigma}^4$ under H_0, one obtains the classic simplification

$$LM = nR^2.$$

The philosophy of BP stands naturally: how significant is the dependence between the squared residuals and some auxiliary regressors $\mathbf{z}_i$ under a linear auxiliary model? Another idea would be: instead of claiming the variance is "linear-ish" in regressors, we suspect it responds to richer nonlinear terms. Then the auxiliary regression changes. We now introduce the White test.

Algorithm 2 White Test

1: **Input**: regressor matrix $\mathbf{X} \in \mathbb{R}^{n\times k}$ and response vector $\mathbf{y} \in \mathbb{R}^n$
2: Null hypothesis H_0: $\{\varepsilon_i\}$ are homoskedastic.
3: Run OLS under $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon$. Obtain residuals $\hat{\varepsilon}_i$.
4: Form auxiliary regressors $\mathbf{z}_i$ that include (some of) the following: original regressors, squares, and cross-products.
5: Run auxiliary regression: $\hat{\varepsilon}_i^2 = \alpha + \mathbf{z}_i^\top\boldsymbol{\gamma} + u_i$.
6: Compute $LM = nR^2$ from the auxiliary regression.
7: Let q be the number of auxiliary slope regressors (excluding intercept). Under H_0, $LM \sim \chi_q^2$.

In practice, we sometimes include only squared terms and omit cross products; then q

changes accordingly. (And yes, the test gets less general when you do that—but your finite sample might thank you.)

4.2 Remedies

So suppose you run the tests, and—unfortunately—they all shout back at you: *heteroskedasticity is here!* What now?

We know OLS remains unbiased, but we've lost efficiency: the usual variance formula $\sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}$ is no longer valid, and we don't know the true $\mathbf{\Omega}$. Still, inference requires variances: without them, we can't construct standard errors, confidence intervals, or tests.

The key idea is therefore: even if we can't recover $\mathbf{\Omega}$ exactly, can we estimate $\text{Var}\left(\hat{\boldsymbol{\beta}} \mid \mathbf{X}\right)$ consistently? If so, we can rescue inference by building robust standard errors (and sometimes regain efficiency through GLS/WLS). That's what we turn to next.

4.2.1 White's Robust Standard Errors

The idea behind White's proposal is simple but powerful: if heteroskedasticity makes the usual variance formula invalid, estimate the variance directly from the data. This gives us robust standard errors without specifying the form of $\mathbf{\Omega}$.

Recall:

$$\text{Var}\left(\hat{\boldsymbol{\beta}} \mid \mathbf{X}\right) = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{\Omega}\,\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}, \qquad \mathbf{\Omega} = \text{Var}\left(\varepsilon \mid \mathbf{X}\right).$$

let's first assume we have heteroskedasticity with conditionally uncorrelated errors (given $\mathbf{X}$). A common structure is $\mathbf{\Omega} = \text{diag}(\sigma_1^2, \ldots, \sigma_n^2)$. Define the population target

$$\mathbf{S} := \mathbb{E}\left[\varepsilon_i^2\,\mathbf{x}_i\mathbf{x}_i^\top\right],$$

and its sample analogue

$$\hat{\mathbf{S}}_n := \frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^2\,\mathbf{x}_i\mathbf{x}_i^\top \quad \xrightarrow{p} \quad \mathbf{S}.$$

Plugging in yields the (HC0) robust covariance estimator:

$$\widehat{\text{Var}\left(\hat{\boldsymbol{\beta}}\right)} = (\mathbf{X}^\top\mathbf{X})^{-1}\Big(\sum_{i=1}^{n}\hat{\varepsilon}_i^2\,\mathbf{x}_i\mathbf{x}_i^\top\Big)(\mathbf{X}^\top\mathbf{X})^{-1}.$$

The diagonal entries (after square roots) are White's robust standard errors.

Notice: we never assumed a functional form for σ_i^2 as a function of regressors. The price is that this is an asymptotic result: in small samples, it can be noisy. Robust S.E. save inference, not efficiency.

4.2.2 Weighted Least Squares (WLS)

Another approach is to re-cast the problem into *OLS on transformed data*. Suppose the error variance is heteroskedastic but structured:

$$\text{Var}\left(\varepsilon_i \mid \mathbf{x}_i\right) = \sigma_i^2 = \sigma^2 f(\mathbf{x}_i),$$

where $f(\mathbf{x}_i) > 0$ is known (or well-estimated).

Divide the regression equation

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_i$$

by $\sqrt{f(\mathbf{x}_i)}$ and define

$$y_i^* := \frac{y_i}{\sqrt{f(\mathbf{x}_i)}}, \quad \mathbf{x}_i^* := \frac{\mathbf{x}_i}{\sqrt{f(\mathbf{x}_i)}}, \quad \varepsilon_i^* := \frac{\varepsilon_i}{\sqrt{f(\mathbf{x}_i)}}.$$

Then

$$y_i^* = \mathbf{x}_i^{*\top} \boldsymbol{\beta} + \varepsilon_i^*, \qquad \operatorname{Var}(\varepsilon_i^* \mid \mathbf{x}_i^*) = \sigma^2,$$

which is now homoskedastic.

Stacking the data, the WLS estimator is

$$\hat{\boldsymbol{\beta}}_{WLS} = (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W} \mathbf{y}, \qquad \mathbf{W} = \operatorname{diag}\Big(\frac{1}{f(\mathbf{x}_1)}, \ldots, \frac{1}{f(\mathbf{x}_n)}\Big).$$

By construction, WLS is efficient *if* $f(\cdot)$ is correctly specified. If not, it can lose efficiency and distort inference. Two cents for free: Assumption-heavy methods always look great right before assumptions betray you.

4.3 Generalized Least Squares (GLS)

We now relax the "no serial correlation" assumption, allowing

$$\operatorname{Var}(\varepsilon \mid \mathbf{X}) = \sigma^2 \boldsymbol{\Psi},$$

where $\boldsymbol{\Psi}$ is symmetric positive definite.

If $\boldsymbol{\Psi}$ were known (up to the scale σ^2), we could transform the model by premultiplying both sides with $\boldsymbol{\Psi}^{-1/2}$:

$$\mathbf{y}^* := \boldsymbol{\Psi}^{-1/2} \mathbf{y}, \quad \mathbf{X}^* := \boldsymbol{\Psi}^{-1/2} \mathbf{X}, \quad \varepsilon^* := \boldsymbol{\Psi}^{-1/2} \varepsilon.$$

Then

$$\mathbf{y}^* = \mathbf{X}^* \boldsymbol{\beta} + \varepsilon^*, \qquad \operatorname{Var}(\varepsilon^* \mid \mathbf{X}) = \sigma^2 \mathbf{I}_n,$$

so OLS on the transformed system yields the GLS estimator:

$$\hat{\boldsymbol{\beta}}_{GLS} = (\mathbf{X}^\top \boldsymbol{\Psi}^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{\Psi}^{-1} \mathbf{y}.$$

If $\mathbb{E}[\varepsilon \mid \mathbf{X}] = \mathbf{0}$ and $\boldsymbol{\Psi}$ is known (up to scale), GLS is BLUE among linear unbiased estimators. However, GLS is rarely feasible in practice because it requires knowing $\boldsymbol{\Psi}$. Estimating an unrestricted $\boldsymbol{\Psi}$ has $\frac{n(n+1)}{2} = O(n^2)$ free parameters, which quickly exhausts degrees of freedom. Thus, GLS is theoretically beautiful, but practically infeasible without additional structure.

Two pragmatic responses: (i) abandon full efficiency and aim for valid inference via robust covariance estimation; (ii) impose structure on $\boldsymbol{\Psi}$ and estimate it, leading to feasible GLS (FGLS).

4.3.1 Cluster-Robust Standard Errors

Since $\mathbf{\Psi}$ is typically unknown and too large to estimate directly, a practical alternative is to use cluster-robust standard errors. The sample is partitioned into G clusters, within which errors may be arbitrarily correlated (both heteroskedastic and serially dependent), while across clusters they are assumed independent.

Let $g = 1, \ldots, G$ index clusters, with data $(\mathbf{y}_g, \mathbf{X}_g)$. The OLS estimator remains

$$\hat{\boldsymbol{\beta}}_{OLS} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y},$$

but its sampling variance can be consistently estimated by

$$\widehat{\mathrm{Var}}\left(\hat{\boldsymbol{\beta}}_{OLS}\right)_{cluster} = (\mathbf{X}^\top \mathbf{X})^{-1} \left(\sum_{g=1}^{G} \mathbf{X}_g^\top \hat{\boldsymbol{\varepsilon}}_g \hat{\boldsymbol{\varepsilon}}_g^\top \mathbf{X}_g \right) (\mathbf{X}^\top \mathbf{X})^{-1},$$

where $\hat{\boldsymbol{\varepsilon}}_g$ denotes the vector of OLS residuals for cluster g.

Within each cluster we permit arbitrary error dependence, but across clusters we rely on independence so that cross terms vanish. As the number of clusters grows, this estimator remains consistent even in the presence of heteroskedasticity and serial correlation. If it helps, one may view the implied covariance matrix as block diagonal after rearranging the observations by cluster.

4.3.2 Feasible GLS (FGLS)

While GLS is efficient under correct $\mathbf{\Psi}$, it requires knowing $\mathbf{\Psi}$. Feasible GLS (FGLS) addresses this by first estimating $\mathbf{\Psi}$ under a parsimonious parametric structure, then applying GLS using $\hat{\mathbf{\Psi}}$.

For instance, suppose errors follow a stationary AR(1) correlation:

$$\mathbf{\Psi} = \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \cdots & \rho^{n-2} \\ \rho^2 & \rho & 1 & \cdots & \rho^{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \cdots & 1 \end{bmatrix}, \qquad \mathbf{\Psi}_{ij} = \rho^{|i-j|}, \quad |\rho| < 1.$$

Estimate $\hat{\rho}$ (e.g. from residual autocorrelation), plug into $\hat{\mathbf{\Psi}}$, and compute

$$\hat{\boldsymbol{\beta}}_{FGLS} = (\mathbf{X}^\top \hat{\mathbf{\Psi}}^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \hat{\mathbf{\Psi}}^{-1} \mathbf{y}.$$

Under correct specification, FGLS is asymptotically more efficient than OLS. If the parametric structure is wrong, inference can be unreliable. Here you see the trade-off between simplicity and practicality again. Cluster methods are assumption-light (consistent under broad misspecification) but may sacrifice efficiency, while FGLS is assumption-heavy.

4.3.3 Seemingly Unrelated Regression Equations (SURE)

Up to now, our discussion of GLS has focused on a single regression equation. But in many applications, analyzing one equation in isolation is not enough. Economic relationships

are intertwined: shocks in one market spill into others, household decisions co-move, and firm outcomes respond to common shocks. If we estimate equations separately, we ignore cross-equation correlation and lose efficiency.

This motivates **Seemingly Unrelated Regression Equations (SURE)**. The name is a bit of a joke: the equations may look "unrelated" (different regressors), but their errors are linked contemporaneously.

Notation switch (for this subsection). We use N observations (indexed by i) and M equations (indexed by m).

Fun Facts 4.3 (Almost Ideal Demand System)**.** The AIDS model (yes, economists are that crazy about acronyms) describes household budget shares of different goods:

$$w_{gi} = \alpha_g + \sum_h \gamma_{gh} \ln p_{hi} + \beta_g \ln\left(\frac{X_i}{P_i}\right) + \varepsilon_{gi}.$$

A shock to household income or preferences shifts spending across multiple goods, so errors $\{\varepsilon_{gi}\}$ are correlated across g. Estimating each share equation separately by OLS ignores this correlation, while SURE exploits it to gain efficiency.

To analyze such cases systematically, we write the system in matrix form.

Definition 4.2 (System of Equations in Matrix Form)**.** Consider M equations observed for N individuals:

$$y_{mi} = \mathbf{x}_{mi}^\top \boldsymbol{\beta}_m + \varepsilon_{mi}, \qquad m = 1, \dots, M, \;\; i = 1, \dots, N.$$

For each m, stack outcomes and regressors:

$$\mathbf{y}_m = \begin{bmatrix} y_{m1} \\ \vdots \\ y_{mN} \end{bmatrix}, \quad \mathbf{X}_m = \begin{bmatrix} \mathbf{x}_{m1}^\top \\ \vdots \\ \mathbf{x}_{mN}^\top \end{bmatrix}, \quad \boldsymbol{\varepsilon}_m = \begin{bmatrix} \varepsilon_{m1} \\ \vdots \\ \varepsilon_{mN} \end{bmatrix}.$$

Stack across equations:

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_M \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{X}_M \end{bmatrix}, \quad \varepsilon = \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \vdots \\ \boldsymbol{\varepsilon}_M \end{bmatrix}.$$

In compact form:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon, \qquad \boldsymbol{\beta} := (\boldsymbol{\beta}_1^\top, \dots, \boldsymbol{\beta}_M^\top)^\top.$$

We assume, for each individual $i = 1, \dots, N$:

(S1) **Linearity.** $y_{mi} = \mathbf{x}_{mi}^\top \boldsymbol{\beta}_m + \varepsilon_{mi}$, with $\mathbf{x}_{mi}$ non-stochastic or strictly exogenous.

(S2) **Independence across individuals.** $\{(\varepsilon_{1i}, \ldots, \varepsilon_{Mi})\}$ are independent across i, but may be correlated across m.

(S3) **Contemporaneous correlation.** Let $\boldsymbol{\varepsilon}_i := (\varepsilon_{1i}, \ldots, \varepsilon_{Mi})^\top$. Then

$$\mathbb{E}\left[\boldsymbol{\varepsilon}_i \mid \mathbf{X}\right] = \mathbf{0}, \qquad \mathrm{Var}\left(\boldsymbol{\varepsilon}_i \mid \mathbf{X}\right) = \boldsymbol{\Sigma},$$

where $\boldsymbol{\Sigma}$ is $M \times M$ positive definite.

(S4) **Homoskedasticity across individuals.** The same $\boldsymbol{\Sigma}$ applies to all i.

Since we are stacking matrices and vectors, you'll find the following algebraic clarification very helpful.

Definition 4.3 (Kronecker Product)**.** For matrices $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $\mathbf{B} \in \mathbb{R}^{p \times q}$, the Kronecker product is

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} & \cdots & a_{1n}\mathbf{B} \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} & \cdots & a_{2n}\mathbf{B} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}\mathbf{B} & a_{m2}\mathbf{B} & \cdots & a_{mn}\mathbf{B} \end{bmatrix}.$$

Given this definition, writing the covariance matrix of the new error vector is much easier.

Proposition 4.4 (Covariance Structure of System Errors)**.** Let $\varepsilon = (\boldsymbol{\varepsilon}_1^\top, \ldots, \boldsymbol{\varepsilon}_M^\top)^\top$ be the stacked system error. Under (S2)–(S4),

$$\mathrm{Var}\left(\varepsilon \mid \mathbf{X}\right) = \mathbf{I}_N \otimes \boldsymbol{\Sigma}.$$

Let

$$\boldsymbol{\Omega}_{SURE} := \mathbf{I}_N \otimes \boldsymbol{\Sigma}.$$

Premultiplying by $\boldsymbol{\Omega}_{SURE}^{-1/2}$ transforms the system into one with spherical errors:

$$\boldsymbol{\Omega}_{SURE}^{-1/2}\mathbf{y} = \boldsymbol{\Omega}_{SURE}^{-1/2}\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\Omega}_{SURE}^{-1/2}\varepsilon, \qquad \mathrm{Var}\left(\boldsymbol{\Omega}_{SURE}^{-1/2}\varepsilon \mid \mathbf{X}\right) = \mathbf{I}_{NM}.$$

This suggests the system GLS estimator:

$$\hat{\boldsymbol{\beta}}_{GLS} = (\mathbf{X}^\top \boldsymbol{\Omega}_{SURE}^{-1}\mathbf{X})^{-1}\mathbf{X}^\top \boldsymbol{\Omega}_{SURE}^{-1}\mathbf{y}.$$

Of course, $\boldsymbol{\Sigma}$ (hence $\boldsymbol{\Omega}_{SURE}$) is unknown. This leads to **feasible SURE**: estimate $\boldsymbol{\Sigma}$ from equation-by-equation OLS residuals, then plug in.

Run OLS separately for each equation $m = 1, \ldots, M$:

$$\hat{\boldsymbol{\beta}}_m^{OLS} = (\mathbf{X}_m^\top \mathbf{X}_m)^{-1}\mathbf{X}_m^\top \mathbf{y}_m, \qquad \hat{\boldsymbol{\varepsilon}}_m = \mathbf{y}_m - \mathbf{X}_m \hat{\boldsymbol{\beta}}_m^{OLS}.$$

For each observation i, stack residuals:

$$\hat{\boldsymbol{\varepsilon}}_i = \begin{bmatrix} \hat{\varepsilon}_{1i} \\ \vdots \\ \hat{\varepsilon}_{Mi} \end{bmatrix}.$$

Estimate the contemporaneous covariance:

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{N}\sum_{i=1}^{N} \hat{\boldsymbol{\varepsilon}}_i \hat{\boldsymbol{\varepsilon}}_i^\top, \qquad \hat{\sigma}_{m\ell} = \frac{1}{N}\sum_{i=1}^{N} \hat{\varepsilon}_{mi}\hat{\varepsilon}_{\ell i}.$$

Construct

$$\hat{\boldsymbol{\Omega}}_{SURE} := \mathbf{I}_N \otimes \hat{\boldsymbol{\Sigma}},$$

and plug into GLS:

$$\hat{\boldsymbol{\beta}}_{SURE} = (\mathbf{X}^\top \hat{\boldsymbol{\Omega}}_{SURE}^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \hat{\boldsymbol{\Omega}}_{SURE}^{-1} \mathbf{y}.$$

Algorithm 3 Feasible GLS Estimation for SURE

1: **Input:** system $\{(\mathbf{y}_m, \mathbf{X}_m)\}_{m=1}^M$ with N observations each.
2: For each m, run OLS and compute residuals $\hat{\boldsymbol{\varepsilon}}_m = \mathbf{y}_m - \mathbf{X}_m \hat{\boldsymbol{\beta}}_m^{OLS}$.
3: Compute $\hat{\boldsymbol{\Sigma}} = \frac{1}{N}\sum_{i=1}^{N} \hat{\boldsymbol{\varepsilon}}_i \hat{\boldsymbol{\varepsilon}}_i^\top$.
4: Form $\hat{\boldsymbol{\Omega}}_{SURE} = \mathbf{I}_N \otimes \hat{\boldsymbol{\Sigma}}$.
5: Output $\hat{\boldsymbol{\beta}}_{SURE} = (\mathbf{X}^\top \hat{\boldsymbol{\Omega}}_{SURE}^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \hat{\boldsymbol{\Omega}}_{SURE}^{-1} \mathbf{y}$.

Fun Facts 4.5. If regressors are identical across equations ($\mathbf{X}_1 = \cdots = \mathbf{X}_M$), then SURE coincides with equation-by-equation OLS (no efficiency gain from cross-equation correlation).

Fun Facts 4.6. Fun fact: Zellner (1962), who introduced SURE, originally described it as a "generalization of Aitken's GLS to several equations." Today, SURE is standard in demand systems, finance, and macroeconomics, but the acronym remains one of the most ironic in econometrics.

4.4 Summary

Once $\mathrm{Var}(\varepsilon \mid \mathbf{X})$ stops being $\sigma^2 \mathbf{I}_n$, OLS does not collapse—it just loses its main selling point: efficiency, and the default standard errors become an act of faith.

There are two broad responses. If you only want inference to be valid, use robust/cluster covariance estimators. If you also want efficiency, you need to model $\boldsymbol{\Omega}$ (WLS, GLS, FGLS, SURE), and then live with the consequences if that model is wrong.

Next we switch to outliers. Covariance misspecification is one way regression misbehaves; a few extreme points is another, and they can break your conclusions even when you did everything "right" above.

5 Outliers

Outliers are observations that do not follow the main pattern in the data. They matter because least squares is, by design, *easy to bully*: a small number of extreme points can materially change the fitted model and the reported standard errors.

Two clarifications up front:

- "Outlier" is not a single phenomenon. There are *response outliers* (unusual y_i given $\mathbf{x}_i$) and *design outliers* (unusual $\mathbf{x}_i$), the latter often showing up as *high leverage*.
- Normality is *not* part of the Gauss–Markov conditions. Gauss–Markov gives unbiasedness and (under homoskedasticity) efficiency among linear unbiased estimators. Normality mainly buys you clean *finite-sample* t/F distributions; without it, large-sample approximations still often work.

5.1 Detecting Outliers

A first pass is purely visual: look at residual plots and histograms. If the residual distribution is strongly skewed or heavy-tailed, it is a warning sign that a small number of observations may be driving the fit (or that the model is simply misspecified).

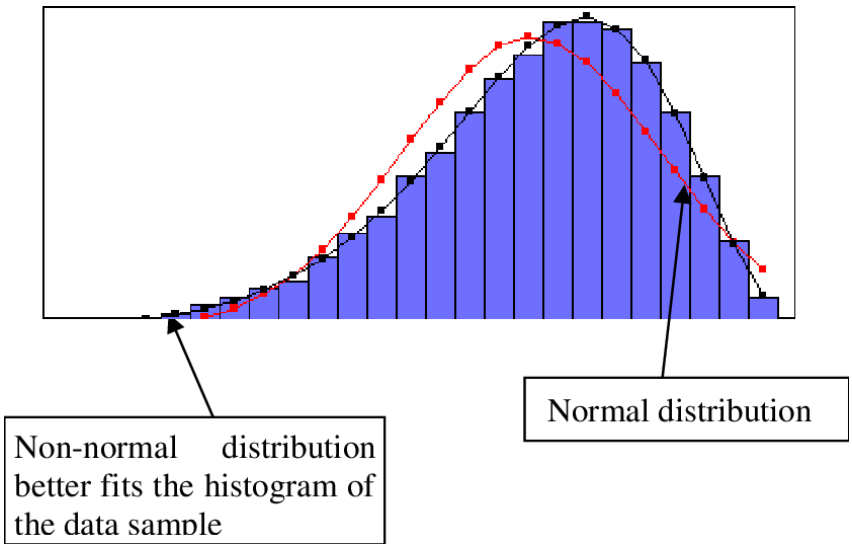

Figure 4: Skewed residual distribution: a visual warning sign

A more informative diagnostic is the **QQ–plot** (Quantile–Quantile plot). The idea is simple: if the sample comes from a reference distribution (say, Normal), then sample quantiles should line up with the corresponding theoretical quantiles.

Concretely:

1. Sort the sample: $y_{(1)} \le \cdots \le y_{(n)}$.
2. For each rank i, compute the theoretical quantile q_i of the reference distribution. For Normal, $q_i = \Phi^{-1}\big(\frac{i-0.5}{n}\big)$.
3. Plot the pairs $(q_i,\, y_{(i)})$ (or $(q_i,\, \hat{\varepsilon}_{(i)})$ if you are QQ-plotting residuals).

If the data follow the reference distribution, the points lie roughly on a straight line. A slope different from 1 typically indicates a scale mismatch (variance differs from the reference), not necessarily non-normality. Departures from linearity are more diagnostic:

- **Heavy tails:** points curve away from the line at both ends.
- **Skewness:** points bend upward/downward systematically.
- **Outliers:** a few points sit far off the line.

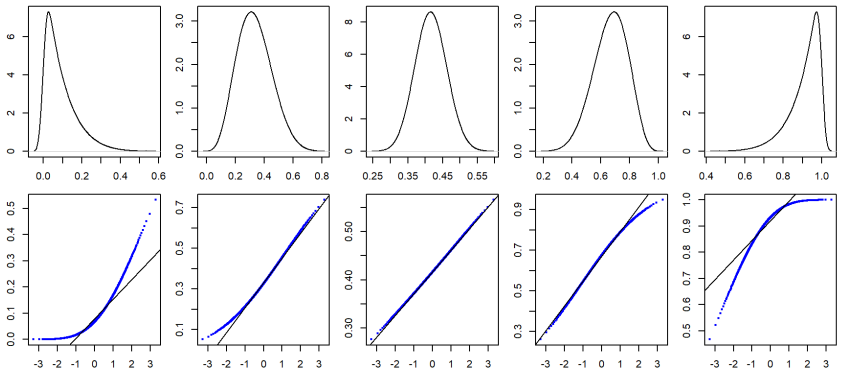

Figure 5: QQ plots: typical departures from normality

If you insist on a formal test of normality, there are many. The **Shapiro–Wilk** test is common in statistics (less common in econometrics lectures), and the **Jarque–Bera** test is a workhorse in econometrics.

Algorithm 4 Shapiro–Wilk Test for Normality

1: **Input:** sample $y_1, \ldots, y_n$.
2: Sort the data: $y_{(1)} \leq \cdots \leq y_{(n)}$.
3: Compute expected normal order statistics $m_i = \Phi^{-1}\left(\frac{i-0.375}{n+0.25}\right)$.
4: Obtain weights a_i (pre-tabulated / computed by software).
5: Compute

$$W = \frac{\left(\sum_{i=1}^{n} a_i y_{(i)}\right)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}.$$

6: Under H_0 (normality), W tends to be close to 1; small W indicates departure.
7: Use software to compute the p–value; reject if $p < \alpha$.

The **Jarque–Bera (JB)** test targets two specific normality implications: skewness 0 and kurtosis 3.

Algorithm 5 Jarque–Bera Test for Normality

1: **Input:** sample $y_1, \dots, y_n$.
2: Compute $\bar{y}$ and $s^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$.
3: Compute sample skewness

$$S = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^3}{s^{3/2}}.$$

4: Compute sample kurtosis

$$K = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^4}{s^2}.$$

5: Form

$$JB = \frac{n}{6}\left(S^2 + \tfrac{1}{4}(K-3)^2\right).$$

6: Under H_0 (normality), $JB \xrightarrow{d} \chi^2(2)$.
7: Reject if $p < \alpha$ (or $JB > \chi^2_{2,1-\alpha}$).

A practical reminder: rejecting normality does not automatically mean "outliers". It can also mean model misspecification, heteroskedasticity, or dependence.

5.2 Leverage and Influence

So far we focused on the *distribution of residuals*. But an observation can have a perfectly ordinary residual and still matter a lot if it sits in an unusual location in regressor space. This brings us to **leverage** and **influence**.

Definition 5.1 (Leverage)**.** Recall the OLS fitted values. Define $\mathbf{H}$ such that:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} = \mathbf{H}\mathbf{y}, \qquad \mathbf{H} := \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top.$$

The **leverage** of observation i is the ith diagonal element of the hat matrix $\mathbf{H}$:

$$h_{ii} := \mathbf{H}_{ii}.$$

It measures how sensitive the fitted value $\hat{y}_i$ is to the observed response vector $\mathbf{y}$. Observations with unusually large h_{ii} are called **high-leverage** points.

Since $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$, we have

$$\hat{y}_i = \sum_{j=1}^{n} h_{ij} y_j,$$

so h_{ii} quantifies how much observation i "helps determine" its own fitted value. One always has $0 \le h_{ii} \le 1$ and $\sum_{i=1}^{n} h_{ii} = \operatorname{tr}(\mathbf{H}) = k$.

Closed form in simple regression. In the univariate regression with intercept,

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

the leverage admits the closed form

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}.$$

Points far from $\bar{x}$ therefore have high leverage, even if their residuals are not dramatic.

Definition 5.2 (Cook's distance)**.** The **Cook's distance** of observation i is the influence diagnostic

$$D_i = \frac{1}{k\,\hat{\sigma}^2}\,\hat{\varepsilon}_i^2 \cdot \frac{h_{ii}}{(1-h_{ii})^2},$$

where k is the number of regressors in $\mathbf{X}$ (including the intercept), $\hat{\varepsilon}_i$ is the OLS residual, $\hat{\sigma}^2$ is the usual OLS residual variance estimate, and

$$h_{ii} = \mathbf{x}_i^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_i.$$

Large D_i indicates that dropping observation i would noticeably change the fitted model.

In practice, you typically look at a small set of diagnostics together: *(i)* large (studentized) residuals $\Rightarrow$ response outliers, *(ii)* large leverage $h_{ii} \Rightarrow$ design outliers, *(iii)* large Cook's distance $D_i \Rightarrow$ influential points.

5.3 Remedies

What to do with outliers depends on why they are there.

- **Fix obvious mistakes.** If an outlier is due to data-entry error or a clear measurement failure, correct it if possible; otherwise exclude it with documentation.
- **Winsorizing / trimming (with caution).** Replacing extreme values by nearby quantiles can reduce sensitivity, but it also changes the estimand. Use it when it matches the scientific question, not as a cosmetic filter.
- **Robust regression.** Methods such as Least Absolute Deviations (LAD) and quantile regression reduce the influence of extreme observations by changing the loss function (from squared loss to absolute loss / check loss).
- **Model the mechanism.** Sometimes "outliers" are the phenomenon (crashes, rare events, tail risk). Then the right response is not to delete them but to use a model that takes tails seriously.

The mature version of outlier handling is boring: inspect, justify, document, and check robustness. The immature version is even more boring: deleting points until the p–values behave.

Next we turn to **endogeneity**. While outliers mostly mess with the precision and stability of our estimates, endogeneity strikes at the heart of unbiasedness, and is often the most serious obstacle in empirical research. In other words, outliers can make you look sloppy, but endogeneity can make you confidently wrong.

6 Endogeneity

So far, OLS has served us faithfully. Under the Gauss–Markov assumptions, it is unbiased, consistent, and even BLUE. But recall that one crucial assumption was hidden inside GM4:

$$\mathbb{E}[\varepsilon \mid \mathbf{X}] = 0.$$

This is the so-called "exogeneity" condition. It states that the error term, the part of y unexplained by regressors, is uncorrelated with the regressors themselves. If this fails, then the regressor is said to be *endogenous*, and the Gauss–Markov world collapses. OLS loses its most precious property: unbiasedness.

Definition 6.1 (Endogeneity)**.** A regressor x_i is called **endogenous** if it is correlated with the error term:

$$\mathbb{E}[\varepsilon_i \mid \mathbf{x}_i] \neq 0.$$

Equivalently, exogeneity means $\mathbb{E}[\varepsilon_i \mid \mathbf{x}_i] = 0$.

Why might this correlation arise? The reasons are, unfortunately, plentiful. Sometimes the disturbance term is serially correlated: today's error depends on yesterday's, and regressors inherit this dependence. Sometimes we measure the wrong thing: the true $\mathbf{x}_i^*$ is hidden, and we only observe $\mathbf{x}_i = \mathbf{x}_i^* + \mathbf{w}_i$ with $\mathbf{w}_i$ a noisy error, thus contaminating the regressor. Sometimes we leave out an important control: the "omitted variable" sneaks into the error term, but is also related to our $\mathbf{x}_i$, causing spurious correlation. And sometimes the relationship is simultaneous: in a supply–demand system, price depends on quantity and quantity depends on price, so regressor and error are entangled by construction.

Fun Facts 6.1 (Structural Form vs. Reduced Form)**.** In simultaneous equations, the **structural form** expresses economic theory directly, with endogenous variables appearing on both sides. For example, in a supply–demand system:

$$Q^d = \alpha_0 + \alpha_1 P + u_d,$$
$$Q^s = \beta_0 + \beta_1 P + u_s,$$

where both Q (quantity) and P (price) are endogenous.
The **reduced form** solves these equations to express each endogenous variable as a function of exogenous variables and disturbances only. For instance,

$$P = \pi_0 + \pi_1 Z + v,$$

where Z is an exogenous shifter (like weather for agriculture). Reduced forms are useful for estimation because they purge simultaneity, but they hide the economic structure that generated them.

No matter the source, the consequence is the same. The OLS estimator is no longer

unbiased. Recall the expression

$$\hat{\boldsymbol{\beta}}_{OLS} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} = \boldsymbol{\beta} + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \varepsilon.$$

Taking expectations (conditional on $\mathbf{X}$) gives

$$\mathbb{E}\left[\hat{\boldsymbol{\beta}}_{OLS} \mid \mathbf{X}\right] = \boldsymbol{\beta} + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbb{E}\left[\varepsilon \mid \mathbf{X}\right].$$

So the condition $\mathbb{E}[\varepsilon \mid \mathbf{X}] = 0$ is exactly what makes OLS unbiased. If it fails, the bias term generally does not vanish, and in large samples the problem usually persists:

$$\operatorname*{plim}_{n\to\infty} \hat{\boldsymbol{\beta}}_{OLS} = \boldsymbol{\beta} + \left(\mathbb{E}\left[\mathbf{x}_i \mathbf{x}_i^\top\right]\right)^{-1} \mathbb{E}\left[\mathbf{x}_i \varepsilon_i\right].$$

so endogeneity typically means inconsistency as well.

This is why endogeneity is not "another mild violation." It makes the target itself ambiguous. OLS can be numerically precise while being wrong about the object of interest.

Instrumental Variables What can be done when regressors are endogenous? Econometricians have devised a powerful remedy: the use of *instrumental variables* (IV). The idea is to find a variable $\mathbf{z}_i$ that moves the endogenous regressor $\mathbf{x}_i$ but is otherwise unrelated to the outcome except through $\mathbf{x}_i$.

Definition 6.2 (Instrument)**.** A variable $\mathbf{z}_i$ is a valid instrument for an endogenous regressor $\mathbf{x}_i$ if it satisfies:

1. **Relevance:** $\operatorname{Cov}(\mathbf{z}_i, \mathbf{x}_i) \neq 0$ (or, in the multivariate case, $\operatorname{rank}(\mathbb{E}\left[\mathbf{Z}^\top \mathbf{X}\right])$ is large enough).
2. **Exogeneity:** $\mathbb{E}[\mathbf{z}_i \varepsilon_i] = 0$ (equivalently $\mathbb{E}[\varepsilon_i \mid \mathbf{z}_i] = 0$ under mild regularity).

Intuitively: $\mathbf{z}_i$ supplies "clean" variation in $\mathbf{x}_i$—variation that is not contaminated by whatever made $\mathbf{x}_i$ endogenous in the first place.

2SLS. A practical way to implement IV in linear models is **two-stage least squares (2SLS)**: first isolate the component of $\mathbf{X}$ explained by instruments, then regress $\mathbf{y}$ on that instrument-predicted component.

Algorithm 6 Two-Stage Least Squares (2SLS)

1: **Stage 1:** Regress the endogenous regressor X on the instrument(s) Z, obtain the fitted values $\hat{X}$.
2: **Stage 2:** Regress the outcome Y on $\hat{X}$. The slope coefficient from this regression is the IV/2SLS estimator.

Formally, letting $\mathbf{P}_\mathbf{Z} := \mathbf{Z}(\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top$ be the projection matrix onto the column space of $\mathbf{Z}$, the 2SLS estimator is

$$\hat{\boldsymbol{\beta}}_{2SLS} = \left(\mathbf{X}^\top \mathbf{P}_\mathbf{Z} \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{P}_\mathbf{Z} \mathbf{y}.$$

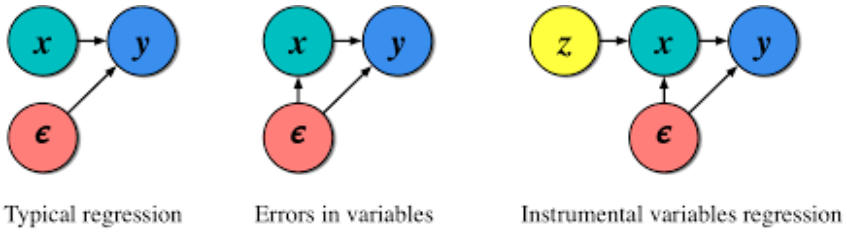

Figure 6: Intuition on 2SLS

To make the discussion less abstract, consider a classic example from labor economics.

Example 6.2. *We want to estimate the effect of education on wages. The model is*

$$\ln(wage_i) = \beta_0 + \beta_1 \, educ_i + \epsilon_i.$$

But education $educ_i$ is endogenous: individuals with high ability may both earn more and seek more education, so ability enters ϵ_i and correlates with $educ_i$.
A clever instrument is distance to college: students who live closer to a college tend to get more education, but distance itself is arguably uncorrelated with individual ability or the wage disturbance. Thus distance shifts $educ_i$ exogenously, allowing us to isolate the causal effect of schooling on wages.

A final, slightly uncomfortable point: endogeneity is rarely a "software issue." It is a design issue. Omitted variables, measurement error, and simultaneity are not algebraic quirks; they are reminders that regressions do not automatically become causal just because the output looks tidy.

In that sense, IV is not a mechanical fix either. It forces you to say *what* your instrument is changing, *why* that change is plausibly as-good-as-random with respect to ε, and *what causal object* you are willing to interpret the resulting coefficient as. When those arguments are credible, IV can rescue identification. When they are not, 2SLS is just OLS with extra steps.

7 Generalized Method of Moments

Up to now, our trajectory has been: asymptotics → OLS → GLS → outliers → endogeneity (fixing the moment condition itself via instruments).

In that last step, **instrumental variables (IV)** reframed estimation as a problem of enforcing *orthogonality restrictions*: at the true parameter, certain functions of the data should have mean zero. OLS can be read the same way. The point is that they are all doing the same conceptual thing: they pick parameters so that sample analogs of some population restrictions are approximately satisfied.

The **Generalized Method of Moments (GMM)** makes this perspective explicit and systematic. Rather than committing to a full likelihood, we assume only that a collection of *population moment conditions* holds at the true parameter value, and we estimate by matching their sample counterparts as closely as possible. The weighting logic you saw in GLS will reappear here: different moment conditions can be more or less informative, and GMM lets us weight them accordingly.

