Advanced Econometrics
05 Maximum Likelihood

Eduard Brüll

Fall 2025

Lecture 5

Advanced Econometrics

Literature: Greene Chapter 14, 17-19

5.1: Intro to Maximum Likelihood Estimation

Is this coin fair?

Experiment

- Flip $n = 10$ times, observe $y = 7$ heads.
- Each flip $Y_i \sim \text{Bernoulli}(p)$.
- Unknown parameter $p = \Pr(Y_i = 1)$.

Goal: Choose p that makes the observed data **most likely**.

Core Principle

This is **Maximum Likelihood Estimation** of parameters θ of a distribution function.

Assumptions for Maximum Likelihood Estimation

Key ingredients of each Maximum Likelihood Problem:

1. **Model specification:** Each Y_i has a well-specified probability density or probability mass function $f(y_i \mid \theta)$
2. **Independence and Identical Distribution (IID) :** $\{Y_i\}_{i=1}^{n}$ are independent and all $Y_i \sim f(y_i \mid \theta)$
3. **Regularity conditions:** Technical assumptions so that the math works
   - Log-likelihood is smooth and information finite (so derivatives/inference valid)
   - Parameters lie in the interior (no weird boundary or pathological cases)

Implication:

With this type of assumptions, we can build and maximize the likelihood of any known or assumed distribution function.

ML2 and ML3: What IID buys us

Independent:

$$\Pr(Y_1, \ldots, Y_n \mid p) = \prod_{i=1}^{n} \Pr(Y_i \mid p)$$

Identically distributed:

$$Y_i \sim \text{Bernoulli}(p) \quad \forall i$$

Payoff: The likelihood to observe our data is just a simple product.

From Bernoulli to "Binomial" Likelihood

Step 1: Start from the Bernoulli model.

Each observation follows a Bernoulli distribution:

$$f(Y_i \mid p) = p^{Y_i}(1-p)^{1-Y_i}, \quad Y_i \in \{0,1\}.$$

Step 2: Use independence.

Since flips are independent,

$$L(p \mid Y_1, \dots, Y_n) = \prod_{i=1}^{n} f(Y_i \mid p) = \prod_{i=1}^{n} p^{Y_i}(1-p)^{1-Y_i}.$$

Step 3: Collect exponents.

$$L(p) = p^{\sum_i Y_i}(1-p)^{n-\sum_i Y_i}.$$

Step 4: Express in terms of observed number of successes.

Let $y = \sum_i Y_i$ be the number of heads (successes):

$$L(p) = p^{y}(1-p)^{n-y}.$$

Likelihood of our sample

$$L(p \mid Y_1, \ldots, Y_n) \;=\; \prod_{i=1}^{n} f(Y_i \mid p) \;=\; p^{y}(1-p)^{n-y}$$

- Independence $\Rightarrow$ product of individual Bernoulli terms.
- Let $y = \sum_i Y_i$ = number of heads.
- This is the **likelihood for the observed sequence**.

Common Trick: Work with the log-likelihood!

$$\ell(p) \;=\; \log L(p) \;=\; y \log p + (n-y)\, \log(1-p)$$

- $\log$ is monotone $\Rightarrow$ same maximizer as L.
- Sums are easier than products; derivatives are simpler.

(If we aggregate over all sequences with y heads instead of just looking at our observed sequence, we add the binomial term $\binom{n}{y}$, but it's constant in p.)

The Maximization Problem for the Coin

Recall:

$$\ell(p) \;=\; \log L(p) \;=\; y \log p + (n-y) \log(1-p)$$

Our First Order Condition:

$$\frac{\partial \ell(p)}{\partial p} \;=\; \frac{y}{p} - \frac{n-y}{1-p} \;=\; 0 \quad \Rightarrow \quad \hat{p} = \frac{y}{n}$$

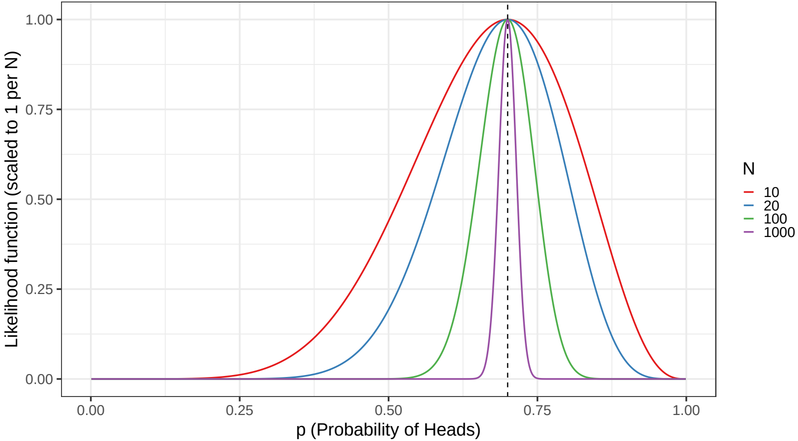
$$\text{Here: } \hat{p} = \frac{7}{10} = 0.7$$

Check curvature to see if its a maximum:

$$\frac{\partial^2 \ell(p)}{\partial p^2} \;=\; -\frac{y}{p^2} - \frac{n-y}{(1-p)^2} \;<\; 0$$

- Negative second derivative ⇒ **unique maximum**.
- Intuition: a sharper peak ⇒ more precise $\hat{p}$.

Likelihood Function for our Coin across Sample Sizes



5.2: MLE Properties

Setting Up the Maximum Likelihood Problem

Goal: Estimate unknown parameter vector $\theta \in \Theta \subseteq \mathbb{R}^k$ that governs the distribution of the observed data $\mathbf{y} = (y_1, \ldots, y_n)$.

Model:

$$f(y_i \mid \theta) \quad \text{for } i = 1, \ldots, n,$$

where $f(\cdot \mid \theta)$ is a known pdf or pmf depending on θ.

Likelihood function:

$$L(\theta \mid \mathbf{y}) = \prod_{i=1}^{n} f(y_i \mid \theta) \quad \text{and} \quad \ell(\theta) = \log L(\theta \mid \mathbf{y}) = \sum_{i=1}^{n} \log f(y_i \mid \theta).$$

Maximum Likelihood Estimator (MLE):

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta \in \Theta} \ \ell(\theta).$$

Interpretation

The MLE chooses parameter values that make the observed data most probable under the assumed model $f(y_i \mid \theta)$.

Properties of MLE

Under regularity conditions MLE has four properties:

M1 **Consistency:** The MLE $\hat{\theta}$ converges in probability to the true parameter value θ_0.

$$\hat{\theta} \xrightarrow{p} \theta_0$$

M2 **Asymptotic Normality:** After scaling by $\sqrt{n}$, the distribution of the estimation error is approximately normal, with variance given by the inverse Fisher information.

$$\sqrt{n}\,(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}\left(0,\, I(\theta_0)^{-1}\right)$$

where $I(\theta_0) = -\mathbb{E}_{\theta_0}\left[\frac{\partial^2}{\partial\theta^2} \ln L(\theta)\right]$ is the Fisher information.

M3 **Asymptotic Efficiency:** Among consistent estimators, the MLE achieves the Cramér–Rao lower bound asymptotically, i.e. it has the smallest possible asymptotic variance.

