

Lecture 13: Fisher information and the Cramér–Rao bound

Lecturer: Yiqiao Zhong[Report an issue](#)*Last update: October 19, 2026*

These notes support STAT 709 at the University of Wisconsin–Madison. Use them alongside the cited textbooks; corrections and questions are welcome.

Priority guide

High Core concepts and calculations to master; directly examined.

Mid Supporting concepts useful in other lectures; in scope for examinations.

Low Enrichment supporting understanding; will not be examined.

Further reading Optional self-study, references, extensions, and historical notes.

A star (*) marks optional material; priority labels give the scope of each block. Lecture-specific exemptions and the optional status of Part VI override the general exam guide.

High priority *Overview*

The score measures how a likelihood responds to a parameter change. Under the usual regularity conditions, its covariance is **Fisher information**; the inverse information bounds the covariance of unbiased estimators of the parameter. We prove the bound, examine equality and failure of equality, and connect it to Gaussian regression. Geometric curvature supports the main route; the KL interpretation and Gauss–Markov extension are optional.

Recap from earlier lectures

From Lectures 3, 4, 8, 11 and 12: likelihood and optimality

Reminder only: this block introduces no new exam requirements.

- **UMVUE and completeness.** An unbiased function of a complete sufficient statistic is the unique UMVUE, up to null sets. Used in Section 4.
- **Exponential families.** A natural-parameter space containing an open set supplies the completeness result under its stated assumptions. Used in Section 5.2.
- **Likelihood asymptotics.** Under regularity, the score fluctuation and Hessian determine the MLE limiting covariance. Used in Section 2 and Section 4.
- **Differentiation under integration.** An integrable bound on parameter derivatives can justify differentiating the normalization integral. Used in Section 1.
- **Cauchy–Schwarz.** The covariance inequality is the elementary engine of the information bound. Used in Section 2.

Learning goals

- Compute a score and Fisher information with regularity assumptions stated.
- Derive and interpret a Cramér–Rao bound.
- Check whether an unbiased estimator attains the bound.

- Use projection geometry and Gaussian linear transformations in a linear model.

High priority Introduction

In this lecture, we will introduce an important lower-bound result for unbiased estimation.

1 Fisher information

High priority Fisher information

Suppose that X is drawn (e.g., single observation or sequence of i.i.d. data) from $P_\theta \in \mathcal{P} = \{P_\theta : \theta \in \Theta\}$, where the parameter space $\Theta \subset \mathbb{R}^k$ is an open set. Also suppose that P_θ has p.d.f. f_θ w.r.t. a dominating measure μ on the observation space for any $\theta \in \Theta$, and f_θ is differentiable in θ .

Definition 1 (Fisher information). *For the parametric family $\{P_\theta : \theta \in \Theta\}$, the Fisher information (matrix) is defined as*

$$I(\theta) = \mathbb{E} \left[\frac{\partial}{\partial \theta} \log f_\theta(X) \left(\frac{\partial}{\partial \theta} \log f_\theta(X) \right)^\top \right],$$

where the expectation is taken over P_θ . We also write $I_X(\theta)$ to emphasize the dependence on X should confusion arise.

Note that by definition, $I(\theta)$ is always positive semidefinite. We call $\frac{\partial}{\partial \theta} \log f_\theta(X)$ the *score function*.

Proposition 1. (i) *If X and Y are independent and their scores have mean zero and finite second moments, then the Fisher information about θ in (X, Y) is*

$$I_{(X,Y)}(\theta) = I_X(\theta) + I_Y(\theta).$$

If $X_1, \dots, X_n$ are i.i.d., then

$$I_{(X_1, \dots, X_n)}(\theta) = nI_{X_1}(\theta)$$

(ii) *Suppose that f_θ is twice differentiable in θ . Under a “regularity condition”,*

$$I(\theta) = -\mathbb{E} \left[\frac{\partial^2}{\partial \theta \partial \theta^\top} \log f_\theta(X) \right].$$

Proof. (i) The joint log density is the sum of the two log densities, so its score is $s_X + s_Y$. Independence and $\mathbb{E} s_X = \mathbb{E} s_Y = 0$ kill both cross terms in $\mathbb{E}(s_X + s_Y)(s_X + s_Y)^\top$, leaving $I_X + I_Y$.

(ii) A useful observation is the following identity.

$$\begin{aligned} \frac{\partial^2}{\partial \theta \partial \theta^\top} \log f_\theta(X) &= \frac{\frac{\partial^2}{\partial \theta \partial \theta^\top} f_\theta(X)}{f_\theta(X)} \\ &\quad - \frac{\partial}{\partial \theta} \log f_\theta(X) \left[\frac{\partial}{\partial \theta} \log f_\theta(X) \right]^\top. \end{aligned}$$

When taking expectation, we get

$$\begin{aligned}
\mathbb{E} \left[\frac{\frac{\partial^2}{\partial \theta \partial \theta^\top} f_\theta(X)}{f_\theta(X)} \right] &= \int \frac{\frac{\partial^2}{\partial \theta \partial \theta^\top} f_\theta(x)}{f_\theta(x)} f_\theta(x) d\mu(x) \\
&= \int \frac{\partial^2}{\partial \theta \partial \theta^\top} f_\theta(x) d\mu(x) \\
&= \frac{\partial^2}{\partial \theta \partial \theta^\top} \int f_\theta(x) d\mu(x) \quad (\text{regularity condition}) \\
&= 0. \quad (\text{since } \int f_\theta(x) d\mu(x) \equiv 1)
\end{aligned}$$

□

The regularity condition is an annoying technicality that ensures the exchangeability of integrals with differentials (again, physicists don't care). A useful sufficient condition is a parameter-independent support and locally integrable envelopes for the first two parameter derivatives. Smoothness alone, or compact support whose boundary moves with the parameter, does not suffice. Regular full exponential families provide important examples.