7.1 Setup and Notation

Consider an i.i.d. sample $\{(y_i, \mathbf{x}_i, \boldsymbol{\xi}_i)\}_{i=1}^N$, where $y_i \in \mathbb{R}$ is the dependent variable, $\mathbf{x}_i \in \mathbb{R}^K$ is a $K \times 1$ vector of regressors, and $\boldsymbol{\xi}_i \in \mathbb{R}^M$ is an $M \times 1$ vector of instruments. Let $\boldsymbol{\beta} \in \mathbb{R}^K$ be the parameter of interest. We translate **zero-mean** and **exogeneity**-type assumptions into a set of *moment conditions*.

Definition 7.1 (Moment Conditions: Exogeneity)**.** Let $\psi_i(\boldsymbol{\beta}) \in \mathbb{R}^M$ denote the moment function for observation i:

$$\psi_i(\boldsymbol{\beta}) = \boldsymbol{\xi}_i\big(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\big).$$

The **population moment condition** is

$$\mathbb{E}\left[\psi_i(\boldsymbol{\beta}_0)\right] = 0,$$

where $\boldsymbol{\beta}_0$ is the true parameter value.

The condition $\mathbb{E}\left[\psi_i(\boldsymbol{\beta}_0)\right] = 0$ says that, at the true parameter, the instruments $\boldsymbol{\xi}_i$ are orthogonal (in expectation) to the structural error $y_i - \mathbf{x}_i^\top \boldsymbol{\beta}_0$. This is the population counterpart of the familiar exogeneity restriction $\mathbb{E}\left[\boldsymbol{\xi}_i \varepsilon_i\right] = 0$ under the structural model $y_i = \mathbf{x}_i^\top \boldsymbol{\beta}_0 + \varepsilon_i$.

Example 7.1 (Connecting Back to OLS and IV)**.** *If $M = K$ and $\boldsymbol{\xi}_i = \mathbf{x}_i$, then the moment condition $\mathbb{E}\left[\mathbf{x}_i(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}_0)\right] = 0$ yields the OLS normal equations (under exogeneity). If instead $\boldsymbol{\xi}_i$ contains valid instruments distinct from $\mathbf{x}_i$, we obtain the classical IV setup.*

Sample analog. The population expectation is not observable, so we replace it with its sample analog:

$$\bar{\psi}_N(\boldsymbol{\beta}) = \frac{1}{N}\sum_{i=1}^{N}\psi_i(\boldsymbol{\beta}) = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\xi}_i\big(y_i - \mathbf{x}_i^\top\boldsymbol{\beta}\big).$$

When the model is **just identified** ($M = K$) and the relevant rank condition holds, we can (in principle) solve the K equations

$$\bar{\psi}_N(\hat{\boldsymbol{\beta}}) = 0$$

for $\hat{\boldsymbol{\beta}}$. When $M > K$, the model is **overidentified**: there are more moment conditions than parameters, so in general there is no $\boldsymbol{\beta}$ that makes *all* sample moments exactly zero. In that case, GMM will choose $\hat{\boldsymbol{\beta}}$ that makes $\bar{\psi}_N(\boldsymbol{\beta})$ "as close to zero as possible" in a weighted least-squares sense (formalized in the next subsection). Finally, if $M < K$, the model is **underidentified**: there is insufficient information to recover $\boldsymbol{\beta}$ uniquely.

Definition 7.2 (Identification in GMM)**.** The parameter $\boldsymbol{\beta}_0$ is **identified** if the mapping $\boldsymbol{\beta} \mapsto \mathbb{E}\left[\psi_i(\boldsymbol{\beta})\right]$ is injective at $\boldsymbol{\beta}_0$, i.e.

$$\mathbb{E}\left[\psi_i(\boldsymbol{\beta})\right] = 0 \quad \Rightarrow \quad \boldsymbol{\beta} = \boldsymbol{\beta}_0.$$

Identification ensures that the population moment conditions uniquely determine the true parameter. Without it, even perfect data would not reveal $\boldsymbol{\beta}_0$.

Fun Facts 7.2. Some notations indexing the notion of 'sizes':

- $K = \dim(\boldsymbol{\beta})$: dimension of the parameter vector,
- $M = \dim(\psi_i(\boldsymbol{\beta})) = \dim(\boldsymbol{\xi}_i)$: number of moment conditions,
- $M = K$: just identified, $M > K$: overidentified, $M < K$: underidentified.

Example 7.3 (Just-identified IV)**.** *Suppose we have one endogenous regressor $\mathbf{x}_i \in \mathbb{R}$ and one valid instrument $\boldsymbol{\xi}_i \in \mathbb{R}$, so $M = K = 1$. The moment condition*

$$\mathbb{E}\left[\boldsymbol{\xi}_i\big(y_i - \mathbf{x}_i\beta_0\big)\right] = 0$$

identifies β_0 as long as $\mathbb{E}\left[\boldsymbol{\xi}_i\mathbf{x}_i\right] \neq 0$. In this case, the estimator is obtained by solving the sample analog:

$$\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\xi}_i\big(y_i - \mathbf{x}_i\hat{\beta}\big) = 0, \qquad \Rightarrow \qquad \hat{\beta} = \frac{\sum_{i=1}^{N}\boldsymbol{\xi}_i y_i}{\sum_{i=1}^{N}\boldsymbol{\xi}_i\mathbf{x}_i}.$$

Example 7.4 (Over-identified IV)**.** *Now suppose we have two instruments $\boldsymbol{\xi}_{1i}, \boldsymbol{\xi}_{2i}$ but still only one endogenous regressor $\mathbf{x}_i$, so $M = 2$ and $K = 1$. The two population restrictions*

$$\mathbb{E}\left[\boldsymbol{\xi}_{1i}(y_i - \mathbf{x}_i\beta_0)\right] = 0, \qquad \mathbb{E}\left[\boldsymbol{\xi}_{2i}(y_i - \mathbf{x}_i\beta_0)\right] = 0$$

need not be simultaneously satisfied by any single β at the sample *level. So instead of insisting on $\bar{\psi}_N(\boldsymbol{\beta}) = 0$, we choose $\hat{\boldsymbol{\beta}}$ that makes the vector of sample moments as small as possible.*

The idea of GMM is therefore simple: if we can describe the population with a set of moment conditions $\mathbb{E}\left[\psi_i(\boldsymbol{\beta}_0)\right] = 0$, then a consistent estimator of $\boldsymbol{\beta}_0$ can be obtained by matching their sample analogs as closely as possible. All the heavy machinery of GMM — weighting matrices, asymptotic variances, and efficiency — will arise naturally from this goal.

7.2 The GMM Objective Function

To formalize this, define the sample average of the moment function:

$$\bar{\psi}_N(\boldsymbol{\beta}) = \frac{1}{N}\sum_{i=1}^{N}\psi_i(\boldsymbol{\beta}).$$

When $M > K$, $\bar{\psi}_N(\boldsymbol{\beta}) \in \mathbb{R}^{M\times 1}$ will generally not equal $\mathbf{0}$ for any $\boldsymbol{\beta}$. We therefore measure the "distance" of $\bar{\psi}_N(\boldsymbol{\beta})$ from zero using a quadratic form.

Definition 7.3 (GMM Objective Function)**.** Let $\mathbf{W}_N \in \mathbb{R}^{M\times M}$ be symmetric and positive definite. The **GMM objective function** is

$$Q_N(\boldsymbol{\beta}) = \bar{\psi}_N(\boldsymbol{\beta})^\top \mathbf{W}_N\, \bar{\psi}_N(\boldsymbol{\beta}).$$

The GMM estimator is

$$\hat{\boldsymbol{\beta}}_{GMM} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^K} Q_N(\boldsymbol{\beta}).$$

Intuitively, $Q_N(\boldsymbol{\beta})$ measures how well the moment conditions are satisfied at $\boldsymbol{\beta}$. If $M = K$ and the system admits a unique solution to $\bar{\psi}_N(\boldsymbol{\beta}) = \mathbf{0}$, then minimizing $Q_N(\boldsymbol{\beta})$ (with any $\mathbf{W}_N \succ 0$) reproduces that solution. If $M > K$, the weighting matrix $\mathbf{W}_N$ determines which moments we treat as more important.

Setting $\mathbf{W}_N = \mathbf{I}_M$ treats all moments equally. That sounds democratic, but it is not always smart: some moments are precise, others are noisy troublemakers. Equal weights say: "I trust every instrument equally, including the one that clearly partied all night before the exam." When moment conditions differ in reliability, weighting is how GMM avoids wasting information.

Example 7.5 (Unequal moment reliability)**.** *Suppose we have two instruments, $\boldsymbol{\xi}_{1i}$ and $\boldsymbol{\xi}_{2i}$, both exogenous but with different relevance. If $\boldsymbol{\xi}_{1i}$ is strongly correlated with $\mathbf{x}_i$ while $\boldsymbol{\xi}_{2i}$ is only weakly correlated, then the first moment condition is typically more informative. Weighting them equally is like giving equal importance to an honest survey and to a coin flip. A better $\mathbf{W}_N$ puts more weight on the informative moment.*

7.3 Optimal $\boldsymbol{\beta}$

Enough justifying the existence of $\mathbf{W}_N$. For now, take this matrix as given. The natural next question is: which $\boldsymbol{\beta}$ minimizes the loss $Q_N(\boldsymbol{\beta})$? This is a straight calculus exercise.

Recall

$$Q_N(\boldsymbol{\beta}) = \bar{\psi}_N(\boldsymbol{\beta})^\top \mathbf{W}_N \bar{\psi}_N(\boldsymbol{\beta}), \qquad \bar{\psi}_N(\boldsymbol{\beta}) \in \mathbb{R}^{M\times 1}, \quad \mathbf{W}_N \in \mathbb{R}^{M\times M}.$$

Let

$$\nabla_{\boldsymbol{\beta}} \bar{\psi}_N(\boldsymbol{\beta}) \in \mathbb{R}^{M\times K}$$

denote the Jacobian of $\bar{\psi}_N(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$. Differentiating the quadratic form gives

$$\nabla_{\boldsymbol{\beta}} Q_N(\boldsymbol{\beta}) = 2\left(\nabla_{\boldsymbol{\beta}} \bar{\psi}_N(\boldsymbol{\beta})\right)^\top \mathbf{W}_N \bar{\psi}_N(\boldsymbol{\beta}).$$

Hence the F.O.C (full of crap, if you like)) for $\hat{\boldsymbol{\beta}}$ is

$$\left(\nabla_{\boldsymbol{\beta}} \bar{\psi}_N(\hat{\boldsymbol{\beta}})\right)^\top \mathbf{W}_N \bar{\psi}_N(\hat{\boldsymbol{\beta}}) = \mathbf{0} \in \mathbb{R}^{K\times 1}.$$

In words: at the minimizer, the sample moments are orthogonal (under the metric induced by $\mathbf{W}_N$) to the directions in which the moments change with $\boldsymbol{\beta}$. This generalizes the normal equations in OLS and the orthogonality condition in IV.

To make the structure explicit, write the Jacobian as the average of per–observation contributions:

$$\nabla_{\boldsymbol{\beta}} \bar{\psi}_N(\boldsymbol{\beta}) = \frac{1}{N}\sum_{i=1}^{N} \nabla_{\boldsymbol{\beta}} \psi_i(\boldsymbol{\beta}),$$

so the F.O.C. can be expanded as

$$\Big(\frac{1}{N}\sum_{i=1}^{N} \nabla_{\boldsymbol{\beta}} \psi_i(\hat{\boldsymbol{\beta}})\Big)^\top \mathbf{W}_N \Big(\frac{1}{N}\sum_{i=1}^{N} \psi_i(\hat{\boldsymbol{\beta}})\Big) = \mathbf{0}.$$

7.4 Optimal Weighting Matrix $\mathbf{W}_N$

We argued informally that not all moments are created equal: some are stable and informative, others are noisy and fickle. The weighting matrix should reflect that. Heuristically: moments with smaller variance (higher "precision") should have more influence in the criterion. We now make this precise.

Proposition 7.6 (Efficient (optimal) weighting)**.** Among quadratic GMM criteria of the form

$$Q_N(\boldsymbol{\beta}) = \bar{\psi}_N(\boldsymbol{\beta})^\top \mathbf{C}\, \bar{\psi}_N(\boldsymbol{\beta}),$$

the choice of weights that delivers the smallest large–sample variance for $\hat{\boldsymbol{\beta}}$ is (up to proportionality)

$$\mathbf{C} \;\propto\; \mathbf{S}^{-1}, \qquad \mathbf{S} \;:=\; \mathbb{E}\left[\,\psi_i(\boldsymbol{\beta}_0)\psi_i(\boldsymbol{\beta}_0)^\top\,\right].$$

Since $\mathbb{E}\left[\psi_i(\boldsymbol{\beta}_0)\right] = \mathbf{0}$, $\mathbf{S}$ is also the covariance matrix of the moment vector at the truth:

$$\mathbf{S} = \mathrm{Cov}\big(\psi_i(\boldsymbol{\beta}_0)\big).$$

In particular, if the moments are uncorrelated and scalar–valued, this reduces to "weight each moment by the inverse of its variance."

Fun Facts 7.7 (Digress: Why symmetry is WLOG)**.** Let $\mathbf{W} \in \mathbb{R}^{M\times M}$ be any (not necessarily symmetric) weighting matrix and let $g \in \mathbb{R}^M$ be any vector (e.g., $g = \bar{\psi}_N(\boldsymbol{\beta})$). Decompose $\mathbf{W}$ into its symmetric and skew–symmetric parts:

$$\mathbf{W} = \mathbf{W}_s + \mathbf{W}_a, \qquad \mathbf{W}_s := \tfrac{1}{2}(\mathbf{W} + \mathbf{W}^\top), \qquad \mathbf{W}_a := \tfrac{1}{2}(\mathbf{W} - \mathbf{W}^\top).$$

Then $\mathbf{W}_s^\top = \mathbf{W}_s$ and $\mathbf{W}_a^\top = -\mathbf{W}_a$. Moreover,

$$g^\top \mathbf{W} g = g^\top \mathbf{W}_s g + g^\top \mathbf{W}_a g, \qquad g^\top \mathbf{W}_a g = (g^\top \mathbf{W}_a g)^\top = g^\top \mathbf{W}_a^\top g = -g^\top \mathbf{W}_a g,$$

so $g^\top \mathbf{W}_a g = 0$ and therefore

$$g^\top \mathbf{W} g = g^\top \mathbf{W}_s g.$$

Conclusion: the quadratic criterion depends only on the symmetric part of $\mathbf{W}$. So we may assume the weighting matrix is symmetric without loss of generality.

Rather than proving Proposition 7.6 in full generality immediately, we start with a concrete two–moment setup that shows the geometry without extra notation.

7.4.1 A two–moment compromise.

Suppose $K = 1$ (one scalar parameter) and $M = 2$ (two valid moment conditions). Write the two sample mean moments as

$$\bar{\psi}_N(\theta) = \begin{bmatrix} \bar{\psi}_{1N}(\theta) \\ \bar{\psi}_{2N}(\theta) \end{bmatrix} = \begin{bmatrix} \bar{X}_N - \theta \\ \bar{Y}_N - \theta \end{bmatrix}, \qquad \mathbb{E}\left[\psi_j(\theta_0)\right] = 0,\ j = 1, 2,$$

and consider the weighted criterion

$$Q_N(\theta) = \begin{bmatrix} \bar{\psi}_{1N}(\theta) & \bar{\psi}_{2N}(\theta) \end{bmatrix} \mathbf{C} \begin{bmatrix} \bar{\psi}_{1N}(\theta) \\ \bar{\psi}_{2N}(\theta) \end{bmatrix},$$

with $\mathbf{C} = \begin{pmatrix} c_{11} & c_{12} \\ c_{12} & c_{22} \end{pmatrix}$ being the weighting matrix symmetric positive definite. Then we can write out the loss function:

$$\begin{aligned}\bar{\psi}_N(\theta)^\top \mathbf{C}\, \bar{\psi}_N(\theta) &= \begin{bmatrix}(\bar{X}_N - \theta) & (\bar{Y}_N - \theta)\end{bmatrix} \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix} \begin{bmatrix} \bar{X}_N - \theta \\ \bar{Y}_N - \theta \end{bmatrix} \\ &= c_{11}(\bar{X}_N - \theta)^2 + (c_{12} + c_{21})(\bar{X}_N - \theta)(\bar{Y}_N - \theta) + c_{22}(\bar{Y}_N - \theta)^2\end{aligned}$$

Taking derivative with respect to θ helps us find the optimal θ first:

$$\frac{\partial Q_N(\theta)}{\partial \theta} = 2c_{11}(\theta - \bar{X}_N) + (c_{12} + c_{21})(2\theta - \bar{X}_N - \bar{Y}_N) + 2c_{22}(\theta - \bar{Y}_N) \equiv 0$$

Rearrange we have

$$(c_{11} + \frac{c_{12} + c_{21}}{2})(\theta - \bar{X}_N) + (c_{22} + \frac{c_{12} + c_{21}}{2})(\theta - \bar{Y}_N)$$

We claim that $\mathbf{C}$ is positive definite and symmetric, and hence the expression above reduce to

$$(c_{11} + c_{12})(\theta - \bar{X}_N) + (c_{22} + c_{21})(\theta - \bar{Y}_N) \to \hat{\theta} = \lambda \bar{X}_N + (1 - \lambda)\bar{Y}_N, \lambda = \frac{c_{11} + c_{12}}{\sum_{i,j} c_{i,j}}$$

Now we will show that the choice of $\mathbf{C}$ may coincide' with the inverse of the CVC of the moment vector $\bar{\psi}_N(\theta)$. Having derived the expression of θ, we can calculate the variance:

$$\text{Var}\left(\hat{\theta}\right) = \lambda^2 \text{Var}\left(\bar{X}_N\right) + (1 - \lambda)^2 \text{Var}\left(\bar{Y}_N\right) + 2\lambda(1 - \lambda)\,\text{Cov}(\bar{X}_N, \bar{Y}_N)$$

An optimal weighting matrix should achieve the minimal variance (recall what we did in proving OLS to be BLUE under GM assumptions!). Again taking derivative with respect to λ gives

$$\frac{\partial \text{Var}\left(\hat{\theta}\right)}{\partial \lambda} = 2\lambda \text{Var}\left(\bar{X}_N\right) - 2(1 - \lambda)\text{Var}\left(\bar{Y}_N\right) + 2(1 - 2\lambda)\,\text{Cov}(\bar{X}_N, \bar{Y}_N)$$

Rearranging get

$$\lambda^* = \frac{\text{Var}\left(\bar{Y}_N\right) - \text{Cov}(\bar{X}_N, \bar{Y}_N)}{\text{Var}\left(\bar{X}_N\right) + \text{Var}\left(\bar{Y}_N\right) - 2\,\text{Cov}(\bar{X}_N, \bar{Y}_N)} := \frac{c_{11} + c_{12}}{\sum_{i,j} c_{ij}}$$

One can then match $c_{11} = \text{Var}\left(\bar{Y}_N\right)$, $c_{22} = \text{Var}\left(\bar{X}_N\right)$ and $c_{12} = c_{21} = -\,\text{Cov}(\bar{X}_N, \bar{Y}_N)$ to get the same expression of λ^*. So now we can rewrite our $\mathbf{C}$ matrix as

$$\mathbf{C} = \begin{bmatrix} \text{Var}\left(\bar{Y}_N\right) & -\,\text{Cov}(\bar{X}_N, \bar{Y}_N) \\ -\,\text{Cov}(\bar{X}_N, \bar{Y}_N) & \text{Var}\left(\bar{X}_N\right) \end{bmatrix}$$

Wait! Did we see this somewhere before? Yes! This is the inverse of the VCV matrix (up to a constant $\det|\mathrm{Var}\left(\bar{\psi}_N(\theta)\right)|$)! The VCV of $\bar{\psi}_N(\theta)$ is

$$\mathrm{Var}\left(\bar{\psi}_N(\theta)\right) = \mathrm{Var}\left(\begin{bmatrix}\bar{X}_N \\ \bar{Y}_N\end{bmatrix}\right) = \begin{bmatrix}\mathrm{Var}\left(\bar{X}_N\right) & \mathrm{Cov}(\bar{X}_N, \bar{Y}_N) \\ \mathrm{Cov}(\bar{X}_N, \bar{Y}_N) & \mathrm{Var}\left(\bar{Y}_N\right)\end{bmatrix}$$

and the inverse is just switching diagonals and negating the off-diagonals. This two–moment compromise is exactly the geometry behind the general claim: the efficient choice is $\mathbf{C} \propto S^{-1}$. In short: *trust precise moments, discount noisy (and redundant) ones.*

The intuition is consistent with common sense:

- If all moments are equally noisy ($\mathbf{S} = c\mathbf{I}_M$), then $\mathbf{S}^{-1} \propto \mathbf{I}_M$ — equal weights.
- If one moment is far more stable than the others, $\mathbf{S}^{-1}$ upweights it automatically.

This is the GMM counterpart of GLS: just as GLS uses Ω^{-1} to weight residuals with unequal variances, efficient GMM uses $\mathbf{S}^{-1}$ to weight moments with unequal reliability.

7.4.2 Estimating Optimal $\mathbf{W}_N$.

In practice we do not know $\mathbf{S}$: it involves the unknown $\boldsymbol{\beta}_0$. A standard procedure is as follow.

Algorithm 7 Two–Step GMM

1: Start with a simple choice, often $\mathbf{W}_N^{(0)} = \mathbf{I}_M$, to obtain an initial consistent estimator $\hat{\boldsymbol{\beta}}^{(0)}$.
2: Estimate the moment covariance using $\hat{\boldsymbol{\beta}}^{(0)}$:

$$\hat{\mathbf{S}} = \frac{1}{N}\sum_{i=1}^{N} \hat{\psi}_i \hat{\psi}_i^\top, \qquad \hat{\psi}_i := \psi_i(\hat{\boldsymbol{\beta}}^{(0)}),$$

and set the updated weight to $\hat{\mathbf{W}}_N := \hat{\mathbf{S}}^{-1}$.
3: Re–estimate $\boldsymbol{\beta}$ using $\hat{\mathbf{W}}_N$ to obtain the efficient two–step GMM estimator.

This refinement makes the final estimator behave as if we had known the optimal $\mathbf{S}$ from the start, achieving the lowest possible asymptotic variance among (regular) GMM estimators.

7.5 Instrumental Variables and 2SLS as GMM

We now return to the instrumental variables (IV) framework and show that the familiar two–stage least squares (2SLS) estimator arises naturally as a special case of GMM. This perspective will be useful when we later discuss efficiency and overidentification.

Consider the linear model

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta}_0 + \varepsilon_i,$$

where some components of $\mathbf{x}_i$ may be endogenous. Let $\boldsymbol{\xi}_i$ denote an $M \times 1$ vector of instruments satisfying the exogeneity condition

$$\mathbb{E}\left[\boldsymbol{\xi}_i \varepsilon_i\right] = 0.$$

The IV assumptions can be written in the GMM form using the moment function

$$\psi_i(\boldsymbol{\beta}) = \boldsymbol{\xi}_i\big(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\big), \qquad \mathbb{E}\left[\psi_i(\boldsymbol{\beta}_0)\right] = 0.$$

The corresponding sample average moment is

$$\bar{\psi}_N(\boldsymbol{\beta}) = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{\xi}_i\big(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\big).$$

Equivalently, stack the instruments row–wise into $\boldsymbol{\Xi} \in \mathbb{R}^{N\times M}$ (with ith row $\boldsymbol{\xi}_i^\top$), and collect outcomes and regressors into $\mathbf{y} \in \mathbb{R}^N$ and $\mathbf{X} \in \mathbb{R}^{N\times K}$. Then

$$\bar{\psi}_N(\boldsymbol{\beta}) = \frac{1}{N}\boldsymbol{\Xi}^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

Consequently, given a positive definite weighting matrix $\mathbf{W}_N \in \mathbb{R}^{M\times M}$, we can define the GMM objective function

$$Q_N(\boldsymbol{\beta}) = \bar{\psi}_N(\boldsymbol{\beta})^\top \mathbf{W}_N \bar{\psi}_N(\boldsymbol{\beta}),$$

and the GMM estimator solves

$$\hat{\boldsymbol{\beta}}_{GMM} = \arg\min_{\boldsymbol{\beta}} Q_N(\boldsymbol{\beta}).$$

Using the stacked form of $\bar{\psi}_N(\boldsymbol{\beta})$, the criterion can be written as

$$Q_N(\boldsymbol{\beta}) = \frac{1}{N^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top \boldsymbol{\Xi}\, \mathbf{W}_N\, \boldsymbol{\Xi}^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

(The factor $1/N^2$ is irrelevant for the minimizer, but let's keep the notation honest.)

If we choose (we will justify this choice below)

$$\mathbf{W}_N = (\boldsymbol{\Xi}^\top\boldsymbol{\Xi})^{-1},$$

then minimizing $Q_N(\boldsymbol{\beta})$ yields the closed–form solution

$$\hat{\boldsymbol{\beta}}_{GMM} = \big(\mathbf{X}^\top\boldsymbol{\Xi}(\boldsymbol{\Xi}^\top\boldsymbol{\Xi})^{-1}\boldsymbol{\Xi}^\top\mathbf{X}\big)^{-1}\mathbf{X}^\top\boldsymbol{\Xi}(\boldsymbol{\Xi}^\top\boldsymbol{\Xi})^{-1}\boldsymbol{\Xi}^\top\mathbf{y},$$

which is exactly the two–stage least squares estimator:

$$\hat{\boldsymbol{\beta}}_{2SLS} = \hat{\boldsymbol{\beta}}_{GMM}.$$

Why choose $\mathbf{W}_N = (\boldsymbol{\Xi}^\top\boldsymbol{\Xi})^{-1}$? This particular choice of weighting matrix is not arbitrary. Recall that the optimal GMM weight is proportional to the inverse variance of the

moment vector:

$$\mathbf{W} \;\propto\; \mathrm{Var}\left(\psi_i(\boldsymbol{\beta}_0)\right)^{-1}.$$

In the linear IV setting,

$$\psi_i(\boldsymbol{\beta}_0) = \boldsymbol{\xi}_i \varepsilon_i, \qquad \mathrm{Var}\left(\psi_i(\boldsymbol{\beta}_0)\right) = \mathbb{E}\left[\boldsymbol{\xi}_i \boldsymbol{\xi}_i^\top \varepsilon_i^2\right].$$

If we impose the simplifying assumption that the structural error is conditionally homoskedastic,

$$\mathbb{E}\left[\varepsilon_i^2 \mid \boldsymbol{\xi}_i\right] = \sigma^2,$$

then

$$\mathrm{Var}\left(\psi_i(\boldsymbol{\beta}_0)\right) = \sigma^2 \mathbb{E}\left[\boldsymbol{\xi}_i \boldsymbol{\xi}_i^\top\right].$$

Up to the irrelevant scalar σ^2, the optimal weight is therefore

$$\mathbf{W} \;\propto\; \mathbb{E}\left[\boldsymbol{\xi}_i \boldsymbol{\xi}_i^\top\right]^{-1}.$$

Replacing the population expectation with its sample analog yields

$$\mathbf{W}_N = \Big(\frac{1}{N}\boldsymbol{\Xi}^\top \boldsymbol{\Xi}\Big)^{-1} \;\propto\; (\boldsymbol{\Xi}^\top \boldsymbol{\Xi})^{-1},$$

and proportionality does not change the minimizer of $Q_N(\boldsymbol{\beta})$.

Hence, two–stage least squares can be viewed as a GMM estimator that adopts a simple plug–in approximation to the optimal weighting matrix under homoskedasticity. More general forms of heteroskedasticity require different choices of $\mathbf{W}_N$, leading to efficient GMM estimators.

7.6 Finite-sample Bias and Consistency of 2SLS

Since we are talking about 2SLS already (and never formally before in this note!), we might as well pause for a second to ask about some properties 2SLS estimator as what we did for OLS estimators. 2SLS is generally *not* unbiased in finite samples, but it is *consistent* under standard IV conditions.

Let $P_{\boldsymbol{\xi}} := \boldsymbol{\Xi}(\boldsymbol{\Xi}^\top \boldsymbol{\Xi})^{-1}\boldsymbol{\Xi}^\top$ denote the projection matrix onto the column span of $\boldsymbol{\Xi}$. The 2SLS estimator can be written compactly as

$$\hat{\boldsymbol{\beta}}_{2SLS} = (\mathbf{X}^\top P_{\boldsymbol{\xi}} \mathbf{X})^{-1} \mathbf{X}^\top P_{\boldsymbol{\xi}} \mathbf{y}.$$

Under the structural model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon}$, we have the decomposition

$$\hat{\boldsymbol{\beta}}_{2SLS} - \boldsymbol{\beta}_0 = (\mathbf{X}^\top P_{\boldsymbol{\xi}} \mathbf{X})^{-1} \mathbf{X}^\top P_{\boldsymbol{\xi}} \boldsymbol{\varepsilon}. \tag{1}$$

Proposition 7.8 (Finite-sample unbiasedness fails for 2SLS)**.** In general, the two–stage least squares estimator is *not* unbiased in finite samples.

Proof. Write the structural equation and reduced form in matrix notation:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon}, \qquad \mathbf{X} = \boldsymbol{\Xi}\boldsymbol{\Delta} + \mathbf{U},$$

where $\boldsymbol{\Xi}$ is the $N \times M$ instrument matrix, $\boldsymbol{\Delta}$ is an $M \times K$ reduced-form coefficient matrix, and $\mathbf{U}$ collects the reduced-form errors. Endogeneity means that, in general, $\mathrm{Cov}(\mathbf{U}, \boldsymbol{\varepsilon}) \neq \mathbf{0}$ (equivalently, at the observation level, $\mathrm{Cov}(u_{ik}, \varepsilon_i) \neq 0$ for at least one regressor component k).

Let

$$P_{\boldsymbol{\xi}} := \boldsymbol{\Xi}(\boldsymbol{\Xi}^\top \boldsymbol{\Xi})^{-1}\boldsymbol{\Xi}^\top$$

be the projection matrix onto $\mathrm{span}(\boldsymbol{\Xi})$. The first-stage fitted regressors are

$$\hat{\mathbf{X}} := P_{\boldsymbol{\xi}}\mathbf{X}.$$

Using the reduced form $\mathbf{X} = \boldsymbol{\Xi}\boldsymbol{\Delta} + \mathbf{U}$ and the idempotence property $P_{\boldsymbol{\xi}}\boldsymbol{\Xi} = \boldsymbol{\Xi}$, we obtain

$$\hat{\mathbf{X}} = P_{\boldsymbol{\xi}}(\boldsymbol{\Xi}\boldsymbol{\Delta} + \mathbf{U}) = \boldsymbol{\Xi}\boldsymbol{\Delta} + P_{\boldsymbol{\xi}}\mathbf{U}. \tag{2}$$

Thus the fitted regressors inherit a projected version of the first-stage error, $P_{\boldsymbol{\xi}}\mathbf{U}$.

In the second stage we regress $\mathbf{y}$ on $\hat{\mathbf{X}}$ by OLS. Finite-sample unbiasedness of this second-stage OLS would require the regressor matrix $\hat{\mathbf{X}}$ to be orthogonal to the structural error, i.e. $\mathbb{E}\left[\hat{\mathbf{X}}^\top \boldsymbol{\varepsilon}\right] = \mathbf{0}$. But by (2),

$$\mathrm{Cov}(\hat{\mathbf{X}}, \boldsymbol{\varepsilon}) = \mathrm{Cov}(\boldsymbol{\Xi}\boldsymbol{\Delta} + P_{\boldsymbol{\xi}}\mathbf{U},\ \boldsymbol{\varepsilon}) = \mathrm{Cov}(P_{\boldsymbol{\xi}}\mathbf{U}, \boldsymbol{\varepsilon}),$$

since $\boldsymbol{\Xi}$ is nonrandom conditional on itself and $\mathrm{Cov}(\boldsymbol{\Xi}\boldsymbol{\Delta}, \boldsymbol{\varepsilon}) = \mathbf{0}$ under instrument exogeneity. Under endogeneity, $\mathrm{Cov}(\mathbf{U}, \boldsymbol{\varepsilon}) \neq \mathbf{0}$, and in general projecting $\mathbf{U}$ onto $\mathrm{span}(\boldsymbol{\Xi})$ does not eliminate its finite-sample correlation with $\boldsymbol{\varepsilon}$; hence typically

$$\mathrm{Cov}(P_{\boldsymbol{\xi}}\mathbf{U}, \boldsymbol{\varepsilon}) \neq \mathbf{0}.$$

Therefore the second-stage OLS orthogonality condition fails in finite samples, and 2SLS is generally biased. □

Heuristically, the fitted regressor still carries the endogenous component u, only projected onto the instrument space; projection does not eliminate correlation with ε in finite samples.

So, validity of instruments, $\mathbb{E}[\xi_i \varepsilon_i] = 0$, only guarantees that the *population* moments are correct. It does *not* imply finite-sample unbiasedness of a ratio estimator such as IV/2SLS. The right takeaway is: *2SLS can be biased in finite samples but asymptotically unbiased under standard conditions.*

This is not the end of the world, of course. We've seen precedents of sacrificing finite-sample unbiasedness in exchange for good large-sample behavior (the Bessel correction for sample variance is a familiar example). And the good news is that despite the finite sample bias, $\hat{\boldsymbol{\beta}}_{\mathrm{2SLS}}$ remains consistency as long as we have large enough sample. In particular, as the sample size N grows, $\hat{\boldsymbol{\beta}}_{2SLS}$ converges in probability to $\boldsymbol{\beta}_0$.

Proposition 7.9 (Consistency of 2SLS)**.** Consider the linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon}$ with instruments $\boldsymbol{\Xi}$. Suppose:

(i) (**IV validity**) $\mathbb{E}[\boldsymbol{\xi}_i \varepsilon_i] = 0$.

(ii) (**Finite nonsingular limits**) As $N \to \infty$,

$$\frac{1}{N}\sum_i \boldsymbol{\xi}^\top y_i \xrightarrow{\mathbb{P}} A, \qquad \frac{1}{N}\sum_i \boldsymbol{\xi}^\top \boldsymbol{\xi}_i \xrightarrow{\mathbb{P}} B, \qquad \frac{1}{N}\sum_i x_i^\top \boldsymbol{\xi}_i \xrightarrow{\mathbb{P}} C$$

where all limiting VCVs has finite metrics (norms, for instance)

(iii) (**Sample moment convergence**)

$$\frac{1}{N}\sum_i \boldsymbol{\xi}_i^\top \varepsilon_i \xrightarrow{\mathbb{P}} 0.$$

Then $\hat{\beta}_{2SLS} \xrightarrow{\mathbb{P}} \beta_0$.

Proof. Write the 2SLS estimator in its closed form:

$$\hat{\boldsymbol{\beta}}_{2SLS} = \left(\mathbf{X}^\top \boldsymbol{\Xi}(\boldsymbol{\Xi}^\top\boldsymbol{\Xi})^{-1}\boldsymbol{\Xi}^\top \mathbf{X}\right)^{-1}\mathbf{X}^\top\boldsymbol{\Xi}(\boldsymbol{\Xi}^\top\boldsymbol{\Xi})^{-1}\boldsymbol{\Xi}^\top\mathbf{y}.$$

Substitute $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon}$ to obtain

$$\hat{\boldsymbol{\beta}}_{2SLS} - \boldsymbol{\beta}_0 = \left(\mathbf{X}^\top \boldsymbol{\Xi}(\boldsymbol{\Xi}^\top\boldsymbol{\Xi})^{-1}\boldsymbol{\Xi}^\top \mathbf{X}\right)^{-1}\mathbf{X}^\top\boldsymbol{\Xi}(\boldsymbol{\Xi}^\top\boldsymbol{\Xi})^{-1}\boldsymbol{\Xi}^\top\boldsymbol{\varepsilon}. \tag{3}$$

Divide by N to match the sample moments:

$$\hat{\boldsymbol{\beta}}_{2SLS} - \boldsymbol{\beta}_0 = \left[\left(\frac{1}{N}\sum_i x_i^\top \boldsymbol{\xi}_i\right)\left(\frac{1}{N}\sum_i \boldsymbol{\xi}_i^\top \boldsymbol{\xi}_i\right)^{-1}\left(\frac{1}{N}\sum_i \boldsymbol{\xi}_i^\top x_i\right)\right]^{-1}\left[\left(\frac{1}{N}\sum_i x_i^\top \boldsymbol{\xi}_i\right)\left(\frac{1}{N}\sum_i \boldsymbol{\xi}_i^\top \boldsymbol{\xi}_i\right)^{-1}\left(\frac{1}{N}\sum_i \boldsymbol{\xi}_i^\top \varepsilon_i\right)\right].$$

By assumption, $\frac{1}{N}\sum_i x_i^\top \boldsymbol{\xi}_i \xrightarrow{\mathbb{P}} \mathbf{C}$ and $\frac{1}{N}\sum_i \boldsymbol{\xi}^\top y_i \xrightarrow{\mathbb{P}} \mathbf{A}$, so $\frac{1}{N}\mathbf{X}^\top\boldsymbol{\Xi} \xrightarrow{\mathbb{P}} \mathbf{A}^\top$. Hence by Slutsky's Theorem, the difference converges to

$$[\mathbf{C}\mathbf{B}^{-1}\mathbf{A}]^{-1}[\mathbf{C}\mathbf{B}^{-1}]\mathbf{0} = \mathbf{0}$$

where $\boldsymbol{\xi}_i\varepsilon_i$ is a $M \times 1$ vector. So is $\mathbf{0}$. □

Fun Facts 7.10 (Reminder: sizes K and M)**.** Let $K = \dim(\beta)$ be the number of explanatory variables in the structural equation (including endogenous and exogenous regressors). Let $M = \dim(\xi_i)$ be the number of instruments / moment conditions. Then $M = K$ corresponds to just identification, while $M > K$ corresponds to overidentification.

