

Point Estimation: properties of estimators

- finite-sample properties (CB 7.3)
- large-sample properties (CB 10.1)



1 FINITE-SAMPLE PROPERTIES

How an estimator performs for finite number of observations n .

Estimator: W

Parameter: θ

Criteria for evaluating estimators:

- **Bias:** does $EW = \theta$?
- Variance of W (you would like an estimator with a smaller variance)

Example: $X_1, \dots, X_n \sim i.i.d. (\mu, \sigma^2)$

Unknown parameters are μ and σ^2 .

Consider:

$\hat{\mu}_n \equiv \frac{1}{n} \sum_i X_i$, estimator of μ

$\hat{\sigma}_n^2 \equiv \frac{1}{n} \sum_i (X_i - \bar{X}_n)^2$, estimator of σ^2 .

Bias: $E\hat{\mu}_n = \frac{1}{n} \cdot n\mu = \mu$. So *unbiased*.

$\text{Var } \hat{\mu}_n = \frac{1}{n^2} n\sigma^2 = \frac{1}{n}\sigma^2$.

$$\begin{aligned} E\hat{\sigma}^2 &= E\left(\frac{1}{n} \sum_i (X_i - \bar{X}_n)^2\right) \\ &= \frac{1}{n} \cdot \sum_i (EX_i^2 - 2EX_i\bar{X}_n + E\bar{X}_n^2) \\ &= \frac{1}{n} \cdot n \left[(\mu^2 + \sigma^2) - 2\left(\mu^2 + \frac{\sigma^2}{n}\right) + \frac{\sigma^2}{n} + \mu^2 \right] \\ &= \frac{n-1}{n}\sigma^2. \end{aligned}$$

Hence it is biased.

To fix this bias, consider the estimator $s_n^2 \equiv \frac{1}{n-1} \sum_i (X_i - \bar{X}_n)^2$, and $Es_n^2 = \sigma^2$ (unbiased).



Mean-squared error (MSE) of W is $E(W - \theta)^2$. Common criterion for comparing estimators.

Decompose: $MSE(W) = VW + (EW - \theta)^2 = \text{Variance} + (\text{Bias})^2$.

Hence, for an unbiased estimator: $MSE(W) = VW$.

Example: $X_1, \dots, X_n \sim U[0, \theta]$. $f(X) = 1/\theta$, $x \in [0, \theta]$.

- Consider estimator $\hat{\theta}_n \equiv 2\bar{X}_n$.
 $E\hat{\theta}_n = 2 \cdot \frac{1}{n} \cdot E \sum_i X_i = 2 \cdot \frac{1}{n} \cdot n \cdot \theta = \theta$. So unbiased
 $MSE(\hat{\theta}_n) = V\hat{\theta}_n = \frac{4}{n^2} \sum_i V X_i = \frac{\theta^2}{3n}$
- Consider estimator $\tilde{\theta}_n \equiv \max(X_1, \dots, X_n)$.

In order to derive moments, start by deriving CDF:

$$\begin{aligned} P(\tilde{\theta}_n \leq z) &= P(X_1 \leq z, X_2 \leq z, \dots, X_n \leq z) \\ &= \prod_{i=1}^n P(X_i \leq z) \\ &= \begin{cases} \left(\frac{z}{\theta}\right)^n & \text{if } z \leq \theta \\ 1 & \text{otherwise} \end{cases} \end{aligned}$$

Therefore $f_{\tilde{\theta}_n}(z) = n \cdot \left(\frac{z}{\theta}\right)^{n-1} \frac{1}{\theta}$, for $0 \leq z \leq \theta$.

$$\begin{aligned} E(\tilde{\theta}_n) &= \int_0^\theta z \cdot n \cdot \left(\frac{z}{\theta}\right)^{n-1} \frac{1}{\theta} dz \\ &= \frac{n}{\theta^n} \int_0^\theta z^n dz = \frac{n}{n+1} \theta. \end{aligned}$$

$$\text{Bias}(\tilde{\theta}_n) = -\theta/(n+1)$$

$$E(\tilde{\theta}_n^2) = \frac{n}{\theta^n} \int_0^\theta z^{n+1} dz = \frac{n}{n+2} \theta^2.$$

$$\text{Hence } V\tilde{\theta}_n = \theta^2 \left(\frac{n}{n+2} - \left(\frac{n}{n+1}\right)^2 \right) = \theta^2 \frac{n}{(n+2)(n+1)^2}.$$

$$\text{Accordingly, } MSE = \frac{2\theta^2}{(n+2)(n+1)}$$



Continue the previous example. Redefine $\tilde{\theta}_n = \frac{n+1}{n} \max(X_1, \dots, X_n)$. Now both estimators $\hat{\theta}_n$ and $\tilde{\theta}_n$ are unbiased.