What “Information” means. Fisher information shares two familiar properties with quantities such as discrete entropy and KL divergence: it is **nonnegative** (positive semidefinite in the matrix case), and it is **additive for independent observations** under the stated mean-zero-score conditions. The additivity follows because independent centered scores have zero cross covariance.

An important difference is its **dependence on parametrization**. Changing the scale of the parameter changes the score and therefore the numerical value of Fisher information. For a scalar change $\theta = \psi(\eta)$,

$$I_\eta(\eta) = [\psi'(\eta)]^2 I_\theta(\psi(\eta)).$$

See Remark 1 for the matrix version and why the variance bound for a fixed target is unchanged.

For comparison, discrete entropy is unchanged by a one-to-one relabeling of outcomes, and KL divergence is unchanged when the same invertible transformation is applied to both distributions. These are changes of the *observations*, whereas the formula above changes the *parameter*. Fisher information for a fixed parameter is also unchanged by an invertible, parameter-independent transformation of the observations. The entropy comparison here concerns discrete entropy; differential entropy for continuous variables is not generally invariant. See the change-of-variables discussion and Fisher-information notes.

High priority *Practice: Information in Bernoulli observations*

Exercise prompt. For n independent Bernoulli(p) observations, $0 < p < 1$, compute the score and information for p .

See Exercise 1 in Section 8 for the full statement and complete solution.

2 Cramér–Rao lower bound

High priority *Cramér–Rao lower bound*

Recall that a symmetric matrix A is positive definite if $u^\top A u > 0$ for all nonzero vectors u , and it is positive semidefinite if $u^\top A u \geq 0$ for all vectors u . For symmetric matrices A and B of the same size, both

$A \geq B$ and $A \succeq B$ will mean that $A - B$ is **positive semidefinite**, equivalently $u^\top (A - B)u \geq 0$ for every vector u . This is the matrix order, not an entrywise comparison. We use $\succeq$ in Theorem 1 to make that meaning explicit.

Let $\nu = g(\theta)$ be the quantity of interest where $g : \mathbb{R}^d \rightarrow \mathbb{R}^k$ is differentiable; in this section $\theta \in \mathbb{R}^d$ and $Dg(\theta)$ is the $k \times d$ Jacobian. Assume T has finite covariance. Suppose that X is drawn from a distribution in $\{P_\theta : \theta \in \Theta\}$ and $T(X)$ is an unbiased estimator of ν . Let f_θ be the density function with respect to a reference measure μ .

Theorem 1. *Suppose that $I(\theta)$ is positive definite, and that for any $\theta \in \Theta$,*

$$\frac{\partial}{\partial \theta} \int h(x) f_\theta(x) d\mu(x) = \int h(x) \frac{\partial}{\partial \theta} f_\theta(x) d\mu(x), \quad (1)$$

holds for $h(x) \equiv 1$ and $h(x) = T(x)$ (regularity conditions). Then,

$$\text{Cov}(T(X)) \succeq Dg(\theta)I(\theta)^{-1}Dg(\theta)^\top.$$

Fisher information provides a lower bound on variance (and thus MSE) of unbiased estimators for a parametric family. It may not always be achievable, but if we can find an estimator that achieves this lower bound, then we certify the optimality of this estimator in the class of unbiased estimators (namely certifying UMVUE). See wiki for the multivariate case.

Proof for only $k = 1$ and $d = 1$; similar for $k > 1$. Our goal is

$$\text{Var}(T(X)) \mathbb{E} \left[\left(\frac{\partial}{\partial \theta} \log f_\theta(X) \right)^2 \right] \geq (g'(\theta))^2. \quad (2)$$

It suffices to show the following two equalities.

$$\begin{aligned} \mathbb{E} \left[\left(\frac{\partial}{\partial \theta} \log f_\theta(X) \right)^2 \right] &= \text{Var} \left(\frac{\partial}{\partial \theta} \log f_\theta(X) \right), \\ g'(\theta) &= \text{Cov} \left(T(X), \frac{\partial}{\partial \theta} \log f_\theta(X) \right). \end{aligned}$$

Once they are proved, Equation 2 follows from Cauchy–Schwarz inequality:

$$\text{Var}(\xi_1)\text{Var}(\xi_2) \geq [\text{Cov}(\xi_1, \xi_2)]^2.$$

Note that

$$\begin{aligned} \mathbb{E} \left[\frac{\partial}{\partial \theta} \log f_\theta(X) \right] &= \int \frac{\frac{\partial}{\partial \theta} f_\theta(x)}{f_\theta(x)} f_\theta(x) d\mu(x) \\ &= \int \frac{\partial}{\partial \theta} f_\theta(x) d\mu(x) \\ &= \frac{\partial}{\partial \theta} \underbrace{\int f_\theta(x) d\mu(x)}_{\equiv 1} = 0, \end{aligned}$$

Also,

$$\begin{aligned}
\mathbb{E} \left[T(X) \frac{\partial}{\partial \theta} \log f_{\theta}(X) \right] &= \int T(x) \frac{\frac{\partial}{\partial \theta} f_{\theta}(x)}{f_{\theta}(x)} f_{\theta}(x) d\mu(x) \\
&= \int T(x) \frac{\partial}{\partial \theta} f_{\theta}(x) d\mu(x) \\
&= \frac{\partial}{\partial \theta} \int T(x) f_{\theta}(x) d\mu(x) \\
&= \frac{\partial}{\partial \theta} g(\theta) \quad (\text{by unbiasedness}) \\
&= g'(\theta).
\end{aligned}$$

□

Again, the differentiation identities require a common support and suitable local domination of the derivatives; decay of the density alone does not control its parameter derivatives.