M4 **Invariance:** If we are interested in a function $g(\theta_0)$ of the parameter, the MLE is obtained simply by applying g to $\hat{\theta}$:

$$\widehat{g(\theta)} = g(\hat{\theta})$$

Finite-Sample Reality of the MLE

Asymptotic theory gives us:

- ✓ $\hat{\theta}_{\text{MLE}}$ is **consistent**: converges to θ_0 as $n \to \infty$.
- ✓ $\hat{\theta}_{\text{MLE}}$ is **asymptotically normal and efficient**.
- ✓ In large samples, likelihood-based inference is straightforward and reliable.

But in finite samples:

- The MLE can be **biased**, especially in nonlinear models or with small n.
- Sampling distributions may be **skewed or heavy-tailed**.
- Standard (asymptotic) confidence intervals may **undercover** the true parameter.

The Score Function

- Log-likelihood:

$$\ell(\theta) = \sum_{i=1}^{n} \ln f(y_i \mid \theta)$$

- **Score vector (gradient):**

$$g(\theta) = \frac{\partial \ell(\theta)}{\partial \theta} = \sum_{i=1}^{n} g_i(\theta), \quad g_i(\theta) = \frac{\partial}{\partial \theta} \ln f(y_i \mid \theta)$$

- Interpretation:
  - $g_i(\theta)$ is the contribution of observation i
  - $g(\theta)$ is the total score

Key Property of the Score

- At the true parameter θ_0:

$$\mathbb{E}[\, g_i(\theta_0) \,] = 0$$

- Therefore:

$$\mathbb{E}[g(\theta_0)] = \mathbb{E}\left[\sum_{i=1}^{n} g_i(\theta_0)\right] = 0$$

- This is the **Likelihood Equation**, key to asymptotic normality.

Intuition

The score is the slope of the log-likelihood. At the true parameter θ_0, the expected slope must vanish, because the model is correctly specified and centered around θ_0.

Second Derivative ⇒ (Fisher) Information

- **Observed information (curvature)** at θ:

$$J(\theta) = -\frac{\partial^2 \ell(\theta)}{\partial \theta^2}$$

- **Fisher information** at θ_0 (expected curvature):

$$I(\theta_0) = \mathbb{E}_{\theta_0}[J(\theta_0)] = -\mathbb{E}_{\theta_0}\left[\left. \frac{\partial^2 \ell(\theta)}{\partial \theta^2} \right|_{\theta=\theta_0} \right]$$

- **Key role:** determines the asymptotic variance

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}\big(0, I(\theta_0)^{-1}\big).$$

Coin Example: Second Derivative

Model: $Y_i \sim \mathrm{Bernoulli}(p)$, independent, with $y = \sum_i Y_i$.

Then

$$\ell(p) = y \log p + (n - y) \log(1 - p).$$

First derivative (score):

$$\frac{\partial \ell(p)}{\partial p} = \frac{y}{p} - \frac{n-y}{1-p}.$$

Second derivative (curvature):

$$\frac{\partial^2 \ell(p)}{\partial p^2} = -\frac{y}{p^2} - \frac{n-y}{(1-p)^2} \; < \; 0.$$

Interpretation

Curvature tells us how sharp or flat the log-likelihood is around p.
More negative $\Rightarrow$ sharper peak $\Rightarrow$ more precise $\hat{p}$.

Expected Curvature and Fisher Information

Step 1: Take expectation at $p = p_0$.

$$\mathbb{E}\left[\frac{\partial^2 \ell(p)}{\partial p^2}\bigg|_{p=p_0}\right] = -\,\mathbb{E}\left[\frac{Y}{p_0^2} + \frac{n-Y}{(1-p_0)^2}\right]$$

Step 2: Use $\mathbb{E}[Y] = np_0$.

$$= -\left(\frac{np_0}{p_0^2} + \frac{n-np_0}{(1-p_0)^2}\right) = -\left(\frac{n}{p_0} + \frac{n}{1-p_0}\right).$$

Step 3: Simplify.

$$\mathbb{E}\left[\frac{\partial^2 \ell(p)}{\partial p^2}\bigg|_{p_0}\right] = -\,\frac{n}{p_0(1-p_0)}.$$

Result (Fisher information):

$$I(p_0) = -\,\mathbb{E}\left[\frac{\partial^2 \ell}{\partial p^2}\right] = \frac{n}{p_0(1-p_0)}.$$

Intuition

The Fisher information is large when the likelihood is steeply curved (near $p_0 \approx 0$ or 1) $\Rightarrow$ variance of $\hat{p}$ is small. Shallow curvature around $p_0 = 0.5 \Rightarrow$ higher variance.

Numeric Variance of the Coin MLE

Recall: For a Bernoulli model with $Y_i \sim \mathrm{Bernoulli}(p)$, the Fisher information is

$$I(p_0) = \frac{n}{p_0(1-p_0)}.$$

Therefore, the asymptotic variance of the MLE:

$$\mathbf{var}(\hat{p}) \ \approx \ \frac{1}{I(p_0)} = \frac{p_0(1-p_0)}{n}.$$

Numerical example:

$$n = 10, \quad \hat{p} = 0.7.$$

$$I(\hat{p}) = \frac{10}{0.7(1-0.7)} = 47.62, \qquad \mathbf{var}(\hat{p}) = \frac{1}{47.62} = 0.021.$$

Interpretation

With $n = 10$ flips, the MLE has an estimated variance of 0.021, or a standard error $\sqrt{0.021} \approx 0.145$. Larger n or more extreme p values $\Rightarrow$ higher information, smaller variance.

Properties of the MLE: Consistency

What we saw: As N increased in the likelihood plots, the peak became narrower and more centered around the true $p_0 = 0.7$.

Formal idea:

$$\frac{1}{N}\ell_N(\theta) = \frac{1}{N}\sum_{i=1}^{N}\log f(y_i \mid \theta) \xrightarrow{p} E[\log f(Y \mid \theta)].$$

Implications:

- The limiting function $E[\log f(Y \mid \theta)]$ is maximized at the true parameter θ_0.
- Therefore, $\hat{\theta}_{MLE} \xrightarrow{p} \theta_0$.
- More data $\longrightarrow$ the likelihood concentrates around θ_0.

Intuition: The likelihood surface becomes less random and more "deterministic" as sample size grows. The MLE stabilizes around the true value. This is <u>consistency.</u>

Information Matrix Equality

Curvature or Variance of the Score:
So far we used the curvature (Hessian) to define Fisher information. But Fisher information can also be written as the variance of the score. The **Information Matrix Equality** says these are the same.

Result (one observation):

$$I(\theta_0) = \mathbf{var}\big[g_i(\theta_0)\big] = -\,\mathbf{E}\big[H_i(\theta_0)\big],$$

where

$$g_i(\theta) = \tfrac{\partial}{\partial\theta} \ln f(y_i \mid \theta), \qquad H_i(\theta) = \tfrac{\partial^2}{\partial\theta\,\partial\theta'} \ln f(y_i \mid \theta).$$

Intuition

- The score $g_i(\theta)$ measures slope (random across samples).
- The Hessian $H_i(\theta)$ measures curvature.
- Their expectations agree $\Rightarrow$ slope-variance and curvature tell the same story about precision.

The Information Matrix in Practice

Information Matrix Equality:

$$I(\theta_0) = \mathbf{var}[g(\theta_0)] = -\ \mathbf{E}[H(\theta_0)]$$

Problem: The expectation $-\ \mathbf{E}[H(\theta_0)]$ is usually not feasible in practice.