Which is better? $V\hat{\theta}_n = \frac{\theta^2}{3n} = O(1/n)$.

$$V\tilde{\theta}_n = \left(\frac{n+1}{n}\right)^2 V(\max(X_1, \dots, X_n)) = \theta^2 \left(\frac{1}{n(n+2)}\right) = O(1/n^2).$$

Hence, for n large enough, $\tilde{\theta}_n$ has a smaller variance, and in this sense it is “better”.



Best unbiased estimator: if you choose the best (in terms of MSE) estimator, and restrict yourself to unbiased estimators, then the best estimator is the one with the *lowest variance*.

A best unbiased estimator is also called the “Uniform minimum variance unbiased estimator” (UMVUE).

Formally: an estimator W is a UMVUE of θ satisfies:

- (i) $E_\theta W = \theta$, for all θ (unbiasedness)
- (ii) $V_\theta W \leq V_\theta \tilde{W}$, for all θ , and all other unbiased estimators \tilde{W} .

The “uniform” condition is crucial, because it is always possible to find estimators which have zero variance for a specific value of θ .



It is difficult in general to verify that an estimator W is UMVUE, since you have to verify condition (ii) of the definition, that VW is smaller than all other unbiased estimators.

Luckily, we have an important result for the *lowest attainable variance* of an estimator.

- **Theorem 7.3.9 (Cramer-Rao Inequality):** Let X_1, \dots, X_n be a sample with joint pdf $f(\vec{X}|\theta)$, and let $W(\vec{X})$ be any estimator satisfying

$$(i) \quad \frac{d}{d\theta} E_\theta W(\vec{X}) = \int \frac{\partial}{\partial \theta} [W(\vec{X}) \cdot f(\vec{X}|\theta)] d\vec{X};$$

$$(ii) \quad V_\theta W(\vec{X}) < \infty.$$

Then

$$V_\theta W(\vec{X}) \geq \frac{\left(\frac{d}{d\theta} E_\theta W(\vec{X})\right)^2}{E_\theta \left(\frac{\partial}{\partial \theta} \log f(\vec{X}|\theta)\right)^2}.$$

The RHS above is called the *Cramer-Rao Lower Bound*.

Proof: CB, pg. 336.

Note: the LHS of condition (i) above is $\frac{d}{d\theta} \int W(\vec{X})f(X|\theta)dX$, so by Leibniz' rule, this condition basically rules out cases where the support of X is dependent on θ .

- The equality

$$\begin{aligned} E_{\theta} \frac{\partial}{\partial \theta} \log f(\vec{X}|\theta) &= \int \frac{1}{f(\vec{X}|\theta)} \frac{\partial f(\vec{X}|\theta)}{\partial \theta} f(\vec{X}|\theta) dx \\ &= \frac{\partial}{\partial \theta} \int f(\vec{X}|\theta) dx \\ &= \frac{\partial}{\partial \theta} \cdot 1 = 0 \end{aligned} \tag{1}$$

is noteworthy, because

$$\frac{\partial}{\partial \theta} \log f(\vec{X}|\theta) = 0$$

is the FOC of maximum likelihood estimation problem. (Alternatively, as in CB, apply condition (i) of CR result, using $W(X) = 1$.)

In the i.i.d. case, this becomes the sample average

$$\frac{1}{n} \sum_i \frac{\partial}{\partial \theta} \log f(x_i|\theta) = 0.$$

And by the LLN:

$$\frac{1}{n} \sum_i \frac{\partial}{\partial \theta} \log f(x_i|\theta) \xrightarrow{p} E_{\theta_0} \frac{\partial}{\partial \theta} \log f(x_i|\theta),$$

where θ_0 is the true value of θ . This shows that maximum likelihood estimation of θ is equivalent to estimation based on the moment condition

$$E_{\theta_0} \frac{\partial}{\partial \theta} \log f(x_i|\theta) = 0$$

which holds only at the true value $\theta = \theta_0$. (Thus MLE is “consistent” for the true value θ_0 , as we’ll see later.)