Remark 1 (Reparameterization). *Let $\theta = \psi(\eta)$ be a differentiable change of coordinates with invertible Jacobian $J = D\psi(\eta)$. The chain rule for column scores gives $s_{\eta} = J^{\top} s_{\theta}$, hence*

$$\begin{aligned}
I_{\eta}(\eta) &= J^{\top} I_{\theta}(\psi(\eta)) J, \\
D_{\eta}(g \circ \psi) &= D_{\theta} g J.
\end{aligned}$$

Substitution into the bound cancels J and J^{-1} . The information matrix changes with coordinates, but the lower bound for the same target does not: we cannot cheat the lower bound by reparameterization.

Remark 2 (Asymptotic optimality of MLE). *Under the regularity and consistency conditions discussed in Lecture 11, the MLE has asymptotic variance $(nI_{X_1}(\theta))^{-1}$, which matches the lower bound asymptotically. This is a key advantage of MLE.*

The unbiased variance bound. The Cramér–Rao inequality compares the derivative of the target with information. Equality requires a centered estimator to be a linear transformation of the score, so the bound need not be attainable.

Used again in the optional Stein phenomenon: going beyond the unbiased class.

High priority *Practice: A bound attained by the Poisson mean*

Exercise prompt. For i.i.d. $\text{Poisson}(\theta)$ data, show that the sample mean attains the Cramér–Rao bound for $\theta > 0$.

See Exercise 2 in Section 8 for the full statement and complete solution.

3 Two interpretations of Fisher information

3.1 Geometric view of Fisher information

Mid priority *Geometric view of Fisher information*

$\log f_{\theta}(x)$ is also known as the log-likelihood (or written as loglik) function. A natural loss that we want to minimize is the following negative loglik.

$$\ell(\theta; X) = -\log f_{\theta}(X).$$



Figure 1: Small $I(\theta)$

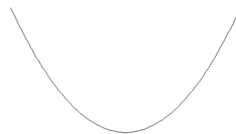


Figure 2: Large $I(\theta)$

If we have i.i.d. data $X = (X_1, \dots, X_n)$, then

$$\ell(\theta; X) = - \sum_{i=1}^n \log f_{\theta}(X_i).$$

Note that Fisher information is related to the expected curvature (in one dimension, the rate of change of tangent vector) of $\ell(\theta; X)$, since

$$I(\theta) = \mathbb{E} \left[\frac{\partial^2}{\partial \theta \partial \theta^{\top}} \ell(\theta; X) \right].$$

Recall that in MLE, we estimate θ by minimizing the negative log-likelihood $\ell(\theta; X)$. We provide some geometric heuristics. An obvious property of MLE is that an interior differentiable minimum corresponds to a stationary point of $\ell(\theta; X)$, which means

$$\frac{\partial}{\partial \theta} \ell(\hat{\theta}; X) = 0,$$

Intuitively, consider the one-dimensional case for simplicity, a large $I(\theta)$ means larger local convexity, so noise perturbation is easier to control. Figures 1 and 2 compare small and large $I(\theta)$. The following is an intuitive understanding of Fisher information.

$$\begin{aligned} \text{large } I(\theta) &\Rightarrow \text{easy to find } \theta \\ &\Rightarrow \text{less uncertainty / smaller lower bound.} \end{aligned}$$

In the Normal mean model, the negative log-likelihood above its minimum is $I(\mu)(\mu - \bar{X})^2/2$, where $I(\mu) = n/\sigma^2$. Quadrupling information halves the horizontal width at a fixed likelihood drop. Compare Lecture 8's likelihood-width example.

Example 1. Let $X_1, X_2, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$, where $\sigma^2 > 0$ is known.

$$\frac{\partial}{\partial \mu} \log f_{\mu}(X) = \sum_{i=1}^n (X_i - \mu)/\sigma^2 \implies I(\mu) = \frac{n}{\sigma^2}.$$

We know $\bar{X}$ is a UMVUE. It is easy to check

$$\text{Var}(\bar{X}) = [I(\mu)]^{-1},$$

attaining the Cramér–Rao lower bound.

Remark 3. $\sigma/\sqrt{n}$ is the “effective noise level”. A large sample size n reduces uncertainty.

Instructor’s Note *Curvature and uncertainty*

Imagine locating the bottom of a bowl after a small tilt. A narrow bowl moves its minimizer only a little; a shallow bowl moves it farther. Expected likelihood curvature plays that role locally. The drawings preserve this intuition, while the score covariance is the precise definition.

3.2 Information theoretic interpretation

Low priority *Information theoretic interpretation*

Let us recall the information notations.

- Mutual information: Defined generally for random variables, unchanged under one-to-one mapping.
- Fisher information: Defined for specific parametric models, changed under reparameterization.