7.7 Simultaneous Equation Systems and System GMM

We now move beyond single–equation models and consider *simultaneous equation systems* (SES), where endogeneity arises not only from omitted variables, but from the joint determination of multiple outcomes. This is a common setup in economics (e.g., general equilibrium): variables that look like "regressors" in one equation are *endogenous objects* determined elsewhere in the system.

7.7.1 An Economic Example: Supply and Demand

What does simultaneity mean? In short, in an equation system, *the regressors in one equation might be responses in other equations.* Hence the notion of simultaneity is hallucination in a one–equation system.

A canonical setting in which simultaneous equation systems arise is a simple *supply and demand* model. Suppose a market is described by the two equations

$$\begin{aligned} \text{Demand:} \quad & Q_i^d = \alpha_0 + \alpha_1 P_i + u_i, \\ \text{Supply:} \quad & Q_i^s = \beta_0 + \beta_1 P_i + v_i, \end{aligned}$$

where P_i is the market price, Q_i is the traded quantity, and u_i, v_i are unobserved demand and supply shocks. Market clearing imposes the equilibrium condition

$$Q_i^d = Q_i^s = Q_i.$$

In equilibrium, both P_i and Q_i are determined *jointly* by the two equations. Solving the system yields

$$P_i = \pi_0 + \pi_1 u_i + \pi_2 v_i, \qquad Q_i = \rho_0 + \rho_1 u_i + \rho_2 v_i,$$

for some constants (π_j, ρ_j). As a result, the endogenous variable P_i is mechanically correlated with both structural errors u_i and v_i.

This source of endogeneity is fundamentally different from the omitted–variable problem. Even if all relevant covariates were observed and included, the equilibrium price P_i would still depend on unobserved shocks through the market–clearing condition. Formally,

$$\operatorname{Cov}(P_i, u_i) \neq 0 \qquad \text{and} \qquad \operatorname{Cov}(P_i, v_i) \neq 0,$$

so ordinary least squares applied to either equation is inconsistent.

To identify the structural parameters, we require variables that shift one equation without directly affecting the other. For example, weather conditions or input costs may shift supply but not demand, while income shifters may affect demand but not supply. These variables serve as instruments that restore exogeneity once simultaneity is accounted for.

7.7.2 System setup

Suppose we observe G equations indexed by $g = 1, \ldots, G$, and N i.i.d. observations indexed by $i = 1, \ldots, N$. Each equation takes the form

$$y_{gi} = x_{gi}^\top \boldsymbol{\beta}_g + \varepsilon_{gi}, \qquad g = 1, \ldots, G,$$

where x_{gi} may include endogenous regressors due to simultaneity across equations.

Stacking equations for a given observation i, define

$$\mathbf{y}_i = \begin{pmatrix} y_{1i} \\ \vdots \\ y_{Gi} \end{pmatrix}, \qquad \boldsymbol{\varepsilon}_i = \begin{pmatrix} \varepsilon_{1i} \\ \vdots \\ \varepsilon_{Gi} \end{pmatrix}.$$

Let $\mathbf{X}_i$ denote the corresponding block–diagonal regressor matrix, and let $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \ldots, \boldsymbol{\beta}_G^\top)^\top$ collect all parameters.

For each equation g, let $\boldsymbol{\xi}_{gi}$ denote a vector of instruments. Collecting instruments across equations yields a stacked instrument vector $\boldsymbol{\xi}_i$.

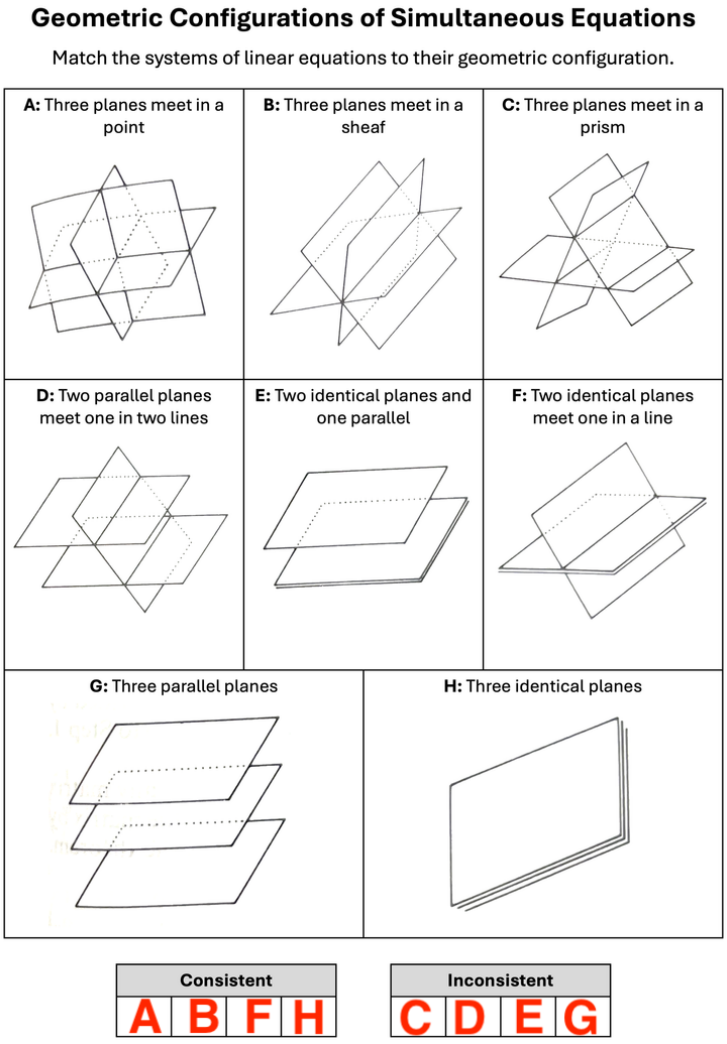

Figure 7: A geometric intuition for SESs. A is a SES with a unique solution. B,F,H are SESs with infinite many solutions. C,D,E,G are inconsistent equation systems, i.e. no solutions.

The system IV assumptions imply the moment restrictions

$$\mathbb{E}\left[\boldsymbol{\xi}_i \boldsymbol{\varepsilon}_i\right] = 0,$$

or equivalently,

$$\mathbb{E}\left[\psi_i(\boldsymbol{\beta}_0)\right] = 0, \qquad \psi_i(\boldsymbol{\beta}) := \boldsymbol{\xi}_i(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}).$$

7.7.3 Validity Conditions in SESs

Before discussing identification, estimation, or efficiency, we must first ensure that the simultaneous equation system and the proposed instrumental variables define a *well–posed* econometric problem. In other words, prior to any optimization or asymptotic analysis, the setup itself must be internally consistent and meaningful.

In simultaneous equation systems, we use the term *validity* to describe precisely this requirement. Specifically, validity holds when the proposed instruments generate moment conditions that are **correct** (exogeneity), **informative** (relevance), and **sufficient** to identify the structural parameters, as ensured by the order and rank conditions (typically implemented through exclusion restrictions).

(1) Order condition. For each equation g, let K_g denote the number of regressors and M_g the number of instruments. A necessary condition for estimation is

$$M_g \geq K_g,$$

so that the number of moment conditions is at least as large as the number of unknown coefficients (just– or over–identified).

(2) Rank condition. Instruments must generate enough independent variation for identification. A common population version is that the $M_g \times K_g$ matrix $\mathbb{E}\left[\boldsymbol{\xi}_{gi} x_{gi}^\top\right]$ has full column rank K_g.

(3) Exogeneity. Instrument exogeneity requires

$$\mathbb{E}\left[\boldsymbol{\xi}_{gi}\, \varepsilon_{gi}\right] = 0 \qquad \text{for each } g,$$

or equivalently (stacked across equations) $\mathbb{E}\left[\boldsymbol{\xi}_i\, \boldsymbol{\varepsilon}_i\right] = 0$.

(4) Relevance. Instruments must be informative about the endogenous regressors. In population terms, this requires a non-degeneracy condition such as

$$\operatorname{rank}\big(\mathbb{E}\left[\boldsymbol{\xi}_{gi} x_{gi}^\top\right]\big) = K_g$$

(or $\mathbb{E}\left[\boldsymbol{\xi}_{gi}^\top x_{gi}\right] \neq 0$ in the scalar case). Without relevance, the parameters are not identified.

(5) Exclusion restrictions. Finally, identification relies on exclusion restrictions: some instruments affect one equation but do not enter others except through endogenous variables.

These restrictions provide the economic content that distinguishes structural systems from reduced-form correlations.

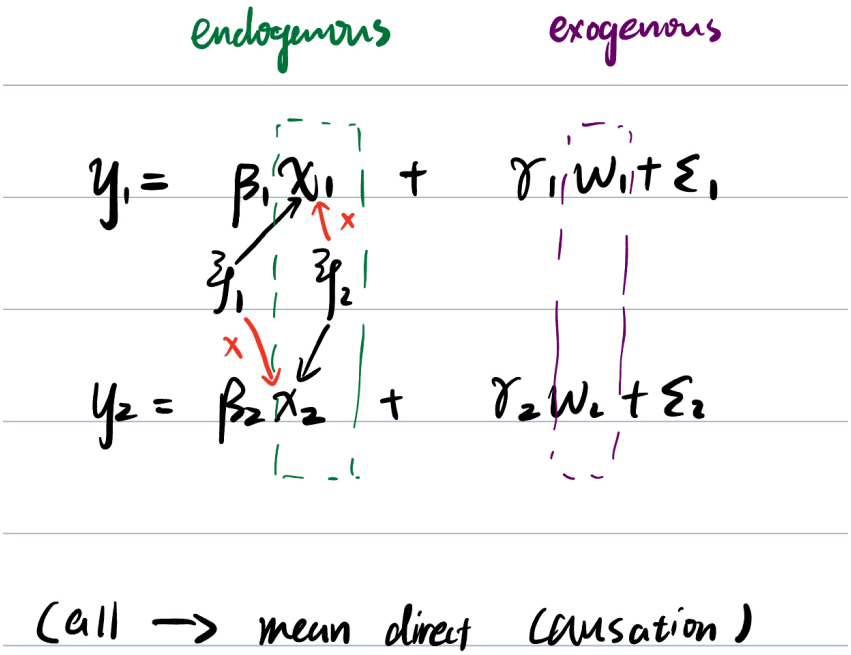

Figure 8: Exclusion. Black direct causations are allowed and red are prohibited. Otherwise, instruments are identical across equations, so the system can behave like everything is driven by a common cause $\boldsymbol{\xi}$, giving little leverage to disentangle $\boldsymbol{\beta}_1$ vs $\boldsymbol{\beta}_2$ (weak/failed system identification unless extra restrictions add separation).

7.7.4 Estimation under Homoskedasticity

We now turn to estimation of simultaneous equation systems under a set of simplifying variance assumptions. These assumptions allow us to derive a concrete form for the optimal weighting matrix and to connect system GMM directly to familiar estimators.

We begin with the benchmark case in which the structural errors are homoskedastic and uncorrelated across equations. Specifically, assume

$$\operatorname{Var}(\boldsymbol{\varepsilon}_i \mid \boldsymbol{\xi}_i) = \sigma^2 \mathbf{I}_G,$$

so that each equation has the same error variance and there is no cross–equation error correlation. Equivalently,

$$\operatorname{Cov}(\varepsilon_{pi}, \varepsilon_{qi} \mid \boldsymbol{\xi}_i) = 0 \qquad \text{for } p \neq q.$$

These assumptions are strong, but they allow us to isolate the role of instrument variation without additional complications from system–wide heteroskedasticity.

Recall that the optimal GMM weighting matrix is proportional to the inverse of the

variance of the moment function:

$$\mathbf{W} \;\propto\; \mathrm{Var}\left(\psi_i(\boldsymbol{\beta}_0)\right)^{-1}.$$

Under the homoskedasticity assumption,

$$\psi_i(\boldsymbol{\beta}_0) = \boldsymbol{\xi}_i \boldsymbol{\varepsilon}_i, \qquad \mathrm{Var}\left(\psi_i(\boldsymbol{\beta}_0)\right) = \mathbb{E}\left[\boldsymbol{\xi}_i\, \mathbb{E}\left[\boldsymbol{\varepsilon}_i \boldsymbol{\varepsilon}_i^\top \mid \boldsymbol{\xi}_i\right]\, \boldsymbol{\xi}_i^\top\right] = \sigma^2 \mathbb{E}\left[\boldsymbol{\xi}_i \boldsymbol{\xi}_i^\top\right].$$

Up to the scalar σ^2, the optimal weight is therefore

$$\mathbf{W} \;\propto\; \mathbb{E}\left[\boldsymbol{\xi}_i \boldsymbol{\xi}_i^\top\right]^{-1}.$$

Replacing the population expectation with its sample analog yields the empirical weighting matrix

$$\mathbf{W}_N = (\boldsymbol{\Xi}^\top \boldsymbol{\Xi})^{-1} = \Big(\sum_{i=1}^{N} \boldsymbol{\xi}_i \boldsymbol{\xi}_i^\top\Big)^{-1},$$

where $\boldsymbol{\Xi}$ stacks $\{\boldsymbol{\xi}_i\}_{i=1}^N$ in the usual way.

With this choice of weighting matrix, the system GMM estimator becomes

$$\hat{\boldsymbol{\beta}}_{SGMM} = \left(\mathbf{X}^\top \boldsymbol{\Xi} (\boldsymbol{\Xi}^\top \boldsymbol{\Xi})^{-1} \boldsymbol{\Xi}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \boldsymbol{\Xi} (\boldsymbol{\Xi}^\top \boldsymbol{\Xi})^{-1} \boldsymbol{\Xi}^\top \mathbf{y}.$$

This estimator coincides with the system two–stage least squares estimator, and reduces to equation–by–equation 2SLS when the system structure is ignored.

Under homoskedasticity and no cross–equation correlation, we recover exactly the 2SLS GMM estimator. System GMM does not exploit additional efficiency gains from joint estimation. In this case, system estimation is essentially an extension of the normal equations for 2SLS to the stacked system setting.

7.7.5 Estimation under General Error Covariance

We now relax the homoskedastic benchmark and allow the structural errors to have a general cross–equation covariance matrix.

Assume that for each observation i,

$$\mathrm{Var}\left(\boldsymbol{\varepsilon}_i \mid \boldsymbol{\xi}_i\right) = \boldsymbol{\Sigma},$$

where $\boldsymbol{\Sigma}$ is a $G \times G$ positive definite matrix.

For each equation g, instrument exogeneity implies

$$\mathbb{E}\left[\boldsymbol{\xi}_{gi}\, \varepsilon_{gi}\right] = 0.$$

Stacking across equations, write the system orthogonality restriction as

$$\mathbb{E}\left[\boldsymbol{\xi}_i\, \boldsymbol{\varepsilon}_i\right] = 0,$$

where $\boldsymbol{\xi}_i$ collects all instruments at observation i and $\boldsymbol{\varepsilon}_i = (\varepsilon_{1i}, \ldots, \varepsilon_{Gi})^\top$.

Define the equation–wise sample moments

$$\psi_{gN}(\boldsymbol{\beta}) := \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{\xi}_{gi}\big(y_{gi} - x_{gi}^{\top}\boldsymbol{\beta}_g\big) \in \mathbb{R}^{M_g},$$

and stack them into the system moment vector

$$\psi_N(\boldsymbol{\beta}) := \begin{pmatrix} \psi_{1N}(\boldsymbol{\beta}) \\ \vdots \\ \psi_{GN}(\boldsymbol{\beta}) \end{pmatrix} \in \mathbb{R}^{M}, \qquad M := \sum_{g=1}^{G} M_g.$$

Equivalently, in compact matrix form,

$$\psi_N(\boldsymbol{\beta}) = \frac{1}{N}\,\boldsymbol{\Xi}^{\top}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

Under the maintained variance assumption,

$$\operatorname{Var}\Big(\sqrt{N}\psi_N(\boldsymbol{\beta}_0)\Big) = \boldsymbol{\Omega}, \qquad \boldsymbol{\Omega} := \mathbb{E}\left[\boldsymbol{\xi}_i \boldsymbol{\Sigma} \boldsymbol{\xi}_i^{\top}\right] \in \mathbb{R}^{M\times M}.$$

In finite samples, a convenient sample analogue is

$$\boldsymbol{\Omega}_N = \frac{1}{N}\boldsymbol{\Xi}^{\top}(\boldsymbol{\Sigma}\otimes \mathbf{I}_N)\boldsymbol{\Xi}, \qquad \boldsymbol{\Omega}_N \in \mathbb{R}^{M\times M}.$$

Hence, in the overidentified case, we estimate $\boldsymbol{\beta}$ by minimizing

$$Q_N(\boldsymbol{\beta}) = \psi_N(\boldsymbol{\beta})^{\top}\mathbf{W}_N\psi_N(\boldsymbol{\beta}), \qquad \mathbf{W}_N \succ 0.$$

The efficient GMM logic suggests choosing

$$\mathbf{W}_N \;\propto\; \operatorname{Var}\Big(\sqrt{N}\,\psi_N(\boldsymbol{\beta}_0)\Big)^{-1},$$

so under the generalized covariance structure we take

$$\mathbf{W}_N = \Omega_N^{-1} = \Big(\frac{1}{N}\boldsymbol{\Xi}^{\top}(\boldsymbol{\Sigma}\otimes \mathbf{I}_N)\boldsymbol{\Xi}\Big)^{-1}.$$

The first–order condition $\nabla_{\boldsymbol{\beta}} Q_N(\hat{\boldsymbol{\beta}}) = 0$ yields the closed form

$$\hat{\boldsymbol{\beta}}_{SGMM} = \big(\mathbf{X}^{\top}\boldsymbol{\Xi}\mathbf{W}_N\boldsymbol{\Xi}^{\top}\mathbf{X}\big)^{-1}\mathbf{X}^{\top}\boldsymbol{\Xi}\mathbf{W}_N\boldsymbol{\Xi}^{\top}\mathbf{y}.$$

Using $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon}$, we can also write

$$\hat{\boldsymbol{\beta}}_{SGMM} = \boldsymbol{\beta}_0 + \big(\mathbf{X}^{\top}\boldsymbol{\Xi}\mathbf{W}_N\boldsymbol{\Xi}^{\top}\mathbf{X}\big)^{-1}\mathbf{X}^{\top}\boldsymbol{\Xi}\mathbf{W}_N\boldsymbol{\Xi}^{\top}\boldsymbol{\varepsilon}.$$

7.7.6 Feasible estimation via two–step weighting.

The practical complication is that Σ (hence $\mathbf{W}_N$) is unknown. A standard feasible procedure is:

Algorithm 8 Feasible two–step System GMM (unknown $\boldsymbol{\Sigma}$)

Require: Data $\{(\mathbf{y}_i, \mathbf{X}_i, \boldsymbol{\xi}_i)\}_{i=1}^N$; initial weight $\mathbf{W}_N^{(0)} = \mathbf{I}_M$

1: Compute an initial consistent estimator

$$\hat{\boldsymbol{\beta}}^{(0)} = \arg\min_{\boldsymbol{\beta}} \ \psi_N(\boldsymbol{\beta})^\top \mathbf{W}_N^{(0)} \psi_N(\boldsymbol{\beta}).$$

2: For $i = 1, \ldots, N$, form residuals

$$\hat{\boldsymbol{\varepsilon}}_i = \mathbf{y}_i - \mathbf{X}_i \hat{\boldsymbol{\beta}}^{(0)}.$$

3: Use $\{\hat{\boldsymbol{\varepsilon}}_i\}_{i=1}^N$ to construct an estimator $\hat{\Sigma}$, then set

$$\hat{\mathbf{W}}_N = \Big(\frac{1}{N} \boldsymbol{\Xi}^\top (\hat{\Sigma} \otimes \mathbf{I}_N) \boldsymbol{\Xi} \Big)^{-1}.$$

4: Re-estimate

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \ \psi_N(\boldsymbol{\beta})^\top \hat{\mathbf{W}}_N \psi_N(\boldsymbol{\beta}).$$

5: **return** $\hat{\boldsymbol{\beta}}$

7.7.7 Properties of System GMM

The formula is ugly, yes. Fortunately, we do not need to love it; we only need to understand its behavior. We therefore study the usual trilogy: finite-sample bias, consistency, and asymptotic normality.

Recall that the (linear) system GMM estimator can be written as

$$\hat{\boldsymbol{\beta}}_{SGMM} = \boldsymbol{\beta}_0 + \big(\mathbf{X}^\top \boldsymbol{\Xi} \mathbf{W}_N \boldsymbol{\Xi}^\top \mathbf{X} \big)^{-1} \mathbf{X}^\top \boldsymbol{\Xi} \mathbf{W}_N \boldsymbol{\Xi}^\top \boldsymbol{\varepsilon},$$

where $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_1^\top, \ldots, \boldsymbol{\varepsilon}_N^\top)^\top$ stacks the system errors, and $\mathbf{W}_N$ is a positive definite weighting matrix.

Finite-sample unbiasedness. In finite samples, system GMM is *not* unbiased in general. The key point is subtle but familiar from single-equation IV.

Although we impose the moment condition

$$\mathbb{E}\left[\boldsymbol{\xi}_{gi} \, \varepsilon_{gi} \right] = 0,$$

this does *not* imply (only the inverse direction holds via LIE)

$$\mathbb{E}\left[\varepsilon_{gi} \mid \boldsymbol{\xi}_{gi} \right] = 0.$$

As a result, the matrix $\big(\mathbf{X}^\top \boldsymbol{\Xi} \mathbf{W}_N \boldsymbol{\Xi}^\top \mathbf{X} \big)^{-1}$ is correlated with the term $\mathbf{X}^\top \boldsymbol{\Xi} \mathbf{W}_N \boldsymbol{\Xi}^\top \boldsymbol{\varepsilon}$ in finite samples. Hence $\mathbb{E}\left[\hat{\boldsymbol{\beta}}_{SGMM} \right] \neq \boldsymbol{\beta}_0$ in general. This mirrors the finite-sample bias of 2SLS: orthogonality of instruments and errors does not eliminate the correlation induced by projection and matrix inversion.

Consistency. Despite finite-sample bias, $\hat{\boldsymbol{\beta}}_{SGMM}$ is consistent under standard regularity conditions.

Write

$$\hat{\boldsymbol{\beta}}_{SGMM} = \boldsymbol{\beta}_0 + \Big[\Big(\frac{1}{N}\mathbf{X}^\top\boldsymbol{\Xi}\Big)\mathbf{W}_N\Big(\frac{1}{N}\boldsymbol{\Xi}^\top\mathbf{X}\Big)\Big]^{-1}\Big(\frac{1}{N}\mathbf{X}^\top\boldsymbol{\Xi}\Big)\mathbf{W}_N\Big(\frac{1}{N}\boldsymbol{\Xi}^\top\boldsymbol{\varepsilon}\Big).$$

Under the maintained assumptions,

$$\frac{1}{N}\mathbf{X}^\top\boldsymbol{\Xi} \xrightarrow{\mathbb{P}} \mathbf{A}, \qquad \mathbf{W}_N \xrightarrow{\mathbb{P}} \mathbf{W}, \qquad \frac{1}{N}\boldsymbol{\Xi}^\top\boldsymbol{\varepsilon} \xrightarrow{\mathbb{P}} \mathbf{0},$$

we obtain

$$\hat{\boldsymbol{\beta}}_{SGMM} \xrightarrow{\mathbb{P}} \boldsymbol{\beta}_0.$$

Thus, system GMM consistently recovers the structural parameters as the sample size grows, even though finite-sample bias remains.

Asymptotic normality. By a multivariate CLT,

$$\sqrt{N}\,\psi_N(\boldsymbol{\beta}_0) = \sqrt{N}\Big(\frac{1}{N}\boldsymbol{\Xi}^\top\boldsymbol{\varepsilon}\Big) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\boldsymbol{\xi}_i\varepsilon_i \xrightarrow{d} \mathcal{N}(\mathbf{0},\Omega),$$

where

$$\Omega := \mathrm{Var}\left(\boldsymbol{\xi}_i\varepsilon_i\right).$$

Let

$$\mathbf{A} := \operatorname{plim}\frac{1}{N}\mathbf{X}^\top\boldsymbol{\Xi}.$$

A standard linearization of the GMM first–order condition yields

$$\sqrt{N}(\hat{\boldsymbol{\beta}}_{SGMM} - \boldsymbol{\beta}_0) = (\mathbf{A}\mathbf{W}\mathbf{A}^\top)^{-1}\mathbf{A}\mathbf{W}\,\sqrt{N}\,\psi_N(\boldsymbol{\beta}_0) + o_p(1).$$

Therefore,

$$\sqrt{N}(\hat{\boldsymbol{\beta}}_{SGMM} - \boldsymbol{\beta}_0) \xrightarrow{d} \mathcal{N}\Big(\mathbf{0},\,(\mathbf{A}\mathbf{W}\mathbf{A}^\top)^{-1}\mathbf{A}\mathbf{W}\Omega\mathbf{W}\mathbf{A}^\top(\mathbf{A}\mathbf{W}\mathbf{A}^\top)^{-1}\Big).$$

If the optimal weighting is used, i.e. $\mathbf{W} \propto \Omega^{-1}$ (and we take $\mathbf{W} = \Omega^{-1}$ for simplicity), the asymptotic variance simplifies to

$$(\mathbf{A}\Omega^{-1}\mathbf{A}^\top)^{-1}.$$

7.8 Three–Stage Least Squares (3SLS)

Three–Stage Least Squares (3SLS) combines system–wide instrumental variables with feasible generalized least squares, exploiting both endogenous regressor structure and cross–equation error correlation.

Algorithm 9 Three–Stage Least Squares (3SLS)

1: **Input:** data $\{(\mathbf{y}_i, \mathbf{X}_i, \boldsymbol{\Xi}_i)\}_{i=1}^{N}$; system specification; instrument matrix $\boldsymbol{\Xi}$.

2: **Stage 1 (First-stage projection):** compute

$$\hat{\mathbf{X}} \;:=\; \boldsymbol{\Xi}(\boldsymbol{\Xi}^\top\boldsymbol{\Xi})^{-1}\boldsymbol{\Xi}^\top\mathbf{X}.$$

3: **Stage 2 (Estimate system error covariance):** obtain a preliminary estimate (e.g. equation-by-equation 2SLS using $\hat{\mathbf{X}}$), form residuals $\hat{\boldsymbol{\varepsilon}}_i$, and estimate

$$\hat{\Sigma} \;:=\; \frac{1}{N}\sum_{i=1}^{N}\hat{\boldsymbol{\varepsilon}}_i\hat{\boldsymbol{\varepsilon}}_i^\top.$$

4: **Stage 3 (Feasible GLS on the system):** compute

$$\hat{\boldsymbol{\beta}}_{3SLS} \;:=\; \big(\hat{\mathbf{X}}^\top(\hat{\Sigma}\otimes\mathbf{I}_N)^{-1}\hat{\mathbf{X}}\big)^{-1}\hat{\mathbf{X}}^\top(\hat{\Sigma}\otimes\mathbf{I}_N)^{-1}\mathbf{y}.$$

5: **Output:** $\hat{\boldsymbol{\beta}}_{3SLS}$.

Here's what we are actually doing. First and foremost, let the reduced–form for the stacked endogenous regressors be

$$\mathbf{X} = \boldsymbol{\Xi}\boldsymbol{\delta} + \mathbf{u},$$

where $\boldsymbol{\Xi}$ collects all instruments used in the system. The first–stage estimator is

$$\hat{\boldsymbol{\delta}} = (\boldsymbol{\Xi}^\top\boldsymbol{\Xi})^{-1}\boldsymbol{\Xi}^\top\mathbf{X}, \qquad \hat{\mathbf{X}} = \boldsymbol{\Xi}(\boldsymbol{\Xi}^\top\boldsymbol{\Xi})^{-1}\boldsymbol{\Xi}^\top\mathbf{X}.$$

We further make the following assumptions on covariances: let $\boldsymbol{\varepsilon}_i = (\varepsilon_{1i}, \ldots, \varepsilon_{Gi})^\top$ denote the vector of structural errors. Assume

$$\mathrm{Cov}(\varepsilon_{gi}, \varepsilon_{hj} \mid \boldsymbol{\Xi}_i) = \begin{cases} \sigma_{gh}, & i = j, \\ 0, & i \neq j, \end{cases}$$

so that

$$\mathrm{Var}\,(\boldsymbol{\varepsilon}_i \mid \boldsymbol{\Xi}_i) = \boldsymbol{\Sigma}, \qquad \mathrm{Var}\,(\boldsymbol{\varepsilon}) = \boldsymbol{\Sigma}\otimes\mathbf{I}_N.$$

So we can apply GLS to the system equation

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

with covariance matrix $\Sigma\otimes\mathbf{I}_N$, yields the 3SLS estimator

$$\hat{\boldsymbol{\beta}}_{3SLS} = \big(\mathbf{X}^\top(\boldsymbol{\Sigma}\otimes\mathbf{I}_N)^{-1}\mathbf{X}\big)^{-1}\mathbf{X}^\top(\boldsymbol{\Sigma}\otimes\mathbf{I}_N)^{-1}\mathbf{y}.$$

Substituting the first–stage fitted regressors gives the operational form

$$\hat{\boldsymbol{\beta}}_{3SLS} = \big(\hat{\mathbf{X}}^\top(\boldsymbol{\Sigma}\otimes\mathbf{I}_N)^{-1}\hat{\mathbf{X}}\big)^{-1}\hat{\mathbf{X}}^\top(\boldsymbol{\Sigma}\otimes\mathbf{I}_N)^{-1}\mathbf{y}.$$

Explicitly, you can write it as

$$\hat{\boldsymbol{\beta}}_{3SLS} = \boldsymbol{\beta}_0 + \left(\hat{\mathbf{X}}^\top (\Sigma \otimes \mathbf{I}_N)^{-1} \hat{\mathbf{X}}\right)^{-1} \hat{\mathbf{X}}^\top (\Sigma \otimes \mathbf{I}_N)^{-1} \boldsymbol{\varepsilon}.$$

As you might recall, the analyses of unbiasedness and consistency both hinges on the behavior of the term

$$\frac{1}{N} \hat{\mathbf{X}}^\top (\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I}_N) \boldsymbol{\varepsilon}.$$

Since $\hat{\mathbf{X}} = \mathbf{P}_{\boldsymbol{\Xi}} \mathbf{X}$ and $\mathbf{P}_{\boldsymbol{\Xi}}$ depends on $\boldsymbol{\Xi}$, it is natural to first analyze the simpler object that carries the exogeneity:

$$\frac{1}{N} \sum_{i=1}^{N} \boldsymbol{\Xi}_i^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\varepsilon}_i. \tag{4}$$

Finite-sample bias. Even if the population orthogonality restriction $\mathbb{E}[\boldsymbol{\Xi}_i \boldsymbol{\varepsilon}_i] = 0$ holds, 3SLS is typically not unbiased in finite samples. The reason is the same "generated regressor" issue as in 2SLS: the GLS score term uses $\hat{\mathbf{X}}$, and $\hat{\mathbf{X}}$ is constructed from the same sample.

Write $\boldsymbol{\Omega} := \boldsymbol{\Sigma} \otimes \mathbf{I}_N$. By definition,

$$\hat{\boldsymbol{\beta}}_{3SLS} = \left(\hat{\mathbf{X}}^\top \boldsymbol{\Omega}^{-1} \hat{\mathbf{X}}\right)^{-1} \hat{\mathbf{X}}^\top \boldsymbol{\Omega}^{-1} \mathbf{y}, \qquad \mathbf{y} = \mathbf{X}\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon}.$$

Substituting $\mathbf{y}$ gives the decomposition

$$\hat{\boldsymbol{\beta}}_{3SLS} - \boldsymbol{\beta}_0 = \left(\hat{\mathbf{X}}^\top \Omega^{-1} \hat{\mathbf{X}}\right)^{-1} \hat{\mathbf{X}}^\top \boldsymbol{\Omega}^{-1} \boldsymbol{\varepsilon}.$$

Taking conditional expectations given $\boldsymbol{\Xi}$,

$$\mathbb{E}\left[\hat{\boldsymbol{\beta}}_{3SLS} \mid \boldsymbol{\Xi}\right] = \boldsymbol{\beta}_0 \quad \Longleftrightarrow \quad \mathbb{E}\left[\left(\hat{\mathbf{X}}^\top \boldsymbol{\Omega}^{-1} \hat{\mathbf{X}}\right)^{-1} \hat{\mathbf{X}}^\top \boldsymbol{\Omega}^{-1} \boldsymbol{\varepsilon} \;\middle|\; \boldsymbol{\Xi}\right] = \mathbf{0}.$$

A sufficient (but not necessary) condition for the right-hand side to be zero would be

$$\mathbb{E}\left[\hat{\mathbf{X}}^\top \Omega^{-1} \boldsymbol{\varepsilon} \mid \boldsymbol{\Xi}\right] = \mathbf{0},$$

but this typically fails in finite samples because $\hat{\mathbf{X}}$ depends on the endogenous components of $\mathbf{X}$.

Indeed, $\hat{\mathbf{X}} = \mathbf{P}_{\boldsymbol{\Xi}} \mathbf{X}$ with

$$\mathbf{P}_{\boldsymbol{\Xi}} := \boldsymbol{\Xi}(\boldsymbol{\Xi}^\top \boldsymbol{\Xi})^{-1} \boldsymbol{\Xi}^\top,$$

and the reduced form for the regressors can be written as $\mathbf{X} = \boldsymbol{\Xi}\boldsymbol{\delta} + \mathbf{U}$, where $\mathbf{U}$ contains the endogenous variation. Hence

$$\hat{\mathbf{X}} = \mathbf{P}_{\boldsymbol{\Xi}} \mathbf{X} = \boldsymbol{\Xi}\boldsymbol{\delta} + \mathbf{P}_{\boldsymbol{\Xi}} \mathbf{U}.$$

Plugging this into the score term yields

$$\hat{\mathbf{X}}^\top \boldsymbol{\Omega}^{-1} \boldsymbol{\varepsilon} = (\boldsymbol{\Xi}\boldsymbol{\delta})^\top \boldsymbol{\Omega}^{-1} \boldsymbol{\varepsilon} + (\mathbf{P}_{\boldsymbol{\Xi}} \mathbf{U})^\top \boldsymbol{\Omega}^{-1} \boldsymbol{\varepsilon}.$$

The first term has conditional mean zero under instrument exogeneity. The second term is the problem: $\mathbf{P}_{\boldsymbol{\Xi}} \mathbf{U}$ is a random function of the same sample and is generally correlated

with $\boldsymbol{\varepsilon}$, so in finite samples

$$\mathbb{E}\left[(\mathbf{P}_{\boldsymbol{\Xi}}\mathbf{U})^\top \Omega^{-1}\boldsymbol{\varepsilon} \mid \boldsymbol{\Xi}\right] \neq \mathbf{0},$$

and therefore $\mathbb{E}\left[\hat{\boldsymbol{\beta}}_{3SLS} \mid \boldsymbol{\Xi}\right] \neq \boldsymbol{\beta}_0$ in general. Intuitively, projecting $\mathbf{X}$ onto instruments does not remove the sampling noise coming from $\mathbf{U}$; it only reshapes it, and GLS then mixes errors across equations via $\boldsymbol{\Omega}^{-1}$.

Consistency. Under standard LLN conditions,

$$\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{\Xi}_i^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\varepsilon}_i \xrightarrow{p} \mathbb{E}\left[\boldsymbol{\Xi}_i^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\varepsilon}_i\right].$$

Hence a sufficient condition for consistency is

$$\mathbb{E}\left[\boldsymbol{\Xi}_i^\top \Sigma^{-1}\boldsymbol{\varepsilon}_i\right] = 0. \tag{5}$$

To see what (5) really requires, we analyze it in a simple two-equation setup. Consider $G = 2$ and suppose the instrument vector is equation-specific, e.g.

$$\boldsymbol{\Xi}_i = \begin{pmatrix} \boldsymbol{\Xi}_{1i} \\ \boldsymbol{\Xi}_{2i} \end{pmatrix}, \qquad \boldsymbol{\varepsilon}_i = \begin{pmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \end{pmatrix}, \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}.$$

Then

$$\boldsymbol{\Sigma}^{-1} = \begin{pmatrix} \omega_{11} & \omega_{12} \\ \omega_{21} & \omega_{22} \end{pmatrix}.$$

Now compute the "mixed" error:

$$\boldsymbol{\Sigma}^{-1}\boldsymbol{\varepsilon}_i = \begin{pmatrix} \omega_{11}\varepsilon_{1i} + \omega_{12}\varepsilon_{2i} \\ \omega_{21}\varepsilon_{1i} + \omega_{22}\varepsilon_{2i} \end{pmatrix}.$$

Therefore, for equation 1's instruments, the moment implied by (5) contains

$$\mathbb{E}\left[\boldsymbol{\Xi}_{1i}\big(\omega_{11}\varepsilon_{1i} + \omega_{12}\varepsilon_{2i}\big)\right] = 0.$$

Even if single-equation exogeneity holds,

$$\mathbb{E}\left[\boldsymbol{\Xi}_{1i}\varepsilon_{1i}\right] = 0,$$

this is not enough when $\omega_{12} \neq 0$, because we also need

$$\mathbb{E}\left[\boldsymbol{\Xi}_{1i}\varepsilon_{2i}\right] = 0.$$

Symmetrically, equation 2 requires both $\mathbb{E}\left[\boldsymbol{\Xi}_{2i}\varepsilon_{2i}\right] = 0$ and $\mathbb{E}\left[\boldsymbol{\Xi}_{2i}\varepsilon_{1i}\right] = 0$ when $\omega_{21} \neq 0$. Because Σ^{-1} generally has nonzero off-diagonal entries, the GLS step in 3SLS forms linear combinations of errors across equations. As a result, to force the key term (4) to have mean zero, it is not sufficient to assume only equation-by-equation exogeneity. Instead, we need the *system exogeneity* requirement:

$$\mathbb{E}\left[\boldsymbol{\Xi}_{gi}\varepsilon_{hi}\right] = 0 \qquad \text{for all pairs } (g, h),$$

i.e., *all instruments must be exogenous with respect to all structural errors in all equations.*

7.9 Testing Moment Conditions

Once estimation is complete, we still need to ask whether the moment restrictions we imposed are actually compatible with the data. Recall the system moment conditions

$$\mathbb{E}\left[\psi_i(\boldsymbol{\beta}_0)\right] = 0, \qquad \psi_i(\boldsymbol{\beta}) = \boldsymbol{\xi}_i\big(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}\big),$$

where $\dim(\psi_i) = M$ and $\dim(\boldsymbol{\beta}) = K$. When $M > K$, the model is *overidentified*. We use the extra moments to construct specification tests.