Two practical alternatives (asymptotically equivalent):

1. **Observed Hessian:**

$$\hat{I}(\hat{\theta}) = -\frac{\partial^2 \ell(\hat{\theta})}{\partial \theta\, \partial \theta'}$$

2. **Outer product of gradients (BHHH):**

$$\tilde{I}(\hat{\theta}) = \sum_{i=1}^{n} g_i(\hat{\theta}) g_i(\hat{\theta})'$$

Intuition

Both estimators approximate the same information. Choice depends on convenience: Hessian requires second derivatives, BHHH uses only first derivatives.

(BHHH = Berndt-Hall-Hall-Hausman, 1974: "Estimation and Inference in Nonlinear Structural Models")

Why the Outer Product Works

Step 1: Information matrix equality (one obs.)

$$I(\theta_0) = \mathbf{E}[g_i(\theta_0) g_i(\theta_0)'] = -\,\mathbf{E}[H_i(\theta_0)].$$

Step 2: Expand variance

$$\mathbf{var}[g_i(\theta_0)] = \mathbf{E}\big[g_i(\theta_0) g_i(\theta_0)'\big] - \mathbf{E}[g_i(\theta_0)]\ \mathbf{E}[g_i(\theta_0)]'.$$

Step 3: Use score property

$$\mathbf{E}[g_i(\theta_0)] = 0 \quad \Rightarrow \quad \mathbf{var}[g_i(\theta_0)] = \mathbf{E}[g_i(\theta_0) g_i(\theta_0)'].$$

In practice:

$$\tilde{I}(\hat{\theta}) = \sum_{i=1}^{n} g_i(\hat{\theta}) g_i(\hat{\theta})'$$

is a sample analogue of the Fisher information, **avoiding second derivatives (BHHH method)**.

The Regularity Conditions we Need

Model and parameter

- Identifiability: $f(y \mid \theta_1) = f(y \mid \theta_2)$ a.s. $\Rightarrow \theta_1 = \theta_2$.
- True parameter θ_0 lies in the interior of a parameter space Θ.

Smoothness & dominance

- $\ell(\theta) = \sum_{i=1}^{n} \log f(y_i \mid \theta)$ is twice continuously differentiable near θ_0.
- Can interchange differentiation and integration; score has mean zero and finite variance.

Information and curvature

- Fisher information $I(\theta_0)$ exists, finite, and is non-singular.

Sampling assumptions

- IID (or weak dependence with LLN/CLT valid); no vanishing information per observation.

Implication Under these, consistency, asymptotic normality, and efficiency results apply.

Invariance of MLE

Property: If $\hat{\theta}$ is the MLE of θ, then for any continuous function $g(\cdot)$:

$$\widehat{g(\theta)} = g(\hat{\theta}).$$

Implications:

- No need to re-maximize likelihood for transformations of parameters.
- Works for nonlinear transformations as well.

Examples:

- **Bernoulli:** if $\hat{p}$ is MLE for success probability, then $1 - \hat{p}$ is MLE for failure probability.
- **Normal:** if $\hat{\sigma}^2$ is MLE for variance, then $\hat{\sigma} = \sqrt{\hat{\sigma}^2}$ is MLE for standard deviation.

Takeaway

MLEs are **automatically invariant** under transformations — a very convenient property.

Efficiency

The precision of the MLE $\hat{\theta}$ is limited by the Fisher information $\mathcal{I}(\theta)$ of the likelihood:

$$\mathbf{var}(\hat{\theta}) \;\geq\; \frac{1}{\mathcal{I}(\theta)}.$$

Interpretation:

- This is the **Cramér–Rao lower bound** for the variance of any regular, asymptotically unbiased estimator of θ.
- For large samples, $\sqrt{n}(\hat{\theta} - \theta_0)$ is asymptotically normal with variance

$$\mathbf{var}(\hat{\theta}) \;\approx\; \frac{1}{n\,\mathcal{I}(\theta_0)}.$$

- Under correct model specification, the MLE achieves this bound asymptotically:

MLE has the smallest asymptotic variance

among all estimators that are root-*n* consistent.

Root-*N* consistency:

$$\sqrt{n}\,(\hat{\theta} - \theta_0) \;\xrightarrow{d}\; \backslash\big(0,\, V(\theta_0)\big),$$

meaning that $\hat{\theta}$ converges to θ_0 at rate $1/\sqrt{n}$. This is the fastest possible rate for regular estimators, and the MLE attains the minimum possible asymptotic variance within this class.

Efficiency Proof (Step 1)

Starting from Unbiasedness:
We begin with the assumption that the estimator $\hat{\theta}$ is unbiased:

$$\mathbb{E}[\hat{\theta} - \theta_0 \mid \theta_0] = 0.$$

In integral form:

$$\int (\hat{\theta} - \theta_0)\, f(y; \theta_0)\, dy = 0,$$

where $f(y; \theta_0)$ is the likelihood function (or probability density).

Idea:

We will differentiate this identity with respect to θ (and evaluate at θ_0) to relate the variance of $\hat{\theta}$ to the information in the data.

Step 2: Differentiate w.r.t. the Parameter

Differentiating the unbiasedness condition with respect to θ and then evaluating at $\theta = \theta_0$:

$$\textbf{FOC:} \quad 0 = \frac{\partial}{\partial \theta} \int (\hat{\theta} - \theta)\, f(y;\theta)\, dy \,\Big|_{\theta=\theta_0}.$$

Applying the product rule inside the integral:
Both $(\hat{\theta} - \theta)$ and $f(y;\theta)$ depend on θ, so when differentiating their product we get:

$$\frac{\partial}{\partial \theta}\big[(\hat{\theta} - \theta) f(y;\theta)\big] = (\hat{\theta} - \theta)\frac{\partial f(y;\theta)}{\partial \theta} + f(y;\theta)\frac{\partial(\hat{\theta} - \theta)}{\partial \theta}.$$

Since $\hat{\theta}$ does not depend on θ, $\frac{\partial(\hat{\theta}-\theta)}{\partial \theta} = -1$.
So:

$$0 = \int \Big[(\hat{\theta} - \theta)\frac{\partial f(y;\theta)}{\partial \theta} - f(y;\theta)\Big] dy.$$

Simplify: Because $\int f(y;\theta)\, dy = 1$ for all θ, differentiating gives $\int \frac{\partial f(y;\theta)}{\partial \theta} dy = 0$. Hence only the first term remains:

$$\int (\hat{\theta} - \theta)\frac{\partial f(y;\theta)}{\partial \theta}\, dy = 0.$$

Evaluating at $\theta = \theta_0$ gives

$$\int (\hat{\theta} - \theta_0)\frac{\partial f(y;\theta)}{\partial \theta}\Big|_{\theta=\theta_0} dy = 0.$$

From the previous step:

$$\int (\hat{\theta} - \theta_0) \, \frac{\partial f(y;\theta)}{\partial \theta}\bigg|_{\theta=\theta_0} dy = 0.$$

Apply the chain rule to the density:

$$\frac{\partial f(y;\theta)}{\partial \theta} = f(y;\theta) \, \frac{\partial \ln f(y;\theta)}{\partial \theta}.$$

This works because differentiating $\ln f$ gives:

$$\frac{\partial \ln f}{\partial \theta} = \frac{1}{f} \frac{\partial f}{\partial \theta} \quad \Rightarrow \quad \frac{\partial f}{\partial \theta} = f \, \frac{\partial \ln f}{\partial \theta}.$$

Substitute back into the integral:

$$\int (\hat{\theta} - \theta_0) \, f(y;\theta_0) \, \frac{\partial \ln f(y;\theta)}{\partial \theta}\bigg|_{\theta=\theta_0} dy = 0.$$

Interpretation: The term $\frac{\partial \ln f(y;\theta)}{\partial \theta}\Big|_{\theta=\theta_0}$ is the score function at the true parameter θ_0. It measures how sensitive the log-likelihood is to changes in θ around θ_0.