Both information quantities measure the amount of information about θ given observation X . Here we give a concrete connection: Under certain “regularity conditions”,

$$D(P_{\theta_0} || P_{\theta_0 + \xi}) = \frac{1}{2} \xi^\top I(\theta_0) \xi + o(\|\xi\|^2); \quad (3)$$

By Taylor expansion at θ_0 , with all parameter derivatives below evaluated at $\theta = \theta_0$,

$$\begin{aligned} \log f_{\theta_0 + \xi}(X) &= \log f_{\theta_0}(X) + \xi^\top \nabla_\theta \log f_\theta(X) \\ &\quad + \frac{1}{2} \xi^\top \nabla_\theta^2 (\log f_\theta(X)) \xi + o(\|\xi\|^2), \end{aligned}$$

Thus, by the definition of KL divergence and the fact $\mathbb{E}_{\theta_0}[\nabla_\theta \log f_\theta(X)] = 0$, we obtain Equation 3.

Thus, Fisher information characterizes how hard it is to distinguish two distributions in a parametric family under the KL divergence. To draw an analogy, Fisher information $I(\theta_0)$ in Equation 3 plays a similar role as local convexity plays in optimization.

4 Cramér–Rao lower bound may not be attained

Mid priority *Cramér–Rao lower bound may not be attained*

Let $X_1, X_2, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$, where $\sigma^2 > 0$ is known, and we are interested in estimating $\nu = g(\mu) = \mu^2$. The following calculation verifies

- (1) $h(\bar{X}) = \bar{X}^2 - \frac{\sigma^2}{n}$ is UMVUE.
- (2) $\text{Var}(h(\bar{X})) = \frac{4\mu^2\sigma^2}{n} + \frac{2\sigma^4}{n^2}$.
- (3) Cramér–Rao lower bound for ν is $\frac{4\mu^2\sigma^2}{n}$.
- (4) If, however, we are interested in estimating μ , then as in Example 1, both $\text{Var}(\bar{X})$ and the Cramér–Rao lower bound are σ^2/n .

Write $\bar{X} = \mu + \tau G$, where $G \sim N(0, 1)$ and $\tau^2 = \sigma^2/n$. Then

$$h(\bar{X}) - \mu^2 = 2\mu\tau G + \tau^2(G^2 - 1).$$

The two terms are uncorrelated because $\mathbb{E} G^3 = \mathbb{E} G = 0$, and $\text{Var}(G^2) = 2$. This gives the stated variance. The rule is unbiased and is a function of the complete sufficient statistic $\bar{X}$, so Lehmann–Scheffé makes it the UMVUE. The information is n/σ^2 and $g'(\mu) = 2\mu$, giving the displayed lower bound. Note that $h(\bar{X})$ matches the leading order term but not the exact lower bound. Why?

- We know that $\bar{X}$ is the MLE for estimating μ with optimal asymptotic variance $V = \sigma^2/n$. By the delta method, the first-order asymptotic variance of $g(\bar{X})$ is $[g'(\mu)]^2 V$ —this expression is the Cramér–Rao variance bound for the transformed target $g(\mu)$ (and equals inverse information when g is a valid local reparameterization), so the asymptotic variance matches the lower bound. However, the asymptotic variance does not need to equal the actual variance of $h(\bar{X})$; nevertheless, it matches the leading-order term in this case.

- The reason for the gap between the variance and the lower bound is that we have a strict inequality in Cauchy–Schwarz inequality in Equation 2. Recall that the equality condition for Cauchy–Schwarz inequality is $T(X) - g(\theta) = c_\theta \frac{\partial}{\partial \theta} \log f_\theta(X)$ a.s. for some constant c_θ . This does not hold for $T(X) = h(\bar{X})$.

Mid priority *Practice: An exponential Normal target*

Exercise prompt. For i.i.d. $N(\mu, 1)$ observations, consider $H = \exp(\bar{X} - 1/(2n))$. Verify unbiasedness for e^μ , compute its variance, and compare with the Cramér–Rao bound.

See Exercise 3 in Section 8 for the full statement and complete solution.

5 Examples of Cramér–Rao lower bound

5.1 Exponential family

Mid priority *Exponential family*

Let $\Theta \subset \mathbb{R}^k$ be an open set. Let $\{f_\theta : \theta \in \Theta\}$ be the exponential family with p.d.f.

$$f_\theta(x) = \exp\{\eta(\theta)^\top T(x) - \xi(\theta)\}c(x).$$

Suppose that X is drawn from a distribution P_θ with corresponding p.d.f. f_θ in this exponential family.

Proposition 2. (i) *Regularity conditions in Equation 1 are satisfied for any measurable h with*

$$\mathbb{E}_\eta |h(X)| < \infty$$

throughout an open neighborhood of the parameter.

(ii) *Consider the natural parameter η . Then*

$$\text{Var}(T(X)) = I(\eta).$$

(iii) *Suppose $I(\eta)$ is positive definite, so the mean parameter is locally invertible, and $\nu = \mathbb{E}[T(X)]$ is the parameter of interest, and $\bar{I}(\nu)$ denotes the Fisher information. Then*

$$\text{Var}(T) = [\bar{I}(\nu)]^{-1}.$$

Proof. Under natural parameters η , we write the exponential family as

$$f_\eta(X) = \exp\{\eta^\top T(X) - \xi(\eta)\}c(X).$$

(i) + (ii) proved in Homework 4. In addition, we proved that $f_\eta(x)$ is C^∞ . Also, from Homework 4,

$$\nu = \frac{\partial}{\partial \eta} \xi(\eta).$$

By reparameterization,

$$I(\eta) = \frac{\partial \nu}{\partial \eta} \bar{I}(\nu) \frac{\partial \nu}{\partial \eta}^\top.$$