7.9.1 Hansen's J–test.

Recall the GMM objective function

$$Q_N(\boldsymbol{\beta}) = \psi_N(\boldsymbol{\beta})^\top \mathbf{C}_N \psi_N(\boldsymbol{\beta}), \qquad \psi_N(\boldsymbol{\beta}) = \frac{1}{N}\sum_{i=1}^{N}\psi_i(\boldsymbol{\beta}).$$

Let $\hat{\boldsymbol{\beta}}$ denote the efficient GMM estimator using the optimal weighting matrix $\mathbf{C}_N$, and define

$$\hat{Q}_N = Q_N(\hat{\boldsymbol{\beta}}).$$

The Hansen J–statistic is

$$J = N\hat{Q}_N.$$

Under the null hypothesis

$$H_0: \quad \mathbb{E}\left[\psi_i(\boldsymbol{\beta}_0)\right] = 0 \qquad \text{(all moment conditions are valid)},$$

we have the asymptotic distribution

$$J \xrightarrow{d} \chi^2_{M-K}.$$

A small J indicates that the moment conditions are mutually consistent with the data. A large J implies that at least one moment condition is violated, though the test does not reveal which one.

7.9.2 The C–test.

Often we are interested in testing only a subset of instruments. Partition the moment vector as

$$\psi = \begin{pmatrix} \psi^{(a)} \\ \psi^{(b)} \end{pmatrix},$$

where $\psi^{(a)}$ are baseline moments and $\psi^{(b)}$ are additional moments to be tested.

Let $\hat{Q}_N^{\text{full}}$ denote the GMM criterion using all moments, and $\hat{Q}_N^{\text{res}}$ the criterion using only $\psi^{(a)}$. Define the C–statistic

$$C = N\big(\hat{Q}_N^{\text{res}} - \hat{Q}_N^{\text{full}}\big).$$

Under the null hypothesis

$$H_0: \quad \mathbb{E}\left[\psi_i^{(b)}(\boldsymbol{\beta}_0)\right] = 0,$$

we have

$$C \xrightarrow{d} \chi^2_{\dim(\psi^{(b)})}.$$

The chi–square limit arises because: (i) the GMM objective is quadratic in sample moments; (ii) the optimal weighting matrix satisfies $\mathbf{C}_N \approx \mathrm{Var}(\sqrt{N}\psi_N)^{-1}$; and (iii) $\psi_N = O_p(N^{-1/2})$, so scaling by N yields a non–degenerate limit.

7.10 Summary

GMM is what you do when you refuse to write down a full likelihood but still want an estimator with a straight face. You assume some moment conditions

$$\mathbb{E}\left[\psi_i(\boldsymbol{\beta}_0)\right] = 0,$$

replace expectations by sample averages,

$$\bar{\psi}_N(\boldsymbol{\beta}) = \frac{1}{N}\sum_{i=1}^{N} \psi_i(\boldsymbol{\beta}),$$

and pick $\boldsymbol{\beta}$ to make $\bar{\psi}_N(\boldsymbol{\beta})$ "as close to zero as possible" in a weighted quadratic sense:

$$Q_N(\boldsymbol{\beta}) = \bar{\psi}_N(\boldsymbol{\beta})^\top \mathbf{W}_N \, \bar{\psi}_N(\boldsymbol{\beta}), \qquad \mathbf{W}_N \succ 0.$$

Identification.

- M<K: underidentified (good luck),
- M=K: just identified (no tests),
- M>K: overidentified (now you can test your life choices).

Weights. $\mathbf{W}_N$ decides which moments get taken seriously when they disagree. Efficiency says the "right" choice is

$$\mathbf{W} \;\propto\; \mathbf{S}^{-1}, \qquad \mathbf{S} := \mathbb{E}\left[\psi_i(\boldsymbol{\beta}_0)\psi_i(\boldsymbol{\beta}_0)^\top\right],$$

so in practice you do two–step GMM: start with something lazy (often $\mathbf{I}_M$), get $\hat{\boldsymbol{\beta}}^{(0)}$, estimate $\mathbf{S}$, set $\hat{\mathbf{W}}_N = \hat{\mathbf{S}}^{-1}$, re–estimate.

IV/2SLS are just GMM in a trench coat. Linear IV uses $\psi_i(\boldsymbol{\beta}) = \boldsymbol{\xi}_i(\mathbf{y}_i - \mathbf{x}_i^\top\boldsymbol{\beta})$. With homoskedasticity, choosing $\mathbf{W}_N = (\boldsymbol{\Xi}^\top\boldsymbol{\Xi})^{-1}$ reproduces 2SLS. Systems just stack everything and make the covariance uglier.

Finite samples vs. asymptotics. No, it's generally not unbiased in finite samples. Yes, it's consistent under standard validity + relevance conditions. Life continues.

Testing (the payoff of $M > K$). Hansen's J–test uses

$$J := NQ_N(\hat{\boldsymbol{\beta}}) \xrightarrow{d} \chi^2_{M-K}$$

under valid moments. If J is large, at least one moment is lying; the test just won't tell you which one. The C–test lets you blame a subset.

Anyway, GMM got us a lot while committing to very little: just a handful of moment restrictions. MLE now takes the opposite bargain: specify a full parametric model for the data and optimize the likelihood. When that model is right, the payoff is sharp efficiency statements (via scores and Fisher information); when it's wrong, at least the failure is systematic.

8 Maximum Likelihood Estimation

In the previous chapter, we approached estimation through moment conditions. OLS, IV, and their generalization via the **Generalized Method of Moments (GMM)** all rely on a common idea: identify parameters by enforcing orthogonality restrictions that should hold in the population, without committing to a full probabilistic model of the data. This is appealing—one can do quite a lot while knowing remarkably little about the data-generating process.

Maximum Likelihood Estimation (MLE) takes the opposite stance. Instead of asking for just a few moments to behave, MLE demands a complete parametric description of the data-generating process and then asks a blunt question: under which parameter values would the observed data be most likely? When the model is correctly specified, this extra commitment pays off—MLE uses all available information and delivers sharp notions of score, Fisher information, and efficiency. When the model is wrong, of course, it pays off differently. In this chapter, we develop the basic principles of MLE, illustrate them through simple examples, and study the resulting large-sample properties.

8.1 A Binomial Example

We begin with a simple example that illustrates the core idea of maximum likelihood estimation in its most transparent form. (For this example we write the sample size as n, in the usual binomial notation; later in the chapter we revert to N.)

Suppose $X_1, \ldots, X_n$ are i.i.d. Bernoulli random variables with success probability $p \in (0, 1)$:

$$X_i \sim \text{Bernoulli}(p), \qquad i = 1, \ldots, n.$$

Equivalently, $\sum_{i=1}^{n} X_i \sim \text{Binomial}(n, p)$.

Given an observed sample $(X_1, \ldots, X_n)$, the likelihood function is the joint pmf evaluated at the realized data, viewed as a function of p:

$$L(p) = \prod_{i=1}^{n} p^{X_i}(1-p)^{1-X_i} = p^{\sum_{i=1}^{n} X_i}(1-p)^{n-\sum_{i=1}^{n} X_i}.$$

It is worth keeping the perspective straight: after we observe the data, $L(p)$ is *not* a probability distribution over p; it is a criterion for comparing which values of p make the observed sample most plausible under the model.

Taking logs,

$$\ell(p) \equiv \log L(p) = \sum_{i=1}^{n} X_i \log p + \left(n - \sum_{i=1}^{n} X_i\right) \log(1-p).$$

Differentiate with respect to p to obtain the score:

$$\frac{\partial \ell(p)}{\partial p} = \frac{\sum_{i=1}^{n} X_i}{p} - \frac{n - \sum_{i=1}^{n} X_i}{1-p}.$$

Setting the score equal to zero gives

$$\frac{\sum_{i=1}^n X_i}{p} = \frac{n - \sum_{i=1}^n X_i}{1-p},$$

and solving for p yields the MLE

$$\hat{p} = \frac{1}{n}\sum_{i=1}^n X_i.$$

Thus, in the Bernoulli model, maximum likelihood estimation reproduces the sample mean. This is a recurring theme: in simple models, MLE often recovers familiar estimators, while in more complex settings it provides a systematic way to build and analyze estimators from a fully specified probabilistic model.

8.2 Principles of MLE

8.2.1 Likelihood and Log-Likelihood

We now formalize the problem setup. Let $X_1, \ldots, X_N$ denote a random sample taking values in a sample space Ω.[1] Suppose the joint distribution of the sample is indexed by an unknown parameter $\theta \in \Theta \subset \mathbb{R}^k$.

Notation convention (total vs. average log-likelihood). It is often useful to distinguish the *total* log-likelihood from its *average*:

$$\mathcal{L}_N(\theta) \equiv f_{X_1,\ldots,X_N}(X_1, \ldots, X_N; \theta), \qquad \ell_N(\theta) \equiv \log \mathcal{L}_N(\theta), \qquad L_N(\theta) \equiv \frac{1}{N}\ell_N(\theta).$$

Maximizing $\ell_N(\theta)$ or $L_N(\theta)$ yields the same maximizer, but asymptotic arguments are typically written in terms of the average objective $L_N(\theta)$.

Definition 8.1 (Likelihood Function)**.** Given $(X_1, \ldots, X_N) \overset{i.i.d.}{\sim} F_X(\theta)$, the *likelihood function* is defined as a mapping

$$\mathcal{L}_N : \Theta \to \mathbb{R}_+,$$

given by the joint density/pmf of the observed sample evaluated at the realized data, viewed as a function of θ:

$$\mathcal{L}_N(\theta) \equiv f_{X_1,\ldots,X_N}(X_1, \ldots, X_N; \theta) = \prod_{i=1}^N f_X(X_i; \theta).$$

It is important to emphasize that, once the data are observed, $\mathcal{L}_N(\theta)$ is *not* a probability distribution over θ. Rather, it is a numerical device for comparing how well different parameter values explain the realized sample under the maintained model.

[1] In the binomial example we used n in the usual binomial notation; throughout the rest of this chapter we use N for the sample size.

Definition 8.2 (Log-Likelihood Function)**.** The *log-likelihood function* is

$$\ell_N(\theta) \equiv \log \mathcal{L}_N(\theta) = \sum_{i=1}^{N} \log f_X(X_i; \theta),$$

and the corresponding *average log-likelihood* is $L_N(\theta) \equiv \ell_N(\theta)/N$.

Maximizing the likelihood and maximizing the log-likelihood are equivalent, since the logarithm is strictly increasing. The log-likelihood replaces products with sums and greatly simplifies computation and analysis, so it is the standard object in likelihood-based inference.

At this point, estimation may look like a straightforward optimization routine: specify a model, write down a likelihood, maximize. However, before asking whether a likelihood can be maximized, we must ask whether different parameter values actually correspond to different distributions of the observable data. This leads to the notion of identification.

8.2.2 Identification

At a conceptual level, identification asks a simple question: do different parameter values generate different distributions of the observed data? If not, then no amount of data, nor clever optimization, can distinguish between them.

Definition 8.3 (Identification)**.** The parameter $\theta_0 \in \Theta$ is said to be *(point) identified* if

$$f_X(\cdot; \theta) = f_X(\cdot; \theta_0) \text{ a.s.} \quad \Longrightarrow \quad \theta = \theta_0.$$

In words: distinct parameter values must induce distinct distributions for a single draw X (and hence, under i.i.d. sampling, for the full sample).

We assume identifiability of our model in this chapter by default. If a model is not identified, the likelihood may be flat or have multiple maximizers, and the notion of a unique maximum likelihood estimator breaks down. In such cases, likelihood methods can at best recover an equivalence class of observationally indistinguishable parameters.

A canonical example arises in latent variable models. Consider a latent regression

$$y_i^* = \mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2),$$

where the latent variable y_i^* is not observed. Instead, we observe only

$$y_i = \mathbf{1}\{y_i^* > 0\}.$$

Then

$$\Pr(y_i = 1 \mid \mathbf{x}_i) = \Pr(\varepsilon_i > -\mathbf{x}_i^\top \boldsymbol{\beta}) = \Phi\left(\frac{\mathbf{x}_i^\top \boldsymbol{\beta}}{\sigma}\right),$$

where $\Phi(\cdot)$ denotes the standard normal CDF (standard context). The right-hand side depends on $(\boldsymbol{\beta}, \sigma)$ only through the ratio $\boldsymbol{\beta}/\sigma$. As a result, $(\boldsymbol{\beta}, \sigma)$ is not uniquely identified: scaling both parameters by the same positive constant leaves the distribution of the observed

data unchanged. Identification can be restored only by imposing a normalization, such as fixing $\sigma = 1$ (equivalently, fixing the scale of the latent index).

8.2.3 Score Function and Hessian

With the likelihood defined and identification clarified, we can now characterize the MLE through standard optimization conditions. Since asymptotic arguments are typically written in terms of the average objective, we define derivatives for $L_N(\theta) = \ell_N(\theta)/N$.

Definition 8.4 (Score Function)**.** The *(average) score function* is the gradient of the average log-likelihood:

$$S_N(\theta) \equiv \frac{\partial L_N(\theta)}{\partial \theta} = \frac{1}{N} \sum_{i=1}^{N} \frac{\partial}{\partial \theta} \log f_X(X_i; \theta).$$

The score is a random $k \times 1$ vector, as it depends on the realized sample. An interior maximizer $\hat{\theta}$ of $L_N(\theta)$ must satisfy the first-order condition

$$S_N(\hat{\theta}) = 0,$$

commonly called the *score equation*. In general, this equation need not admit a closed-form solution, and the MLE is often obtained numerically.

First-order conditions alone do not guarantee that a solution corresponds to a maximum. To characterize local curvature, we introduce the second derivative.

Definition 8.5 (Hessian)**.** The *Hessian matrix* of the average log-likelihood is

$$\mathcal{H}_N(\theta) \equiv \frac{\partial^2 L_N(\theta)}{\partial \theta \, \partial \theta^\top}.$$

A candidate solution $\hat{\theta}$ satisfying the score equation is a local maximizer if $\mathcal{H}_N(\hat{\theta})$ is negative definite. This second-order condition ensures the objective is locally concave around $\hat{\theta}$.

8.2.4 Invariance of the MLE

Before moving on to probabilistic properties of the MLE, it is useful to address a basic conceptual question: how does maximum likelihood behave under reparameterization? In empirical work, the same model is often expressed in different but equivalent parameterizations—for example, variance versus precision, scale versus log-scale, or structural parameters versus transformed quantities of interest. A reasonable estimator should not depend on how we choose to label the parameter.

One appealing feature of maximum likelihood estimation is a strong invariance property: estimating first and transforming later yields the same result as transforming first and estimating directly.

Theorem 8.1 (Invariance of the MLE)**.** Let $\theta \in \Theta$ be the original parameter, and let $\eta = g(\theta)$ be a reparameterization, where $g : \Theta \to \mathcal{H}$ is a one-to-one mapping. If $\hat{\theta}$ is the maximum likelihood estimator of θ, then the MLE of η is

$$\hat{\eta} = g(\hat{\theta}).$$

Proof. Under the reparameterization, the likelihood as a function of η can be written as

$$\mathcal{L}_N^{\eta}(\eta) = \mathcal{L}_N^{\theta}(g^{-1}(\eta)).$$

Since g is one-to-one, maximizing $\mathcal{L}_N^{\eta}(\eta)$ over η is equivalent to maximizing $\mathcal{L}_N^{\theta}(\theta)$ over θ. Therefore, if $\hat{\theta}$ maximizes $\mathcal{L}_N^{\theta}(\theta)$, then $\hat{\eta} = g(\hat{\theta})$ maximizes $\mathcal{L}_N^{\eta}(\eta)$. □

The takeaway is simple: for any one-to-one transformation g, the MLE of $g(\theta)$ is obtained by plugging the MLE of θ into g. Below is a standard example showing the equivalence of parameterizing a Gaussian likelihood using variance or precision.

Example 8.2 (Variance and Precision Parameterizations)**.** *Consider a Gaussian model*

$$X_1, \ldots, X_N \overset{i.i.d.}{\sim} \mathcal{N}(\mu, \sigma^2),$$

where the parameter of interest is the variance $\sigma^2 > 0$. An equivalent parameterization replaces variance by precision

$$\gamma \equiv \frac{1}{\sigma^2}.$$

Under the variance parameterization, the log-likelihood is

$$\ell_N(\mu, \sigma^2) = -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(X_i - \mu)^2.$$

Alternatively, expressed in terms of (μ, γ), it becomes

$$\ell_N(\mu, \gamma) = \frac{N}{2}\log\gamma - \frac{\gamma}{2}\sum_{i=1}^{N}(X_i - \mu)^2 - \frac{N}{2}\log(2\pi).$$

Maximizing either expression yields the same estimator for the underlying quantity. If $\hat{\sigma}^2$ denotes the MLE of σ^2, then the MLE of the precision parameter is

$$\hat{\gamma} = \frac{1}{\hat{\sigma}^2},$$

in accordance with invariance.

8.3 Asymptotics of Score Vectors and Fisher Information

Up to this point, the score and the Hessian were treated as finite-sample, data-dependent objects that characterize the maximizer of the likelihood. To conduct inference, however, we

must understand their probabilistic behavior under the true data-generating process. This leads naturally to studying expectations and variances of likelihood derivatives, and to the concept of Fisher information.

Throughout this subsection, let the *per-observation* score be

$$S_i(\theta) \equiv \frac{\partial}{\partial \theta} \log f_X(X_i; \theta),$$

and recall that we defined the *average* score and Hessian as derivatives of the average log-likelihood $L_N(\theta) = \frac{1}{N} \sum_{i=1}^{N} \log f_X(X_i; \theta)$:

$$S_N(\theta) = \frac{\partial L_N(\theta)}{\partial \theta} = \frac{1}{N} \sum_{i=1}^{N} S_i(\theta), \qquad \mathcal{H}_N(\theta) = \frac{\partial^2 L_N(\theta)}{\partial \theta \, \partial \theta^\top} = \frac{1}{N} \sum_{i=1}^{N} \mathcal{H}_i(\theta),$$

where $\mathcal{H}_i(\theta) \equiv \partial^2_{\theta\theta^\top} \log f_X(X_i; \theta)$.

We begin with a basic but fundamental property of the score.

Theorem 8.3 (Mean of score vectors)**.** Suppose regularity conditions hold so that differentiation under the integral sign is valid at θ_0 (e.g., Leibniz' rule applies and boundary terms vanish).[a] Then

$$\mathbb{E}\left[S_i(\theta_0)\right] = 0, \qquad \text{and hence} \qquad \mathbb{E}\left[S_N(\theta_0)\right] = 0.$$

[a] For instance, in a one-dimensional continuous-support case with support $[A(\theta), B(\theta)]$, one sufficient condition is $A'(\theta_0) = B'(\theta_0) = 0$ together with integrability conditions that justify exchanging differentiation and integration.

Proof. We prove the per-observation statement and then use linearity to pass to S_N.

By definition,

$$\begin{aligned}
\mathbb{E}\left[S_i(\theta_0)\right] &= \int \frac{\partial}{\partial \theta} \log f_X(x; \theta)\Big|_{\theta=\theta_0} f_X(x; \theta_0) dx \\
&= \int \frac{1}{f_X(x; \theta_0)} \frac{\partial f_X(x; \theta)}{\partial \theta}\Big|_{\theta=\theta_0} f_X(x; \theta_0) dx \\
&= \int \frac{\partial f_X(x; \theta)}{\partial \theta}\Big|_{\theta=\theta_0} dx.
\end{aligned}$$

Since $\int f_X(x; \theta)\, dx = 1$ for all θ, regularity conditions imply we may differentiate under the integral sign at θ_0:

$$\int \frac{\partial f_X(x; \theta)}{\partial \theta}\Big|_{\theta=\theta_0} dx = \frac{\partial}{\partial \theta} \int f_X(x; \theta)\, dx\Big|_{\theta=\theta_0} = \frac{\partial}{\partial \theta}(1)\Big|_{\theta=\theta_0} = 0.$$

Therefore $\mathbb{E}\left[S_i(\theta_0)\right] = 0$.

Finally,

$$\mathbb{E}\left[S_N(\theta_0)\right] = \mathbb{E}\left[\frac{1}{N} \sum_{i=1}^{N} S_i(\theta_0)\right] = \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}\left[S_i(\theta_0)\right] = 0.$$

□

This result implies that, although the score is random, it is centered at zero when evaluated at the true parameter. Its variability therefore captures how sensitive the log-likelihood is to local perturbations of θ.

Definition 8.6 (Fisher Information)**.** The Fisher information for one observation is

$$\mathcal{I}(\theta) \equiv \mathrm{Var}\left(S_i(\theta) \right) = \mathbb{E}\left[S_i(\theta) S_i(\theta)^\top \right],$$

where the second equality uses $\mathbb{E}\left[S_i(\theta)\right] = 0$ under the same regularity conditions. For the sample average score,

$$\mathrm{Var}\left(S_N(\theta) \right) = \mathrm{Var}\left(\frac{1}{N} \sum_{i=1}^{N} S_i(\theta) \right) = \frac{1}{N} \mathcal{I}(\theta) \quad \text{under i.i.d. sampling.}$$

(Equivalently, for the *total* score $\sum_{i=1}^{N} S_i(\theta)$, the variance is $N\,\mathcal{I}(\theta)$.)

Under the same regularity conditions, Fisher information admits an equivalent representation in terms of the curvature of the log-likelihood.

Proposition 8.4 (Information equality)**.** Under regularity conditions,

$$\mathcal{I}(\theta_0) = -\mathbb{E}\left[\mathcal{H}_i(\theta_0)\right] \qquad \text{and hence} \qquad \mathcal{I}(\theta_0) = -\mathbb{E}\left[\mathcal{H}_N(\theta_0)\right].$$

Proof. Since $\mathbb{E}\left[S_i(\theta)\right] = 0$ for all θ in a neighborhood of θ_0 (under the same regularity conditions), differentiate both sides with respect to $\theta^\top$ and then evaluate at θ_0:

$$0 = \frac{\partial}{\partial \theta^\top} \mathbb{E}\left[S_i(\theta)\right] \Big|_{\theta=\theta_0}.$$

Regularity allows exchanging differentiation and expectation, so

$$0 = \mathbb{E}\left[\frac{\partial}{\partial \theta^\top} S_i(\theta) \right] \Big|_{\theta=\theta_0} + \mathbb{E}\left[S_i(\theta_0) S_i(\theta_0)^\top \right].$$

The first term is $\mathbb{E}\left[\mathcal{H}_i(\theta_0)\right]$ and the second term is $\mathcal{I}(\theta_0)$, hence

$$\mathbb{E}\left[\mathcal{H}_i(\theta_0)\right] + \mathcal{I}(\theta_0) = 0,$$

which gives $\mathcal{I}(\theta_0) = -\mathbb{E}\left[\mathcal{H}_i(\theta_0)\right]$. Averaging over i yields $-\mathbb{E}\left[\mathcal{H}_N(\theta_0)\right] = \mathcal{I}(\theta_0)$. □

As a result, Fisher information plays a dual role: it measures (i) how sharply the likelihood is peaked around the true parameter and (ii) how much information the data carry about θ_0. These properties of the score and Fisher information form the foundation of likelihood-based inference. In particular, they will determine the asymptotic distribution of the MLE and its variance, which we study next.

8.4 Asymptotics of MLE

We now study the large-sample behavior of the maximum likelihood estimator. The key insight underlying essentially all asymptotic results is that the MLE maximizes a *sample* objective function that converges to a deterministic *population* criterion. As N grows, the behavior of the estimator is governed by (i) the geometry of the population objective and (ii) stochastic fluctuations of the sample objective around it. We begin with a heuristic route to consistency to see what assumptions are needed, and then make the argument rigorous using the extremum (M-) estimator framework.

8.4.1 A heuristic route to consistency

Our goal is to show $\hat{\theta}_N \xrightarrow{p} \theta_0$. In general, the MLE has no closed-form solution, so we define it as the maximizer of the *average* log-likelihood:

$$\hat{\theta}_N \in \arg\max_{\theta\in\Theta} L_N(\theta), \qquad L_N(\theta) \equiv \frac{1}{N}\sum_{i=1}^{N} \log f_X(X_i;\theta).$$

By definition of $\hat{\theta}_N$, we have the one-sided inequality

$$L_N(\hat{\theta}_N) \geq L_N(\theta_0).$$

To obtain the reverse inequality "in the limit," we compare the *population* (expected) log-likelihoods. Define

$$Q(\theta) \equiv \mathbb{E}\left[\log f_X(X;\theta)\right],$$

where the expectation $\mathbb{E}[\cdot]$ is taken under the true distribution indexed by θ_0.

Consider the log-likelihood ratio $\log \frac{f_X(X;\theta)}{f_X(X;\theta_0)}$. By Jensen's inequality,

$$\mathbb{E}\left[\log \frac{f_X(X;\theta)}{f_X(X;\theta_0)}\right] \leq \log \mathbb{E}\left[\frac{f_X(X;\theta)}{f_X(X;\theta_0)}\right].$$

Moreover,

$$\mathbb{E}\left[\frac{f_X(X;\theta)}{f_X(X;\theta_0)}\right] = \int \frac{f_X(x;\theta)}{f_X(x;\theta_0)} f_X(x;\theta_0)\,dx = \int f_X(x;\theta)\,dx = 1,$$

so we obtain

$$Q(\theta) = \mathbb{E}\left[\log f_X(X;\theta)\right] \leq \mathbb{E}\left[\log f_X(X;\theta_0)\right] = Q(\theta_0).$$

Equivalently,

$$Q(\theta_0) - Q(\theta) = \mathbb{E}\left[\log \frac{f_X(X;\theta_0)}{f_X(X;\theta)}\right] \geq 0,$$

which is the (expected) log-likelihood gap.

To conclude that θ_0 is the *unique* maximizer of $Q(\theta)$, we need an identification condition. Intuitively, this requires that different parameter values correspond to genuinely different data-generating distributions. Under such an assumption, equality $Q(\theta) = Q(\theta_0)$ can only occur when $\theta = \theta_0$ (equivalently, when $f_X(\cdot;\theta) = f_X(\cdot;\theta_0)$ almost surely). Without identification, the population objective cannot single out the true parameter, and consistency

cannot be expected.

Next, since $L_N(\theta)$ is an average of i.i.d. log-likelihood contributions, it is natural to expect that for each fixed θ,

$$L_N(\theta) = \frac{1}{N}\sum_{i=1}^{N} \log f_X(X_i;\theta) \;\approx\; \mathbb{E}\left[\log f_X(X;\theta)\right] = Q(\theta) \quad \text{for large } N.$$

For this approximation to be useful for locating the maximizer, however, it must hold *uniformly* over $\theta \in \Theta$. A uniform law of large numbers ensures that the entire sample log-likelihood surface is close to its population counterpart. Without uniform convergence, pointwise convergence alone may still allow the maximizer of $L_N(\theta)$ to drift with N, preventing stabilization near θ_0.

Putting these pieces together: if (i) $Q(\theta)$ has a unique maximizer at θ_0 and (ii) $L_N(\theta)$ converges uniformly to $Q(\theta)$, then heuristically

$$L_N(\hat{\theta}_N) \approx Q(\hat{\theta}_N) \leq Q(\theta_0) \approx L_N(\theta_0).$$

Combined with $L_N(\hat{\theta}_N) \geq L_N(\theta_0)$, this suggests that $\hat{\theta}_N$ must concentrate near θ_0, i.e. $\hat{\theta}_N \xrightarrow{p} \theta_0$.

We now make this argument fully rigorous by treating MLE as a special case of an extremum (M-) estimator and invoking an argmax consistency theorem.

8.4.2 Consistency of M-Estimators

We formalize the consistency argument using the framework of M-estimators. Let $Q_N(\theta)$ denote a sample objective function of the form

$$Q_N(\theta) \equiv \frac{1}{N}\sum_{i=1}^{N} q(X_i,\theta),$$

where $q(\cdot,\theta)$ is a measurable criterion function. The corresponding population objective is

$$Q^*(\theta) \equiv \mathbb{E}\left[q(X,\theta)\right],$$

where the expectation is taken under the true data-generating process indexed by θ_0. In the MLE case,

$$q(X,\theta) = \log f_X(X;\theta), \qquad \text{so} \qquad Q^*(\theta) = \mathbb{E}\left[\log f_X(X;\theta)\right].$$

An M-estimator is defined as

$$\hat{\theta}_N \in \arg\max_{\theta\in\Theta} Q_N(\theta).$$

Assumptions for consistency. To formalize the proof, we impose the following conditions.

(A1) Compactness. The parameter space Θ is compact.

(A2) Uniform convergence in probability.

$$\sup_{\theta\in\Theta}\left|Q_N(\theta)-Q^*(\theta)\right| \xrightarrow{p} 0.$$

(A3) Continuity. $Q^*(\theta)$ is continuous on Θ.

(A4) Existence of a maximizer. There exists $\theta_0 \in \Theta$ such that

$$\theta_0 \in \arg\max_{\theta\in\Theta} Q^*(\theta).$$

(A5) Identification (unique global maximizer).

$$Q^*(\theta_0) > Q^*(\theta) \qquad \forall\, \theta \neq \theta_0.$$

Theorem 8.5 (Consistency of M-estimators)**.** Under Assumptions **(A1)**–**(A5)**,

$$\hat{\theta}_N \xrightarrow{p} \theta_0.$$

Proof. Fix $\eta > 0$ and define the complement of an η-ball around θ_0:

$$B_\eta^c \equiv \{\theta \in \Theta :\ d(\theta,\theta_0) \geq \eta\},$$

where d is any metric on Θ. By **(A1)**, B_η^c is compact; by **(A3)**, Q^* is continuous. Therefore Q^* attains a maximum over B_η^c; let

$$\tilde{\theta}_\eta \in \arg\max_{\theta\in B_\eta^c} Q^*(\theta).$$

By uniqueness in **(A5)**, $\tilde{\theta}_\eta \neq \theta_0$ and hence

$$\Delta_\eta \equiv Q^*(\theta_0) - Q^*(\tilde{\theta}_\eta) > 0. \tag{1}$$

Now consider the event

$$\mathcal{E}_N(\eta) \equiv \left\{\sup_{\theta\in\Theta} |Q_N(\theta) - Q^*(\theta)| < \frac{\Delta_\eta}{3}\right\}.$$

By **(A2)**, $\Pr(\mathcal{E}_N(\eta)) \to 1$.

On $\mathcal{E}_N(\eta)$ we have, first,

$$Q_N(\theta_0) \geq Q^*(\theta_0) - \frac{\Delta_\eta}{3}, \tag{2}$$

and second, for any $\theta \in B_\eta^c$,

$$Q_N(\theta) \leq Q^*(\theta) + \frac{\Delta_\eta}{3} \leq Q^*(\tilde{\theta}_\eta) + \frac{\Delta_\eta}{3}. \tag{3}$$

Combining (1)–(3),

$$\sup_{\theta\in B_\eta^c} Q_N(\theta) \leq Q^*(\tilde{\theta}_\eta) + \frac{\Delta_\eta}{3} = Q^*(\theta_0) - \Delta_\eta + \frac{\Delta_\eta}{3} = Q^*(\theta_0) - \frac{2\Delta_\eta}{3} < Q^*(\theta_0) - \frac{\Delta_\eta}{3} \leq Q_N(\theta_0).$$

Thus, on $\mathcal{E}_N(\eta)$, every $\theta \in B_\eta^c$ yields a strictly smaller sample objective than θ_0, so any maximizer $\hat{\theta}_N \in \arg\max_\Theta Q_N(\theta)$ must satisfy $\hat{\theta}_N \notin B_\eta^c$, i.e. $d(\hat{\theta}_N, \theta_0) < \eta$.

Therefore,

$$\Pr(d(\hat{\theta}_N, \theta_0) < \eta) \;\geq\; \Pr(\mathcal{E}_N(\eta)) \;\to\; 1.$$

Since $\eta > 0$ was arbitrary, $\hat{\theta}_N \xrightarrow{p} \theta_0$. □

In the case of maximum likelihood estimation, $Q_N(\theta) = L_N(\theta)$ with $q(X, \theta) = \log f_X(X; \theta)$. Identification corresponds to strict maximization of the expected log-likelihood (heuristically justified earlier via Jensen's inequality). Uniform convergence follows from a uniform law of large numbers under standard regularity conditions. Thus, consistency of the MLE follows as a special case of the general M-estimator consistency theorem.

8.4.3 Asymptotic Normality

We now study the limiting distribution of the MLE. Throughout this section, assume $\hat{\theta}_N \xrightarrow{p} \theta_0$ has already been established, and that θ_0 lies in the interior of Θ so that first-order conditions apply without constraints.[2]

Recall the average score and Hessian (derivatives of L_N):

$$S_N(\theta) \equiv \frac{\partial L_N(\theta)}{\partial \theta} = \frac{1}{N}\sum_{i=1}^{N} S_i(\theta), \qquad \mathcal{H}_N(\theta) \equiv \frac{\partial^2 L_N(\theta)}{\partial \theta\, \partial \theta^\top} = \frac{1}{N}\sum_{i=1}^{N} \mathcal{H}_i(\theta),$$

where

$$S_i(\theta) = \frac{\partial}{\partial \theta} \log f_X(X_i; \theta), \qquad \mathcal{H}_i(\theta) = \frac{\partial^2}{\partial \theta\, \partial \theta^\top} \log f_X(X_i; \theta).$$

[2] If Θ imposes constraints and θ_0 lies on the boundary, one needs a constrained optimization/KKT version of the argument.

Theorem 8.6 (Asymptotic normality of the MLE)**.** Suppose $\{X_i\}_{i=1}^N$ are i.i.d. with density $f_X(\cdot;\theta)$, and let

$$\hat{\theta}_N \in \arg\max_{\theta\in\Theta} L_N(\theta), \qquad L_N(\theta) = \frac{1}{N}\sum_{i=1}^{N} \log f_X(X_i;\theta).$$

Assume:

(i) (**Consistency**) $\hat{\theta}_N \xrightarrow{p} \theta_0$.

(ii) (**Smoothness**) $L_N(\theta)$ is twice continuously differentiable in a neighborhood of θ_0 (with probability approaching one).

(iii) (**Score regularity**)

$$\mathbb{E}\left[S_i(\theta_0)\right] = 0, \qquad \mathrm{Var}\left(S_i(\theta_0)\right) = \mathcal{I}(\theta_0),$$

with $\mathcal{I}(\theta_0)$ finite and nonsingular.

(iv) (**Hessian convergence**)

$$-\mathcal{H}_N(\tilde{\theta}_N) \xrightarrow{p} \mathcal{I}(\theta_0)$$

for any sequence $\tilde{\theta}_N \xrightarrow{p} \theta_0$.

Then

$$\sqrt{N}\left(\hat{\theta}_N - \theta_0\right) \xrightarrow{d} \mathcal{N}\big(0,\ \mathcal{I}(\theta_0)^{-1}\big)\,.$$

Proof. By the first-order condition for an interior maximizer,

$$S_N(\hat{\theta}_N) = 0.$$

Apply the mean value theorem to the map $\theta \mapsto S_N(\theta)$ along the line segment joining θ_0 and $\hat{\theta}_N$. Then there exists a random intermediate point $\tilde{\theta}_N$ on this segment such that

$$S_N(\hat{\theta}_N) - S_N(\theta_0) = \mathcal{H}_N(\tilde{\theta}_N)\left(\hat{\theta}_N - \theta_0\right).$$

Using $S_N(\hat{\theta}_N) = 0$, we obtain

$$\sqrt{N}\left(\hat{\theta}_N - \theta_0\right) = -\Big[\mathcal{H}_N(\tilde{\theta}_N)\Big]^{-1} \sqrt{N}\, S_N(\theta_0). \tag{1}$$

Since $\mathbb{E}\left[S_i(\theta_0)\right] = 0$ and $\mathrm{Var}\left(S_i(\theta_0)\right) = \mathcal{I}(\theta_0)$,

$$\sqrt{N}\, S_N(\theta_0) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} S_i(\theta_0) \xrightarrow{d} \mathcal{N}(0,\ \mathcal{I}(\theta_0))\,. \tag{2}$$

By consistency, $\tilde{\theta}_N \xrightarrow{p} \theta_0$. By Assumption **(iv)**,

$$-\mathcal{H}_N(\tilde{\theta}_N) \xrightarrow{p} \mathcal{I}(\theta_0).$$

Since $\mathcal{I}(\theta_0)$ is nonsingular, matrix inversion is continuous in a neighborhood of $\mathcal{I}(\theta_0)$, so

$$-\Big[\mathcal{H}_N(\tilde{\theta}_N)\Big]^{-1} \xrightarrow{p} \mathcal{I}(\theta_0)^{-1}. \tag{3}$$

Combine (1), (2), and (3) and apply Slutsky's theorem:

$$\sqrt{N}\,(\hat{\theta}_N - \theta_0) \xrightarrow{d} \mathcal{N}\big(0,\ \mathcal{I}(\theta_0)^{-1}\mathcal{I}(\theta_0)\mathcal{I}(\theta_0)^{-1}\big) = \mathcal{N}\big(0,\ \mathcal{I}(\theta_0)^{-1}\big)\,.$$

□

Under correct specification and standard regularity conditions, the MLE is asymptotically normal with covariance matrix $\mathcal{I}(\theta_0)^{-1}/N$. Equivalently, $\mathcal{I}(\theta_0)$ measures local curvature of the population log-likelihood and therefore the amount of information in the data about θ_0.