(However, note that Eq. (1) holds at *all* values of θ , not just θ_0 .)

What if model is “misspecified”, in the sense that true density of \vec{X} is $g(\vec{x})$, and that for all $\theta \in \Theta$, $f(\vec{x}|\theta) \neq g(\vec{x})$ (that is, there is no value of the parameter θ such that the postulated model f coincides with the true model g)? Does Eq. (1) still hold? What is MLE looking for?

- In the iid case, the CR lower bound can be simplified

Corollary 7.3.10: if $X_1, \dots, X_n \sim i.i.d. f(X|\theta)$, then

$$V_{\theta}W(\vec{X}) \geq \frac{\left(\frac{d}{d\theta}E_{\theta}W(\vec{X})\right)^2}{n \cdot E_{\theta}\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)^2}.$$

- Up to this point, Cramer-Rao results not that operational for us to find a “best” estimator, because the estimator $W(\vec{X})$ is on both sides of the inequality. However, for an unbiased estimator, how can you simplify the expression further?
- **Example:** $X_1, \dots, X_n \sim i.i.d. N(\mu, \sigma^2)$.

What is CRLB for an unbiased estimator of μ ?

Unbiased \rightarrow numerator =1.

$$\begin{aligned}\log f(x|\theta) &= \log \sqrt{2\pi} - \log \sigma - \frac{1}{2} \left(\frac{x - \mu}{\sigma}\right)^2 \\ \frac{\partial}{\partial\mu} \log f(x|\theta) &= - \left(\frac{x - \mu}{\sigma}\right) \cdot \left(\frac{-1}{\sigma}\right) = \frac{x - \mu}{\sigma^2} \\ E \left(\frac{\partial}{\partial\theta} \log f(X|\theta)\right)^2 &= E \left((X - \mu)^2 \sigma^{-4}\right) = \frac{1}{\sigma^4} V X = \frac{1}{\sigma^2}.\end{aligned}$$

Hence the CRLB = $\frac{1}{n \cdot \frac{1}{\sigma^2}} = \frac{\sigma^2}{n}$.

This is the variance of the sample mean, so that the sample mean is a UMVUE for μ .

- Sometimes we can simplify the denominator of the CRLB further:

Lemma 7.3.11 (Information inequality): if $f(X|\theta)$ satisfies

$$(*) \quad \frac{d}{d\theta}E_{\theta} \left(\frac{\partial}{\partial\theta} \log f(X|\theta)\right) = \int \frac{\partial}{\partial\theta} \left[\left(\frac{\partial}{\partial\theta} \log f(X|\theta)\right) f(X|\theta) \right] dx,$$

then

$$E_{\theta} \left(\frac{\partial}{\partial\theta} \log f(X|\theta)\right)^2 = -E_{\theta} \left(\frac{\partial^2}{\partial\theta^2} \log F(X|\theta)\right).$$

Rough proof:

LHS of (*): Using Eq. (1) above, we get that LHS of (*) =0.

RHS of (*):

$$\begin{aligned} & \int \frac{\partial}{\partial \theta} \left[\left(\frac{\partial}{\partial \theta} \log f \right) f \right] dx \\ &= \int \frac{\partial^2 \log f}{\partial \theta^2} f dx + \int \frac{1}{f} \left(\frac{\partial f}{\partial \theta} \right)^2 dx \\ &= E \frac{\partial^2 \log f}{\partial \theta^2} + E \left(\frac{\partial \log f}{\partial \theta} \right)^2. \end{aligned}$$

Putting the LHS and RHS together yields the desired result. ■

Note: the LHS of the above condition (*) is just $\frac{d}{d\theta} \int \left(\frac{\partial}{\partial \theta} \log f(X|\theta) \right) f(X|\theta) dX$, so that, by Leibniz' rule, the condition (*) just states that the bounds of the integration (i.e., the support of X) does not depend on θ . Normal distribution satisfies this (support is always $(-\infty, \infty)$), but $U[0, \theta]$ does not.