Note that

$$\frac{\partial \nu}{\partial \eta} = \frac{\partial^2}{\partial \eta \partial \eta^\top} \xi(\eta) = \text{Var}(T) = I(\eta).$$

So,

$$\bar{I}(\nu) = [I(\eta)]^{-1} = [\text{Var}(T)]^{-1}.$$

□

5.2 Linear model

High priority *Linear model*

Basics. Data $(X_1, Z_1), \dots, (X_n, Z_n)$, where $X_i \in \mathbb{R}$, $Z_i \in \mathbb{R}^p$. Often independent across $i = 1, 2, \dots, n$, and they follow

$$X_i = \beta^\top Z_i + \epsilon_i, \quad (4)$$

where

- X_i is the response, observation, or outcome.
- β is the coefficient vector.
- Z_i is the vector of explanatory variables (or covariates, features, attributes, regressors, predictors).
- ϵ_i is the random error (or noise).

Here Z_i can be random or nonrandom. For simplicity, we will treat Z_i as fixed (If Z_i is random and independent of the other random variables, we condition on Z_i and conclusions remain the same.) We can also write the matrix form.

$$\begin{matrix} X \\ n \times 1 \end{matrix} = \begin{matrix} Z \\ n \times p \end{matrix} \cdot \begin{matrix} \beta \\ p \times 1 \end{matrix} + \begin{matrix} \epsilon \\ n \times 1 \end{matrix}.$$

Note: Equation 4 subsumes the case with an intercept term: $X_i = \beta_0 + \beta^\top Z_i + \epsilon_i$. In the matrix form, we include a column of ones in Z for the intercept.

Least squares estimator (LSE). LSE $\hat{\beta} \in \mathbb{R}^p$ satisfies

$$\|X - Z\hat{\beta}\|^2 = \min_{b \in \mathbb{R}^p} \|X - Zb\|^2. \quad (5)$$

For any $\ell \in \mathbb{R}^p$, $\ell^\top \hat{\beta}$ is called an LSE of $\ell^\top \beta$. As we already know, LSE admits an explicit formula.

- If Z has full column rank, then

$$\hat{\beta} = (Z^\top Z)^{-1} Z^\top X.$$

- If Z does not have full column rank, then there are infinitely many solutions to Equation 5, and the minimum ℓ_2 -norm solution is unique and given by

$$\hat{\beta} = (Z^\top Z)^+ Z^\top X,$$

where $(\cdot)^+$ denotes the (Moore–Penrose) pseudoinverse.

Least squares as a projection. The vectors Zb , as b varies, fill the **column space** $\mathcal{C}(Z) \subseteq \mathbb{R}^n$. Least squares finds the point in this space closest to the observation vector X . Write P_Z for the orthogonal projection onto $\mathcal{C}(Z)$. The **fitted vector** and **residual vector** are

$$\hat{X} = Z\hat{\beta} = P_Z X, \quad r = X - \hat{X} = (I_n - P_Z)X.$$

The residual is perpendicular to every column of Z : $Z^\top r = 0$. For full column rank, $P_Z = Z(Z^\top Z)^{-1}Z^\top$; the geometric projection is still unique when the coefficient vector is not.

Reading the projection diagram. The shaded plane represents the column space of the design matrix. The fitted vector lies in this plane; the residual joins the fitted vector to the observation and is perpendicular to the whole plane.

The diagram uses y for our observation vector X , X for our design matrix Z , and $\hat{\varepsilon}$ for our residual r . Thus its $X\hat{\beta}$ is our $Z\hat{\beta} = \hat{X}$.