8.5 Likelihood-based Inference

So far we have treated MLE primarily as an estimator. This subsection is about what you actually do with it: standard errors, confidence intervals, and hypothesis tests that are *likelihood-based* (or at least likelihood-adjacent).

8.5.1 Standard errors from (observed) information

Under correct specification and regularity conditions,

$$\sqrt{N}(\hat{\theta}_N - \theta_0) \ \xrightarrow{d}\ \mathcal{N}\big(0,\ \mathcal{I}(\theta_0)^{-1}\big),$$

so asymptotically

$$\mathrm{Var}\left(\hat{\theta}_N\right) \approx \frac{1}{N}\,\mathcal{I}(\theta_0)^{-1}.$$

A practical plug-in estimator replaces $\mathcal{I}(\theta_0)$ with an estimate of Fisher information.

Definition 8.7 (Observed information and plug-in covariance)**.** The *observed information* is

$$J_N(\hat{\theta}_N) \equiv -\frac{\partial^2 \ell_N(\theta)}{\partial\theta\,\partial\theta^\top}\Big|_{\theta=\hat{\theta}_N}.$$

Equivalently, in average form $-\,\mathcal{H}_N(\hat{\theta}_N) = J_N(\hat{\theta}_N)/N$. A standard covariance estimator is

$$\widehat{\mathrm{Var}}(\hat{\theta}_N) \ \equiv\ \frac{1}{N}\Big[-\mathcal{H}_N(\hat{\theta}_N)\Big]^{-1} \ =\ \Big[J_N(\hat{\theta}_N)\Big]^{-1}.$$

In words: the likelihood is locally quadratic near the maximizer; the curvature (information) tells you how sharp that peak is; sharper peak ⇒ smaller standard errors.

8.5.2 Three classical tests: Wald, likelihood ratio, and score

We now consider testing restrictions on θ. Let the null impose q restrictions through a smooth map $r : \Theta \to \mathbb{R}^q$:

$$H_0 : \ r(\theta_0) = 0, \qquad H_1 : \ r(\theta_0) \neq 0,$$

and write $R(\theta) \equiv \partial r(\theta)/\partial \theta^\top$.

Let

$$\hat{\theta}_N \in \arg\max_{\theta \in \Theta} \ell_N(\theta), \qquad \tilde{\theta}_N \in \arg\max_{\theta \in \Theta:\ r(\theta)=0} \ell_N(\theta)$$

denote the unrestricted and restricted MLEs.

Wald test. Wald uses the idea: "estimate first, then see how far the estimate violates the null." Under H_0,

$$W_N \equiv N\, r(\hat{\theta}_N)^\top \Big[R(\hat{\theta}_N) \widehat{\mathrm{Var}}(\hat{\theta}_N) R(\hat{\theta}_N)^\top \Big]^{-1} r(\hat{\theta}_N) \ \xrightarrow{d} \ \chi^2_q.$$

Likelihood ratio (LR) test. LR compares how much log-likelihood you lose by forcing the null to hold:

$$LR_N \equiv 2\big(\ell_N(\hat{\theta}_N) - \ell_N(\tilde{\theta}_N)\big) \ \xrightarrow{d} \ \chi^2_q \qquad (H_0).$$

Equivalently, using averages,

$$LR_N = 2N\big(L_N(\hat{\theta}_N) - L_N(\tilde{\theta}_N)\big).$$

Score (LM) test. Score tests the null by checking whether the score wants to "push you away" from the restricted fit:

$$LM_N \equiv U_N(\tilde{\theta}_N)^\top \Big[J_N(\tilde{\theta}_N) \Big]^{-1} U_N(\tilde{\theta}_N) \ \xrightarrow{d} \ \chi^2_q \qquad (H_0),$$

where $U_N(\theta) = \partial_\theta \ell_N(\theta)$ and $J_N(\theta)$ is information (observed, or expected if you prefer). Under H_0, the score vanishes at the *unrestricted* maximizer, not necessarily at the restricted one; that mismatch is the whole point.

8.5.3 Likelihood-based confidence intervals and regions

Wald-type confidence region. From asymptotic normality, an approximate $(1-\alpha)$ confidence region is

$$\mathcal{C}^{\mathrm{Wald}}_{1-\alpha} \equiv \left\{ \theta \in \Theta : N(\theta - \hat{\theta}_N)^\top \Big[\widehat{\mathrm{Var}}(\hat{\theta}_N) \Big]^{-1} (\theta - \hat{\theta}_N) \leq \chi^2_{k,1-\alpha} \right\}.$$

Likelihood ratio confidence region. A likelihood-based alternative is

$$\mathcal{C}^{\mathrm{LR}}_{1-\alpha} \equiv \left\{ \theta \in \Theta : 2\big(\ell_N(\hat{\theta}_N) - \ell_N(\theta)\big) \leq \chi^2_{k,1-\alpha} \right\}.$$

This is attractive because it is invariant to reparameterization (it depends only on likelihood values) and often behaves better when the likelihood is asymmetric.

Profile likelihood for one component. Suppose $\theta = (\psi, \lambda)$ where ψ is a scalar parameter of interest and λ is nuisance. Define the profile log-likelihood

$$\ell_N^p(\psi) \equiv \max_{\lambda} \ \ell_N(\psi, \lambda).$$

Then an approximate $(1-\alpha)$ confidence interval for ψ is

$$\left\{\psi : 2\big(\ell_N(\hat{\psi}, \hat{\lambda}) - \ell_N^p(\psi)\big) \le \chi^2_{1,1-\alpha}\right\}.$$

This is the likelihood version of "optimize out the nuisance and only then do inference."

8.6 Summary

This chapter treated maximum likelihood estimation as the moment-method's more demanding cousin: instead of asking for a few orthogonality conditions, MLE asks you to fully specify a parametric model and then live with the consequences.

Here's what we've wrestled with:

- **Likelihood vs. probability.** $\mathcal{L}_N(\theta)$ (and $\ell_N(\theta)$) is the joint density of the observed sample *viewed as a function of* θ. Once data are realized, it is not a distribution over θ.
- **Identification first, optimization second.** If distinct θ generate the same distribution of observables, the likelihood cannot distinguish them. No optimizer can fix that.
- **Local geometry.** The score $S_N(\theta)$ and Hessian $\mathcal{H}_N(\theta)$ encode local slope and curvature:
$$S_N(\hat{\theta}_N) = 0, \qquad \mathcal{H}_N(\hat{\theta}_N) \prec 0 \ \text{ (interior maximum).}$$
- **Information.** Under regularity conditions, the score is centered at zero at the truth and Fisher information measures its variability:
$$\mathcal{I}(\theta_0) = \mathbb{E}\left[S_i(\theta_0) S_i(\theta_0)^\top\right] = -\,\mathbb{E}\left[\mathcal{H}_i(\theta_0)\right].$$
- **Asymptotics.** Consistency is an extremum-estimator story (uniform convergence + unique population maximizer). Asymptotic normality is a Taylor expansion story:
$$\sqrt{N}\,(\hat{\theta}_N - \theta_0) \ \xrightarrow{d} \ \mathcal{N}\big(0, \mathcal{I}(\theta_0)^{-1}\big).$$
Information = curvature = (asymptotic) precision.
- **Inference.** Wald, LR, and score tests are three ways to say the same thing asymptotically (they disagree only when N is not yet pretending to be infinite).

Everything above is cleanest under correct specification. When the model is wrong, the MLE typically targets a pseudo-true parameter and the variance becomes a sandwich.

9 Binary Choice Models

In many empirical settings, the outcome of interest is binary: someone works or does not, a loan defaults or survives, a policy is adopted or ignored. When y_i only takes values in $\{0, 1\}$, modeling the conditional mean is no longer just a matter of fitting a regression line, but more about understanding how covariates shift the probability of an event.

Binary choice models tackle this problem head-on. We start with the Linear Probability Model (LPM), which is blunt, linear, and unapologetically simple. It does a surprisingly decent job at describing probabilities and remains popular largely because it is easy to estimate and easy to interpret. But its simplicity comes at a cost. To see what goes wrong(and how we fix it) we then move to nonlinear probability models, namely Probit and Logit, which keep predicted probabilities in check and admit a clean latent-variable interpretation grounded in economic decision-making. Along the way, we study estimation by maximum likelihood, marginal effects, inference, and hypothesis testing, before closing with a discussion of endogeneity in binary choice models and a full-information likelihood approach.

9.1 Linear Probability Model

Let $y_i \in \{0, 1\}$ denote a binary outcome and $\mathbf{x}_i$ a $k \times 1$ vector of covariates. The Linear Probability Model (LPM) specifies the conditional expectation of y_i as a linear function of $\mathbf{x}_i$:

$$\mathbb{E}\left[y_i \mid \mathbf{x}_i\right] = \mathbf{x}_i^\top \boldsymbol{\beta}.$$

Since y_i is binary, this conditional mean coincides with the conditional probability

$$\mathbb{E}\left[y_i \mid \mathbf{x}_i\right] = \Pr(y_i = 1 \mid \mathbf{x}_i).$$

Equivalently, the model can be written as

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_i,$$

where the error term satisfies

$$\mathbb{E}\left[\varepsilon_i \mid \mathbf{x}_i\right] = 0, \qquad \varepsilon_i = y_i - \mathbf{x}_i^\top \boldsymbol{\beta}.$$

A distinctive feature of the LPM is that the conditional variance of the error term is inherently heteroskedastic:

$$\begin{aligned}
\operatorname{Var}\left(\varepsilon_i \mid \mathbf{x}_i\right) &= \operatorname{Var}\left(y_i \mid \mathbf{x}_i\right) \\
&= \mathbb{E}\left[y_i^2 \mid \mathbf{x}_i\right] - \mathbb{E}\left[y_i \mid \mathbf{x}_i\right]^2 \\
&= \Pr(y_i = 1 \mid \mathbf{x}_i) - \Pr(y_i = 1 \mid \mathbf{x}_i)^2 \\
&= \mathbf{x}_i^\top \boldsymbol{\beta} - (\mathbf{x}_i^\top \boldsymbol{\beta})^2 \\
&= \mathbf{x}_i^\top \boldsymbol{\beta}\big(1 - \mathbf{x}_i^\top \boldsymbol{\beta}\big).
\end{aligned}$$

Thus, even under correct specification of the conditional mean, homoskedasticity fails by

construction.

Despite its limitations, the LPM has several attractive features:

- **Linear specification.** Estimation and interpretation are straightforward. Each coefficient β_k measures the marginal effect of x_{ik} on the probability $\Pr(y_i = 1 \mid \mathbf{x}_i)$.
- **Compatibility with IV methods.** When regressors are endogenous, the model can be estimated using linear instrumental variables techniques such as 2SLS.

At the same time, the linear specification **does not** constrain $\mathbf{x}_i^\top \boldsymbol{\beta}$ to lie in the unit interval $[0, 1]$, so fitted values may fail to correspond to valid probabilities. This motivates the nonlinear probability models that follow.

9.2 Probit and Logit Models

The main problem with the Linear Probability Model is not subtle: nothing in $\mathbb{E}[y_i \mid \mathbf{x}_i] = \mathbf{x}_i^\top \boldsymbol{\beta}$ prevents the right-hand side from wandering outside $[0, 1]$. Interpreting values below 0 or above 1 as probabilities requires a certain amount of optimism. Rather than policing the fitted values ex post, a more principled approach is to build the unit-interval restriction directly into the model.

The idea behind nonlinear probability models is simple. We retain the index structure $\mathbf{x}_i^\top \boldsymbol{\beta}$, but pass it through a smooth, monotone function that maps $\mathbb{R}$ into $[0, 1]$:

$$\Pr(y_i = 1 \mid \mathbf{x}_i) = F(\mathbf{x}_i^\top \boldsymbol{\beta}),$$

where $F(\cdot)$ is a cumulative distribution function. Different choices of F give rise to different binary choice models. We focus on the two most common choices.

Definition 9.1 (Probit model)**.** The *Probit* model specifies the conditional success probability as

$$\Pr(y_i = 1 \mid \mathbf{x}_i) = \Phi(\mathbf{x}_i^\top \boldsymbol{\beta}),$$

where $\Phi(\cdot)$ is the standard normal CDF,

$$\Phi(t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{s^2}{2}\right) ds.$$

Definition 9.2 (Logit model)**.** The *Logit* model specifies the conditional success probability as

$$\Pr(y_i = 1 \mid \mathbf{x}_i) = \Lambda(\mathbf{x}_i^\top \boldsymbol{\beta}),$$

where $\Lambda(\cdot)$ is the logistic CDF,

$$\Lambda(t) = \frac{\exp(t)}{1 + \exp(t)} = \frac{1}{1 + \exp(-t)}.$$

Both models ensure that predicted probabilities lie strictly between 0 and 1 for any finite value of $\mathbf{x}_i^\top \boldsymbol{\beta}$. Compared to the LPM, this comes at the cost of nonlinearity, but the payoff is immediate: fitted values are always interpretable as probabilities.

Importantly, the role of the function $F(\cdot)$ goes beyond mechanically enforcing the unit-interval restriction. Each choice of F corresponds to a different economic interpretation of the decision process. In particular, Probit and Logit models emerge naturally from latent-variable formulations in which individuals compare an unobserved utility index to a threshold, with the functional form of F determined by assumptions on the distribution of unobserved shocks. From this perspective, nonlinear probability models are not merely ad hoc fixes to the LPM, but structural models of discrete choice. We make this interpretation explicit in the next subsection.

9.3 Latent Variable Interpretation

A more natural way to understand the choice of link functions is through a latent-variable representation. Suppose there exists an unobserved continuous variable y_i^* that captures the net utility or propensity associated with choosing $y_i = 1$. We assume

$$y_i^* = \mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_i,$$

where ε_i represents unobserved factors affecting the decision. The observed binary outcome is generated according to the threshold rule

$$y_i = \mathbf{1}\{y_i^* > 0\}.$$

Only the sign of y_i^* matters for the observed outcome, not its magnitude. As a result, the latent index is only identified up to *location* and *scale* normalizations. Concretely: (i) adding a constant to y_i^* is observationally equivalent to shifting the threshold, so one typically normalizes the threshold to zero (equivalently, absorbs the constant into the intercept in $\mathbf{x}_i^\top \boldsymbol{\beta}$); and (ii) multiplying the latent equation by a positive constant leaves the sign unchanged, so the scale is not identified and must be fixed by normalizing the dispersion of ε_i (e.g. setting $\mathrm{Var}(\varepsilon_i) = 1$) through distributional assumptions.

Under this formulation,

$$\Pr(y_i = 1 \mid \mathbf{x}_i) = \Pr(y_i^* > 0 \mid \mathbf{x}_i) = \Pr(\varepsilon_i > -\mathbf{x}_i^\top \boldsymbol{\beta} \mid \mathbf{x}_i).$$

Let $F_\varepsilon(t) := \Pr(\varepsilon_i \le t)$ denote the CDF of ε_i. Then

$$\Pr(\varepsilon_i > -\mathbf{x}_i^\top \boldsymbol{\beta} \mid \mathbf{x}_i) = 1 - F_\varepsilon(-\mathbf{x}_i^\top \boldsymbol{\beta}).$$

For the common symmetric choices used in Probit/Logit (where $F_\varepsilon(-t) = 1 - F_\varepsilon(t)$), this simplifies to

$$\Pr(y_i = 1 \mid \mathbf{x}_i) = F_\varepsilon(\mathbf{x}_i^\top \boldsymbol{\beta}).$$

Thus the link function $F(\cdot)$ in $\Pr(y_i = 1 \mid \mathbf{x}_i) = F(\mathbf{x}_i^\top \boldsymbol{\beta})$ then implies assumptions about the distribution of unobserved utility shocks. In the Probit model, ε_i is assumed to follow

a standard normal distribution, leading to $F = \Phi$. In the Logit model, ε_i follows a logistic distribution, yielding $F = \Lambda$.

9.4 Alternative Interpretations of Logit and Probit

The latent-variable formulation provides a convenient and intuitive way to motivate binary choice models, but it is not the only interpretation under which Probit and Logit arise. More generally, these models can be viewed as *single-index* probability models: the conditional probability $\Pr(y_i = 1 \mid \mathbf{x}_i)$ depends on $\mathbf{x}_i$ only through the scalar index $\mathbf{x}_i^\top \boldsymbol{\beta}$, transformed by a smooth, monotone link function.

Log-odds for Logit. A distinctive interpretation of the Logit model is built around the concept of *log-odds*. While probabilities themselves are bounded between 0 and 1, odds compare the likelihood of an event occurring to the likelihood of it not occurring:

$$\text{odds}(y_i = 1 \mid \mathbf{x}_i) = \frac{\Pr(y_i = 1 \mid \mathbf{x}_i)}{1 - \Pr(y_i = 1 \mid \mathbf{x}_i)}.$$

Odds are unbounded and strictly positive, which makes them convenient for describing relative likelihoods. Outside econometrics, they are commonly used in betting and games of chance (such as poker) where decisions are often framed in terms of how many times more likely one outcome is relative to another.

The Logit model goes one step further by applying a logarithmic transformation to the odds:

$$\log \text{odds}(y_i = 1 \mid \mathbf{x}_i).$$

Taking logs maps the positive real line to $\mathbb{R}$ and converts multiplicative changes in odds into additive ones. The Logit specification imposes a linear structure on this transformed quantity,

$$\log \text{odds}(y_i = 1 \mid \mathbf{x}_i) = \mathbf{x}_i^\top \boldsymbol{\beta},$$

which implies that covariates shift the log-odds additively.

Exponentiating both sides yields

$$\frac{\Pr(y_i = 1 \mid \mathbf{x}_i)}{1 - \Pr(y_i = 1 \mid \mathbf{x}_i)} = \exp(\mathbf{x}_i^\top \boldsymbol{\beta}),$$

and solving for the probability gives the logistic CDF form

$$\Pr(y_i = 1 \mid \mathbf{x}_i) = \frac{\exp(\mathbf{x}_i^\top \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i^\top \boldsymbol{\beta})} = \Lambda(\mathbf{x}_i^\top \boldsymbol{\beta}).$$

Under this formulation, a one-unit increase in x_{ik} multiplies the odds of the event occurring by $\exp(\beta_k)$, holding other covariates fixed.

Probit and CLT. An alternative and particularly natural interpretation of the Probit model is based on the aggregation of unobserved heterogeneity. Suppose the binary outcome reflects whether an underlying index exceeds a threshold, but the unobserved shock consists

of the sum of many small, (approximately) independent components:

$$y_i = \mathbf{1}\{\mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_i > 0\}, \qquad \varepsilon_i = \sum_{j=1}^{J} u_{ij},$$

where each u_{ij} captures a minor unmodeled influence on the decision. If no single component dominates and J is large, then informally the Central Limit Theorem suggests that the aggregate disturbance ε_i is approximately normally distributed (after an appropriate scaling). This leads naturally to the Probit link,

$$\Pr(y_i = 1 \mid \mathbf{x}_i) \approx \Phi(\mathbf{x}_i^\top \boldsymbol{\beta}).$$

This perspective is often appealing in empirical applications, where unobserved factors are numerous, individually negligible, and difficult to model explicitly.

9.5 Marginal Effects

A convenient way to compare binary choice models is through *marginal effects*, i.e., the partial derivative of the conditional probability with respect to a regressor. Fix a coordinate k and consider a continuous regressor x_{ik} (the kth component of $\mathbf{x}_i$).

Definition 9.3 (Marginal effects in the LPM)**.** Suppose the Linear Probability Model implies

$$\Pr(y_i = 1 \mid \mathbf{x}_i) = \mathbb{E}\left[y_i \mid \mathbf{x}_i\right] = \mathbf{x}_i^\top \boldsymbol{\beta}.$$

For a continuous regressor x_{ik}, the marginal effect is

$$\mathrm{ME}_{ik}^{L} := \frac{\partial \mathbb{E}\left[y_i \mid \mathbf{x}_i\right]}{\partial x_{ik}} = \beta_k,$$

which is constant across observations.

Definition 9.4 (Marginal effects in the Probit model)**.** Suppose the Probit model implies

$$\Pr(y_i = 1 \mid \mathbf{x}_i) = \Phi(\mathbf{x}_i^\top \boldsymbol{\beta}).$$

For a continuous regressor x_{ik}, the marginal effect is

$$\mathrm{ME}_{ik}^{P} := \frac{\partial \Phi(\mathbf{x}_i^\top \boldsymbol{\beta})}{\partial x_{ik}} = \phi(\mathbf{x}_i^\top \boldsymbol{\beta})\, \beta_k = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(\mathbf{x}_i^\top \boldsymbol{\beta})^2}{2}\right) \beta_k,$$

where $\phi(\cdot)$ is the standard normal density.

Definition 9.5 (Marginal effects in the Logit model)**.** Suppose the Logit model implies

$$\Pr(y_i = 1 \mid \mathbf{x}_i) = \Lambda(\mathbf{x}_i^\top \boldsymbol{\beta}) = \frac{\exp(\mathbf{x}_i^\top \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i^\top \boldsymbol{\beta})}.$$

For a continuous regressor x_{ik}, the marginal effect is

$$\mathrm{ME}_{ik}^{Lgt} := \frac{\partial \Lambda(\mathbf{x}_i^\top \boldsymbol{\beta})}{\partial x_{ik}} = \Lambda(\mathbf{x}_i^\top \boldsymbol{\beta})\big(1 - \Lambda(\mathbf{x}_i^\top \boldsymbol{\beta})\big)\beta_k = \frac{\exp(\mathbf{x}_i^\top \boldsymbol{\beta})}{\big(1 + \exp(\mathbf{x}_i^\top \boldsymbol{\beta})\big)^2}\,\beta_k.$$

Because Probit and Logit marginal effects depend on the value of the index $\mathbf{x}_i^\top \boldsymbol{\beta}$, they are not constant across observations. As a result, applied work typically reports marginal effects in one of two ways.

The first approach reports *marginal effects at a representative covariate value*, most commonly at the sample mean $\bar{\mathbf{x}}$:

$$\mathrm{ME}_k(\bar{\mathbf{x}}) = \left.\frac{\partial \Pr(y_i = 1 \mid \mathbf{x})}{\partial x_k}\right|_{\mathbf{x}=\bar{\mathbf{x}}}.$$

This provides a single summary measure evaluated at a typical point in the covariate space.

The second approach reports *average marginal effects (AMEs)*, defined as the sample average of individual marginal effects:

$$\mathrm{AME}_k = \frac{1}{N}\sum_{i=1}^{N} \frac{\partial \Pr(y_i = 1 \mid \mathbf{x}_i)}{\partial x_{ik}}.$$

Average marginal effects capture the mean impact of a covariate change across the observed covariate distribution and are often preferred when the sample exhibits substantial heterogeneity.

9.6 Estimation by Maximum Likelihood

9.6.1 Likelihood Setup

For binary choice models, once the conditional probability

$$\Pr(y_i = 1 \mid \mathbf{x}_i) = F(\mathbf{x}_i^\top \boldsymbol{\beta})$$

is specified, the conditional distribution of $y_i \mid \mathbf{x}_i$ is Bernoulli with success probability $F(\mathbf{x}_i^\top \boldsymbol{\beta})$. Let $\eta_i := \mathbf{x}_i^\top \boldsymbol{\beta}$ and $F_i := F(\eta_i)$. The likelihood contribution of observation i is

$$\mathcal{L}_i(\boldsymbol{\beta}) = F_i^{y_i}(1 - F_i)^{1-y_i}.$$

Assuming conditional independence across i given $\{\mathbf{x}_i\}_{i=1}^N$, the sample likelihood is

$$\mathcal{L}(\boldsymbol{\beta}) = \prod_{i=1}^{N} F(\mathbf{x}_i^\top \boldsymbol{\beta})^{y_i}\big[1 - F(\mathbf{x}_i^\top \boldsymbol{\beta})\big]^{1-y_i}.$$

Taking logs yields the log-likelihood

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{N} \Big[y_i \log F(\mathbf{x}_i^\top \boldsymbol{\beta}) + (1 - y_i) \log\big(1 - F(\mathbf{x}_i^\top \boldsymbol{\beta})\big) \Big] =: \sum_{i=1}^{N} \ell_i(\boldsymbol{\beta}).$$

9.6.2 First Order Conditions

The maximum likelihood estimator $\hat{\boldsymbol{\beta}}$ satisfies $\nabla_{\boldsymbol{\beta}} \ell(\hat{\boldsymbol{\beta}}) = 0$. For a generic link F with density $f = F'$, the score can be written as

$$\begin{aligned} \nabla_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}) &= \sum_{i=1}^{N} \left[\frac{y_i}{F(\eta_i)} \, f(\eta_i) \, \mathbf{x}_i + \frac{1 - y_i}{1 - F(\eta_i)} \big(-f(\eta_i)\big) \, \mathbf{x}_i \right] \\ &= \sum_{i=1}^{N} \frac{y_i - F_i}{F_i(1 - F_i)} \, f(\eta_i) \, \mathbf{x}_i. \end{aligned} \tag{1}$$

9.6.3 Hessian and curvature.

To understand whether the criterion is globally concave (so that any solution to the first-order condition is a global maximizer), we examine the Hessian. Write $f_i := f(\eta_i) = F'(\eta_i)$. For Logit, $F = \Lambda$ and $f(\eta) = \Lambda(\eta)\big(1 - \Lambda(\eta)\big)$; for Probit, $F = \Phi$ and $f = \phi$.

LPM. The LPM is estimated by OLS:

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_i, \qquad \mathbb{E}\left[\varepsilon_i \mid \mathbf{x}_i\right] = 0.$$

With the (half) sum of squared errors objective

$$Q(\boldsymbol{\beta}) := \frac{1}{2} \sum_{i=1}^{N} (y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2,$$

we have

$$\nabla_{\boldsymbol{\beta}} Q(\boldsymbol{\beta}) = -\mathbf{X}^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \qquad \nabla^2_{\boldsymbol{\beta}} Q(\boldsymbol{\beta}) = \mathbf{X}^\top \mathbf{X},$$

so the LPM objective is globally convex (Hessian p.s.d., and p.d. if $\mathbf{X}$ has full column rank).

Logit. For Logit, $F_i = \Lambda(\eta_i) = \exp(\eta_i)/(1 + \exp(\eta_i))$. From (1),

$$\frac{\partial \ell_i}{\partial \eta_i} = \frac{y_i - F_i}{F_i(1 - F_i)} \, f_i.$$

Since $f_i = F_i(1 - F_i)$ for the logistic CDF, this simplifies to

$$\frac{\partial \ell_i}{\partial \eta_i} = y_i - F_i, \qquad \nabla_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}) = \sum_{i=1}^{N} (y_i - F_i) \mathbf{x}_i.$$

Differentiating again,

$$\nabla^2_{\boldsymbol{\beta}}\ell(\boldsymbol{\beta}) = -\sum_{i=1}^{N} f_i\,\mathbf{x}_i\mathbf{x}_i^\top = -\sum_{i=1}^{N} \Lambda(\eta_i)\big(1-\Lambda(\eta_i)\big)\,\mathbf{x}_i\mathbf{x}_i^\top,$$

which is negative semidefinite.

Probit. For Probit, $F_i = \Phi(\eta_i)$ and $f_i = \phi(\eta_i)$, so

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{N}\Big[y_i \log\Phi(\eta_i) + (1-y_i)\log\big(1-\Phi(\eta_i)\big)\Big].$$

Define the selected probability

$$\bar{\Phi}_i := y_i\Phi(\eta_i) + (1-y_i)\big(1-\Phi(\eta_i)\big),$$

and the scalar score factor

$$\xi_i := (2y_i-1)\frac{\phi(\eta_i)}{\bar{\Phi}_i}.$$

Then

$$\frac{\partial \ell_i}{\partial \eta_i} = \xi_i, \qquad \nabla_{\boldsymbol{\beta}}\ell(\boldsymbol{\beta}) = \sum_{i=1}^{N}\xi_i\,\mathbf{x}_i.$$

Using $\phi'(\eta) = -\eta\,\phi(\eta)$, one obtains

$$\frac{d\xi_i}{d\eta_i} = -\eta_i\xi_i - \xi_i^2,$$

and therefore the Hessian can be written compactly as

$$\nabla^2_{\boldsymbol{\beta}}\ell(\boldsymbol{\beta}) = \sum_{i=1}^{N}\left(\frac{d\xi_i}{d\eta_i}\right)\mathbf{x}_i\mathbf{x}_i^\top = -\sum_{i=1}^{N}\xi_i(\eta_i+\xi_i)\,\mathbf{x}_i\mathbf{x}_i^\top.$$

Both Logit and Probit log-likelihood Hessians take a weighted least-squares form,

$$\nabla^2_{\boldsymbol{\beta}}\ell(\boldsymbol{\beta}) = -\sum_{i=1}^{N} w_i(\boldsymbol{\beta})\,\mathbf{x}_i\mathbf{x}_i^\top \;=\; -\mathbf{X}^\top W(\boldsymbol{\beta})\mathbf{X},$$

for a suitable diagonal weight matrix $W(\boldsymbol{\beta}) = \operatorname{diag}(w_1(\boldsymbol{\beta}),\dots,w_N(\boldsymbol{\beta}))$. For Logit, the weights are

$$w_i(\boldsymbol{\beta}) = \Lambda(\eta_i)\big(1-\Lambda(\eta_i)\big) \geq 0.$$

For Probit, the weights are

$$w_i(\boldsymbol{\beta}) = \xi_i(\eta_i+\xi_i), \qquad \xi_i = (2y_i-1)\frac{\phi(\eta_i)}{\bar{\Phi}_i}, \qquad \bar{\Phi}_i = y_i\Phi(\eta_i) + (1-y_i)\big(1-\Phi(\eta_i)\big).$$

Under standard regularity conditions, these weights are nonnegative so that the log-likelihood is concave and the Hessian is negative semidefinite.[3]

[3]One way to justify concavity for Probit is that the standard normal CDF Φ is log-concave, which implies

9.7 Inference on $\hat{\boldsymbol{\beta}}$

Recall our discussion on M-estimators in the previous chapter. We can treat $\hat{\boldsymbol{\beta}}$ as an M-estimator: it maximizes the sample criterion

$$\hat{\boldsymbol{\beta}} \in \arg\max_{\boldsymbol{\beta}\in\mathbb{R}^k} \ \ell_N(\boldsymbol{\beta}), \qquad \ell_N(\boldsymbol{\beta}) := \frac{1}{N}\sum_{i=1}^{N} \ell_i(\boldsymbol{\beta}),$$

or equivalently solves the first-order condition

$$\frac{1}{N}\sum_{i=1}^{N} S_i(\boldsymbol{\beta}) = 0, \qquad S_i(\boldsymbol{\beta}) := \nabla_{\boldsymbol{\beta}} \ell_i(\boldsymbol{\beta}).$$

Under standard regularity conditions (as in the general M-estimator theory developed earlier), we have

$$\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \ \xrightarrow{d}\ \mathcal{N}\Big(0, \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}\Big),$$

where

$$\mathbf{A} := -\mathbb{E}\big[\nabla_{\boldsymbol{\beta}} S_i(\boldsymbol{\beta}_0)\big] = -\mathbb{E}\big[\nabla^2_{\boldsymbol{\beta}} \ell_i(\boldsymbol{\beta}_0)\big], \qquad \mathbf{B} := \mathbb{E}\big[S_i(\boldsymbol{\beta}_0) S_i(\boldsymbol{\beta}_0)^\top\big].$$

This is the familiar *sandwich* form for asymptotic variance.

If the likelihood is correctly specified so that the information matrix identity holds, then

$$\mathbf{B} = \mathbf{A},$$

and the asymptotic variance simplifies to

$$\text{A.Var}(\hat{\boldsymbol{\beta}}) = \mathbf{A}^{-1}.$$

In practice, one uses the sample analogs evaluated at $\hat{\boldsymbol{\beta}}$, such as the observed-information estimator

$$\widehat{\text{Var}}(\hat{\boldsymbol{\beta}}) = \left[-\sum_{i=1}^{N} \nabla^2_{\boldsymbol{\beta}} \ell_i(\hat{\boldsymbol{\beta}})\right]^{-1}.$$

(Equivalently, since ℓ_N is defined as an average, one may write $[-N\nabla^2_{\boldsymbol{\beta}} \ell_N(\hat{\boldsymbol{\beta}})]^{-1}$; the scaling is the same.)

For Logit, recall

$$-\nabla^2_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}) = \sum_{i=1}^{N} \Lambda(\eta_i)\big(1 - \Lambda(\eta_i)\big)\, \mathbf{x}_i \mathbf{x}_i^\top = \mathbf{X}^\top W(\boldsymbol{\beta}) \mathbf{X},$$

so the observed-information estimator becomes

$$\widehat{\text{Var}}(\hat{\boldsymbol{\beta}}) = \big(\mathbf{X}^\top W(\hat{\boldsymbol{\beta}})\mathbf{X}\big)^{-1}.$$

$\log \Phi(\cdot)$ and $\log(1 - \Phi(\cdot))$ are concave; composing with the linear index preserves concavity.

For Probit, the same form holds with weights $w_i(\boldsymbol{\beta}) = \xi_i(\eta_i + \xi_i)$:

$$-\nabla^2_{\boldsymbol{\beta}}\ell(\boldsymbol{\beta}) = \sum_{i=1}^{N} \xi_i(\eta_i + \xi_i)\, \mathbf{x}_i\mathbf{x}_i^\top = \mathbf{X}^\top W(\boldsymbol{\beta})\mathbf{X} \qquad \Rightarrow \qquad \widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}) = \big(\mathbf{X}^\top W(\hat{\boldsymbol{\beta}})\mathbf{X}\big)^{-1}.$$

Even when the model is misspecified, the M-estimator asymptotic variance remains valid in sandwich form by M-estimator theory:

$$\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}) = \left(\sum_{i=1}^{N} -\nabla^2_{\boldsymbol{\beta}}\ell_i(\hat{\boldsymbol{\beta}})\right)^{-1} \left(\sum_{i=1}^{N} S_i(\hat{\boldsymbol{\beta}})S_i(\hat{\boldsymbol{\beta}})^\top\right) \left(\sum_{i=1}^{N} -\nabla^2_{\boldsymbol{\beta}}\ell_i(\hat{\boldsymbol{\beta}})\right)^{-1}.$$

9.8 Inference on Marginal Effects

Inference for $\boldsymbol{\beta}$ in Probit and Logit models follows directly from the general M-estimator (or MLE) theory developed above: once we have $\hat{\boldsymbol{\beta}}$ and an estimator of its asymptotic variance, standard errors and hypothesis tests for components of $\boldsymbol{\beta}$ are immediate.

However, in nonlinear binary choice models, coefficients are often not the primary object of interest. Unlike the LPM, where β_k is itself the marginal effect of x_{ik} on $\Pr(y_i = 1 \mid x_i)$, in Probit and Logit the effect of a covariate on the probability depends on the index $\mathbf{x}_i^\top\boldsymbol{\beta}$. As a result, the economically meaningful quantities are typically *marginal effects*:

$$\mathrm{ME}_{ik}(\boldsymbol{\beta}) := \frac{\partial \Pr(y_i = 1 \mid \mathbf{x}_i)}{\partial x_{ik}} = \frac{\partial F(\mathbf{x}_i^\top\boldsymbol{\beta})}{\partial x_{ik}},$$

as well as their summaries such as $\mathrm{ME}_k(\bar{\mathbf{x}})$ and AME_k introduced earlier.

Since marginal effects are nonlinear functions of $\boldsymbol{\beta}$, inference on marginal effects is not obtained by directly reading off standard errors of $\hat{\boldsymbol{\beta}}$. Instead, we treat the marginal effect as a smooth function of the estimated parameter and apply the *delta method*. Concretely, for any target functional $m(\boldsymbol{\beta})$ (e.g. $m(\boldsymbol{\beta}) = \mathrm{ME}_k(\bar{\mathbf{x}})$ or $m(\boldsymbol{\beta}) = \mathrm{AME}_k$), we estimate it by $m(\hat{\boldsymbol{\beta}})$ and obtain its asymptotic variance from a first-order Taylor expansion around $\boldsymbol{\beta}_0$.