Step 4: Review of the Cauchy–Schwarz Inequality

Statement (Expectation Form):

Cauchy–Schwarz Inequality

For any random variables U and V with finite second moments,

$$\left| \mathbb{E}[UV] \right|^2 \le \mathbb{E}[U^2]\,\mathbb{E}[V^2].$$

Equivalent Integral Form: When $U = u(y)$ and $V = v(y)$ under density $f(y) > 0$,

$$\left| \int u(y)v(y)\, f(y)\, dy \right|^2 \le \left(\int u(y)^2 f(y)\, dy \right) \left(\int v(y)^2 f(y)\, dy \right).$$

Equality holds if and only if

$$u(y) = c\, v(y) \quad \text{for some constant } c.$$

In our context (at θ_0):

$$u(y) = \hat{\theta} - \theta_0, \qquad v(y) = \frac{\partial \ln f(y;\theta)}{\partial \theta}\Big|_{\theta=\theta_0},$$

with $f(y;\theta_0)$ as the weighting function (i.e. the probability density under the true parameter).

Step 5: Applying the Cauchy–Schwarz Inequality

Apply the Cauchy–Schwarz inequality to

$$0 = \int (\hat{\theta} - \theta_0)\, f(y;\theta_0)\, \frac{\partial \ln f(y;\theta)}{\partial \theta}\Big|_{\theta=\theta_0} dy.$$

By the Cauchy–Schwarz inequality:

$$0^2 \le \left[\int (\hat{\theta} - \theta_0)^2 f(y;\theta_0)\, dy\right] \left[\int \left(\frac{\partial \ln f(y;\theta)}{\partial \theta}\Big|_{\theta=\theta_0}\right)^2 f(y;\theta_0)\, dy\right].$$

This inequality is trivially true, but equality holds only if

$$\hat{\theta} - \theta_0 = c\, \frac{\partial \ln f(y;\theta)}{\partial \theta}\Big|_{\theta=\theta_0} \quad \text{for some constant } c.$$

Interpretation: The efficient estimator (the one achieving equality in the Cauchy–Schwarz bound) is proportional to the score function. This insight motivates the next step, where we normalize the proportionality constant to obtain the Cramér–Rao lower bound:

$$\mathbf{var}_{\theta_0}(\hat{\theta}) \ge \frac{1}{\mathcal{I}(\theta_0)}, \qquad \mathcal{I}(\theta_0) = \mathbb{E}_{\theta_0}\left[\left(\frac{\partial \ln f(y;\theta)}{\partial \theta}\Big|_{\theta=\theta_0}\right)^2\right].$$

Step 6: Deriving the Cramér–Rao Lower Bound

So far we have:

$$\mathbb{E}_{\theta_0}\left[(\hat{\theta} - \theta_0)\frac{\partial \ln f(y;\theta)}{\partial \theta}\Big|_{\theta=\theta_0}\right] = 0.$$

This holds directly from the unbiasedness condition. To obtain a meaningful lower bound, we differentiate the unbiasedness condition itself:

$$\mathbb{E}_\theta[\hat{\theta}] = \theta \quad \Longrightarrow \quad \frac{\partial}{\partial \theta}\mathbb{E}_\theta[\hat{\theta}] = 1.$$

Expanding the derivative:

$$\frac{\partial}{\partial \theta}\mathbb{E}_\theta[\hat{\theta}] = \int \hat{\theta}\,\frac{\partial f(y;\theta)}{\partial \theta}\,dy = \int \hat{\theta}\,f(y;\theta)\,\frac{\partial \ln f(y;\theta)}{\partial \theta}\,dy.$$

Subtracting θ times the derivative of $\int f(y;\theta)\,dy = 1$ gives:

$$\mathbb{E}_{\theta_0}\left[(\hat{\theta} - \theta_0)\frac{\partial \ln f(y;\theta)}{\partial \theta}\Big|_{\theta_0}\right] = 1.$$

Now apply Cauchy–Schwarz:

$$1^2 \leq \mathbb{E}_{\theta_0}[(\hat{\theta} - \theta_0)^2]\,\mathbb{E}_{\theta_0}\left[\left(\frac{\partial \ln f(y;\theta)}{\partial \theta}\Big|_{\theta_0}\right)^2\right].$$

Step 7: Finishing up

$$1^2 \leq \mathbb{E}_{\theta_0}\big[(\hat{\theta} - \theta_0)^2\big]\, \mathbb{E}_{\theta_0}\left[\left(\frac{\partial \ln f(y;\theta)}{\partial \theta}\Big|_{\theta_0}\right)^2\right].$$

Hence:

$$\boxed{\mathbf{var}_{\theta_0}(\hat{\theta}) \geq \frac{1}{\mathcal{I}(\theta_0)}, \qquad \mathcal{I}(\theta_0) = \mathbb{E}_{\theta_0}\left[\left(\frac{\partial \ln f(y;\theta)}{\partial \theta}\Big|_{\theta_0}\right)^2\right].}$$

Equality condition:

$$\hat{\theta} - \theta_0 = c\, \frac{\partial \ln f(y;\theta)}{\partial \theta}\Big|_{\theta_0}, \quad c = \frac{1}{\mathcal{I}(\theta_0)}.$$

Numerical Computation of the MLE

Many likelihoods (e.g. logit, Poisson, probit) have no closed-form solution.

Instead, we solve the first-order condition $g(\theta) = \frac{\partial \ell(\theta)}{\partial \theta} = 0$ iteratively.

Newton–Raphson update:

$$\theta^{(k+1)} = \theta^{(k)} - \left[H(\theta^{(k)})\right]^{-1} g(\theta^{(k)}),$$

where

$$g(\theta) = \frac{\partial \ell(\theta)}{\partial \theta}, \qquad H(\theta) = \frac{\partial^2 \ell(\theta)}{\partial \theta \partial \theta'}.$$

Interpretation:

- Move in the direction of steepest ascent (gradient g),
- scaled by curvature (H^{-1}) — a local quadratic approximation.
- Repeat until change in $\ell(\theta)$ or θ is negligible.

In 1D illustration:

$$\theta^{(k+1)} = \theta^{(k)} - \frac{\ell'(\theta^{(k)})}{\ell''(\theta^{(k)})}.$$

Visually: tangent to $\ell(\theta)$ at $\theta^{(k)}$ intersects the axis — next iterate.

5.3: Likelihood-based Tests

Repetition of Distributions (B.4.2)

Chi-squared distribution:

- If $z \sim \mathcal{N}(0,1)$, then $z^2 \sim \chi^2[1]$ (one d.f.)
- And $\sum_{i=1}^{n} z_i^2 \sim \chi^2[n]$ (with n d.f.)

  Note: variables must be independent.