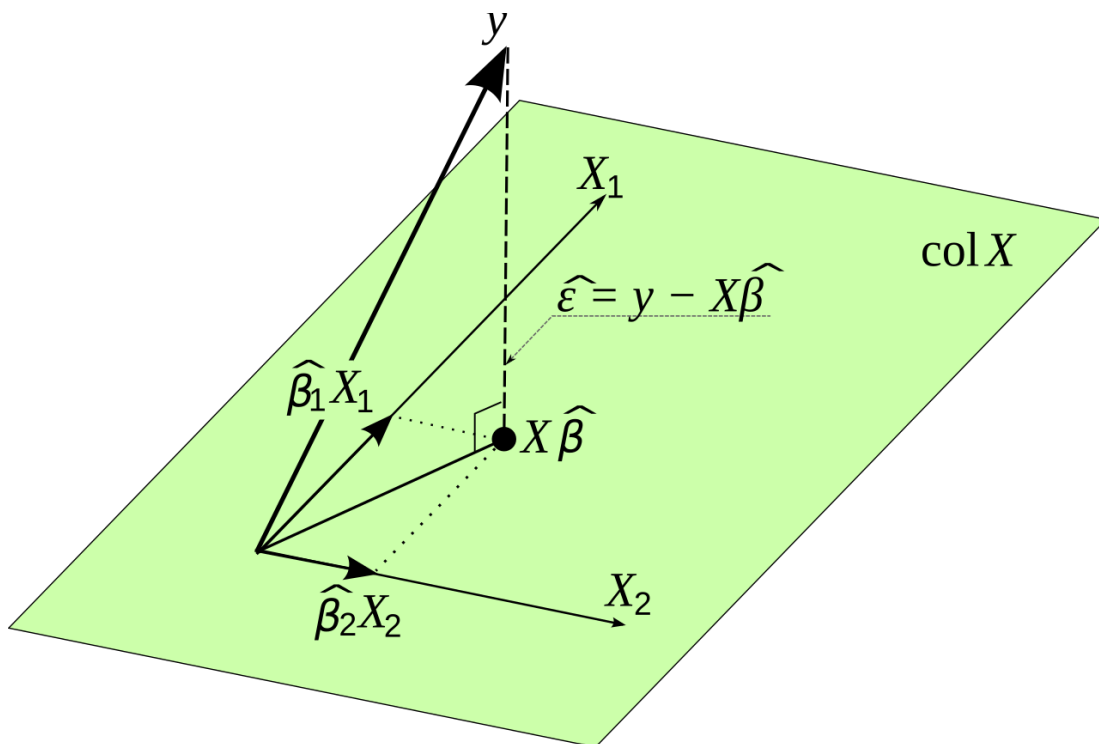


Figure 3: Least squares projects the observation onto the model space. The fitted vector is a linear combination of the design columns; the perpendicular residual is what the model cannot explain. Figure: Stpasha, *OLS geometric interpretation* (public domain).

Why the projection is closest. For any coefficient b , the two terms in $X - Zb = r + (\hat{X} - Zb)$ are orthogonal. Pythagoras gives

$$\|X - Zb\|^2 = \|r\|^2 + \|\hat{X} - Zb\|^2 \geq \|r\|^2.$$

For a numerical illustration, take $Z = (1, 1)^\top$ and $X = (1, 3)^\top$. Then $\hat{\beta} = (X_1 + X_2)/2 = 2$, and the squared error is $2 + 2(b - 2)^2$: its minimum is 2, compared with 4 at $b = 1$. The fit captures the component along the model space; the residual is what remains in the perpendicular directions. For another visual explanation, see Strang's MIT lecture on projections onto subspaces.

From orthogonality to Gaussian independence. The geometry does not require Gaussian noise. To anticipate Theorem 3, now suppose $X = Z\beta + \epsilon$, with fixed full-column-rank Z , $n > p$, and $\epsilon \sim N(0, \sigma^2 I_n)$. Recall from Lecture 1: affine transformations of Gaussian vectors that a linear transformation of a Gaussian vector is still Gaussian. In particular, $P_Z X$ and $(I_n - P_Z)X$ are *jointly* Gaussian, and their cross-covariance is

$$\text{Cov}(P_Z X, (I_n - P_Z)X) = \sigma^2 P_Z (I_n - P_Z) = 0.$$

Joint Gaussianity turns this zero cross-covariance into independence. Orthogonality alone would not give independence for arbitrary noise.

Reading Theorem 3 in this basis. Choose orthonormal columns Q_1 spanning $\mathcal{C}(Z)$ and complete them with Q_2 to an orthogonal matrix $Q = (Q_1, Q_2)$. For $G = \epsilon/\sigma$, the vector $Q^\top G \sim N(0, I_n)$ has independent standard Normal coordinates. The fitted vector depends only on $Q_1^\top G$ (and the fixed mean), while the residual depends only on $Q_2^\top G$. Thus any function of the fit, including $\ell^\top \hat{\beta}$, is independent of any function of the residual, including $\hat{\sigma}^2$. Moreover,

$$\frac{\|r\|^2}{\sigma^2} = \|Q_2^\top G\|^2 = \sum_{j=p+1}^n (Q^\top G)_j^2 \sim \chi_{n-p}^2.$$

This explains both the independence and the $n - p$ degrees of freedom in Theorem 3.

Optimality of LSE. We make the following parametric assumption: $\epsilon \sim N(0, \sigma^2 I_n)$, where $\sigma^2 > 0$ is unknown.

Theorem 2. *Under the above assumption, if Z has full column rank and $n > p$, then*

- (i) $\ell^\top \hat{\beta}$ is UMVUE of $\ell^\top \beta$.
- (ii) $\hat{\sigma}^2 = \frac{1}{n-p} \|X - Z\hat{\beta}\|^2$ is UMVUE of σ^2 .

Proof.

$$\begin{aligned} \|X - Z\beta\|^2 &= \|X - Z\hat{\beta}\|^2 - 2\beta^\top Z^\top X \\ &\quad + \|Z\beta\|^2 + \|Z\hat{\beta}\|^2. \end{aligned}$$

The joint p.d.f. of X is

$$\begin{aligned} p_{\beta, \sigma^2}(x) &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp \left\{ \frac{\beta^\top Z^\top x}{\sigma^2} \right. \\ &\quad \left. - \frac{\|x - Z\hat{\beta}\|^2 + \|Z\hat{\beta}\|^2}{2\sigma^2} - \frac{\|Z\beta\|^2}{2\sigma^2} \right\}. \end{aligned}$$

We can use the property about exponential family (Proposition 1 from Lecture 12) to check that $(Z\hat{\beta}, \|X - Z\hat{\beta}\|^2)$ is complete and sufficient for $\theta = (\beta, \sigma^2)$. Note that $\ell^\top \hat{\beta}$, $\hat{\sigma}^2$ are functions of $Z\hat{\beta}$, $\|X - Z\hat{\beta}\|^2$ respectively. Further, it is easy to verify unbiasedness (see Exercise 7)

$$\mathbb{E}[\ell^\top \hat{\beta}] = \ell^\top \beta, \quad \mathbb{E}[\hat{\sigma}^2] = \sigma^2.$$

Therefore, the Lehmann–Scheffé theorem implies UMVUE. $\square$

In fact, we can characterize the distribution of the LSE and residual explicitly, which is useful for deriving statistical tests for LSE.