9.8.1 Delta Method

Let $b \in \mathbb{R}^k$ denote an estimator of the true parameter vector $\boldsymbol{\beta} \in \mathbb{R}^k$. Let $g : \mathbb{R}^k \to \mathbb{R}^m$ be a *continuously differentiable* function, and define its Jacobian matrix

$$C(b) := \frac{\partial g(b)}{\partial b^\top} = \left[\frac{\partial g_j(b)}{\partial b_\ell}\right]_{j=1,\dots,m;\ \ell=1,\dots,k}.$$

Assume $b \xrightarrow{p} \boldsymbol{\beta}$. By Slutsky's theorem and continuity of derivatives, we also have

$$C(b) \xrightarrow{p} C(\boldsymbol{\beta}).$$

Now consider the first-order Taylor expansion of $g(b)$ around $\boldsymbol{\beta}$:

$$g(b) = g(\boldsymbol{\beta}) + C(\boldsymbol{\beta})(b - \boldsymbol{\beta}) + o\big(\|b - \boldsymbol{\beta}\|\big).$$

Under $b \xrightarrow{p} \boldsymbol{\beta}$, the higher order terms are of smaller order than $\|b - \boldsymbol{\beta}\|$, so

$$g(b) - g(\boldsymbol{\beta}) \approx C(\boldsymbol{\beta})(b - \boldsymbol{\beta}).$$

If, moreover, b is asymptotically normal,

$$\sqrt{N}(b - \boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}(0, \boldsymbol{\Sigma}_b),$$

then by the delta method,

$$\sqrt{N}\big(g(b) - g(\boldsymbol{\beta})\big) \xrightarrow{d} \mathcal{N}\big(0,\ C(\boldsymbol{\beta})\boldsymbol{\Sigma}_b C(\boldsymbol{\beta})^\top\big),$$

equivalently,

$$g(b) \mathrel{\dot\sim} \mathcal{N}\left(g(\boldsymbol{\beta}), \frac{1}{N} C(\boldsymbol{\beta})\Sigma_b C(\boldsymbol{\beta})^\top\right).$$

In the MLE case, we typically take $b = \hat{\boldsymbol{\beta}}$ and $\boldsymbol{\Sigma}_b$ to be the (estimated) asymptotic covariance of $\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$, e.g. the inverse Fisher information (or its sample analog), or the robust sandwich analog under misspecification.

The term *delta* is just shorthand for a small change. In the Taylor expansion,

$$g(b) - g(\boldsymbol{\beta}) \approx C(\boldsymbol{\beta})(b - \boldsymbol{\beta}),$$

the quantity $(b - \boldsymbol{\beta})$ is a small perturbation of $\boldsymbol{\beta}$, and the delta method says that the induced perturbation in $g(\cdot)$ is approximately linear in that small change. The Jacobian $C(\boldsymbol{\beta})$ acts as the local conversion factor.

9.8.2 Delta Method for Marginal Effects

Fix an observation i. For a smooth link F, define

$$\boldsymbol{\pi}_i(\boldsymbol{\beta}) := f(\mathbf{x}_i^\top \boldsymbol{\beta})\, \boldsymbol{\beta} \in \mathbb{R}^k, \qquad f(\cdot) = F'(\cdot),$$

so that the marginal effects (with respect to all coordinates of $\mathbf{x}_i$) are stacked in the vector $\boldsymbol{\pi}_i(\boldsymbol{\beta})$. The plug-in estimator is

$$\hat{\boldsymbol{\pi}}_i := \boldsymbol{\pi}_i(\hat{\boldsymbol{\beta}}) = f(\mathbf{x}_i^\top \hat{\boldsymbol{\beta}})\, \hat{\boldsymbol{\beta}}.$$

By the delta method,

$$\text{A.Var}(\hat{\boldsymbol{\pi}}_i) = \left(\frac{\partial \boldsymbol{\pi}_i}{\partial \boldsymbol{\beta}^\top}\right) \text{A.Var}(\hat{\boldsymbol{\beta}}) \left(\frac{\partial \boldsymbol{\pi}_i}{\partial \boldsymbol{\beta}^\top}\right)^\top,$$

with the Jacobian evaluated at $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ and in practice at $\hat{\boldsymbol{\beta}}$. Let $\eta_i := \mathbf{x}_i^\top \boldsymbol{\beta}$, and write $f_i := f(\eta_i)$ and $f_i' := \frac{d}{d\eta} f(\eta)\big|_{\eta = \eta_i}$. Using the product rule,

$$\frac{\partial \boldsymbol{\pi}_i}{\partial \boldsymbol{\beta}^\top} = f_i \mathbf{I}_k + \boldsymbol{\beta}\, \frac{\partial f_i}{\partial \boldsymbol{\beta}^\top}.$$

Since $\frac{\partial f_i}{\partial \boldsymbol{\beta}^\top} = f_i' \frac{\partial \eta_i}{\partial \boldsymbol{\beta}^\top} = f_i' \mathbf{x}_i^\top$, we obtain the compact form

$$\frac{\partial \boldsymbol{\pi}_i}{\partial \boldsymbol{\beta}^\top} = f_i \mathbf{I}_k + f_i' \, \boldsymbol{\beta} \mathbf{x}_i^\top.$$

Therefore,

$$\text{A.Var}(\hat{\boldsymbol{\pi}}_i) = \left(f_i \mathbf{I}_k + f_i' \, \boldsymbol{\beta} \mathbf{x}_i^\top \right) \text{A.Var}(\hat{\boldsymbol{\beta}}) \left(f_i \mathbf{I}_k + f_i' \, \boldsymbol{\beta} \mathbf{x}_i^\top \right)^\top.$$

Logit. For Logit, $F(\eta) = \Lambda(\eta)$ and

$$\Lambda_i := \Lambda(\eta_i), \qquad f_i = \Lambda_i(1 - \Lambda_i).$$

Differentiate f_i w.r.t. η_i:

$$f_i' = \frac{d}{d\eta} \big[\Lambda(\eta)(1 - \Lambda(\eta)) \big]_{\eta = \eta_i} = \Lambda_i'(1 - \Lambda_i) - \Lambda_i \Lambda_i' = \Lambda_i'(1 - 2\Lambda_i).$$

Since $\Lambda_i' = f_i$, we have

$$f_i' = f_i(1 - 2\Lambda_i).$$

Plugging into the generic Jacobian gives

$$\frac{\partial \boldsymbol{\pi}_i}{\partial \boldsymbol{\beta}^\top} = f_i \mathbf{I}_k + f_i(1 - 2\Lambda_i) \, \boldsymbol{\beta} \mathbf{x}_i^\top = f_i \Big[\mathbf{I}_k + (1 - 2\Lambda_i) \boldsymbol{\beta} \mathbf{x}_i^\top \Big],$$

and hence

$$\text{A.Var}(\hat{\boldsymbol{\pi}}_i) = f_i^2 \Big[\mathbf{I}_k + (1 - 2\Lambda_i) \boldsymbol{\beta} \mathbf{x}_i^\top \Big] \text{A.Var}(\hat{\boldsymbol{\beta}}) \Big[\mathbf{I}_k + (1 - 2\Lambda_i) \boldsymbol{\beta} \mathbf{x}_i^\top \Big]^\top.$$

Probit. For Probit, $F(\eta) = \Phi(\eta)$ and $f(\eta) = \phi(\eta)$, so

$$f_i = \phi_i := \phi(\eta_i).$$

Once again using $\phi'(\eta) = -\eta \, \phi(\eta)$,

$$f_i' = \phi_i' = -\eta_i \phi_i.$$

Therefore,

$$\frac{\partial \boldsymbol{\pi}_i}{\partial \boldsymbol{\beta}^\top} = \phi_i \mathbf{I}_k - \eta_i \phi_i \boldsymbol{\beta} \mathbf{x}_i^\top = \phi_i \Big[\mathbf{I}_k - \eta_i \boldsymbol{\beta} \mathbf{x}_i^\top \Big],$$

and

$$\text{A.Var}(\hat{\boldsymbol{\pi}}_i) = \phi_i^2 \Big[\mathbf{I}_k - \eta_i \boldsymbol{\beta} \mathbf{x}_i^\top \Big] \text{A.Var}(\hat{\boldsymbol{\beta}}) \Big[\mathbf{I}_k - \eta_i \boldsymbol{\beta} \mathbf{x}_i^\top \Big]^\top.$$

9.9 Hypothesis Testing

At this point we have all the ingredients needed for classical large-sample tests: the likelihood $\ell(\beta)$, the score $\nabla_\beta \ell(\beta)$, and an asymptotic variance estimator for $\hat{\beta}$. We focus on

testing parametric restrictions on β, either linear or nonlinear. Throughout, the large-sample justification comes from

$$\sqrt{N}(\hat{\beta} - \beta_0) \xrightarrow{d} \mathcal{N}(0, \mathbf{\Sigma}),$$

where $\mathbf{\Sigma}$ denotes the asymptotic covariance of $\sqrt{N}(\hat{\beta} - \beta_0)$ (estimated in practice by the inverse information under correct specification, or by the sandwich form under misspecification).

9.9.1 Wald Test

As discussed before, the Wald test asks: *is the unrestricted estimate $\hat{\beta}$ close enough to satisfying the restriction?* It uses (i) a plug-in estimate of the restriction evaluated at $\hat{\beta}$ and (ii) the asymptotic variance of that plug-in quantity (via the delta method). Computationally, Wald is cheap: it only needs the unrestricted estimate and a variance estimator.

Algorithm 10 Wald test for $H_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{r}$ in binary choice MLE (Logit/Probit)

Require: Data $\{(y_i, \mathbf{x}_i)\}_{i=1}^N$, link F (Logit: Λ, Probit: Φ), score $S_i(\boldsymbol{\beta}) = \nabla_{\boldsymbol{\beta}} \ell_i(\boldsymbol{\beta})$, Hessian $\mathcal{H}_i(\boldsymbol{\beta}) = \nabla^2_{\boldsymbol{\beta}} \ell_i(\boldsymbol{\beta})$, restriction matrix $\mathbf{R} \in \mathbb{R}^{q \times k}$ (full row rank), vector $\mathbf{r} \in \mathbb{R}^q$, significance level α.

1: Estimate $\hat{\boldsymbol{\beta}}$ by MLE:

$$\hat{\boldsymbol{\beta}} \in \arg\max_{\boldsymbol{\beta}} \ \ell(\boldsymbol{\beta}), \qquad \ell(\boldsymbol{\beta}) = \sum_{i=1}^{N} \Big[y_i \log F(\mathbf{x}_i^\top \boldsymbol{\beta}) + (1 - y_i) \log\big(1 - F(\mathbf{x}_i^\top \boldsymbol{\beta})\big) \Big].$$

2: Compute an asymptotic covariance estimator for $\hat{\boldsymbol{\beta}}$ using Hessians:

$$\widehat{\mathbf{\Sigma}} := \left(\frac{1}{N} \sum_{i=1}^{N} -\mathcal{H}_i(\hat{\boldsymbol{\beta}}) \right)^{-1}.$$

3: If allowing misspecification, use the sandwich form:

$$\widehat{\mathbf{\Sigma}} := \left(\frac{1}{N} \sum_{i=1}^{N} -\mathcal{H}_i(\hat{\boldsymbol{\beta}}) \right)^{-1} \left(\frac{1}{N} \sum_{i=1}^{N} S_i(\hat{\boldsymbol{\beta}}) S_i(\hat{\boldsymbol{\beta}})^\top \right) \left(\frac{1}{N} \sum_{i=1}^{N} -\mathcal{H}_i(\hat{\boldsymbol{\beta}}) \right)^{-1}.$$

4: Define the restriction function $\mathbf{g}(\boldsymbol{\beta}) := \mathbf{R}\boldsymbol{\beta} - \mathbf{r}$ and compute the plug-in estimate $\hat{\mathbf{g}} := \mathbf{g}(\hat{\boldsymbol{\beta}}) = \mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}$.

5: The Jacobian is $\nabla_{\boldsymbol{\beta}} \mathbf{g}(\boldsymbol{\beta}) = \mathbf{R}$, so

$$\widehat{\mathbf{\Omega}} := \mathbf{R}\, \widehat{\mathbf{\Sigma}}\, \mathbf{R}^\top$$

estimates the asymptotic covariance of $\sqrt{N}\hat{\mathbf{g}}$.

6: Compute the Wald statistic:

$$W := N \hat{\mathbf{g}}^\top \widehat{\mathbf{\Omega}}^{-1} \hat{\mathbf{g}}.$$

7: Compute p-value: $p := 1 - F_{\chi^2_q}(W)$.

8: Reject H_0 if $p < \alpha$ (equivalently, if $W > \chi^2_{q,\,1-\alpha}$).

Algorithm 11 Wald test for $H_0 : h(\boldsymbol{\beta}) = \mathbf{0}$ in binary choice MLE (Logit/Probit)

Require: Data $\{(y_i, \mathbf{x}_i)\}_{i=1}^N$, link F (Logit: Λ, Probit: Φ), score $S_i(\boldsymbol{\beta}) = \nabla_{\boldsymbol{\beta}} \ell_i(\boldsymbol{\beta})$, Hessian $\mathcal{H}_i(\boldsymbol{\beta}) = \nabla^2_{\boldsymbol{\beta}} \ell_i(\boldsymbol{\beta})$, restriction function $h : \mathbb{R}^k \to \mathbb{R}^q$ (continuously differentiable), Jacobian $\mathbf{H}(\boldsymbol{\beta}) := \partial h(\boldsymbol{\beta}) / \partial \boldsymbol{\beta}^\top$, significance level α.

1: Estimate $\hat{\boldsymbol{\beta}}$ by MLE:

$$\hat{\boldsymbol{\beta}} \in \arg\max_{\boldsymbol{\beta}} \ \ell(\boldsymbol{\beta}), \qquad \ell(\boldsymbol{\beta}) = \sum_{i=1}^N \Big[y_i \log F(\mathbf{x}_i^\top \boldsymbol{\beta}) + (1 - y_i) \log\big(1 - F(\mathbf{x}_i^\top \boldsymbol{\beta})\big) \Big].$$

2: Compute an asymptotic covariance estimator for $\hat{\boldsymbol{\beta}}$ using Hessians (information-based):

$$\widehat{\boldsymbol{\Sigma}} := \left(\frac{1}{N} \sum_{i=1}^N -\mathcal{H}_i(\hat{\boldsymbol{\beta}}) \right)^{-1}.$$

3: If allowing misspecification, use the sandwich form:

$$\widehat{\boldsymbol{\Sigma}} := \left(\frac{1}{N} \sum_{i=1}^N -\mathcal{H}_i(\hat{\boldsymbol{\beta}}) \right)^{-1} \left(\frac{1}{N} \sum_{i=1}^N S_i(\hat{\boldsymbol{\beta}}) S_i(\hat{\boldsymbol{\beta}})^\top \right) \left(\frac{1}{N} \sum_{i=1}^N -\mathcal{H}_i(\hat{\boldsymbol{\beta}}) \right)^{-1}.$$

4: Evaluate the restriction at the estimate: $\hat{\mathbf{h}} := h(\hat{\boldsymbol{\beta}}) \in \mathbb{R}^q$.

5: Evaluate the Jacobian at the estimate: $\hat{\mathbf{H}} := \mathbf{H}(\hat{\boldsymbol{\beta}}) \in \mathbb{R}^{q \times k}$.

6: Compute the estimated covariance of $\sqrt{N}\, \hat{\mathbf{h}}$ (delta method):

$$\widehat{\boldsymbol{\Omega}} := \hat{\mathbf{H}}\, \widehat{\boldsymbol{\Sigma}}\, \hat{\mathbf{H}}^\top.$$

7: Compute the Wald statistic:

$$W := N\, \hat{\mathbf{h}}^\top \widehat{\boldsymbol{\Omega}}^{-1} \hat{\mathbf{h}}.$$

8: Compute p-value: $p := 1 - F_{\chi^2_q}(W)$.

9: Reject H_0 if $p < \alpha$ (equivalently, if $W > \chi^2_{q,\,1-\alpha}$).

9.9.2 Likelihood Ratio (LR) Tests

The LR test asks a question similar in spirit: *how much does the best possible fit worsen when we force the restriction to hold?* It compares the maximized log-likelihood under the unrestricted model to the maximized log-likelihood under the restricted model. The LR test is likelihood-native and does not require a separate variance estimator, but it does require solving the restricted optimization problem.

Algorithm 12 Likelihood ratio test for H_0 imposing q restrictions in binary choice MLE (Logit/Probit)

Require: Data $\{(y_i, \mathbf{x}_i)\}_{i=1}^N$, link F (Logit: Λ, Probit: Φ), log-likelihood

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{N} \Big[y_i \log F(\mathbf{x}_i^\top \boldsymbol{\beta}) + (1 - y_i) \log\big(1 - F(\mathbf{x}_i^\top \boldsymbol{\beta})\big) \Big],$$

unrestricted MLE $\hat{\boldsymbol{\beta}}$, restricted parameter set Θ_0 (encoding H_0), significance level α.

1: Compute the unrestricted maximized log-likelihood: $\ell_U := \ell(\hat{\boldsymbol{\beta}})$.

2: Compute the restricted MLE:

$$\hat{\boldsymbol{\beta}}_0 \in \arg\max_{\boldsymbol{\beta} \in \Theta_0} \ \ell(\boldsymbol{\beta}).$$

3: Compute the restricted maximized log-likelihood: $\ell_R := \ell(\hat{\boldsymbol{\beta}}_0)$.

4: Compute the LR statistic:

$$LR := 2\big(\ell_U - \ell_R\big).$$

5: Compute p-value: $p := 1 - F_{\chi^2_q}(LR)$.

6: Reject H_0 if $p < \alpha$ (equivalently, if $LR > \chi^2_{q,\,1-\alpha}$).

9.9.3 Lagrange Multiplier (LM) / Score Tests

What if we ask: *if we force the restriction to hold, does the likelihood still want to move away from it?* That is what LM test is doing. It evaluates the score at the restricted estimate. If H_0 is true, the score (properly scaled) should be close to zero at the restricted optimum. Computationally, LM avoids solving the unrestricted problem (helpful when the unrestricted model is expensive), but it does require computing the score and an information matrix at the restricted estimate.

Algorithm 13 LM / score test for H_0 imposing q restrictions in binary choice MLE (Logit/Probit)

Require: Data $\{(y_i, \mathbf{x}_i)\}_{i=1}^N$, link F (Logit: Λ, Probit: Φ), log-likelihood

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{N} \Big[y_i \log F(\mathbf{x}_i^\top \boldsymbol{\beta}) + (1 - y_i) \log\big(1 - F(\mathbf{x}_i^\top \boldsymbol{\beta})\big) \Big],$$

individual score $S_i(\boldsymbol{\beta}) := \nabla_{\boldsymbol{\beta}} \ell_i(\boldsymbol{\beta})$, individual Hessian $\mathcal{H}_i(\boldsymbol{\beta}) := \nabla^2_{\boldsymbol{\beta}} \ell_i(\boldsymbol{\beta})$, restricted parameter set Θ_0 (encoding H_0), significance level α.

1: Compute the restricted MLE:

$$\hat{\boldsymbol{\beta}}_0 \in \arg\max_{\boldsymbol{\beta} \in \Theta_0} \ \ell(\boldsymbol{\beta}).$$

2: Evaluate the (sample) score at the restriction:

$$\hat{\mathbf{S}} := \nabla_{\boldsymbol{\beta}} \ell(\hat{\boldsymbol{\beta}}_0) = \sum_{i=1}^{N} S_i(\hat{\boldsymbol{\beta}}_0).$$

3: Compute an information matrix estimator at the restriction (observed information):

$$\widehat{\mathbf{I}}_0 := -\nabla^2_{\boldsymbol{\beta}} \ell(\hat{\boldsymbol{\beta}}_0) = \sum_{i=1}^{N} \big(-\mathcal{H}_i(\hat{\boldsymbol{\beta}}_0)\big).$$

4: If using expected information, replace $\widehat{\mathbf{I}}_0$ by its sample analog.

5: Compute the LM statistic:

$$LM := \hat{\mathbf{S}}^\top \widehat{\mathbf{I}}_0^{-1} \hat{\mathbf{S}}.$$

6: Compute p-value: $p := 1 - F_{\chi^2_q}(LM)$.

7: Reject H_0 if $p < \alpha$ (equivalently, if $LM > \chi^2_{q,\,1-\alpha}$).

All three tests are asymptotically equivalent under H_0 (same χ^2_q limit), but they differ in what they require computationally: Wald uses only the unrestricted estimate and $\widehat{\boldsymbol{\Sigma}}$, LR requires both unrestricted and restricted maximization, and LM requires only the restricted estimate plus the score/information evaluated at the restriction.

9.10 Goodness of Fit

In linear regression, goodness of fit is often summarized by R^2. For Logit/Probit models estimated by maximum likelihood, the usual variance-decomposition interpretation does not apply, so fit is typically assessed using likelihood-based measures.

Let $\hat{\boldsymbol{\beta}}$ denote the unrestricted MLE, and let $\hat{\boldsymbol{\beta}}_0$ denote the MLE from the intercept-only ("null") model. Write $\ell(\hat{\boldsymbol{\beta}})$ and $\ell(\hat{\boldsymbol{\beta}}_0)$ for the corresponding maximized *log-likelihood* values, where

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{N} \Big[y_i \log F(\mathbf{x}_i^\top \boldsymbol{\beta}) + (1 - y_i) \log\big(1 - F(\mathbf{x}_i^\top \boldsymbol{\beta})\big) \Big].$$

9.10.1 Pseudo R^2

Define the log-likelihood gain (relative to the null model)

$$\Delta\ell := \ell(\hat{\boldsymbol{\beta}}) - \ell(\hat{\boldsymbol{\beta}}_0).$$

A likelihood-based pseudo-R^2 is

$$R^2_{\text{pseudo}} := 1 - \frac{1}{1 + \frac{2\Delta\ell}{N}} = 1 - \frac{1}{1 + \frac{2\left(\ell(\hat{\boldsymbol{\beta}}) - \ell(\hat{\boldsymbol{\beta}}_0)\right)}{N}}.$$

Equivalently,

$$R^2_{\text{pseudo}} = \frac{\frac{2\Delta\ell}{N}}{1 + \frac{2\Delta\ell}{N}} = \frac{\Delta\ell}{\frac{N}{2} + \Delta\ell}.$$

9.10.2 McFadden's R^2

McFadden's pseudo-R^2 is defined as

$$R^2_{\text{McF}} := 1 - \frac{\ell(\hat{\boldsymbol{\beta}})}{\ell(\hat{\boldsymbol{\beta}}_0)}.$$

It compares the maximized log-likelihood under the full model to that under the intercept-only baseline.

Both measures are likelihood-based and should not be interpreted as "fraction of variance explained." They are best viewed as summaries of how much the model improves fit relative to the null specification, and their numerical scale typically differs from the OLS R^2.

9.11 Endogeneity and Full-Information MLE

Up to now, Probit and Logit models looked pleasantly self-contained: specify a link function F, estimate $\boldsymbol{\beta}$ by MLE, and do inference using the information matrix and the delta method. That entire pipeline, however, leans on *conditional exogeneity*. When an important regressor is endogenous, the likelihood we maximized is built on the wrong conditional distribution, and the resulting $\hat{\boldsymbol{\beta}}$ is not merely imprecise; it is biased for the structural parameter we care about. The usual "just instrument it" instinct is not straightforward in nonlinear probability models. So we sketch one standard remedy: specify a joint model for the endogenous regressor and the latent error and estimate all parameters by *full-information maximum likelihood* (FIML).

Suppose the binary outcome is generated by a latent index model

$$y_i = \mathbf{1}\{y_i^* > 0\}, \qquad y_i^* = \mathbf{x}_i^\top \boldsymbol{\beta} + \delta w_i + \varepsilon_i,$$

where w_i is *endogenous*. In this case, the core binary-choice assumption behind single-equation Probit/Logit,

$$\varepsilon_i \mid (\mathbf{x}_i, w_i) \;\sim\; F,$$

where F is a known distribution, is not credible: endogeneity means ε_i and w_i are not

independent (equivalently, $\mathbb{E}[\varepsilon_i \mid w_i] \neq 0$), so the usual single-equation likelihood is misspecified.

9.11.1 Attempting 2SLS

A naive thought is to instrument w_i and run 2SLS. In a *linear* model, this instinct is correct: the first stage projects w_i onto the instrument space, and the fitted value $\hat{w}_i$ is (by construction) the component of w_i that lives in the span of valid instruments. Geometrically, IV is a projection argument in a linear inner-product space: the fitted part is the "orthogonal-basis-safe" component, and the residual is orthogonal to the instruments.

In a nonlinear binary choice model, this geometry does not survive the link function. The probability is

$$\Pr(y_i = 1 \mid \mathbf{x}_i, w_i) = F(\mathbf{x}_i^\top \boldsymbol{\beta} + \delta w_i),$$

so a two-step replacement $w_i \mapsto \hat{w}_i$ produces $F(\mathbf{x}_i^\top \boldsymbol{\beta} + \delta \hat{w}_i)$, not $F(\mathbf{x}_i^\top \boldsymbol{\beta} + \delta w_i)$. Even if $\hat{w}_i$ is the best linear projection of w_i onto instruments, the nonlinear transformation $F(\cdot)$ generally reintroduces dependence between the transformed index and the structural error. Put differently: linear IV works because orthogonality is preserved under linear operations; a nonlinear map can bend the instrument-projected variation in ways that destroy the clean "orthogonal decomposition" intuition (one can think of this as a form of *concurvity*).

The deeper issue is twofold: (i) credible excluded instruments are often difficult to justify in binary choice applications, and (ii) even when instruments $\mathbf{z}_i$ are available, exploiting them typically requires additional structure (e.g. a joint model for (ε_i, u_i)) rather than a direct two-stage plug-in.

9.11.2 A Moment-Condition Attempt

One might try to mimic linear IV logic by writing a moment condition such as

$$\mathbb{E}\Big[(y_i^* - \mathbf{x}_i^\top \boldsymbol{\beta} - \delta w_i)\, \mathbf{z}_i\Big] = 0,$$

with instruments $\mathbf{z}_i$. The problem is immediate: y_i^* is *unobservable*.

Example 9.1. *Suppose for simplicity that $\mathbf{z}_i$ is a scalar instrument and the latent index is*

$$y_i^* = \delta w_i + \varepsilon_i, \qquad y_i = \mathbf{1}\{y_i^* > 0\}.$$

If we could observe y_i^*, the IV moment $\mathbb{E}[(y_i^* - \delta w_i)\mathbf{z}_i] = 0$ would reduce to $\mathbb{E}[\varepsilon_i \mathbf{z}_i] = 0$, and we could proceed as in the linear case. But we only observe the sign of y_i^* through y_i. Knowing $y_i = 1$ tells us that $y_i^* > 0$, i.e. $\varepsilon_i > -\delta w_i$, but it does not tell us the magnitude of y_i^*. Many different values of (w_i, ε_i) produce the same y_i. As a result, there is no observable residual of the form $(y_i^* - \delta w_i)$ to multiply by $\mathbf{z}_i$.

One way around the unobserved y_i^* is to treat it as missing data and integrate it out ("data augmentation").

Example 9.2 (A Probit Example)**.** *Consider the Probit latent-index model*

$$y_i = \mathbf{1}\{y_i^* > 0\}, \qquad y_i^* = \mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0,1),$$

with $\{(y_i, \mathbf{x}_i)\}_{i=1}^N$ *observed. If we could observe* y_i^**, the model would be a Gaussian linear regression with known error variance, and estimation of* $\boldsymbol{\beta}$ *would reduce to least squares. Data augmentation makes this idea operational by treating* y_i^* *as missing data.*

The latent regression implies the joint density:

$$p(y_i, y_i^* \mid \mathbf{x}_i, \boldsymbol{\beta}) = p(y_i \mid y_i^*)\, p(y_i^* \mid \mathbf{x}_i, \boldsymbol{\beta}), \qquad p(y_i \mid y_i^*) = \mathbf{1}\{y_i = \mathbf{1}(y_i^* > 0)\},$$

since $y_i^* > 0 \Rightarrow y_i = 1$ *is deterministic and vice versa, and*

$$y_i^* \mid \mathbf{x}_i, \boldsymbol{\beta} \sim \mathcal{N}(\mathbf{x}_i^\top \boldsymbol{\beta}, 1).$$

Conditioning on y_i *enforces truncation. Let* $\mu_i := \mathbf{x}_i^\top \boldsymbol{\beta}$*. Then*

$$y_i^* \mid (y_i = 1, \mathbf{x}_i, \boldsymbol{\beta}) \sim \mathcal{N}(\mu_i, 1) \text{ truncated to } (0, \infty),$$

and

$$y_i^* \mid (y_i = 0, \mathbf{x}_i, \boldsymbol{\beta}) \sim \mathcal{N}(\mu_i, 1) \text{ truncated to } (-\infty, 0].$$

Equivalently, the density is

$$p(y_i^* \mid y_i, \mathbf{x}_i, \boldsymbol{\beta}) = \frac{\phi(y_i^* - \mu_i)}{\Pr(y_i \mid \mathbf{x}_i, \boldsymbol{\beta})}\, \mathbf{1}\{y_i^* \in A_i\},$$

where $A_i = (0, \infty)$ *if* $y_i = 1$ *and* $A_i = (-\infty, 0]$ *if* $y_i = 0$*, and*

$$\Pr(y_i = 1 \mid \mathbf{x}_i, \boldsymbol{\beta}) = \Phi(\mu_i), \qquad \Pr(y_i = 0 \mid \mathbf{x}_i, \boldsymbol{\beta}) = 1 - \Phi(\mu_i).$$

Once the latent variables $\mathbf{y}^* := (y_1^*, \ldots, y_N^*)^\top$ *are treated as observed, the model becomes*

$$\mathbf{y}^* = \mathbf{X}\boldsymbol{\beta} + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, I_N),$$

where $\mathbf{X} = [\mathbf{x}_1^\top; \ldots; \mathbf{x}_N^\top]$*. Hence the likelihood for* $\boldsymbol{\beta}$ *is*

$$p(\mathbf{y}^* \mid \mathbf{X}, \boldsymbol{\beta}) \propto \exp\left(-\frac{1}{2}(\mathbf{y}^* - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y}^* - \mathbf{X}\boldsymbol{\beta})\right).$$

Now impose the conjugate Gaussian prior

$$\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{b}_0, \mathbf{B}_0), \qquad \mathbf{B}_0 \succ 0.$$

Example 9.3 (A Probit Example (cont.))**.** *By Bayes' rule, the posterior kernel is the product:*

$$p(\boldsymbol{\beta} \mid \mathbf{y}^*, \mathbf{X}) \;\propto\; p(\mathbf{y}^* \mid \mathbf{X}, \boldsymbol{\beta})\, p(\boldsymbol{\beta}).$$

Taking logs and dropping constants (terms not involving $\boldsymbol{\beta}$),

$$\begin{aligned}\log p(\boldsymbol{\beta} \mid \mathbf{y}^*, \mathbf{X}) &= -\frac{1}{2}(\mathbf{y}^* - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y}^* - \mathbf{X}\boldsymbol{\beta}) - \frac{1}{2}(\boldsymbol{\beta} - \mathbf{b}_0)^\top \mathbf{B}_0^{-1}(\boldsymbol{\beta} - \mathbf{b}_0) + \textit{const} \\ &= -\frac{1}{2}\Big[\boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{X}\boldsymbol{\beta} - 2\boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{y}^* + \boldsymbol{\beta}^\top \mathbf{B}_0^{-1}\boldsymbol{\beta} - 2\boldsymbol{\beta}^\top \mathbf{B}_0^{-1}\mathbf{b}_0\Big] + \textit{const.}\end{aligned}$$

Let

$$\mathbf{B}_1 := (\mathbf{B}_0^{-1} + \mathbf{X}^\top \mathbf{X})^{-1}, \qquad \mathbf{b}_1 := \mathbf{B}_1(\mathbf{B}_0^{-1}\mathbf{b}_0 + \mathbf{X}^\top \mathbf{y}^*).$$

Then

$$\boldsymbol{\beta} \mid (\mathbf{y}^*, \mathbf{X}) \sim \mathcal{N}(\mathbf{b}_1, \mathbf{B}_1).$$

Side Notes: *If we take an increasingly diffuse prior (formally $\mathbf{B}_0^{-1} \to 0$), then $\mathbf{B}_1 \to (\mathbf{X}^\top \mathbf{X})^{-1}$ and*

$$\mathbf{b}_1 \to (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{y}^*,$$

which is the OLS estimator from regressing the augmented latent outcome $\mathbf{y}^*$ on $\mathbf{X}$ (when $\mathbf{X}^\top \mathbf{X}$ is invertible).
Data augmentation alternates between:

(A) *Draw y_i^* independently from the truncated normal*

$$y_i^* \mid (y_i, \mathbf{x}_i, \boldsymbol{\beta}) \sim \mathcal{N}(\mathbf{x}_i^\top \boldsymbol{\beta}, 1) \textit{ truncated to } (0, \infty) \textit{ if } y_i = 1,$$

$$y_i^* \mid (y_i, \mathbf{x}_i, \boldsymbol{\beta}) \sim \mathcal{N}(\mathbf{x}_i^\top \boldsymbol{\beta}, 1) \textit{ truncated to } (-\infty, 0] \textit{ if } y_i = 0;$$

(B) *Draw $\boldsymbol{\beta}$ from its conditional Gaussian distribution*

$$\boldsymbol{\beta} \mid (\mathbf{y}^*, \mathbf{X}) \sim \mathcal{N}(\mathbf{b}_1, \mathbf{B}_1).$$

Iterating (A)–(B) generates a Markov chain whose stationary distribution is the joint posterior $p(\boldsymbol{\beta}, \mathbf{y}^* \mid \mathbf{y}, \mathbf{X})$.

The augmentation is a computational device; it does not change the implied likelihood for the observed data. Indeed,

$$\Pr(y_i = 1 \mid \mathbf{x}_i, \boldsymbol{\beta}) = \Pr(y_i^* > 0 \mid \mathbf{x}_i, \boldsymbol{\beta}) = \Pr(\varepsilon_i > -\mathbf{x}_i^\top \boldsymbol{\beta}) = \Phi(\mathbf{x}_i^\top \boldsymbol{\beta}),$$

and similarly $\Pr(y_i = 0 \mid \mathbf{x}_i, \boldsymbol{\beta}) = 1 - \Phi(\mathbf{x}_i^\top \boldsymbol{\beta})$. Thus integrating out $\mathbf{y}^*$ recovers the usual Probit likelihood

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{N} \Phi(\mathbf{x}_i^\top \boldsymbol{\beta})^{y_i} \big(1 - \Phi(\mathbf{x}_i^\top \boldsymbol{\beta})\big)^{1 - y_i}.$$

9.11.3 A Frequentist Fix: Joint Modeling and Full-Information MLE

The data-augmentation route above is a natural Bayesian response to the missing latent variable y_i^*: introduce it explicitly and sample (or integrate) it out. In applied econometrics, however, sampling is expensive and endogeneity is often treated in a *frequentist* way: rather than placing priors and augmenting latent variables, we *change the likelihood* to reflect the endogeneity mechanism. Concretely, we specify a reduced form for the endogenous regressor and a joint distribution for the structural error and the first-stage error. This turns the problem into a likelihood model for the *observables* (y_i, w_i) given $(\mathbf{x}_i, \mathbf{z}_i)$, and we estimate all parameters by maximum likelihood. We illustrate this using the following Probit example.

Assume the latent index is

$$y_i = \mathbf{1}\{y_i^* > 0\}, \qquad y_i^* = \mathbf{x}_i^\top \boldsymbol{\beta} + \delta w_i + \varepsilon_i, \qquad \mathrm{Var}\,(\varepsilon_i) = 1,$$

where w_i is endogenous. Introduce a first-stage (reduced-form) equation

$$w_i = \mathbf{z}_i^\top \boldsymbol{\alpha} + u_i, \qquad \mathrm{Var}\,(u_i) = \sigma_u^2,$$

with $\mathbf{z}_i$ excluded instruments (and/or exogenous controls). Endogeneity is captured by allowing (ε_i, u_i) to be correlated. Following the notes, assume a bivariate normal disturbance:

$$\begin{pmatrix} \varepsilon_i \\ u_i \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho\sigma_u \\ \rho\sigma_u & \sigma_u^2 \end{pmatrix} \right),$$

so that $\mathrm{Cov}(\varepsilon_i, u_i) = \rho\sigma_u$ and $\mathrm{Corr}(\varepsilon_i, u_i) = \rho$.

Let $u_i = w_i - \mathbf{z}_i^\top \boldsymbol{\alpha}$. A standard property of the bivariate normal implies

$$\varepsilon_i \mid u_i \sim \mathcal{N}\left(\frac{\rho}{\sigma_u} u_i,\ 1 - \rho^2 \right).$$

Therefore,

$$\begin{aligned} \Pr(y_i = 1 \mid \mathbf{x}_i, w_i, \mathbf{z}_i) &= \Pr\big(y_i^* > 0 \mid \mathbf{x}_i, w_i, \mathbf{z}_i\big) \\ &= \Pr\big(\mathbf{x}_i^\top \boldsymbol{\beta} + \delta w_i + \varepsilon_i > 0 \mid u_i\big) \\ &= \Pr\big(\varepsilon_i > -\mathbf{x}_i^\top \boldsymbol{\beta} - \delta w_i \mid u_i\big) \\ &= \Phi\left(\frac{\mathbf{x}_i^\top \boldsymbol{\beta} + \delta w_i + \frac{\rho}{\sigma_u} u_i}{\sqrt{1-\rho^2}} \right) \\ &= \Phi\left(\frac{\mathbf{x}_i^\top \boldsymbol{\beta} + \delta w_i + \frac{\rho}{\sigma_u} (w_i - \mathbf{z}_i^\top \boldsymbol{\alpha})}{\sqrt{1-\rho^2}} \right). \end{aligned}$$

Define the Probit index appearing above as

$$\zeta_i := \frac{\mathbf{x}_i^\top \boldsymbol{\beta} + \delta w_i + \frac{\rho}{\sigma_u} (w_i - \mathbf{z}_i^\top \boldsymbol{\alpha})}{\sqrt{1-\rho^2}}.$$

Under $w_i = \mathbf{z}_i^\top \boldsymbol{\alpha} + u_i$ with $u_i \sim \mathcal{N}(0, \sigma_u^2)$,

$$w_i \mid \mathbf{z}_i \sim \mathcal{N}(\mathbf{z}_i^\top \boldsymbol{\alpha}, \sigma_u^2), \qquad f(w_i \mid \mathbf{z}_i) = \frac{1}{\sigma_u}\phi\left(\frac{w_i - \mathbf{z}_i^\top \boldsymbol{\alpha}}{\sigma_u}\right).$$

Let

$$\nu_i := \frac{w_i - \mathbf{z}_i^\top \boldsymbol{\alpha}}{\sigma_u}, \qquad \text{so that} \qquad f(w_i \mid \mathbf{z}_i) = \frac{1}{\sigma_u}\phi(\nu_i).$$

The likelihood for the *observables* (y_i, w_i) given $(\mathbf{x}_i, \mathbf{z}_i)$ is

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{i=1}^{N} \Big[\Phi(\zeta_i)\Big]^{y_i} \Big[1 - \Phi(\zeta_i)\Big]^{1-y_i} \cdot \frac{1}{\sigma_u}\phi(\nu_i),$$

where the parameter vector is

$$\boldsymbol{\theta} := (\boldsymbol{\beta}, \delta, \boldsymbol{\alpha}, \sigma_u, \rho).$$

Maximizing $\mathcal{L}(\boldsymbol{\theta})$ over $\boldsymbol{\theta}$ yields the so-called **full-information MLE** (FIML).