Also, the information inequality depends crucially on the equality $E_\theta \frac{\partial}{\partial \theta} \log f(X|\theta) = 0$, which depends on the correct specification of the model. Thus information inequality can be used as basis of "specification test". (How?)

- **Example:** for the previous example, consider CRLB for unbiased estimator of σ^2 .

We can use the information inequality, because condition (*) is satisfied for the normal distribution. Hence:

$$\begin{aligned} \frac{\partial}{\partial \sigma^2} \log f(x|\theta) &= \frac{-1}{2\sigma^2} + \frac{1}{2} \frac{(x - \mu)^2}{\sigma^4} \\ \frac{\partial}{\partial \sigma^2} \left(\frac{\partial}{\partial \sigma^2} \log f(x|\theta) \right) &= \frac{1}{2\sigma^4} - \frac{(x - \mu)^2}{\sigma^6} \\ -E \left(\frac{\partial}{\partial \sigma^2} \left(\frac{\partial}{\partial \sigma^2} \log f(x|\theta) \right) \right) &= - \left(\frac{1}{2\sigma^4} - \frac{1}{\sigma^4} \right) = \frac{1}{2\sigma^4}. \end{aligned}$$

Hence the CRLB is $\frac{2\sigma^4}{n}$.

- **Example:** $X_1, \dots, X_n \sim U[0, \theta]$. Check conditions for CRLB for an unbiased estimator $W(\vec{X})$ of θ .

$\frac{d}{d\theta} EW(\vec{X}) = 1$ (because it is unbiased)

$$\int \frac{\partial}{\partial \theta} \left[W(\vec{X}) f(\vec{X}|\theta) \right] d\vec{X} = \int W(\vec{X}) \cdot \left(\frac{-1}{\theta^2} \right) d\vec{X} \neq \frac{d}{d\theta} EW(\vec{X}) = 1$$

Hence, condition (i) of theorem not satisfied.

- But when can CRLB (if it exists) be attained?

Corollary 7.3.15: $X_1, \dots, X_n \sim i.i.d. f(X|\theta)$, satisfying the conditions of CR theorem.

- Let $L(\theta|\vec{X}) = \prod_{i=1}^n f(X|\theta)$ denote the likelihood function.
- Estimator $W(\vec{X})$ unbiased for θ
- $W(\vec{X})$ attains CRLB iff you can write

$$\frac{\partial}{\partial \theta} \log L(\theta|\vec{X}) = a(\theta) [W(\vec{X}) - \theta]$$

for some function $a(\theta)$.

- **Example:** $X_1, \dots, X_n \sim i.i.d. N(\mu, \sigma^2)$

Consider estimating μ :

$$\begin{aligned} \frac{\partial}{\partial \mu} \log L(\theta|\vec{X}) &= \frac{\partial}{\partial \mu} \left(\sum_{i=1}^n \log f(X|\theta) \right) \\ &= \frac{\partial}{\partial \mu} \left(\sum_{i=1}^n -\log \sqrt{2\pi} - \log \sigma - \frac{1}{2} \left(\frac{(X - \mu)^2}{\sigma^2} \right) \right) \\ &= \sum_{i=1}^n \left(\frac{X - \mu}{\sigma^2} \right) \\ &= \frac{n}{\sigma^2} \cdot (\bar{X}_n - \mu). \end{aligned}$$

Hence, CRLB can be attained (in fact, we showed earlier that CRLB attained by \bar{X}_n)



Loss function optimality

Let $\vec{X} \sim f(\vec{X}|\theta)$.

Consider a *loss function* $\mathcal{L}(\theta, W(\vec{X}))$, taking values in $[0, +\infty)$, which penalizes you when your $W(\vec{X})$ estimator is “far” from the true parameter θ . Note that $\mathcal{L}(\theta, W(\vec{X}))$ is a random variable, since \vec{X} (and $W(\vec{X})$) are random.

Consider estimators which *minimize expected loss*: that is

$$\min_{W(\dots)} E_{\theta} \mathcal{L}(\theta, W(\vec{X})) \equiv \min_{W(\dots)} R(\theta, W(\dots))$$

where $R(\theta, W(\dots))$ is the *risk function*. (Note: the risk function is not a random variable, because \vec{X} has been integrated out.)