Theorem 3. Under the same assumption as in Theorem 2, $\ell^\top \hat{\beta}$ is independent of $\hat{\sigma}^2$. Specifically,

$$\begin{aligned}\ell^\top \hat{\beta} &\sim \mathcal{N}(\ell^\top \beta, \sigma^2 \ell^\top (Z^\top Z)^+ \ell), \\ \hat{\sigma}^2 \frac{n-p}{\sigma^2} &\sim \chi_{n-p}^2.\end{aligned}$$

Proof sketch. Denote $(P_Z, P_Z^\perp)$ to be the projection matrices to the column space of Z and its orthogonal complement in $\mathbb{R}^n$, respectively. Then,

$$Z\hat{\beta} = P_Z(X), \quad X - Z\hat{\beta} = P_Z^\perp(X).$$

After subtracting $Z\beta$ and dividing by σ , the noise is isotropic¹. An orthogonal transformation preserves its standard Gaussian distribution. Thus,

$$\begin{aligned}P_Z(X) \text{ is independent of } P_Z^\perp(X), \\ \frac{1}{\sigma^2} \|P_Z^\perp(X)\|^2 \sim \chi_{n-p}^2.\end{aligned}$$

□

There is a geometric intuition behind this proof: the best linear fit is the orthogonal projection of the observation vector X onto the subspace spanned by the columns of Z . See the earlier projection picture and orthonormal-coordinate explanation.

Fisher information. We consider the parametric assumption $\epsilon \sim N(0, \sigma^2 I_n)$ and the unknown parameters are $\theta = (\beta, \sigma^2)$.

$$I(\theta) = \frac{1}{\sigma^2} \begin{pmatrix} Z^\top Z & 0 \\ 0 & \frac{n}{2\sigma^2} \end{pmatrix}.$$

If Z is not of full rank, $I(\theta)$ may not be invertible. Indeed, in the null space $\text{Ker}(Z)$, two distributions $P_\theta, P_{\theta'}$ are not distinguishable.

It is clear that LSE is the MLE for estimating β under our assumption, and it attains the Cramér–Rao lower bound; the corresponding question for σ^2 is the linked exercise below.

Regression and information. For fixed full-column-rank design Z with Gaussian noise, $I_{\beta\beta} = Z^\top Z / \sigma^2$. Least squares attains its inverse covariance, while unbiased residual variance uses the remaining $n - p$ dimensions.

High priority Practice: Does residual variance attain its bound?

Exercise prompt. In full-rank Gaussian regression with $n > p \geq 1$, compare the variance of $\hat{\sigma}^2 = \|X - Z\hat{\beta}\|^2 / (n - p)$ with the information bound for σ^2 .

See Exercise 4 in Section 8 for the full statement and complete solution.

6 Further reading

Further reading Further reading

The information perspective of Fisher information can be found in many sources, e.g., Section 2.6.2 of [1]. Fisher information is a central quantity in classical mathematical statistics, and in STAT 710 you will revisit this concept, especially when studying MLE.

¹Here isotropic means mean zero and covariance equal to the identity. In particular, every unit-direction projection has variance one.

Least squares estimator. The discovery of least squares has an interesting history. A nice reference is the Wikipedia page.

Least squares estimators can be derived using either MLE or the method of moments, both of which will be covered in STAT 709 and STAT 710. The practical side of least squares is covered in STAT 849 and 850. Some key related notions:

- collinearity,
- basis expansion,
- least absolute deviations.

For solving nonlinear least squares, a classical approach is *iteratively reweighted least squares*.

In the modern context where dimension grows proportionally with sample size, least squares estimator is inconsistent; in over-parametrized settings, least squares estimator is not uniquely defined, but min-norm solution is usually expected and enjoys good statistical properties.

Optimality of LSE under weaker conditions We make the following assumptions: $\mathbb{E}[\epsilon] = 0$, $\text{Var}(\epsilon) = \sigma^2 I_n$ where $\sigma^2 > 0$ is unknown.

Theorem 4 (Gauss–Markov theorem). *Under the above assumptions, if Z has full column rank and $n > p$, then $\ell^\top \hat{\beta}$ is the best linear (linear in X) unbiased estimator (BLUE) of $\ell^\top \beta$.*

We make weaker assumptions, and accordingly the optimality is stated in a smaller space of candidate estimators.

Key ideas

- Fisher information measures local likelihood sensitivity and depends on parametrization.
- The Cramér–Rao bound requires regularity and can fail to be attained.
- Complete sufficient statistics and information bounds give different routes to optimality.
- Least-squares projection explains fitted values and residuals; Gaussianity additionally yields independence.

7 Detailed reference

High priority *What “Information” means*

Fisher information is nonnegative, adds across independent observations under the stated regularity conditions, and depends on the parametrization. Changing parameter coordinates is different from an invertible recoding of the observations. Return to the earlier explanation.

High priority *The unbiased variance bound*

The Cramér–Rao inequality compares the derivative of the target with information. Equality requires a centered estimator to be a linear transformation of the score, so the bound need not be attainable. Return to the earlier explanation.

High priority *Regression and information*

For fixed full-column-rank design Z with Gaussian noise, $I_{\beta\beta} = Z^\top Z / \sigma^2$. Least squares attains its inverse covariance, while unbiased residual variance uses the remaining $n - p$ dimensions. Return to the earlier explanation.

Looking ahead

Cramér–Rao bounds the variance of unbiased estimators. Lecture 14 asks how likely a large deviation is at a fixed sample size, using exponential tail bounds rather than only a variance comparison. Continue to Lecture 14.

You have finished Part IV—check yourself. Preview Part V.