9.12 From Binary to Multinomial Choice

Binary choice models (Logit/Probit) cover the case $y_i \in \{0, 1\}$. In many applications, however, the outcome has more than two *unordered* categories, e.g. $y_i \in \{1, \ldots, J\}$. The goal is then to model the full vector of conditional choice probabilities

$$p_{ij}(\boldsymbol{\beta}) := \Pr(y_i = j \mid \mathbf{x}_i), \qquad j = 1, \ldots, J, \qquad \sum_{j=1}^{J} p_{ij}(\boldsymbol{\beta}) = 1.$$

The logic of binary choice generalizes cleanly: we keep an index structure in utilities, map it to probabilities via a link implied by a distributional assumption, estimate by MLE, and conduct inference using the same M-estimator / delta-method pipeline.

9.12.1 Latent-utility setup and normalization

A standard structural starting point is a latent utility representation. For each alternative $j \in \{1, \ldots, J\}$, let

$$y_{ij}^* = \mathbf{x}_i^\top \boldsymbol{\beta}_j + \varepsilon_{ij}, \qquad y_i = \arg \max_{j \in \{1,\ldots,J\}} y_{ij}^*.$$

Only *differences* in utilities matter for the argmax. Hence the model is not identified without a normalization: adding the same constant to all y_{ij}^* leaves y_i unchanged. A convenient normalization is to select a baseline category, say $j = 1$, and set

$$\boldsymbol{\beta}_1 = \mathbf{0}, \qquad \text{so the free parameters are } (\boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_J).$$

9.12.2 Multinomial Logit (MNL)

The multinomial analogue of Logit arises when the utility shocks $\{\varepsilon_{ij}\}_{j=1}^{J}$ are i.i.d. Type-I extreme value. Then the choice probabilities take the so-called softmax form:

$$\Pr(y_i = j \mid \mathbf{x}_i) = p_{ij}(\boldsymbol{\beta}) = \frac{\exp(\mathbf{x}_i^\top \boldsymbol{\beta}_j)}{\sum_{\ell=1}^{J} \exp(\mathbf{x}_i^\top \boldsymbol{\beta}_\ell)}, \qquad j = 1, \ldots, J,$$

with the normalization $\boldsymbol{\beta}_1 = \mathbf{0}$ implicitly understood. This is the direct generalization of binary Logit: instead of mapping one index through $\Lambda(\cdot)$, we map the vector of indices $\{\mathbf{x}_i^\top \boldsymbol{\beta}_j\}_{j=1}^J$ through a softmax that enforces the simplex restriction automatically.

A key implication of MNL is the **IIA property** (Independence of Irrelevant Alternatives). For any pair of alternatives j, ℓ,

$$\frac{\Pr(y_i = j \mid \mathbf{x}_i)}{\Pr(y_i = \ell \mid \mathbf{x}_i)} = \exp\big(\mathbf{x}_i^\top (\boldsymbol{\beta}_j - \boldsymbol{\beta}_\ell)\big),$$

which does not depend on the presence or attributes of other alternatives. This can be reasonable when alternatives are well-separated, but it can be restrictive when some options are "close substitutes."

9.12.3 Estimation by MLE

Given $p_{ij}(\boldsymbol{\beta})$, the conditional likelihood contribution is

$$\mathcal{L}_i(\boldsymbol{\beta}) = \prod_{j=1}^J p_{ij}(\boldsymbol{\beta})^{\mathbf{1}\{y_i = j\}}, \qquad \ell_i(\boldsymbol{\beta}) = \sum_{j=1}^J \mathbf{1}\{y_i = j\} \log p_{ij}(\boldsymbol{\beta}),$$

so the sample log-likelihood is

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^N \ell_i(\boldsymbol{\beta}) = \sum_{i=1}^N \sum_{j=1}^J \mathbf{1}\{y_i = j\} \log p_{ij}(\boldsymbol{\beta}).$$

The MLE is defined as

$$\hat{\boldsymbol{\beta}} \in \arg\max_{\boldsymbol{\beta}} \ \ell(\boldsymbol{\beta}),$$

where $\boldsymbol{\beta}$ stacks the free blocks $(\boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_J)$ under the chosen normalization. Inference proceeds exactly as in binary choice: treat $\hat{\boldsymbol{\beta}}$ as an M-estimator / MLE, use the (observed) information matrix or the sandwich form for misspecification, and apply Wald/LR/LM tests as needed.

9.12.4 Marginal Effects

In multinomial models, changing a covariate shifts *all* choice probabilities. For a continuous regressor x_{ik}, the marginal effect on alternative j is

$$\mathrm{ME}_{ijk}(\boldsymbol{\beta}) := \frac{\partial \Pr(y_i = j \mid \mathbf{x}_i)}{\partial x_{ik}} = \frac{\partial p_{ij}(\boldsymbol{\beta})}{\partial x_{ik}}.$$

For MNL, this derivative has a compact closed form. Let β_{jk} denote the kth coordinate of $\boldsymbol{\beta}_j$. Then

$$\frac{\partial p_{ij}(\boldsymbol{\beta})}{\partial x_{ik}} = p_{ij}(\boldsymbol{\beta}) \left(\beta_{jk} - \sum_{\ell=1}^J p_{i\ell}(\boldsymbol{\beta})\, \beta_{\ell k} \right), \qquad j = 1, \ldots, J.$$

Thus the effect on option j is "own slope minus a probability-weighted average slope," scaled by p_{ij}. As in binary Logit/Probit, marginal effects depend on $\mathbf{x}_i$ through $p_{ij}(\boldsymbol{\beta})$, so applied work typically reports marginal effects at representative covariates or average

marginal effects:

$$\text{AME}_{jk} := \frac{1}{N}\sum_{i=1}^{N} \text{ME}_{ijk}(\hat{\boldsymbol{\beta}}).$$

Because $\text{ME}_{ijk}(\boldsymbol{\beta})$ is a smooth function of $\boldsymbol{\beta}$, inference on marginal effects follows by the delta method, using the estimated asymptotic variance of $\hat{\boldsymbol{\beta}}$.

If you need richer substitution patterns than MNL (i.e. to relax IIA), one can move to models such as multinomial Probit (correlated normal utility shocks) or mixed/nested Logit. The basic workflow—index model → implied probabilities → MLE/M-estimation → delta-method inference—remains the same; what changes is the probability formula and the computational burden.

9.13 Summary

Binary choice models were our reminder that once you leave the warm, linear world, the objects you *estimate* and the objects you *care about* stop being the same thing. In particular:

- **Coefficients vs. effects.** In the LPM, $\boldsymbol{\beta}_k$ *is* the marginal effect. In Logit/Probit, $\boldsymbol{\beta}_k$ is just a slope inside an index, and the economically meaningful effect is a nonlinear function of $(\mathbf{x}_i, \boldsymbol{\beta})$, hence the delta method shows up to keep everyone honest.
- **Identification vs. computation.** Maximizing $\ell(\boldsymbol{\beta})$ is an algorithm; getting a structural parameter out of the data requires exogeneity/orthogonality conditions. The optimizer does not check your assumptions for you (tragically).
- **Distributional assumptions are doing real work.** Choosing F is not mere aesthetics: it pins down a likelihood (and therefore an estimator) by turning qualitative restrictions (monotonicity, latent index, threshold crossing) into quantitative probabilities.

Now comes the plot twist: linear panel models bring back the linear conditional mean (so the algebra becomes friendly again), but they keep the hardest part. The threat is no longer "your fitted probability wandered outside $[0, 1]$." The threat is that instead of bending probabilities with a nonlinear link, we add an individual-specific component to the error that you *cannot* ignore because it is very good at being correlated with $\mathbf{x}_{it}$.

The good news is that nothing we built was wasted: the same orthogonality logic and "what exactly is identified under which exogeneity assumption" mindset carries over directly. The bad news is that panels will try to gaslight you with notation.

In the next chapter, we study the canonical linear panel model, make the exogeneity conditions explicit (strict vs. sequential), and derive the two workhorse solutions: within/differencing for fixed effects, and GLS-style estimation for random effects.

Hang in there. We are literally one chapter away from the end. (Yes, the last chapter is a bit annoying. But still: one chapter.)

10 Linear Panel Models

The previous chapter studied models where the outcome is nonlinear (or enters through a nonlinear link). A recurring difficulty there is that unobserved heterogeneity—individual-specific components that persist over time—typically enters the model nonlinearly as well.

In this chapter we take a different route: we focus on *linear* panel models. The payoff is tractability. Linearity lets us remove time-invariant unobserved heterogeneity by simple transformations (demeaning or differencing), leading to estimators with closed forms and identification conditions that are easy to see and hard to fake.

This also sets up the canonical policy-evaluation application: difference-in-differences (DiD), which can be viewed as a special case of differencing (and, in multi-period settings, two-way fixed effects).

10.1 Specifications and POLS

Assume we observe a panel $\{(y_{it}, \mathbf{x}_{it}) : i = 1, \dots, N,\ t = 1, \dots, T\}$, where $y_{it} \in \mathbb{R}$ and $\mathbf{x}_{it} \in \mathbb{R}^k$. A very general linear panel specification writes

$$y_{it} = \mathbf{x}_{it}^\top \boldsymbol{\beta}_{it} + \varepsilon_{it},$$

where the disturbance ε_{it} may encode individual-specific and/or time-series structure, and we allow the slope to vary across both individuals and time.

Fun Facts 10.1. As a digress, several restrictions on $\boldsymbol{\beta}_{it}$ appear frequently:

$$\begin{aligned} \boldsymbol{\beta}_{it} &= \boldsymbol{\beta} \quad \text{(constant slope)} \\ \boldsymbol{\beta}_{it} &= \boldsymbol{\beta}_i \quad \text{(individual-specific slope)} \\ \boldsymbol{\beta}_{it} &= \boldsymbol{\beta}_t \quad \text{(time-specific slope).} \end{aligned}$$

In this chapter we mainly focus on the constant-slope case, and model cross-sectional/time heterogeneity through additive unobservables (e.g. unit effects c_i and time effects d_t introduced below).

A first (and very tempting) idea is to *ignore* the panel structure, treat the NT observations as if they were a single pooled sample, and run one big OLS regression.

Definition 10.1 (Pooled OLS (POLS))**.** Let $\mathbf{y} \in \mathbb{R}^{NT}$ and $\mathbf{X} \in \mathbb{R}^{NT \times k}$ denote the stacked outcome vector and regressor matrix formed by stacking all pairs (i, t). The pooled OLS (POLS) estimator is

$$\hat{\boldsymbol{\beta}}_{\text{POLS}} := (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y},$$

whenever $\mathbf{X}^\top \mathbf{X}$ is invertible.

This approach is computationally simple and uses all observations, but it is also *naive* in typical panel settings: observations within the same individual are often dependent over

time, and there may be time-invariant unobserved effects c_i correlated with $\mathbf{x}_{it}$. If such panel structure is present and ignored, POLS can be biased/inconsistent, and its usual standard errors can be misleading.

To make pooled regression valid, one needs conditions that effectively rule out (or neutralize) these panel-specific complications. The next lemma records the baseline benchmark: if the pooled sample behaves like an ordinary cross section, then POLS behaves like ordinary OLS.

Lemma 10.1 (Unbiasedness and consistency of POLS in a pooled-regression benchmark)**.** Consider the pooled linear regression model

$$y_{it} = \mathbf{x}_{it}^\top \boldsymbol{\beta} + \varepsilon_{it}, \qquad i = 1, \dots, N,\ t = 1, \dots, T,$$

and the POLS estimator $\hat{\boldsymbol{\beta}}_{\text{POLS}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$. Assume:

(A1) $\mathbb{E}\left[\mathbf{x}_{it}\mathbf{x}_{it}^\top\right]$ exists, is finite, and nonsingular, and $\mathbb{E}\left[\varepsilon_{it}^2\right] < \infty$.

(A2) **Contemporaneous exogeneity.** $\mathbb{E}\left[\varepsilon_{it} \mid \mathbf{x}_{it}\right] = 0$ for all (i, t).

(A3) As $NT \to \infty$, a law of large numbers applies so that

$$\frac{1}{NT}\mathbf{X}^\top \mathbf{X} \xrightarrow{p} \mathbb{E}\left[\mathbf{x}_{it}\mathbf{x}_{it}^\top\right], \qquad \frac{1}{NT}\mathbf{X}^\top \varepsilon \xrightarrow{p} \mathbb{E}\left[\mathbf{x}_{it}\varepsilon_{it}\right].$$

Then:

(i) If $(\mathbf{X}^\top \mathbf{X})^{-1}$ exists, then $\mathbb{E}\left[\hat{\boldsymbol{\beta}}_{\text{POLS}} \mid \mathbf{X}\right] = \boldsymbol{\beta}$, hence $\mathbb{E}\left[\hat{\boldsymbol{\beta}}_{\text{POLS}}\right] = \boldsymbol{\beta}$.

(ii) $\hat{\boldsymbol{\beta}}_{\text{POLS}} \xrightarrow{p} \boldsymbol{\beta}$ as $NT \to \infty$.

Proof. Stack the model as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon,$$

so

$$\hat{\boldsymbol{\beta}}_{\text{POLS}} - \boldsymbol{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \varepsilon.$$

We first show unbiasedness. Conditioning on $\mathbf{X}$,

$$\mathbb{E}\left[\hat{\boldsymbol{\beta}}_{\text{POLS}} - \boldsymbol{\beta} \mid \mathbf{X}\right] = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbb{E}\left[\varepsilon \mid \mathbf{X}\right].$$

By (A2), $\mathbb{E}\left[\varepsilon_{it} \mid \mathbf{x}_{it}\right] = 0$ for each (i, t), hence $\mathbb{E}\left[\varepsilon \mid \mathbf{X}\right] = \mathbf{0}$ and therefore $\mathbb{E}\left[\hat{\boldsymbol{\beta}}_{\text{POLS}} \mid \mathbf{X}\right] = \boldsymbol{\beta}$.

Next show consistency. Rewrite

$$\hat{\boldsymbol{\beta}}_{\text{POLS}} - \boldsymbol{\beta} = \left(\frac{1}{NT}\mathbf{X}^\top \mathbf{X}\right)^{-1} \left(\frac{1}{NT}\mathbf{X}^\top \varepsilon\right).$$

By (A3), $\frac{1}{NT}\mathbf{X}^\top \mathbf{X} \xrightarrow{p} \mathbb{E}\left[\mathbf{x}_{it}\mathbf{x}_{it}^\top\right]$, which is nonsingular by (A1), hence by continuity of inversion,

$$\left(\frac{1}{NT}\mathbf{X}^\top \mathbf{X}\right)^{-1} \xrightarrow{p} \mathbb{E}\left[\mathbf{x}_{it}\mathbf{x}_{it}^\top\right]^{-1}.$$

Also by (A2),

$$\mathbb{E}\left[\mathbf{x}_{it}\varepsilon_{it}\right] = \mathbb{E}\left[\mathbb{E}\left[\mathbf{x}_{it}\varepsilon_{it} \mid \mathbf{x}_{it}\right]\right] = \mathbb{E}\left[\mathbf{x}_{it}\mathbb{E}\left[\varepsilon_{it} \mid \mathbf{x}_{it}\right]\right] = \mathbf{0},$$

so (A3) gives $\frac{1}{NT}\mathbf{X}^\top\varepsilon \xrightarrow{p} \mathbf{0}$. Slutsky's theorem implies $\hat{\boldsymbol{\beta}}_{\text{POLS}} - \boldsymbol{\beta} \xrightarrow{p} \mathbf{0}$. □

The proof above leans heavily on the contemporaneous exogeneity assumption. This is very likely not the case given a panel data structure. As the data tracks over time, correlations can be passed down through t and POLS easily become biased. We model this formally next via **Linear Unobserved Effects**.

10.2 Linear unobserved effects, FE vs. RE, and strict exogeneity

Definition 10.2 (Linear unobserved-effects (UE) models)**.** In the constant-slope case, a standard way to represent cross-sectional and time heterogeneity is through additive unobservables:

$$\begin{aligned} y_{it} &= c_i + \mathbf{x}_{it}^\top\boldsymbol{\beta} + u_{it} \qquad \text{(individual effects)},\\ y_{it} &= d_t + \mathbf{x}_{it}^\top\boldsymbol{\beta} + u_{it} \qquad \text{(time effects)},\\ y_{it} &= c_i + d_t + \mathbf{x}_{it}^\top\boldsymbol{\beta} + u_{it} \qquad \text{(two-way effects)}. \end{aligned}$$

Here c_i is an unobserved component constant within individual i, d_t is an unobserved component constant within period t, and u_{it} is the idiosyncratic disturbance.

A useful way to read Definition 10.2 is: c_i and d_t carry the *persistent* structure (time-invariant or common time shocks), while u_{it} is the *residual* short-run shock left after accounting for that structure. For intuition and for baseline inference formulas, it is common to treat $\{u_{it}\}$ as "approximately white noise" within each i once c_i (and possibly d_t) have been removed: centered, often homoskedastic, and with no strong time-persistent dependence.

The key econometric fork appears immediately: do we allow c_i to be correlated with the regressor path $\{\mathbf{x}_{it}\}_{t=1}^T$ (fixed effects), or do we impose an orthogonality restriction linking c_i and $\{\mathbf{x}_{it}\}$ (random effects)? Before choosing sides, we need the baseline exogeneity condition that gives $\boldsymbol{\beta}$ its usual meaning.

Definition 10.3 (Strict exogeneity)**.** In the unobserved-effects model, $\mathbf{x}_{it}$ is said to be *strictly exogenous* (relative to the idiosyncratic shock u_{it} and unit effect c_i) if

$$\mathbb{E}\left[u_{it} \mid \mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}, c_i\right] = 0 \qquad \text{for all } t = 1, \ldots, T.$$

Equivalently: after conditioning on the entire regressor path and the individual effect, the remaining innovation has zero mean. In particular, u_{it} cannot feed back into the future evolution of $\{\mathbf{x}_{is}\}_{s=1}^T$.

This is a strong condition: it rules out both "feedback" (shocks affecting future regressors) and many dynamic adjustment stories. Two common settings where strict exogeneity is typically violated or simply inappropriate are:

- **Binary choice / nonlinear outcome models:** outcomes are generated through a nonlinear latent-index structure, and the disturbance enters in a way that generally does not align with the linear conditional-mean restriction above. One usually replaces strict exogeneity with exogeneity conditions tailored to the nonlinear model.
- **Dynamic panels:** if the regressor includes lagged outcomes, then past shocks are mechanically embedded in the regressor path. For example, in

$$y_{it} = c_i + \rho\, y_{i,t-1} + u_{it},$$

the regressor $y_{i,t-1}$ is a function of $u_{i,t-1}$ (and c_i), so conditioning on the full path $\{y_{i,s-1}\}_{s=1}^{T}$ typically violates the strict-exogeneity requirement.

10.3 Fixed effects

We start from the individual-effects model

$$y_{it} = c_i + \mathbf{x}_{it}^{\top}\boldsymbol{\beta} + u_{it}, \qquad i = 1, \ldots, N,\ t = 1, \ldots, T,$$

where c_i collects everything about individual i that does not change over time (ability, preferences, baseline productivity, etc.). The point of fixed effects is not that c_i is small, but that it can be *correlated* with the regressors. If individuals choose $\mathbf{x}_{it}$ based on stable traits we do not observe, then treating c_i as part of the error (random effects) is exactly the wrong move. FE takes the opposite stance: it allows c_i to be arbitrarily related to the regressor path and tries to estimate $\boldsymbol{\beta}$ anyway.

To interpret $\boldsymbol{\beta}$ in this setting, we still need that the remaining shock u_{it} behaves like an innovation once we condition on what the individual looks like (through c_i) and on the regressor history. The standard benchmark is strict exogeneity (Definition 10.3).

10.3.1 LSDV estimation.

Definition 10.4 (Least Squares Dummy Variable (LSDV) estimator)**.** A literal way to implement fixed effects is to treat $\{c_i\}$ as N nuisance parameters and estimate them by OLS along with $\boldsymbol{\beta}$. Write

$$y_{it} = \alpha_i + \mathbf{x}_{it}^{\top}\boldsymbol{\beta} + u_{it}, \qquad \alpha_i \equiv c_i,$$

and represent α_i using individual dummies. Let $D_{ij} = \mathbf{1}\{i = j\}$ be the indicator that observation (i, t) belongs to individual j. Then the FE model can be written as the pooled regression

$$y_{it} = \sum_{j=1}^{N} \alpha_j D_{ij} + \mathbf{x}_{it}^{\top}\boldsymbol{\beta} + u_{it},$$

with one normalization (e.g. omit one dummy or impose $\sum_{j=1}^{N} \alpha_j = 0$). Running OLS on this regression is the *LSDV* estimator.

It uses the same identifying idea as FE: each individual gets their own intercept, so all time-invariant heterogeneity is absorbed before $\boldsymbol{\beta}$ is estimated.

10.3.2 Within estimaton.

While LSDV is conceptually transparent, it is often more convenient to eliminate $\{\alpha_i\}$ algebraically by demeaning over time, which leads to the within estimator. Because c_i does not vary with t, we can remove it by comparing an individual to themself over time.

Fix an individual i and define time averages

$$\bar{y}_i := \frac{1}{T}\sum_{t=1}^{T} y_{it}, \qquad \bar{\mathbf{x}}_i := \frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_{it}, \qquad \bar{u}_i := \frac{1}{T}\sum_{t=1}^{T} u_{it}.$$

Averaging the FE model gives

$$\bar{y}_i = c_i + \bar{\mathbf{x}}_i^\top \boldsymbol{\beta} + \bar{u}_i,$$

so subtracting yields the within regression

$$y_{it} - \bar{y}_i = (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)^\top \boldsymbol{\beta} + (u_{it} - \bar{u}_i).$$

To make the algebra compact, stack observations for individual i into $T \times 1$ and $T \times k$ objects:

$$\mathbf{y}_i := \begin{pmatrix} y_{i1} \\ \vdots \\ y_{iT} \end{pmatrix}, \qquad \mathbf{X}_i := \begin{pmatrix} \mathbf{x}_{i1}^\top \\ \vdots \\ \mathbf{x}_{iT}^\top \end{pmatrix}, \qquad \mathbf{u}_i := \begin{pmatrix} u_{i1} \\ \vdots \\ u_{iT} \end{pmatrix}, \qquad \mathbf{e} := \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \in \mathbb{R}^T.$$

Then the FE model becomes

$$\mathbf{y}_i = c_i \mathbf{e} + \mathbf{X}_i \boldsymbol{\beta} + \mathbf{u}_i.$$

Definition 10.5 (Within transformation and FE estimator)**.** Define the demeaning matrix

$$\mathbf{Q} := \mathbf{I}_T - \frac{1}{T}\mathbf{e}\mathbf{e}^\top.$$

Then $\mathbf{Q}\mathbf{e} = 0$, $\mathbf{Q}^\top = \mathbf{Q}$, and $\mathbf{Q}^2 = \mathbf{Q}$. Premultiplying by $\mathbf{Q}$ yields the within-transformed model

$$\mathbf{Q}\mathbf{y}_i = \mathbf{Q}\mathbf{X}_i \boldsymbol{\beta} + \mathbf{Q}\mathbf{u}_i,$$

since $\mathbf{Q}(c_i\mathbf{e}) = 0$. The within (FE) estimator is the pooled OLS coefficient in this transformed regression:

$$\hat{\boldsymbol{\beta}}_{\mathrm{FE}} = \left(\sum_{i=1}^{N} \mathbf{X}_i^\top \mathbf{Q} \mathbf{X}_i\right)^{-1} \left(\sum_{i=1}^{N} \mathbf{X}_i^\top \mathbf{Q} \mathbf{y}_i\right).$$

Equivalently, with $\ddot{y}_{it} := y_{it} - \bar{y}_i$ and $\ddot{\mathbf{x}}_{it} := \mathbf{x}_{it} - \bar{\mathbf{x}}_i$,

$$\hat{\boldsymbol{\beta}}_{\mathrm{FE}} = \left(\sum_{i=1}^{N}\sum_{t=1}^{T} \ddot{\mathbf{x}}_{it} \ddot{\mathbf{x}}_{it}^\top\right)^{-1} \left(\sum_{i=1}^{N}\sum_{t=1}^{T} \ddot{\mathbf{x}}_{it} \ddot{y}_{it}\right).$$

Rank condition. For $\hat{\boldsymbol{\beta}}_{\mathrm{FE}}$ to be well-defined, the transformed regressor matrix must have full column rank:

$$\operatorname{rank}\left(\sum_{i=1}^{N} \mathbf{X}_i^\top \mathbf{Q} \mathbf{X}_i\right) = k.$$

Intuitively, each regressor must have enough within-individual variation over time; if a component of $\mathbf{x}_{it}$ is constant in t for every i, then $\mathbf{Q}\mathbf{X}_i$ removes it completely and its coefficient cannot be identified under FE.

10.3.3 Asymptotics of the within estimator

For fixed effects, the usual large-N panel asymptotics treat T as fixed (or at least not growing fast) and let $N \to \infty$. Intuitively, FE identifies $\boldsymbol{\beta}$ by comparing each individual to themself over time; once we have transformed the data, we can think of each individual i as contributing one "block" of T observations, and consistency comes from having many such blocks.

Recall the within-transformed regression

$$\mathbf{Q}\mathbf{y}_i = \mathbf{Q}\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Q}\mathbf{u}_i, \qquad \mathbf{Q} = \mathbf{I}_T - \frac{1}{T}\mathbf{e}\mathbf{e}^\top.$$

A natural orthogonality restriction in this transformed model is

$$\mathbb{E}\left[\mathbf{X}_i^\top \mathbf{Q}\mathbf{u}_i\right] = 0,$$

or, more explicitly,

$$\mathbb{E}\left[\ddot{\mathbf{x}}_{it}(u_{it} - \bar{u}_i)\right] = 0.$$

This is the within analogue of the usual OLS mean-orthogonality condition: once we remove c_i by demeaning, the remaining regressor variation should be uncorrelated with the remaining error variation.

Lemma 10.2 (Unbiasedness of within estimators)**.** Let $\mathbf{Q} := \mathbf{I}_T - \frac{1}{T}\mathbf{e}\mathbf{e}^\top$. Assume $\sum_{i=1}^{N} \mathbf{X}_i^\top \mathbf{Q}\mathbf{X}_i$ is nonsingular (in the realized sample) and that the *within orthogonality* condition holds:

$$\mathbb{E}\left[\mathbf{X}_i^\top \mathbf{Q}\mathbf{u}_i \mid \mathbf{X}_i\right] = \mathbf{0}.$$

Then the within (FE) estimator

$$\hat{\boldsymbol{\beta}}_{\mathrm{FE}} = \left(\sum_{i=1}^{N} \mathbf{X}_i^\top \mathbf{Q}\mathbf{X}_i\right)^{-1} \left(\sum_{i=1}^{N} \mathbf{X}_i^\top \mathbf{Q}\mathbf{y}_i\right)$$

is conditionally unbiased given the regressors:

$$\mathbb{E}\left[\hat{\boldsymbol{\beta}}_{\mathrm{FE}} \mid \mathbf{X}_1, \ldots, \mathbf{X}_N\right] = \boldsymbol{\beta}.$$

Proof. Premultiplying by $\mathbf{Q}$ eliminates the fixed effect:

$$\mathbf{Q}\mathbf{y}_i = \mathbf{Q}\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Q}\mathbf{u}_i.$$

Plug into the FE formula:

$$\hat{\boldsymbol{\beta}}_{\mathrm{FE}} = \boldsymbol{\beta} + \left(\sum_{i=1}^{N} \mathbf{X}_i^\top \mathbf{Q}\mathbf{X}_i\right)^{-1} \left(\sum_{i=1}^{N} \mathbf{X}_i^\top \mathbf{Q}\mathbf{u}_i\right).$$

Condition on $\mathbf{X}_1, \ldots, \mathbf{X}_N$ and take expectations. By the assumed within orthogonality, $\mathbb{E}\left[\mathbf{X}_i^\top \mathbf{Q}\mathbf{u}_i \mid \mathbf{X}_i\right] = \mathbf{0}$, hence

$$\mathbb{E}\left[\hat{\boldsymbol{\beta}}_{\mathrm{FE}} - \boldsymbol{\beta} \mid \mathbf{X}_1, \ldots, \mathbf{X}_N\right] = \left(\sum_{i=1}^{N} \mathbf{X}_i^\top \mathbf{Q}\mathbf{X}_i\right)^{-1} \sum_{i=1}^{N} \mathbb{E}\left[\mathbf{X}_i^\top \mathbf{Q}\mathbf{u}_i \mid \mathbf{X}_1, \ldots, \mathbf{X}_N\right] = \mathbf{0}.$$

The (i,t) statement follows from $\mathbf{X}_i^\top \mathbf{Q}\mathbf{u}_i = \sum_{t=1}^{T} \ddot{\mathbf{x}}_{it}\ddot{u}_{it}$. □

Consistency is valid under the same type of argument we did for OLS.

Lemma 10.3 (Consistency of the within estimator)**.** Recall the within (FE) estimator

$$\hat{\boldsymbol{\beta}}_{\mathrm{FE}} = \left(\sum_{i=1}^{N} \mathbf{X}_i^\top \mathbf{Q}\mathbf{X}_i\right)^{-1} \left(\sum_{i=1}^{N} \mathbf{X}_i^\top \mathbf{Q}\mathbf{y}_i\right).$$

Assume:

(A1) $\mathbb{E}\left[\mathbf{X}_i^\top \mathbf{Q}\mathbf{u}_i\right] = 0$.

(A2) $\mathbb{E}\left[\|\mathbf{X}_i^\top \mathbf{Q}\mathbf{X}_i\|\right] < \infty$ and $\mathbb{E}\left[\|\mathbf{X}_i^\top \mathbf{Q}\mathbf{u}_i\|\right] < \infty$, and a law of large numbers applies so that

$$\frac{1}{N}\sum_{i=1}^{N} \mathbf{X}_i^\top \mathbf{Q}\mathbf{X}_i \xrightarrow{p} \mathbf{A} := \mathbb{E}\left[\mathbf{X}_i^\top \mathbf{Q}\mathbf{X}_i\right], \qquad \frac{1}{N}\sum_{i=1}^{N} \mathbf{X}_i^\top \mathbf{Q}\mathbf{u}_i \xrightarrow{p} \mathbb{E}\left[\mathbf{X}_i^\top \mathbf{Q}\mathbf{u}_i\right] = 0.$$

(A3) $\mathbf{A}$ is nonsingular.

Then

$$\hat{\boldsymbol{\beta}}_{\mathrm{FE}} \xrightarrow{p} \boldsymbol{\beta} \qquad (N \to \infty).$$

Proof. Premultiply the model by $\mathbf{Q}$ to eliminate the fixed effect:

$$\mathbf{Q}\mathbf{y}_i = \mathbf{Q}\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Q}\mathbf{u}_i,$$

since $\mathbf{Q}(c_i\mathbf{e}) = c_i(\mathbf{Q}\mathbf{e}) = \mathbf{0}$. Plugging this into the definition of $\hat{\boldsymbol{\beta}}_{\mathrm{FE}}$ gives

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{\mathrm{FE}} &= \left(\sum_{i=1}^{N} \mathbf{X}_i^\top \mathbf{Q}\mathbf{X}_i\right)^{-1} \left(\sum_{i=1}^{N} \mathbf{X}_i^\top (\mathbf{Q}\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Q}\mathbf{u}_i)\right) \\ &= \boldsymbol{\beta} + \left(\sum_{i=1}^{N} \mathbf{X}_i^\top \mathbf{Q}\mathbf{X}_i\right)^{-1} \left(\sum_{i=1}^{N} \mathbf{X}_i^\top \mathbf{Q}\mathbf{u}_i\right).\end{aligned}$$

Rewrite this as

$$\hat{\boldsymbol{\beta}}_{\mathrm{FE}} - \boldsymbol{\beta} = \left(\frac{1}{N} \sum_{i=1}^{N} \mathbf{X}_i^\top \mathbf{Q} \mathbf{X}_i \right)^{-1} \left(\frac{1}{N} \sum_{i=1}^{N} \mathbf{X}_i^\top \mathbf{Q} \mathbf{u}_i \right).$$

By (A2), the first term converges in probability to $\mathbf{A}$ and the second converges in probability to $\mathbf{0}$. By (A3), $\mathbf{A}$ is nonsingular, hence by continuity of matrix inversion,

$$\left(\frac{1}{N} \sum_{i=1}^{N} \mathbf{X}_i^\top \mathbf{Q} \mathbf{X}_i \right)^{-1} \xrightarrow{p} \mathbf{A}^{-1}.$$

Slutsky's theorem then implies $\hat{\boldsymbol{\beta}}_{\mathrm{FE}} - \boldsymbol{\beta} \xrightarrow{p} \mathbf{0}$. □

The route to asymptotic normality follows the textbook proof as well.

Lemma 10.4 (Asymptotic normality of the within estimator)**.** Under the setup of Lemma 10.3, suppose in addition that:

(B1) $\{(\mathbf{X}_i, \mathbf{u}_i)\}_{i=1}^{N}$ are i.i.d. across i.

(B2) $\mathbb{E}\left[\|\mathbf{X}_i^\top \mathbf{Q} \mathbf{u}_i\|^2\right] < \infty$.

Let

$$\mathbf{A} := \mathbb{E}\left[\mathbf{X}_i^\top \mathbf{Q} \mathbf{X}_i\right], \qquad \boldsymbol{\Omega} := \mathbb{E}\left[\mathbf{X}_i^\top \mathbf{Q} \mathbf{u}_i \, \mathbf{u}_i^\top \mathbf{Q} \mathbf{X}_i\right].$$

Then

$$\sqrt{N}\,(\hat{\boldsymbol{\beta}}_{\mathrm{FE}} - \boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}\big(\mathbf{0},\ \mathbf{A}^{-1} \boldsymbol{\Omega} \mathbf{A}^{-1}\big).$$

Moreover, under the homoskedastic "within white-noise" benchmark

$$\mathbb{E}\left[\mathbf{u}_i \mathbf{u}_i^\top \mid \mathbf{X}_i\right] = \sigma_u^2 \mathbf{I}_T,$$

we have $\boldsymbol{\Omega} = \sigma_u^2 \mathbf{A}$ and thus

$$\sqrt{N}\,(\hat{\boldsymbol{\beta}}_{\mathrm{FE}} - \boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}\big(\mathbf{0},\ \sigma_u^2 \mathbf{A}^{-1}\big).$$

Proof. Start from the decomposition (already derived in the proof of Lemma 10.3):

$$\hat{\boldsymbol{\beta}}_{\mathrm{FE}} - \boldsymbol{\beta} = \left(\frac{1}{N} \sum_{i=1}^{N} \mathbf{X}_i^\top \mathbf{Q} \mathbf{X}_i \right)^{-1} \left(\frac{1}{N} \sum_{i=1}^{N} \mathbf{X}_i^\top \mathbf{Q} \mathbf{u}_i \right).$$

Define the $k \times 1$ random vector

$$\mathbf{g}_i := \mathbf{X}_i^\top \mathbf{Q} \mathbf{u}_i.$$

By (A1), $\mathbb{E}[\mathbf{g}_i] = \mathbf{0}$. By (B1)–(B2), $\{\mathbf{g}_i\}$ are i.i.d. with finite second moments, so the multivariate CLT yields

$$\sqrt{N} \left(\frac{1}{N} \sum_{i=1}^{N} \mathbf{g}_i \right) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \mathbf{g}_i \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega}), \qquad \boldsymbol{\Omega} = \mathbb{E}\left[\mathbf{g}_i \mathbf{g}_i^\top\right] = \mathbb{E}\left[\mathbf{X}_i^\top \mathbf{Q} \mathbf{u}_i \, \mathbf{u}_i^\top \mathbf{Q} \mathbf{X}_i\right].$$

Meanwhile, by the LLN in (A2),

$$\frac{1}{N}\sum_{i=1}^{N}\mathbf{X}_i^\top\mathbf{Q}\mathbf{X}_i \xrightarrow{p} \mathbf{A},$$

and by (A3) and continuity of inversion,

$$\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{X}_i^\top\mathbf{Q}\mathbf{X}_i\right)^{-1} \xrightarrow{p} \mathbf{A}^{-1}.$$

Combine these with Slutsky's theorem to obtain

$$\sqrt{N}\,(\hat{\boldsymbol{\beta}}_{\mathrm{FE}}-\boldsymbol{\beta}) = \left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{X}_i^\top\mathbf{Q}\mathbf{X}_i\right)^{-1}\sqrt{N}\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{X}_i^\top\mathbf{Q}\mathbf{u}_i\right) \xrightarrow{d} \mathcal{N}\big(\mathbf{0},\ \mathbf{A}^{-1}\boldsymbol{\Omega}\mathbf{A}^{-1}\big).$$

For the homoskedastic benchmark, take conditional expectations:

$$\begin{aligned}
\boldsymbol{\Omega} &= \mathbb{E}\left[\mathbf{X}_i^\top\mathbf{Q}\mathbf{u}_i\mathbf{u}_i^\top\mathbf{Q}\mathbf{X}_i\right] \\
&= \mathbb{E}\left[\mathbf{X}_i^\top\mathbf{Q}\mathbb{E}\left[\mathbf{u}_i\mathbf{u}_i^\top \mid \mathbf{X}_i\right]\mathbf{Q}\mathbf{X}_i\right] \\
&= \mathbb{E}\left[\mathbf{X}_i^\top\mathbf{Q}(\sigma_u^2\mathbf{I}_T)\mathbf{Q}\mathbf{X}_i\right] \\
&= \sigma_u^2\mathbb{E}\left[\mathbf{X}_i^\top\mathbf{Q}^2\mathbf{X}_i\right] \\
&= \sigma_u^2\mathbb{E}\left[\mathbf{X}_i^\top\mathbf{Q}\mathbf{X}_i\right] \\
&= \sigma_u^2\mathbf{A},
\end{aligned}$$

using $\mathbf{Q}^2=\mathbf{Q}$. Substituting into the general limit variance gives the stated simplification. □

For the estimation of variances, it is enough to record the simple homoskedastic benchmark in most cases.