Loss function optimality is a more general criterion than minimum MSE. In fact, because $MSE(W(\vec{X})) = E_{\theta} \left(W(\vec{X}) - \theta \right)^2$, the MSE is actually the risk function associated with the *quadratic loss function* $\mathcal{L}(\theta, W(\vec{X})) = \left(W(\vec{X}) - \theta \right)^2$.

Other examples of loss functions:

- Absolute error loss: $|W(\vec{X}) - \theta|$
- Relative quadratic error loss: $\frac{(W(\vec{X}) - \theta)^2}{|\theta| + 1}$

The exercise of minimizing risk takes a given value of θ as given, so that the minimized risk of an estimator depends on whichever value of θ you are considering. You might be interested in an estimator which does well regardless of which value of θ you are considering. (Analogous to the focus on the *uniform* minimal variance.)

For this different problem, you want to consider a notion of risk which does not depend on θ . Two criteria which have been considered are:

- “Average” risk:

$$\min_{W(\dots)} \int R(\theta, W(\dots)) h(\theta) d\theta.$$

where $h(\theta)$ is some weighting function across θ . (In a Bayesian interpretation, $h(\theta)$ is a prior density over θ .)

- Minmax criterion:

$$\min_{W(\dots)} \max_{\theta} R(\theta, W(\dots)).$$

Here you choose the estimator $W(\dots)$ to minimize the maximum risk = $\max_{\theta} R(\theta, W(\dots))$, where θ is set to the “worse” value. So minmax optimizer is the best that can be achieved in a “worst-case” scenario.

Example: $X_1, \dots, X_n \sim i.i.d. N(\mu, \sigma^2)$. Sample mean \bar{X}_n is:
 unbiased
 minimum MSE
 UMVUE
 attains CRLB
 minimizes expected quadratic loss

2 LARGE SAMPLE PROPERTIES OF ESTIMATORS

It can be difficult to compute MSE, risk functions, etc., for some estimators, especially when estimator does not resemble a sample average.

Large-sample properties: exploit LLN, CLT

Consider data $\{X_1, X_2, \dots\}$ by which we construct a sequence of estimators $W_n \equiv \{W(\vec{X}_1), W(\vec{X}_2), \dots\}$. W_n is a random sequence.

Define: we say that W_n is **consistent** for a parameter θ iff the random sequence W_n converges (in some stochastic sense) to θ . *Strong consistency* obtains when $W_n \xrightarrow{as} \theta$. *Weak consistency* obtains when $W_n \xrightarrow{p} \theta$.

For estimators like sample-means, consistency (either weak or strong) follows easily using a LLN. Consistency can be thought of as the large-sample version of unbiasedness.



Define: an **M-estimator** is an estimator of θ which a maximizer of an objective function $Q_n(\theta)$.

Examples:

- MLE: $Q_n(\theta) = \frac{1}{n} \sum_{i=1}^n \log f(x_i|\theta)$
- Least squares: $Q_n(\theta) = \sum_{i=1}^n [y_i - g(x_i; \theta)]^2$. OLS is special case when $g(x_i; \theta) = \alpha + X_i' \beta$.
- GMM: $Q_n(\theta) = G_n(\theta)' W_n(\theta) G_n(\theta)$ where

$$G_n(\theta) = \left[\frac{1}{n} \sum_{i=1}^n m_1(x_i; \theta), \frac{1}{n} \sum_{i=1}^n m_2(x_i; \theta), \dots, \frac{1}{n} \sum_{i=1}^n m_M(x_i; \theta) \right]',$$

an $M \times 1$ vector of sample moment conditions, and W_n is an $M \times M$ weighting matrix.

Notation: Denote the limit objective function $Q_0(\theta) = \text{plim}_{n \rightarrow \infty} Q_n(\theta)$ (at each θ). Define $\theta_0 \equiv \text{argmax}_{\theta} Q_0(\theta)$.