F-distribution:

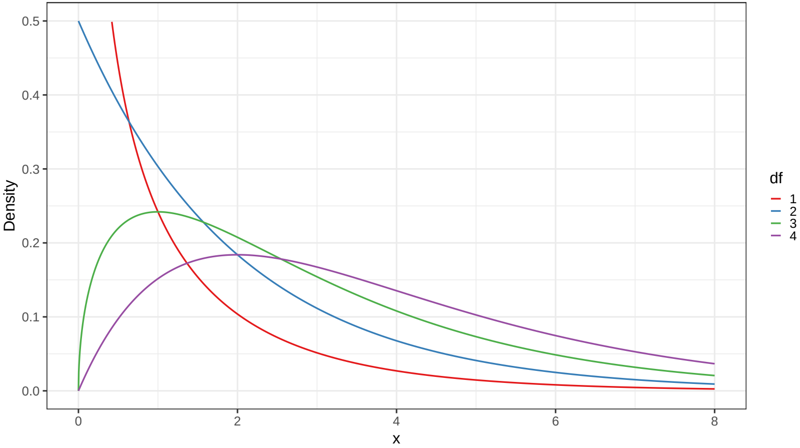
- $$\frac{\chi^2_{\nu_1}/\nu_1}{\chi^2_{\nu_2}/\nu_2} \sim F[\nu_1, \nu_2]$$

  Note: numerators/denominators independent.

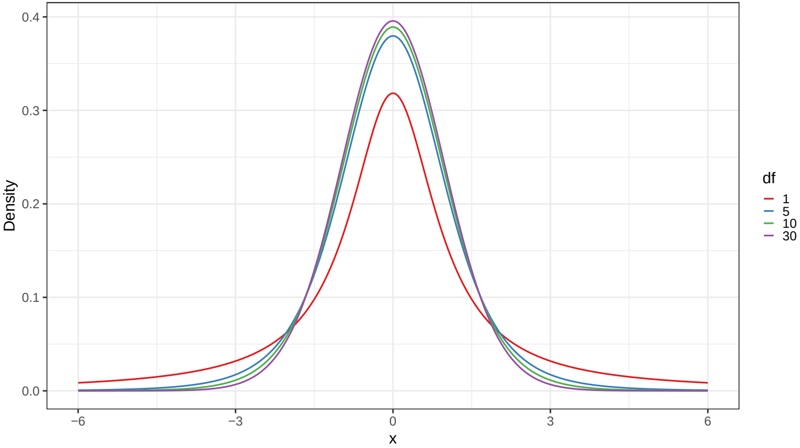
t-distribution:

- $$\frac{z}{\sqrt{\chi^2_\nu/\nu}} \sim t[\nu]$$
- If $t \sim t[\nu]$, then $t^2 \sim F[1, \nu]$

Chi-Square Densities across Degrees of Freedom



t Densities across Degrees of Freedom



Review: Restrictions & How We Test Them

What do we mean by restrictions? (recap of Lecture 4)

Linear restrictions (OLS):

$$H_0 : R\beta = q, \qquad R \in \mathbb{R}^{J\times(K+1)},\ q \in \mathbb{R}^J,\ J = \text{ \# of restrictions.}$$

Examples:

- Zero restrictions (joint significance): $\beta_2 = \cdots = \beta_K = 0$
- Equality restrictions: $\beta_1 = \beta_3$ or $\beta_2 + \beta_3 = 1$

Nonlinear restrictions (MLE world):

$$H_0 : c(\theta) = 0 \quad \text{with } c : \mathbb{R}^p \to \mathbb{R}^J.$$

Examples: odds-ratio equalities, elasticities at a point, variance components equal, or $p = 0.5$ in a Bernoulli model.

Why this review now?

In MLE we assess H_0 with likelihood-based tests (Wald/LR/Score). They generalize the OLS t/F logic you saw in Lecture 4.

Unrestricted OLS: $\hat{\beta}_{UR} = (X'X)^{-1}X'y$, residuals e_{UR}, $\text{SSR}_{UR} = e'_{UR}e_{UR}$.
Restricted OLS: impose $R\beta = q$,

$$\hat{\beta}_R = \hat{\beta}_{UR} - (X'X)^{-1}R'\left[R(X'X)^{-1}R'\right]^{-1}(R\hat{\beta}_{UR} - q),$$

residuals e_R, $\text{SSR}_R = e'_R e_R$.

Loss of fit & the F test (exact in classical homoskedastic normal model):

$$F \;=\; \frac{(\text{SSR}_R - \text{SSR}_{UR})/J}{\text{SSR}_{UR}/(n-K)} \;\sim\; F[J, n-K].$$

Intuition

If H_0 is true, enforcing the restriction barely worsens fit $\Rightarrow$ small loss of fit $\Rightarrow$ small F.

Examples of $c(\theta)$ in Practice

Reminder: We test restrictions of the form

$$H_0 : c(\theta) = 0, \quad c : \mathbb{R}^p \to \mathbb{R}^J.$$

Typical examples across models:

Model / Context	Restriction $c(\theta)$ and Interpretation
Linear Model	$c(\theta) = R\theta - q$: e.g. $R = [0, 1, -1]$, $q = 0$ tests $\beta_2 = \beta_3$.
Bernoulli	$c(p) = p - 0.5$: tests fairness of a coin ($p = 0.5$).
Cobb–Douglas	$c(\theta) = \alpha + \beta - 1$: constant returns to scale.
Elasticity restriction	$c(\theta) = x_0'\beta - 1$: elasticity at x_0 equals 1.
Nonlinear example	$c(\theta) = \theta_1\theta_2 - 1$: product of parameters equals 1.

Key idea

$c(\theta)$ can express any relationship among parameters — from simple linear equalities to nonlinear or cross-equation constraints.

Three Likelihood-Based Tests

Setup: Test $H_0 : c(\theta) = 0$.

- **Wald test:** Estimate model without restriction. Check if $c(\hat{\theta})$ is "far" from zero given its variance.
- **Likelihood Ratio test:** Compare log-likelihoods with and without restriction.
$$-2\left(\ell(\hat{\theta}_R) - \ell(\hat{\theta}_U)\right) \quad \rightarrow \quad \chi^2_J$$
- **Score (LM) test:** Estimate model under restriction. Test if slope of log-likelihood at $\hat{\theta}_R$ is near zero.

What are restrictions?

Think of restrictions similar to our F-Test example in the OLS lectures. Typically we use some linear restrictions on estimated parameters and specify them using the likelihood, the score or the variance.

Goal: Approximate the nonlinear restriction $c(\theta)$ near the true parameter θ_0.

When $c(\cdot)$ is differentiable, a first-order Taylor expansion gives:

$$c(\hat{\theta}) \approx c(\theta_0) + \frac{\partial c(\theta_0)}{\partial \theta'}(\hat{\theta} - \theta_0) = c(\theta_0) + G(\theta_0)(\hat{\theta} - \theta_0),$$

where $G(\theta_0)$ is the Jacobian of $c(\theta)$ at θ_0.
Under $H_0 : c(\theta_0) = 0$, the approximation simplifies to:

$$c(\hat{\theta}) \approx G(\theta_0)(\hat{\theta} - \theta_0).$$

Intuition

If $\hat{\theta}$ is close to θ_0, $c(\hat{\theta})$ changes almost linearly with $\hat{\theta}$. The matrix $G(\theta_0)$ maps parameter uncertainty into uncertainty about the restrictions.