8 Exercises and complete solutions

High priority *Information in Bernoulli observations*

Exercise 1 (Information in Bernoulli observations).

For n independent Bernoulli(p) observations, $0 < p < 1$, compute the score and information for p .

Solution.

The score is $\sum_i (X_i - p)/[p(1 - p)]$. It has mean zero and variance $np(1 - p)/[p^2(1 - p)^2] = n/[p(1 - p)]$. Independence yields the factor n . The endpoints are excluded because the regular interior model and its score formula change there. □

High priority *A bound attained by the Poisson mean*

Exercise 2 (A bound attained by the Poisson mean).

For i.i.d. Poisson(θ) data, show that the sample mean attains the Cramér–Rao bound for $\theta > 0$.

Solution.

The score is $\sum_i (X_i/\theta - 1)$, with information n/θ . Since $\bar{X}$ is unbiased and has variance θ/n , it attains $1/I(\theta)$ and is UMVUE among unbiased square-integrable estimators. □

Mid priority *An exponential Normal target*

Exercise 3 (An exponential Normal target).

For i.i.d. $N(\mu, 1)$ observations, consider $H = \exp(\bar{X} - 1/(2n))$. Verify unbiasedness for e^μ , compute its variance, and compare with the Cramér–Rao bound.

Solution.

The Normal MGF gives $\mathbb{E} H = e^\mu$ and $\mathbb{E} H^2 = e^{2\mu+1/n}$. Thus $\text{Var} H = e^{2\mu}(e^{1/n} - 1)$. Information is n and the target derivative is e^μ , giving bound $e^{2\mu}/n$. Since $e^x > 1 + x$ for $x > 0$, the bound is strict at every finite n . □

High priority *Does residual variance attain its bound?*

Exercise 4 (Does residual variance attain its bound?).

In full-rank Gaussian regression with $n > p \geq 1$, compare the variance of $\hat{\sigma}^2 = \|X - Z\hat{\beta}\|^2/(n - p)$ with the information bound for σ^2 .

Solution.

The chi-square law gives variance $2\sigma^4/(n - p)$. The information matrix is block diagonal, with variance-parameter information $n/(2\sigma^4)$, so its bound is $2\sigma^4/n$. The former is strictly larger for $p > 0$. Being UMVUE does not imply attaining the Cramér–Rao bound.

□

Mid priority *Why Uniform support breaks the regular bound*

Exercise 5 (Why Uniform support breaks the regular bound).

For one $\text{Uniform}(0, \theta)$ observation, explain why differentiating the log density on its interior does not justify the regular Cramér–Rao bound.

Solution.

For $0 < x < \theta$, the formal score is $-1/\theta$, whose mean is not zero. The derivative of the integral $\int_0^\theta \theta^{-1} dx = 1$ includes a moving-boundary contribution. Interchanging the derivative with the integral over the interior alone misses that term. Therefore the regular score identities and the stated CR theorem do not apply.

□

High priority *The vector information bound*

Exercise 6 (The vector information bound).

Prove the vector Cramér–Rao bound from $\mathbb{E} s = 0$, $\mathbb{E} s s^\top = I$ and $\mathbb{E}[(T - g)s^\top] = Dg$.

Solution.

Let $R = T - g - Dg I^{-1} s$. Expanding its covariance and using the three identities gives $\text{Cov}(R) = \text{Cov}(T) - Dg I^{-1} Dg^\top$. Every covariance matrix is positive semidefinite, proving the claim. Equality holds exactly when $R = 0$ almost surely.

□

High priority *Unbiasedness of the regression estimators*

Exercise 7 (Unbiasedness of the regression estimators).

Prove $\mathbb{E} \hat{\beta} = \beta$ and $\mathbb{E} \|X - Z\hat{\beta}\|^2 = (n - p)\sigma^2$ under the stated fixed-design assumptions.

Solution.

The formula gives $\hat{\beta} = \beta + (Z^\top Z)^{-1} Z^\top \epsilon$, with mean β . Let $P = Z(Z^\top Z)^{-1} Z^\top$. The residual is $(I - P)\epsilon$, so its expected squared norm is $\text{Tr}[(I - P)\sigma^2 I] = \sigma^2(n - p)$, since P is a rank- p orthogonal projector. This expectation argument needs only the first two noise moments.

□

Low priority *Gauss–Markov by an orthogonal decomposition*

Exercise 8 (Gauss–Markov by an orthogonal decomposition).

Prove the BLUE claim for $a^\top X$ estimating $\ell^\top \beta$ when $Z^\top a = \ell$.

Solution.

The least-squares coefficient is $a_0 = Z(Z^\top Z)^{-1} \ell$. Any unbiased coefficient has $a = a_0 + d$ with $Z^\top d = 0$. Since a_0 belongs to the column space of Z , $a_0^\top d = 0$. Therefore $\text{Var}(a^\top X) = \sigma^2(\|a_0\|^2 + \|d\|^2) \geq \sigma^2\|a_0\|^2$.

□

References

- [1] Yury Polyanskiy and Yihong Wu. *Information Theory: From Coding to Learning*. Cambridge University Press, 2024.

What changed

October 19, 2026: Corrected notation, wording and cross-references after the course-wide errata review; refreshed the linked HTML and PDF editions.

October 14, 2026. Added issue-reporting links to the webpage and PDF. The mathematics and priority assignments are unchanged in this update.

October 14, 2026 (figure follow-up). Restored the original Fisher-information sketches and the credited least-squares projection figure, with its notation key.