Consistency of M-estimators Make the following assumptions:

1. $Q_0(\theta)$ is uniquely maximized at some value θ_0 (“identification”)
2. Parameter space Θ is a compact subset of \mathbb{R}^K .
3. $Q_0(\theta)$ is continuous in θ
4. $Q_n(\theta)$ converges uniformly in probability to $Q_0(\theta)$; that is:

$$\sup_{\theta \in \Theta} |Q_n(\theta) - Q_0(\theta)| \xrightarrow{P} 0.$$

(Interpretation: *uniform LLN*) **Theorem: (Consistency of M-Estimator)** Under assumption 1,2,3,4, $\theta_n \xrightarrow{P} \theta_0$.

Proof: We need to show: for any arbitrarily small neighborhood \mathcal{N} containing θ_0 , $P(\theta_n \in \mathcal{N}) \rightarrow 1$.

For n large enough, the uniform convergence conditions that, for all $\epsilon, \delta > 0$,

$$P\left(\sup_{\theta \in \Theta} |Q_n(\theta) - Q_0(\theta)| < \epsilon/2\right) > 1 - \delta.$$

The event “ $\sup_{\theta \in \Theta} |Q_n(\theta) - Q_0(\theta)| < \epsilon/2$ ” implies

$$Q_n(\theta_n) - Q_0(\theta_n) < \epsilon/2 \Leftrightarrow Q_0(\theta_n) > Q_n(\theta_n) - \epsilon/2 \tag{2}$$

Similarly,

$$Q_n(\theta_0) - Q_0(\theta_0) > -\epsilon/2 \Rightarrow Q_n(\theta_0) > Q_0(\theta_0) - \epsilon/2. \tag{3}$$

Since $\theta_n = \operatorname{argmax}_{\theta} Q_n(\theta)$, Eq. (2) implies

$$Q_0(\theta_n) > Q_n(\theta_n) - \epsilon/2. \tag{4}$$

Hence, adding Eqs. (3) and (4), we have

$$Q_0(\theta_n) > Q_0(\theta_0) - \epsilon. \tag{5}$$

So we have shown that

$$\begin{aligned} & \sup_{\theta \in \Theta} |Q_n(\theta) - Q_0(\theta)| < \epsilon/2 \implies Q_0(\theta_n) > Q_0(\theta_0) - \epsilon \\ \Leftrightarrow & P(Q_0(\theta_n) > Q_0(\theta_0) - \epsilon) \geq P\left(\sup_{\theta \in \Theta} |Q_n(\theta) - Q_0(\theta)| < \epsilon/2\right) \rightarrow 1. \end{aligned}$$

Now define \mathcal{N} as any open neighborhood of \mathbb{R}^K , which contains θ_0 , and $\bar{\mathcal{N}}$ is the complement of \mathcal{N} in \mathbb{R}^K . Then $\Theta \cap \bar{\mathcal{N}}$ is compact, so that $\max_{\theta \in \Theta \cap \bar{\mathcal{N}}} Q_0(\theta)$ exists. Set $\epsilon = Q_0(\theta_0) - \max_{\theta \in \Theta \cap \bar{\mathcal{N}}} Q_0(\theta)$. Then

$$\begin{aligned} Q_0(\theta_n) > Q_0(\theta_0) - \epsilon &\Rightarrow Q_0(\theta_n) > \max_{\theta \in \Theta \cap \bar{\mathcal{N}}} Q_0(\theta) \\ &\Rightarrow \theta_n \in \mathcal{N} \\ \Leftrightarrow P(\theta_n \in \mathcal{N}) &\geq P(Q_0(\theta_n) > Q_0(\theta_0) - \epsilon) \rightarrow 1. \end{aligned}$$

Since the argument above holds for any arbitrarily small neighborhood of θ_0 , we are done. ■

- In general, the limit objective function $Q_0(\theta) = \text{plim}_{n \rightarrow \infty} Q_n(\theta)$ may not be that straightforward to determine. But in many cases, $Q_n(\theta)$ is a sample average of some sort:

$$Q_n(\theta) = \frac{1}{n} \sum_i q(x_i | \theta)$$

(eg. least squares, MLE). Then by a law of large numbers, we conclude that (for all θ)

$$Q_0(\theta) = \text{plim} \frac{1}{n} \sum_i q(x_i | \theta) = E_{x_i} q(x_i | \theta)$$

where E_{x_i} denote expectation *with respect to the true (but unobserved) distribution of x_i* .