Note: By MLE invariance, $c(\hat{\theta})$ is the MLE of $c(\theta)$, so it inherits asymptotic normality from $\hat{\theta}$. This justifies the next step on the next slide.

Step 1: Use Asymptotic Normality of the MLE

Under standard regularity conditions, the MLE satisfies

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, V_\theta),$$

where $V_\theta = I(\theta_0)^{-1}$ is the asymptotic covariance matrix given by the inverse of the Fisher information.

Interpretation:

$\hat{\theta}$ is approximately normal around θ_0 with sampling variance V_θ/n.

Linearization:

$$c(\hat{\theta}) \approx c(\theta_0) + G(\theta_0)(\hat{\theta} - \theta_0), \quad \text{where } G(\theta_0) = \frac{\partial c(\theta_0)}{\partial \theta'}.$$

Under $H_0 : c(\theta_0) = 0$, this simplifies to

$$c(\hat{\theta}) \approx G(\theta_0)(\hat{\theta} - \theta_0).$$

Multiply by $\sqrt{n}$:

$$\sqrt{n}\, c(\hat{\theta}) \approx G(\theta_0)\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}\big(0,\ G(\theta_0)V_\theta G(\theta_0)'\big).$$

*(This is an application of the **delta method**. Or equivalently, note that a linear transformation of a multivariate normal vector is itself normal: if $X \sim \mathcal{N}(0, V)$ and A is a matrix, then $AX \sim \mathcal{N}(0,\ AVA')$. The covariance matrix is premultiplied and postmultiplied by the transformation matrix.)*

Step 3: From Linearized Restrictions to the Wald Test

From Step 2:

$$\sqrt{n}\, c(\hat{\theta}) \xrightarrow{d} \mathcal{N}\big(0,\, \Sigma_c\big), \quad \Sigma_c = G(\theta_0)V_\theta G(\theta_0)'.$$

Idea: To test $H_0 : c(\theta_0) = 0$, we measure how far the estimated restrictions $c(\hat{\theta})$ are from zero in standard deviation units.

Define the standardized quadratic form:

$$W_n = n\, c(\hat{\theta})'\, \Sigma_c^{-1}\, c(\hat{\theta}).$$

Wald statistic:

$$W_n = n\, c(\hat{\theta})'\, [G(\theta_0)\, \widehat{V}_\theta\, G(\theta_0)']^{-1} c(\hat{\theta}) \xrightarrow{d} \chi^2_J.$$

Why is this χ^2 dsitributed?

If $Z \sim \mathcal{N}(0, I_J)$, then $Z'Z \sim \chi^2_J$.
Here, $Z = \Sigma_c^{-1/2}\sqrt{n}\, c(\hat{\theta})$ is asymptotically standard normal, so its quadratic form is asymptotically χ^2_J.

$$W = nc(\hat{\theta})' \left[G(\hat{\theta})\, \widehat{V}_\theta\, G(\hat{\theta})'\right]^{-1} c(\hat{\theta}) \xrightarrow{d} \chi^2_J$$

Decomposition:

- $c(\hat{\theta})$ = *Estimated restriction*: how far the fitted model is from satisfying H_0.
- $G(\hat{\theta})\, \widehat{V}_\theta\, G(\hat{\theta})'$ = *Sampling variance of* $c(\hat{\theta})$ from the Delta method.
- Quadratic form = *Standardizes* the restriction by its variance and sums across J restrictions.

Special case: for one restriction ($J = 1$),

$$W = \frac{[c(\hat{\theta})]^2}{\widehat{\mathbf{var}}[c(\hat{\theta})]} = z^2.$$

Key idea

The Wald test is a multivariate and non-linear generalization of the familiar *"estimate divided by SE"* logic.

(*Note:* The n appears in the general form because $\widehat{V}_\theta$ is the variance of $\sqrt{n}(\hat{\theta} - \theta_0)$. In the $J = 1$ case, this scaling is already built into $\widehat{\mathbf{var}}[c(\hat{\theta})]$, so no explicit n is needed.)

In general, the Wald test is based on nonlinear restrictions $c(\theta) = 0$.
After linearization:

$$c(\hat{\theta}) \approx G(\theta_0)(\hat{\theta} - \theta_0).$$

For **linear restrictions**, this approximation is exact:

$$c(\theta) = R\theta - q \quad \Longrightarrow \quad G(\theta) = R, \quad c(\hat{\theta}) = R\hat{\theta} - q.$$

Then the Wald statistic simplifies to

$$W = n(R\hat{\theta} - q)'[R\,\widehat{V}_\theta\,R']^{-1}(R\hat{\theta} - q).$$

Wald Test: Coin Example

$$\ell(p) = y \log p + (n - y) \log(1 - p).$$

Null:

$$H_0 : \; p = p_0 \qquad (J = 1 \text{ Restrictions})$$

Here p_0 denotes the hypothesized probability of success under H_0

1. **Unrestricted MLE:** $\hat{p} = y/n$.
2. **Variance:** $\widehat{\mathbf{var}}(\hat{p}) = \dfrac{\hat{p}(1 - \hat{p})}{n}$.
3. **Wald statistic:**

$$W \;=\; \frac{(\hat{p} - p_0)^2}{\widehat{\mathbf{var}}(\hat{p})} \;=\; \frac{n(\hat{p} - p_0)^2}{\hat{p}(1 - \hat{p})} \;\xrightarrow{d}\; \chi^2_1.$$

   Equivalent z-form: $z = \dfrac{\hat{p} - p_0}{\sqrt{\hat{p}(1 - \hat{p})/n}}$ with $W = z^2$.

4. **Decision:** Reject H_0 if $W > \chi^2_{1,1-\alpha}$.

Note: Wald evaluates the variance at the unrestricted estimate $\hat{p}$.

Wald Test: Coin Example (Numeric)

Data: $n = 10, y = 7 \Rightarrow \hat{p} = 0.7.$

Null hypothesis: $H_0 : p_0 = 0.5.$ (Our coin is fair)

Variance of MLE:

$$\widehat{\mathbf{var}}(\hat{p}) = \frac{\hat{p}(1-\hat{p})}{n} = \frac{0.7 \times 0.3}{10} = 0.021.$$

Wald statistic:

$$W = \frac{(\hat{p} - p_0)^2}{\widehat{\mathbf{var}}(\hat{p})} = \frac{(0.7 - 0.5)^2}{0.021} = 1.90.$$

Decision: Compare with $\chi^2_{1,0.95} = 3.84.$

$$W = 1.90 < 3.84 \quad \Rightarrow \quad \text{Fail to reject } H_0.$$

Intuition

With only 10 flips, we do not have enough evidence to reject fairness.

Guess What? You Already Know the Wald Test

From the OLS section: we test

$$H_0 : \; R\beta = q.$$

Using:

$$t_j = \frac{\hat{\beta}_j - 0}{\widehat{\text{se}}(\hat{\beta}_j)}, \qquad F = \frac{(R\hat{\beta} - q)'[R\widehat{\mathbf{var}}(\hat{\beta})R']^{-1}(R\hat{\beta} - q)}{J}.$$

In fact: these are Wald tests under the normal likelihood.

$$\underbrace{(R\hat{\beta} - q)'[R\widehat{\mathbf{var}}(\hat{\beta})R']^{-1}(R\hat{\beta} - q)}_{\text{Wald statistic } W} \sim \chi^2_J, \quad t^2 = W \text{ when } J = 1.$$

Takeaway

Your familiar t- and F-tests are special cases of the **general Wald test**. All we're doing now is extending this logic to any likelihood model.