Definition 10.6 (Homoskedastic benchmark variance estimator for within estimator)**.** Under the homoskedastic "within white-noise" benchmark, a degrees-of-freedom estimator of the idiosyncratic variance is

$$\hat{\sigma}_u^2 = \frac{1}{NT-N-k}\sum_{i=1}^{N}\sum_{t=1}^{T}\hat{u}_{it}^2,$$

and the corresponding plug-in variance estimator for $\hat{\boldsymbol{\beta}}_{\mathrm{FE}}$ is

$$\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}_{\mathrm{FE}}) = \hat{\sigma}_u^2\left(\sum_{i=1}^{N}\sum_{t=1}^{T}\ddot{\mathbf{x}}_{it}\ddot{\mathbf{x}}_{it}^\top\right)^{-1}.$$

This is the direct analogue of the usual OLS variance formula, applied to the within-transformed regression. If one wants to allow for arbitrary heteroskedasticity and serial correlation within individuals, replace it with a cluster-robust (by i) variance estimator.

10.3.4 First differences

Within transformation removed c_i by subtracting an individual's time average. First differencing achieves the same goal in the most direct way: subtract period $t-1$ from period t. Because c_i is constant over time, it vanishes automatically, leaving a regression in changes.

Definition 10.7 (First-difference transformation and estimator)**.** For any scalar or vector process m_{it} (e.g. $m = y, \mathbf{x}, u$), define

$$\Delta m_{it} := m_{it} - m_{i,t-1}, \qquad t = 2, \dots, T.$$

Differencing the individual-effects model yields

$$\Delta y_{it} = \Delta \mathbf{x}_{it}^\top \boldsymbol{\beta} + \Delta u_{it}, \qquad t = 2, \dots, T,$$

since $\Delta c_i = 0$. The FD estimator is the OLS coefficient from regressing Δy_{it} on $\Delta \mathbf{x}_{it}$ over $i = 1, \dots, N$ and $t = 2, \dots, T$:

$$\hat{\boldsymbol{\beta}}_{\mathrm{FD}} = \left(\sum_{i=1}^{N} \sum_{t=2}^{T} \Delta \mathbf{x}_{it} \Delta \mathbf{x}_{it}^\top \right)^{-1} \left(\sum_{i=1}^{N} \sum_{t=2}^{T} \Delta \mathbf{x}_{it} \Delta y_{it} \right).$$

To justify OLS on the differenced regression, we impose three benchmark conditions:

(FD1) **Mean orthogonality in differences.** For each $t = 2, \dots, T$,

$$\mathbb{E}\left[\Delta \mathbf{x}_{it}\, \Delta u_{it}\right] = 0.$$

(FD2) **Rank variation.** The differenced regressors contain enough variation to identify $\boldsymbol{\beta}$:

$$\operatorname{rank}\!\left(\mathbb{E}\left[\Delta \mathbf{x}_{it} \Delta \mathbf{x}_{it}^\top\right]\right) = k.$$

In particular, any regressor that is time-invariant for each individual disappears after differencing and cannot be identified.

(FD3) **Homoskedastic "white-noise".** Let $\Delta \mathbf{u}_i := (\Delta u_{i2}, \dots, \Delta u_{iT})^\top$. Assume

$$\mathbb{E}\left[\Delta \mathbf{u}_i \Delta \mathbf{u}_i^\top \mid \mathbf{X}_i\right] = \sigma_{\Delta u}^2 \mathbf{I}_{T-1}.$$

Note this is an assumption about Δu_{it} directly: i.i.d. u_{it} would induce negative serial correlation in $\Delta u_{it} = u_{it} - u_{i,t-1}$, while "$\Delta u_{it}$ is white noise" corresponds to u_{it} behaving like a random walk.

Asymptotics. Similar to the within estimator's asymptotics, under the same type of LLN/CLT across i as in the within case, $\hat{\boldsymbol{\beta}}_{\mathrm{FD}}$ is consistent as $N \to \infty$, and its asymptotic variance takes the familiar OLS form applied to the differenced regression; in the homoskedastic benchmark above, it simplifies to a constant times the inverse second-moment matrix of $\Delta \mathbf{x}_{it}$.

10.3.5 Policy evaluation, first differences, and DiD

When $T = 2$ and $\mathbf{x}_{it}$ encodes treatment status, the FD regression is exactly the two-period difference-in-differences setup: differencing removes c_i, and the treatment effect is identified from changes over time.

To keep one canonical policy-evaluation picture in mind, imagine two periods: $t = 1$ (pre) and $t = 2$ (post). In period 1, nobody is treated; in period 2, a subset of units participates in a program. A convenient specification is

$$y_{it} = c_i + \beta_1 D_{2t} + \beta_2 \text{Prog}_{it} + u_{it},$$

where $D_{2t} = \mathbf{1}\{t = 2\}$ is a time dummy, and Prog_{it} indicates participation.

First differencing removes the nuisance c_i automatically:

$$\Delta y_i = y_{i2} - y_{i1} = \beta_1 + \beta_2 \, \Delta\text{Prog}_i + \Delta u_i, \qquad \Delta\text{Prog}_i = \text{Prog}_{i2} - \text{Prog}_{i1}.$$

So the differenced regression says: compare how much outcomes change for participants versus non-participants, net of the common time shift β_1.

Definition 10.8 (Two-period DiD estimand)**.** In the two-period setup with $\Delta y_i := y_{i2} - y_{i1}$ and treatment indicator $D_i := \Delta\text{Prog}_i \in \{0, 1\}$, the population DiD estimand is

$$\delta_{\text{DiD}} := \mathbb{E}\left[\Delta y_i \mid D_i = 1\right] - \mathbb{E}\left[\Delta y_i \mid D_i = 0\right].$$

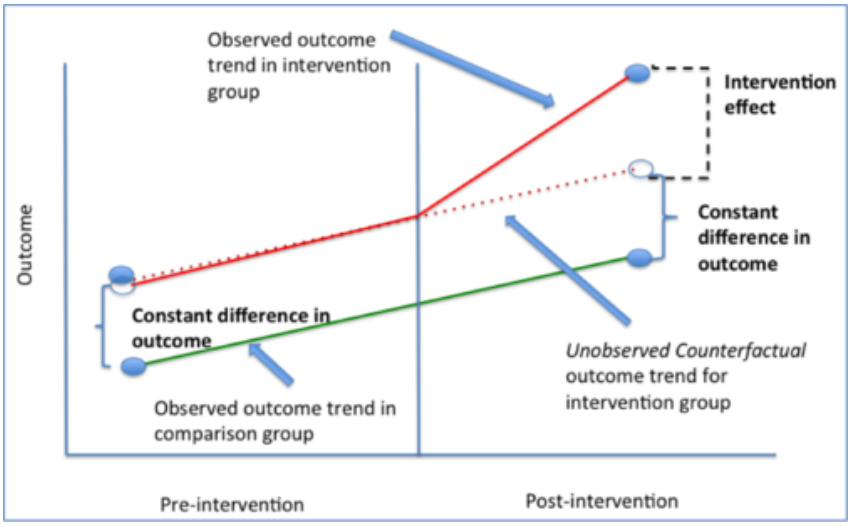

Figure 9: The idea of DID.

The same logic is often written using group and time indicators in a single regression:

$$y_{it} = \alpha + \beta \text{Treat}_i + \gamma \text{Post}_t + \delta(\text{Treat}_i \times \text{Post}_t) + \mathbf{x}_{it}^\top \boldsymbol{\eta} + u_{it}.$$

Here Treat_i flags the treated group and Post_t flags the post-policy period. The interaction

coefficient δ is the DiD effect: untreated units shift by γ from pre to post, treated units shift by $\gamma + \delta$, and the difference between these shifts is δ.

With more than two periods, it is convenient to replace explicit group and time dummies with unit and time fixed effects:

$$y_{it} = \delta(\text{Treat}_i \times \text{Post}_t) + c_i + d_t + \mathbf{x}_{it}^{\top}\boldsymbol{\eta} + u_{it}.$$

The unit effect c_i absorbs time-invariant group membership (so a standalone Treat_i term becomes redundant), and the time effects d_t absorb shocks common to all units in each period (so a standalone Post_t term becomes redundant). What remains identified is the interaction: the differential post-period shift for treated units.

If treatment timing varies across units, one often expands the interaction into a sequence of time-specific effects (an event-study specification):

$$y_{it} = \alpha + \sum_t \delta_t\big(\text{Treat}_i \times D_t\big) + c_i + d_t + \mathbf{x}_{it}^{\top}\boldsymbol{\eta} + u_{it},$$

where D_t denotes a time dummy and one period must be omitted as a reference. Plotting the estimated $\{\delta_t\}$ with confidence intervals gives the event-study diagnostic: pre-treatment coefficients provide a visual check of parallel trends, while post-treatment coefficients trace dynamic treatment effects over time.

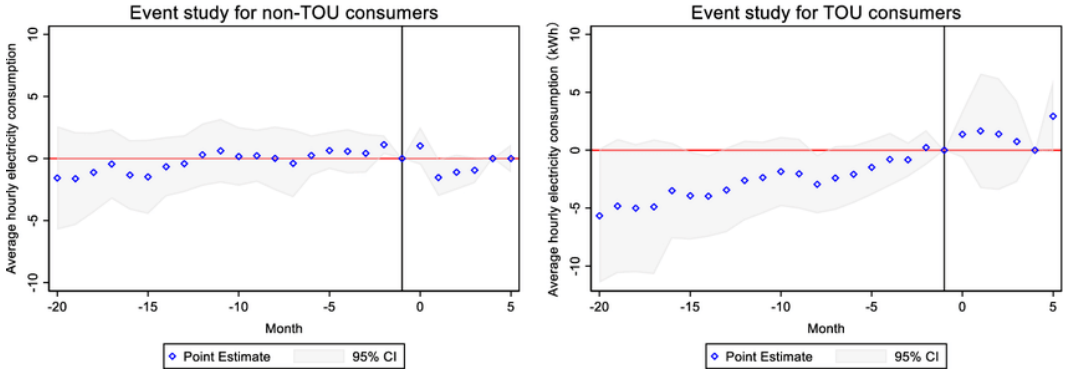

Figure 10: Visual diagnostic from [3].Each panel plots estimated month-by-month treatment effects (blue markers) relative to an omitted pre-treatment reference month, with 95% confidence bands (shaded); the vertical line marks the treatment/adoption date ($t = 0$) and the horizontal line marks zero. Parallel trends is supported when the pre-treatment coefficients (months < 0) are statistically indistinguishable from zero and show no systematic drift; post-treatment coefficients (months ≥ 0) trace the dynamic response after adoption. The left panel (non-TOU consumers) shows little evidence of pre-trends, while the right panel (TOU consumers) displays a pronounced pre-period patterns.

As a closing remark, fixed effects identifies β only from *within-individual* movements over time. Any regressor that is constant in t for a given i is annihilated by the within transformation, so its coefficient is not identified under FE. More generally, if a regressor exhibits little within variation relative to its cross-sectional variation, the FE estimate of its

effect will be noisy: you are asking the estimator to learn from the small "wiggles" left after subtracting each unit's mean. This is the core tradeoff. Individual FE buys robustness to arbitrary correlation between c_i and $\{\mathbf{x}_{it}\}_{t=1}^{T}$ by *projecting out* all time-invariant differences across units—so it refuses to learn from purely cross-sectional comparisons, even when those comparisons would be informative.

10.4 Random effects

Fixed effects is robust because it allows c_i to be arbitrarily correlated with the regressors, but it pays for this robustness by discarding all between-individual variation. Random effects (RE) takes a different stance: it treats the individual component as part of the stochastic error structure and tries to use both within- and between-individual variation. The price is an additional orthogonality restriction linking c_i and the regressor path.

Definition 10.9 (Random-effects model and RE exogeneity)**.** Consider the individual-effects specification

$$y_{it} = c_i + \mathbf{x}_{it}^{\top}\boldsymbol{\beta} + u_{it}, \qquad i = 1, \dots, N,\ t = 1, \dots, T,$$

and define the composite disturbance $v_{it} := c_i + u_{it}$. The *random-effects* approach views c_i as a random variable (a permanent shock), typically with $\mathbb{E}[c_i] = 0$ after absorbing the mean into an intercept. The key identifying restriction is orthogonality between c_i and the regressor path:

$$\mathbb{E}[c_i \mid \mathbf{x}_{i1}, \dots, \mathbf{x}_{iT}] = 0,$$

along with strict exogeneity of the idiosyncratic shock conditional on $(\mathbf{x}_{i1}, \dots, \mathbf{x}_{iT}, c_i)$:

$$\mathbb{E}[u_{it} \mid \mathbf{x}_{i1}, \dots, \mathbf{x}_{iT}, c_i] = 0 \qquad \text{for all } t.$$

Together these imply the RE moment condition

$$\mathbb{E}[v_{it}\,\mathbf{x}_{is}] = 0 \qquad \text{for all } s, t. \tag{\textbf{RE.1}}$$

Even if $\{u_{it}\}$ is serially uncorrelated over t, the composite error $\{v_{it}\}$ is mechanically correlated within i because the same c_i appears in every period. This is the central econometric feature of RE: under **RE.1**, pooled OLS is unbiased/consistent for $\boldsymbol{\beta}$, but it is typically inefficient, and its naive i.i.d. standard errors are wrong because v_{it} is not independent over t within individual.

Definition 10.10 (Variance-components benchmark for RE)**.** A standard benchmark assumes

$$\mathbb{E}\left[c_i^2\right] = \sigma_c^2, \qquad \mathbb{E}\left[u_{it}^2\right] = \sigma_u^2, \qquad \mathbb{E}\left[u_{it}u_{is}\right] = 0 \ (t \neq s),$$

and independence between c_i and $\{u_{it}\}_{t=1}^T$. Then, for a fixed i (conditionally on X_i),

$$\operatorname{Var}\left(v_i \mid X_i\right) = \operatorname{Var}\left(c_i e + u_i \mid X_i\right) = \sigma_c^2 ee^\top + \sigma_u^2 \mathbf{I}_T. \tag{RE.2}$$

10.4.1 Random effects via GLS

Because v_i is not spherical, the efficient estimator under the benchmark **RE.2** is GLS with weight matrix

$$\boldsymbol{\Omega} := \operatorname{Var}\left(v_i \mid \mathbf{X}_i\right) = \sigma_u^2 \mathbf{I}_T + \sigma_c^2 ee^\top.$$

The RE–GLS estimator is

$$\hat{\boldsymbol{\beta}}_{\mathrm{RE}} = \left(\sum_{i=1}^N \mathbf{X}_i^\top \boldsymbol{\Omega}^{-1} \mathbf{X}_i\right)^{-1} \left(\sum_{i=1}^N \mathbf{X}_i^\top \boldsymbol{\Omega}^{-1} \mathbf{y}_i\right).$$

Substituting $\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + v_i$ yields

$$\hat{\boldsymbol{\beta}}_{\mathrm{RE}} - \boldsymbol{\beta} = \left(\sum_{i=1}^N \mathbf{X}_i^\top \boldsymbol{\Omega}^{-1} \mathbf{X}_i\right)^{-1} \left(\sum_{i=1}^N \mathbf{X}_i^\top \boldsymbol{\Omega}^{-1} v_i\right),$$

so the large-sample arguments mirror OLS/FE with weighted moments.

For RE–GLS, the identifying orthogonality can be written compactly as

$$\mathbb{E}\left[\mathbf{X}_i^\top \boldsymbol{\Omega}^{-1} v_i\right] = 0,$$

and the usual rank condition is

$$\mathbf{A} := \mathbb{E}\left[\mathbf{X}_i^\top \boldsymbol{\Omega}^{-1} \mathbf{X}_i\right] \text{ is nonsingular.} \tag{RE.3}$$

Under an LLN/CLT across i, this yields $\hat{\boldsymbol{\beta}}_{\mathrm{RE}} \xrightarrow{p} \boldsymbol{\beta}$ and a CLT for $\sqrt{N}(\hat{\boldsymbol{\beta}}_{\mathrm{RE}} - \boldsymbol{\beta})$.

10.4.2 Feasible GLS for random effects (FGLS)

In practice $\boldsymbol{\Omega} = \sigma_u^2 \mathbf{I}_T + \sigma_c^2 ee^\top$ is unknown. Feasible GLS replaces $\boldsymbol{\Omega}$ by an estimate $\hat{\boldsymbol{\Omega}}$ and runs GLS with $\hat{\boldsymbol{\Omega}}^{-1}$.

Run pooled OLS on

$$y_{it} = \mathbf{x}_{it}^\top \boldsymbol{\beta} + v_{it}$$

and form residuals

$$\hat{v}_{it} := y_{it} - \mathbf{x}_{it}^\top \hat{\boldsymbol{\beta}}_{\mathrm{POLS}}.$$

A natural estimator of the composite variance $\sigma_v^2 := \mathrm{Var}\,(v_{it}) = \sigma_u^2 + \sigma_c^2$ is

$$\hat{\sigma}_v^2 := \frac{1}{NT-k}\sum_{i=1}^{N}\sum_{t=1}^{T}\hat{v}_{it}^2.$$

Under the benchmark, for $t \neq s$ within the same individual,

$$\mathrm{Cov}(v_{it}, v_{is}) = \sigma_c^2,$$

motivating the within-i cross-time estimator

$$\hat{\sigma}_c^2 := \frac{1}{N\binom{T}{2}}\sum_{i=1}^{N}\sum_{2\le t\le T}\sum_{1\le s<t}\hat{v}_{it}\hat{v}_{is}.$$

Then set

$$\hat{\sigma}_u^2 := \hat{\sigma}_v^2 - \hat{\sigma}_c^2, \qquad \hat{\Omega} := \hat{\sigma}_u^2\mathbf{I}_T + \hat{\sigma}_c^2 ee^\top,$$

and run feasible GLS:

$$\hat{\boldsymbol{\beta}}_{\text{RE-FGLS}} = \left(\sum_{i=1}^{N}\mathbf{X}_i^\top\hat{\boldsymbol{\Omega}}^{-1}\mathbf{X}_i\right)^{-1}\left(\sum_{i=1}^{N}\mathbf{X}_i^\top\hat{\boldsymbol{\Omega}}^{-1}\mathbf{y}_i\right).$$

10.4.3 Random effects as quasi-demeaning

GLS can be viewed as OLS after a whitening transformation. In the RE variance-components model, the whitening step becomes a simple *partial demeaning* (quasi-demeaning) that shrinks each unit's time series toward its own mean by an amount governed by σ_c^2/σ_u^2.

Lemma 10.5 (Quasi-demeaning for RE–GLS)**.** Let $e \in \mathbb{R}^T$ be the vector of ones and define $\mathbf{P} := \frac{1}{T}ee^\top$ and $\mathbf{Q} := \mathbf{I}_T - \frac{1}{T}ee^\top$. Consider the RE covariance $\boldsymbol{\Omega} = \sigma_u^2\mathbf{I}_T + \sigma_c^2 ee^\top$, where $\sigma_u^2 > 0$ and $\sigma_c^2 \ge 0$. Then:

(i) **Diagonalization in $(\mathbf{P}, \mathbf{Q})$.** $\boldsymbol{\Omega} = \sigma_u^2\mathbf{Q} + (\sigma_u^2 + T\sigma_c^2)\mathbf{P}, \boldsymbol{\Omega}^{-1} = \frac{1}{\sigma_u^2}\mathbf{Q} + \frac{1}{\sigma_u^2+T\sigma_c^2}\mathbf{P}$.

(ii) **Quasi-demeaning representation.** Define $\eta := \frac{\sigma_u^2}{\sigma_u^2+T\sigma_c^2} \in (0,1], \lambda := 1 - \eta^{1/2} \in [0,1)$. Up to an irrelevant overall scalar factor, premultiplication by $\boldsymbol{\Omega}^{-1/2}$ is equivalent to $(\mathbf{I}_T - \lambda\mathbf{P})$. Hence RE–GLS is numerically equal to OLS in the quasi-demeaned regression

$$\mathbf{y}_i^* := (\mathbf{I}_T - \lambda\mathbf{P})\mathbf{y}_i, \qquad \mathbf{X}_i^* := (\mathbf{I}_T - \lambda\mathbf{P})\mathbf{X}_i,$$

i.e.

$$y_{it}^* = y_{it} - \lambda\bar{y}_i, \qquad \mathbf{x}_{it}^* = \mathbf{x}_{it} - \lambda\bar{\mathbf{x}}_i.$$

Proof. First, $\mathbf{P}$ and $\mathbf{Q}$ are orthogonal idempotent projections with $\mathbf{P} + \mathbf{Q} = \mathbf{I}_T$ and $ee^\top = T\mathbf{P}$. Then

$$\boldsymbol{\Omega} = \sigma_u^2(\mathbf{P}+\mathbf{Q}) + \sigma_c^2(T\mathbf{P}) = \sigma_u^2\mathbf{Q} + (\sigma_u^2 + T\sigma_c^2)\mathbf{P},$$

and the candidate inverse $\boldsymbol{\Omega}^{-1}$ follows by multiplying out and using $\mathbf{PQ} = \mathbf{QP} = 0$.

Let $\eta = \sigma_u^2/(\sigma_u^2 + T\sigma_c^2)$ so that $\mathbf{\Omega} = (\sigma_u^2 + T\sigma_c^2)(\mathbf{P} + \eta\mathbf{Q})$. Because $\mathbf{P}$ and $\mathbf{Q}$ diagonalize the space, we have

$$(\mathbf{P} + \eta\mathbf{Q})^{-1/2} = \mathbf{P} + \eta^{-1/2}\mathbf{Q},$$

hence (up to an overall scalar factor),

$$\mathbf{P} + \eta^{-1/2}\mathbf{Q} = \mathbf{P} + \eta^{-1/2}(\mathbf{I}_T - \mathbf{P}) = \eta^{-1/2}\big(\mathbf{I}_T - (1 - \eta^{1/2})\mathbf{P}\big) = \eta^{-1/2}(\mathbf{I}_T - \lambda\mathbf{P}).$$

Finally, $\mathbf{P}\mathbf{y}_i = \bar{y}_i\, e$ and similarly $\mathbf{P}\mathbf{X}_i = \bar{\mathbf{x}}_i\, e^\top$, which yields the quasi-demeaned variables. □

Intuitively, the quasi-demeaning factor

$$\lambda = 1 - \left(\frac{\sigma_u^2}{\sigma_u^2 + T\sigma_c^2}\right)^{1/2}$$

controls how much of each unit's time average is subtracted. If $\sigma_c^2 \to 0$, then $\lambda \to 0$ and RE collapses to pooled OLS (no demeaning). If T is large and/or $\sigma_c^2 \gg \sigma_u^2$, then $\lambda \to 1$ and RE becomes increasingly FE-like (almost full demeaning).

10.5 Hausman test (FE vs. RE)

FE and RE rely on different identifying restrictions. FE remains valid even when c_i is correlated with the regressor path; RE is more efficient if c_i is orthogonal to the regressors, but becomes inconsistent if that orthogonality fails. The Hausman test turns this into a comparison of two estimators that target the same $\boldsymbol{\beta}$ under the null.

The null is the RE orthogonality restriction

$$H_0 : \ \mathbb{E}\left[c_i \mid \mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}\right] = 0.$$

Let $\hat{\boldsymbol{\beta}}_{\mathrm{FE}}$ be the within estimator and $\hat{\boldsymbol{\beta}}_{\mathrm{RE}}$ the RE–GLS (or RE–FGLS) estimator, and define

$$\Delta\hat{\boldsymbol{\beta}} := \hat{\boldsymbol{\beta}}_{\mathrm{FE}} - \hat{\boldsymbol{\beta}}_{\mathrm{RE}}.$$

Under H_0, $\Delta\hat{\boldsymbol{\beta}}$ is asymptotically centered at 0, and the Hausman statistic is

$$H = (\Delta\hat{\boldsymbol{\beta}})^\top \left[\widehat{\mathrm{Var}}(\Delta\hat{\boldsymbol{\beta}})\right]^{-1} (\Delta\hat{\boldsymbol{\beta}}), \qquad H \ \xrightarrow{d}\ \chi_k^2.$$

A common plug-in choice is

$$\widehat{\mathrm{Var}}(\Delta\hat{\boldsymbol{\beta}}) = \widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}_{\mathrm{FE}}) - \widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}_{\mathrm{RE}}),$$

motivated by the efficiency of RE under H_0.

In finite samples, the difference $\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}_{\mathrm{FE}}) - \widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}_{\mathrm{RE}})$ may fail to be positive semidefinite. Software typically handles this by using an alternative estimator of $\mathrm{Var}(\Delta\hat{\boldsymbol{\beta}})$ or a generalized inverse.

Algorithm 14 Hausman test for FE vs. RE

Require: Panel data $\{(y_{it}, \mathbf{x}_{it})\}_{i=1,t=1}^{N,T}$; significance level α.

1: Compute $\hat{\boldsymbol{\beta}}_{\mathrm{FE}}$ and $\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}_{\mathrm{FE}})$.
2: Compute $\hat{\boldsymbol{\beta}}_{\mathrm{RE}}$ (GLS or FGLS) and $\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}_{\mathrm{RE}})$.
3: Form $\Delta\hat{\boldsymbol{\beta}} \leftarrow \hat{\boldsymbol{\beta}}_{\mathrm{FE}} - \hat{\boldsymbol{\beta}}_{\mathrm{RE}}$.
4: Set $\widehat{\mathrm{Var}}(\Delta\hat{\boldsymbol{\beta}}) \leftarrow \widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}_{\mathrm{FE}}) - \widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}_{\mathrm{RE}})$.
5: Compute $H \leftarrow (\Delta\hat{\boldsymbol{\beta}})^{\top}[\widehat{\mathrm{Var}}(\Delta\hat{\boldsymbol{\beta}})]^{-1}(\Delta\hat{\boldsymbol{\beta}})$ and $p \leftarrow 1 - F_{\chi^2_k}(H)$.
6: **if** $p < \alpha$ **then**
7: Reject H_0; evidence that c_i is correlated with regressors; prefer FE as baseline.
8: **else**
9: Fail to reject H_0; RE is plausible (and typically more efficient under the null).
10: **end if**

10.6 Summary

Panel data are not "more observations." They are the same observations with personalities. Each individual drags around a persistent component c_i, and pretending it isn't there is how pooled OLS becomes confidently wrong.

Everything starts from

$$y_{it} = c_i + d_t + x_{it}^{\top}\beta + u_{it}.$$

The only question that really matters is whether c_i is related to the regressor path.

To tackle it, we introduced several estimators (a.k.a. how much you are willing to subtract away)

- **Pooled OLS:** delete nothing. Fast, tempting, and only valid if the panel structure is somehow harmless.
- **FE (within):** delete the entire individual mean. Works even if c_i and x_{it} are best friends, but needs within-i variation.
- **FD:** delete c_i by differencing. Great when T is small and the action is in changes.
- **RE (GLS/FGLS):** delete *some* of the mean (quasi-demeaning). Efficient if c_i is orthogonal to x_{it}; otherwise it's efficient at estimating the wrong thing.

Strict exogeneity buys you identification after the FE/FD transformation. It also breaks in the usual suspects: dynamic panels and nonlinear/binary choice settings.

If RE's orthogonality story is true, FE and RE should agree up to sampling noise. If they don't, the Hausman test says: "cute assumption," and you go back to FE. That's the chapter: choose how much unobserved heterogeneity you're willing to pretend away, and then live with the consequences.

Epilogue

If you have read this far and retained the thread, you have witnessed a slow, methodical dismantling of the idea that a regression is just a regression.

A Gentle Rewind

We started with asymptotics because honesty requires admitting that we are never sure about finite samples. We proved that sample averages stabilize (LLN) and that they wobble in a predictable, Gaussian way (CLT). Then we applied this to OLS. Under the Gauss-Markov assumptions, OLS is not just a line through a cloud of points; it is the Best Linear Unbiased Estimator. It is unbiased, it is efficient, and its standard errors come from a formula so clean you could frame it. This is the econometric equivalent of a childhood memory: simplified, comforting, and not entirely true.

Then we admitted that the errors are never spherical. They are heteroskedastic (their variance changes with the regressors) or correlated within clusters. The elegant $\sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}$ becomes a lie. We learned to diagnose this (Breusch-Pagan, White) and to fix it—either by robust standard errors (White's magic trick: square the residuals and pray) or by modelling the covariance explicitly (GLS, FGLS, SURE). The message: your coefficients are probably fine, but your confidence intervals are a work of fiction unless you adjust.

Then came endogeneity. This is not a mere violation; it is the moment when OLS stops estimating what you think it estimates. The regressor is correlated with the error, and the bias does not go away as $n \to \infty$. We reached for instruments: variables that move the regressor but are otherwise unrelated to the outcome. This led us to IV and then to GMM, which is just the formal admission that we have more equations (moment conditions) than unknowns. We learned to weight them optimally, to test them (Hansen's J-test: a large statistic means your instruments are either invalid or your model is wrong, but the test won't tell you which), and to accept that 2SLS is biased in finite samples but consistent asymptotically. It was the moment we stopped believing in unbiasedness and started believing in large samples.

MLE asked us to stop being coy with moments and just specify the whole distribution. It is the method of maximum honesty and maximum liability. When the model is right, you get efficiency, Fisher information, and the three classical tests (Wald, LR, LM) that all agree asymptotically but squabble in finite samples. When the model is wrong, you get a sandwich estimator and a vague sense of unease.

Binary choice models reminded us that outcomes are not always continuous. The LPM is simple but will predict probabilities outside [0,1] without shame. Probit and Logit fix this by squashing the linear index through a CDF. But they also introduced the delta method: if you want a standard error for a marginal effect, you cannot just read it off the output; you have to differentiate, multiply by the covariance, and hold your breath. And if you have endogeneity in a binary model, you cannot just run 2SLS because the nonlinearity breaks the geometry. You have to model the joint distribution of the endogenous regressor and the error, which is either heroic or foolhardy, depending on your referee.

Finally, panels. Here, the problem is not just correlation with the error, but correlation with a persistent, unobserved individual effect c_i. Pooled OLS ignores it and is biased. Fixed effects eliminates it by demeaning (or differencing) and is consistent even if c_i is correlated

with the regressors—but it discards all cross-sectional variation and cannot estimate time-invariant coefficients. Random effects keeps the cross-sectional variation by treating c_i as part of the error, but it requires c_i to be uncorrelated with the regressors. The Hausman test asks whether the two estimators agree; if they don't, you are supposed to believe FE, because it is robust and RE is not. In practice, the test often rejects, and you go back to FE, muttering about the loss of efficiency.

If there is a single thread, it is this: we are trying to learn something about the world from data that never cooperates. We start with OLS because it is the honest baseline. Then we adjust for heteroskedasticity, because the world is not constant. Then we instrument, because the world is not exogenous. Then we go nonlinear, because the world is not linear. Then we add fixed effects, because the world is not independent across time. By the end, we are running GMM on a system of equations with cluster-robust standard errors and a Hansen J-statistic that we pray is insignificant. And if someone asks what we are doing, we say we are "doing inference"—which is a polite way of saying we are trying to tell a story that the data cannot easily disprove.

Acknowledgement

These notes were written for Professor Ruochen Wu's course `ECON130277H` **Cross-Sectional and Panel Data Analysis**, and they owe him far more than a polite line at the end. Most of what follows is edited (and occasionally rescued) from the notes I took in his lectures.

Professor Wu is sharp enough to teach technical material cleanly without slides—which is both impressive and mildly unfair to the rest of us—and kind enough to make questions feel welcome even when they arrive in their most confused form. He has a rare combination of clarity, rigor, and patience: the kind that makes theorems feel inevitable *after* he explains them, and makes my earlier confusion feel like it had no excuse.

I am sincerely grateful for his teaching and for the time he spent answering questions, correcting my misunderstandings, and encouraging me while I was turning messy class notes into something readable. His support made this write-up possible, and it has also helped me beyond the course—including my graduate school application. Any remaining errors are mine: they are the parts where I stopped being supervised and started being confident.

Miscellany

To be clear, these notes were not authored by a calm, well-rested adult with a stable sleep schedule. They were written by someone who treats stress like a productivity hack, who insists that "one more section" counts as a bedtime routine, and who once decided that writing sixteen pages in five hours was a sensible plan—the kind of decision that feels heroic at 3:00 a.m. and starts to look like a small, preventable tragedy by 9:00 a.m. I also take "short breaks" the way instrumental variables are "exogenous": mostly as an article of faith, occasionally as a slogan, and rarely as an observable event. Please do not emulate this strategy unless you enjoy negotiating with your circadian rhythm like it's a hostile referee.

And yes: there are almost certainly mistakes. Some are typos. Some are "I swear this made sense when I wrote it." And some are the inevitable consequence of letting an overcaffeinated undergraduate play editor, typesetter, and theorist at the same time. If

you catch anything wrong (or merely suspicious enough to deserve a raised eyebrow), I'd genuinely appreciate a note—I can be reached at `hmfang22@m.fudan.edu.cn`

Tactical Advice for the Final

As far as I can recall, the exam is not asking whether you have a deep spiritual relationship with asymptotics. It is asking whether, under time pressure, you can reconstruct the standard pipeline without panicking: start from the goal, back out what must be true for the goal to even make sense, and then push the machinery until an estimator and its large-sample behavior fall out.

Concretely, the workflow is usually: state the target parameter (so we know what game we are playing), specify the population object that pins it down (so it is actually identified), write the sample objective or estimating equation that defines the estimator (so we have something to compute), and then list the assumptions that make the asymptotics go through (so you are allowed to take limits without committing crimes). If you can produce that chain on demand, you are basically holding the exam's hand while it tries to trick you.

So in practice, it is checking whether you can:

1. Write down the object being estimated (parameter, estimand, or functional of the DGP);
2. State the population condition that identifies it (moment condition, orthogonality, likelihood, or rank condition);
3. Write the estimator as a sample analog (minimizer/maximizer/solution to an equation);
4. List the assumptions that justify the limit operations (LLN, CLT, differentiability/regularity, non-singularity);
5. Not confuse "uncorrelated" with "independent" or "exogeneity" with "I feel like it should be true" when the clock is ticking.

Everything else is garnish. Delicious garnish, but garnish nonetheless.

A Final Word

If you are about to take Prof. Wu's final: I hope your brain produces clean derivations, your pen never runs out of ink, and your estimators converge faster than your anxiety. May your fixed effects actually wipe out the fixed effects. May your instruments be relevant and exogenous (a rare and magical two-for-one deal). And may you never again write "strict exogeneity" without first checking what it actually means.

If the exam starts feeling personal: remember, it isn't. (Well, it might be. But you should proceed as if it isn't.) Take a breath, write down the model, label the assumptions, and move. This is econometrics: you do not need to win every battle; you just need to keep your identification alive long enough to finish the proof.

When you walk out of that room, remember: you did real work. You built a mental model for how empirical claims are justified or debunked, which is more valuable than any single grade will ever admit. Please do one healthy thing: Sleep. Eat. Go outside. Rejoin

society. The Gauss–Markov theorem will not be offended if you ignore it for twelve hours. The shiny little badge—whether it's a grade on `jwfw.com` or a *"We are delighted to inform you..."*—will arrive as life promised.

> *"My point is, while you're spending all this time on your own, building computers or practicing your cello, what you're really doing is becoming interesting. When people finally do notice you, they're gonna find someone a lot cooler than they thought... Congratulations."*
>
> — *Leonard Hofstadter, Ph.D.*

References

[1] Hugh A. Chipman, Edward I. George, and Robert E. McCulloch. Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1), March 2010.

[2] Angus Deaton and John Muellbauer. An almost ideal demand system. *The American Economic Review*, 70(3):312–326, 1980.

[3] Jing Liang, Yueming Qiu, and Bo Xing. Social versus private benefits of energy efficiency under time-of-use and increasing block pricing. *Environmental and Resource Economics*, 78:1–33, 01 2021.

[4] Lucas Mentch and Giles Hooker. Quantifying uncertainty in random forests via confidence intervals and hypothesis tests, 2015.