- Most of the time, θ_0 can be interpreted as a “true value”. But if model is misspecified, then this interpretation doesn’t hold (indeed, under misspecification, not even clear what the “true value” is). So a more cautious way to interpret the consistency result is that

$$\theta_n \xrightarrow{P} \text{argmax}_{\theta} Q_0(\theta)$$

which holds (given the conditions) no matter whether model is correctly specified.

- ** Of the four assumptions above, the most “high-level” one is the uniform convergence condition. Sufficient conditions for this conditions are:

1. Pointwise convergence: For each $\theta \in \Theta$, $Q_n(\theta) - Q_0(\theta) = o_p(1)$.

2. $Q_n(\theta)$ is *stochastically equicontinuous*: for every $\epsilon > 0, \eta > 0$ there exists a sequence of random variable $\Delta_n(\epsilon, \eta)$ and $n^*(\epsilon, \eta)$ such that for all $n > n^*$, $P(|\Delta_n| > \epsilon) < \eta$ and for each θ there is an open set \mathcal{N} containing θ with

$$\sup_{\tilde{\theta} \in \mathcal{N}} |Q_n(\tilde{\theta}) - Q_n(\theta)| \leq \Delta_n, \quad \forall n > n^*.$$

Note that both Δ_n and n^* *do not* depend on θ : it is uniform result.

In a deterministic context, this is an “in probability” version of the notion of uniform equicontinuity: we say a sequence of deterministic functions $R_n(\theta)$ is uniformly equicontinuous if, for every $\epsilon > 0$ there exists $\delta(\epsilon)$ and $n^*(\epsilon)$ such that for all θ

$$\sup_{\tilde{\theta}: \|\tilde{\theta} - \theta\| < \delta} |R_n(\tilde{\theta}) - R_n(\theta)| \leq \epsilon, \quad \forall n > n^*.$$

Asymptotic normality for M-estimators Define the “score vector”

$$\nabla_{\tilde{\theta}} Q_n(\theta) = \left[\frac{\partial Q_n(\theta)}{\partial \theta_1} \Big|_{\theta=\tilde{\theta}}, \dots, \frac{\partial Q_n(\theta)}{\partial \theta_K} \Big|_{\theta=\tilde{\theta}} \right]'$$

Similarly, define the $K \times K$ Hessian matrix

$$[\nabla_{\tilde{\theta}\tilde{\theta}} Q_n(\theta)]_{i,j} = \frac{\partial^2 Q_n(\theta)}{\partial \theta_i \partial \theta_j} \Big|_{\theta=\tilde{\theta}}, \quad 1 \leq i, j \leq K.$$

Note that the Hessian is symmetric.

Make the following assumptions:

1. $\theta_n = \operatorname{argmax}_{\theta} Q_n(\theta) \xrightarrow{P} \theta_0$
2. $\theta_0 \in \operatorname{interior}(\Theta)$
3. $Q_n(\theta)$ is twice continuously differentiable in a neighborhood \mathcal{N} of θ_0 .
4. $\sqrt{n} \nabla_{\theta_0} Q_n(\theta) \xrightarrow{d} N(0, \Sigma)$
5. Uniform convergence of Hessian: there exists the matrix $H(\theta)$ which is continuous at θ_0 and $\sup_{\theta \in \mathcal{N}} \|\nabla_{\theta\theta} Q_n(\theta) - H(\theta)\| \xrightarrow{P} 0$. (Interpretation: another uniform LLN).
6. $H(\theta_0)$ is nonsingular

Theorem (Asymptotic normality for M-estimator): Under assumptions 1,2,3,4,5,

$$\sqrt{n}(\theta_n - \theta_0) \xrightarrow{d} N(0, H_0^{-1}\Sigma H_0^{-1})$$

where $H_0 \equiv H(\theta_0)$.