Why the t- and F-Tests Are Wald Tests

Model: $y = X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I)$

Wald statistic under normal likelihood:

$$W = (R\hat{\beta} - q)'[R(\hat{\sigma}^2(X'X)^{-1})R']^{-1}(R\hat{\beta} - q) \sim \chi^2_J.$$

Textbook small sample adjustments:

$$s^2 = \frac{1}{n - (K+1)}(y - X\hat{\beta})'(y - X\hat{\beta})$$

leads to exact t/F distributions in finite samples.

Takeaway

As $n \to \infty$, $t^2 \to W$ and $F \cdot J \to W$. Same logic, different parameterization.

Background:

- Born in Cluj (then Austro-Hungarian Empire, now Romania) to a German-speaking Jewish family.
- Studied mathematics in Vienna; emigrated to the U.S. in 1938 and joined Columbia University.
- Developed the **Wald test**, **sequential analysis**, and fundamental results on **MLE efficiency**.
- His WWII work on aircraft damage led to the famous “missing bullet holes” example of selection bias.

Likelihood Ratio (LR): Principle

Goal: Test $H_0 : c(\theta) = 0$ (J restrictions).

MLEs: $\hat{\theta}_U$ (unrestricted), $\hat{\theta}_R$ (restricted).

Test statistic (sometimes named Wilks-Statistic):

$$LR \; = \; -2\Big(\ell(\hat{\theta}_R) - \ell(\hat{\theta}_U) \Big) \;\; \xrightarrow{d} \;\; \chi^2_J.$$

Interpretation: How much does imposing H_0 reduce the best attainable fit?

Key idea

If H_0 is true, the restricted optimum is close to the unrestricted one, so the log-likelihood drop is small (and LR is near 0).

Likelihood Ratio: Coin Example (Step 1)

Goal: Test $H_0 : p = p_0$ against $H_1 : p \neq p_0$.

Log-likelihood:

$$\ell(p) = y \log p + (n - y) \log(1 - p)$$

Unrestricted MLE:

$$\hat{p} = \frac{y}{n} \quad \Rightarrow \quad \ell(\hat{p}) = y \log \hat{p} + (n - y) \log(1 - \hat{p})$$

Restricted under H_0:

$$p = p_0 \quad \Rightarrow \quad \ell(p_0) = y \log p_0 + (n - y) \log(1 - p_0)$$

Test statistic:

$$LR = -2\big[\ell(p_0) - \ell(\hat{p})\big]$$

Likelihood Ratio: Coin Example (Step 2)

Expand:

$$\ell(\hat{p}) - \ell(p_0) = \left[y \log \hat{p} + (n-y)\log(1-\hat{p})\right] - \left[y \log p_0 + (n-y)\log(1-p_0)\right]$$

Group like terms:

$$\ell(\hat{p}) - \ell(p_0) = y(\log \hat{p} - \log p_0) + (n-y)\left[\log(1-\hat{p}) - \log(1-p_0)\right]$$

Simplify using log rules:

$$\ell(\hat{p}) - \ell(p_0) = y \log \frac{\hat{p}}{p_0} + (n-y) \log \frac{1-\hat{p}}{1-p_0}$$

Likelihood Ratio: Coin Example (Step 3)

Plug into the LR definition:

$$LR = 2\left[\ell(\hat{p}) - \ell(p_0)\right] = 2\left[y \log \frac{\hat{p}}{p_0} + (n-y)\log \frac{1-\hat{p}}{1-p_0}\right]$$

With $\hat{p} = y/n$:

$$LR = 2\left[y \log \frac{y/n}{p_0} + (n-y)\log \frac{1-y/n}{1-p_0}\right]$$

Asymptotic distribution:

$$LR \xrightarrow{d} \chi^2_1$$

Decision rule: Reject H_0 if $LR > \chi^2_{1,1-\alpha}$.

LR is invariant to reparameterizations; no variance estimator needed.

Likelihood Ratio Test: Numeric Example

Data: $n = 10, y = 7 \Rightarrow \hat{p} = 0.7$. **Null:** $H_0 : p_0 = 0.5$.

Compute:

$$LR = 2\left[y \log \frac{\hat{p}}{p_0} + (n - y) \log \frac{1 - \hat{p}}{1 - p_0}\right].$$

$$LR = 2\left[7 \log \frac{0.7}{0.5} + 3 \log \frac{0.3}{0.5}\right] = 2(7 \times 0.3365 + 3 \times (-0.511)) = 2(1.525) = 3.05.$$

Decision: Compare with $\chi^2_{1,0.95} = 3.84$.

$$LR = 3.05 < 3.84 \quad \Rightarrow \quad \text{Fail to reject } H_0.$$

Interpretation

Log-likelihood drops only slightly when $p = 0.5$ is imposed. The data are consistent with a fair coin at 5% level.

Historical Note: Samuel S. Wilks (1906–1964)

Background:

- American statistician, professor at Princeton University.
- Introduced the **Likelihood Ratio (LR) test** in the 1930s.
- Proved **Wilks' theorem:**

$$-2(\ell_{\text{restricted}} - \ell_{\text{unrestricted}}) \xrightarrow{d} \chi^2_J.$$

- This result underlies virtually all likelihood-based model comparison tests.
- the American Statistical Assoication named the Wilks Memorial Award in his honor.

Score (LM): Principle

Goal: Test $H_0 : c(\theta) = 0$ (J restrictions) using only the **restricted** fit.

Score and information at the restricted estimate $\hat{\theta}_R$:

$$S(\theta) = \frac{\partial \ell(\theta)}{\partial \theta}, \qquad I(\theta) = \mathrm{E}\left[-\frac{\partial^2 \ell(\theta)}{\partial \theta\, \partial \theta'}\right].$$

Test statistic:

$$LM \;=\; S(\hat{\theta}_R)'\, I(\hat{\theta}_R)^{-1}\, S(\hat{\theta}_R) \;\;\xrightarrow{d}\;\; \chi^2_J.$$

Key idea

If H_0 is true, the slope (score) at the restricted optimum should be near zero, once scaled by its information.

Score (LM): Coin Example

1. **Restricted point:** evaluate at p_0.
2. **Score at p_0:**

$$S(p_0) = \frac{\partial \ell(p)}{\partial p}\bigg|_{p_0} = \frac{y - np_0}{p_0(1-p_0)} = \frac{n(\hat{p} - p_0)}{p_0(1-p_0)}.$$

3. **Fisher information at p_0:**

$$I(p_0) = \frac{n}{p_0(1-p_0)}.$$

4. **LM statistic:**

$$LM = \frac{S(p_0)^2}{I(p_0)} = \frac{n(\hat{p} - p_0)^2}{p_0(1-p_0)} \xrightarrow{d} \chi^2_1.$$

5. **Decision:** Reject H_0 if $LM > \chi^2_{1,1-\alpha}$.

Note: LM uses only the restricted fit (no unrestricted optimization needed).

Score (LM) Test: Numeric Example

Data: $n = 10, y = 7 \Rightarrow \hat{p} = 0.7$. **Null:** $H_0 : p_0 = 0.5$.

Score at p_0:

$$S(p_0) = \frac{n(\hat{p} - p_0)}{p_0(1 - p_0)} = \frac{10(0.7 - 0.5)}{0.5 \times 0.5} = 8.$$

Fisher information at p_0:

$$I(p_0) = \frac{n}{p_0(1 - p_0)} = \frac{10}{0.25} = 40.$$

LM statistic:

$$LM = \frac{S(p_0)^2}{I(p_0)} = \frac{8^2}{40} = 1.6.$$

Decision: $LM = 1.6 < 3.84 \Rightarrow$ fail to reject H_0.

Takeaway

All three tests (Wald, LR, LM) lead to the same qualitative conclusion.

Historical Note: Calyampudi Radhakrishna Rao (1920–2023)

Background:

- Indian statistician and one of the most influential figures in 20th-century statistics.
- Developed the **Score (Rao)** test, later known as the **Lagrange Multiplier (LM)** test.
- Based on the **score function:**

$$s(\theta) = \frac{\partial \ell(\theta)}{\partial \theta}.$$

- Rao showed that testing H_0 can rely on how large the score is at the restricted MLE:

$$LM = s(\hat{\theta}_0)' I(\hat{\theta}_0)^{-1} s(\hat{\theta}_0) \xrightarrow{d} \chi^2_J.$$

5.4: OLS as Maximum Likelihood Problem

Linear Normal Model & Likelihood Setup

Model: $y = X\beta + \epsilon$, with $\epsilon \mid X \sim \mathcal{N}(0, \sigma^2 I_n)$.

Parameters: $\theta = (\beta, \sigma^2) \quad (\beta \in \mathbb{R}^{K+1}, \sigma^2 > 0)$.

Joint density (conditional on X):

$$f(y \mid X; \theta) = (2\pi\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta) \right\}.$$

Log-likelihood:

$$\ell(\beta, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta).$$

Goal

Maximize $\ell(\beta, \sigma^2)$ w.r.t. (β, σ^2).

Our OLS Assumptions are hiding here in plain sight.
Can you spot them?

FOC for β: OLS Normal Equations

Score for β:

$$\frac{\partial \ell}{\partial \beta} = -\frac{1}{2\sigma^2} \cdot (-2X'(y - X\beta)) = \frac{1}{\sigma^2} X'(y - X\beta).$$

Set to zero:

$$X'(y - X\hat{\beta}) = 0 \quad \Longleftrightarrow \quad X'X\,\hat{\beta} = X'y \quad \Longrightarrow \quad \hat{\beta} = (X'X)^{-1}X'y.$$

- This is **exactly** the OLS estimator.
- Requires full column rank: $\text{rank}(X) = K + 1$ so $(X'X)^{-1}$ exists.

FOC for σ^2: MLE of the Error Variance

Score for σ^2:

$$\frac{\partial \ell}{\partial \sigma^2} = -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2(\sigma^2)^2}(y - X\beta)'(y - X\beta).$$

Set to zero and plug in *b*:

$$\hat{\sigma}^2_{\text{MLE}} = \frac{1}{n}(y - X\hat{\beta})'(y - X\hat{\beta}) = \frac{1}{n}\sum_{i=1}^{n} e_i^2.$$

- Note the **denominator is** n (finite-sample MLE). The usual OLS unbiased estimator uses $\frac{1}{n-(K+1)}\sum e_i^2$. We did not correct for $\hat{\beta}$ being estimated!

Concentrated Likelihood & Equivalence

Concentrate out σ^2:

$$\tilde{\ell}(\beta) = \ell\big(\beta,\ \hat{\sigma}^2(\beta)\big) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\left(\frac{(y - X\beta)'(y - X\beta)}{n}\right) - \frac{n}{2}.$$

Maximizing $\tilde{\ell}(\beta)$ is equivalent to minimizing

$$S(\beta) = (y - X\beta)'(y - X\beta) \quad \Rightarrow \quad \text{OLS normal equations.}$$

Takeaway

Under normality and linearity, **MLE for β equals OLS**. The difference shows up only in the small-sample estimator of σ^2.

When MLE $\neq$ OLS (and What Replaces It)

If errors are **non-spherical**:

$$\varepsilon \mid X \sim \mathcal{N}(0,\ \sigma^2\Omega), \quad \Omega \neq I_n,$$

then the log-likelihood maximizer for β is

$$b_{\text{GLS}} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y,$$

not OLS. Special case with known diagonal Ω gives **WLS**.

Takeaway:

OLS is the MLE if $\varepsilon \mid X$ is $\mathcal{N}(0, \sigma^2 I_n)$. With heteroskedasticity or autocorrelation, **GLS** is the MLE analogue.

MLE, OLS, and GLS: Connecting the Dots

Implications for inference:

- Wald-type tests remain valid once the covariance is replaced by $\widehat{Var}(\hat{\beta}_{GLS}) = \hat{\sigma}^2(X'\Omega^{-1}X)^{-1}$.
- With unknown Ω, use an estimate $\hat{\Omega}$
  $\longrightarrow$ **Feasible GLS (FGLS)**.
- Alternatively, use **robust (sandwich)** standard errors for OLS if efficiency is less important.

Takeaway:

Non-spherical errors do not invalidate the likelihood framework: They simply change the MLE from OLS to GLS.

Why Not Just Use Robust SEs Instead of GLS?

Valid point:
If the goal is inference on β, OLS with robust (sandwich) SEs already works:

$$\widehat{Var}(\hat{\beta}_{OLS}) = (X'X)^{-1}X'\hat{\Omega}X(X'X)^{-1}.$$

$\longrightarrow$ **Correct inference, same point estimate.**

Then why study GLS?

- GLS is the **MLE analogue** under non-spherical errors.
- If Ω is correctly specified, GLS (or FGLS) has **lower sampling variability**:
$$Var(\hat{\beta}_{GLS}) \preceq Var(\hat{\beta}_{OLS}).$$
- Smaller sampling variance $\longrightarrow$ **tighter confidence intervals and higher power.**
- Also improves **predictions and fitted values** when Ω captures real dependence.

Wald Tests and Robust Covariances

$$W = (R\hat{\beta} - q)'[R\,\widehat{\text{Var}}(\hat{\beta})\,R']^{-1}(R\hat{\beta} - q) \xrightarrow{d} \chi^2_J$$

Robust covariance estimate (Huber–Eicker–White)

$$\widehat{\text{Var}}(\hat{\beta}) = (X'X)^{-1}X'\widehat{\Omega}X(X'X)^{-1}, \quad \widehat{\Omega} = \text{diag}(\hat{e}_i^2)$$

- Same Wald logic as before, but robust to heteroskedasticity.
- Classical t and F tests are finite–sample Wald tests under normal, homoskedastic errors.

Key idea

Inference = Wald + consistent variance estimate.

Unified View: OLS, GLS, and Wald Testing

Assumptions on errors	Estimator	Covariance matrix	Wald test uses
Normal, spherical	OLS (MLE)	$\sigma^2(X'X)^{-1}$	t, F tests
Non-spherical, known Ω	GLS (MLE)	$\sigma^2(X'\Omega^{-1}X)^{-1}$	General Wald
Heteroskedastic, unknown form	OLS + HEW SEs	$(X'X)^{-1}X'\hat{\Omega}X(X'X)^{-1}$	Robust Wald

Takeaway

OLS, GLS, and robust regression are all **MLE-inspired**.
Inference is unified through the **Wald principle**.