Proof: (sketch) By Assumptions 1,2,3, $\nabla_{\theta_n} Q_n(\theta) = 0$ (this is FOC of maximization problem). Then using mean-value theorem (with $\bar{\theta}$ denoting mean value):

$$\begin{aligned} 0 &= \nabla_{\theta_n} Q_n(\theta) = \nabla_{\theta_0} Q_n(\theta) + \nabla_{\bar{\theta}\bar{\theta}} Q_n(\theta)(\theta_n - \theta_0) \\ \Rightarrow \underbrace{\nabla_{\bar{\theta}\bar{\theta}} Q_n(\theta)}_{\xrightarrow{p} H_0 \text{ (using A5)}} \sqrt{n}(\theta_n - \theta_0) &= - \underbrace{\sqrt{n} \nabla_{\theta_0} Q_n(\theta)}_{\xrightarrow{d} N(0, \Sigma) \text{ (using A4)}} \\ \Leftrightarrow \sqrt{n}(\theta_n - \theta_0) &\xrightarrow{d} -H(\theta_0)^{-1} N(0, \Sigma) = N(0, H_0^{-1}\Sigma H_0^{-1}). \quad \blacksquare \end{aligned}$$

Explore: ordinary least squares



2.1 Maximum likelihood estimation

The consistency of MLE can follow by application of the theorem above for consistency of M-estimators.

Essentially, as we noted above, what the consistency theorem showed above was that, for any M-estimator sequence θ_n :

$$\text{plim}_{n \rightarrow \infty} \theta_n = \text{argmax}_{\theta} Q_0(\theta).$$

For MLE, there is an argument due to Wald (1949), who shows that, in the i.i.d. case, the “limiting likelihood function” (corresponding to $Q_0(\theta)$) is indeed globally maximized at θ_0 , the “true value”. Thus, we can directly confirm the identification assumption of the M-estimator consistency theorem. This argument is of interest by itself.

Argument: (summary of Amemiya, pp. 141–142)

- Define $\hat{\theta}_n^{MLE} \equiv \text{argmax}_{\theta} \frac{1}{n} \sum_i \log f(x_i|\theta)$. Let θ_0 denote the true value.
- By LLN: $\frac{1}{n} \sum_i \log f(x_i|\theta) \xrightarrow{p} E_{\theta_0} \log f(x_i|\theta)$, for all θ (not necessarily the true θ_0).
- By Jensen’s inequality: $E_{\theta_0} \log \left(\frac{f(x|\theta)}{f(x|\theta_0)} \right) < \log E_{\theta_0} \left(\frac{f(x|\theta)}{f(x|\theta_0)} \right)$

- But $E_{\theta_0} \left(\frac{f(x|\theta)}{f(x|\theta_0)} \right) = \int \left(\frac{f(x|\theta)}{f(x|\theta_0)} \right) f(x|\theta_0) = 1$, since $f(x|\theta)$ is a density function, for all θ .¹
- Hence:

$$\begin{aligned}
 E_{\theta_0} \log \left(\frac{f(x|\theta)}{f(x|\theta_0)} \right) &< 0, \quad \forall \theta \\
 \implies E_{\theta_0} \log f(x|\theta) &< E_{\theta_0} \log f(x|\theta_0), \quad \forall \theta \\
 \implies E_{\theta_0} \log f(x|\theta) &\text{ is maximized at the true } \theta_0.
 \end{aligned}$$

- At this point, we have shown that the limiting likelihood function $Q_0(\theta)$ is globally maximized at the true value θ_0 : thus assumption 1 of the M-estimator consistency theorem above is satisfied. That consistency proof can then be used to show that the sequence of finite-sample θ_n 's converge in probability to θ_0 .



Now we consider another idea, **efficiency**, which can be thought of as the large-sample analogue of the “minimum variance” concept.

For the sequence of estimators W_n , suppose that

$$k(n)(W_n - \theta) \xrightarrow{d} N(0, \sigma^2)$$

where $k(n)$ is a polynomial in n . Then σ^2 is denoted the *asymptotic variance* of W_n .

In “usual” cases, $k(n) = \sqrt{n}$. For example, by the CLT, we know that $\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2)$. Hence, σ^2 is the asymptotic variance of the sample mean \bar{X}_n .

Definition 10.1.11: An estimator sequence W_n is *asymptotically efficient* for θ if

- $\sqrt{n}(W_n - \theta) \xrightarrow{d} N(0, v(\theta))$, where
- the asymptotic variance $v(\theta) = \frac{1}{E_{\theta_0} \left(\frac{\partial}{\partial \theta} \log f(X|\theta) \right)^2}$ (the CRLB)



By asymptotic normality result for M-estimator, we know what the asymptotic distribution for the MLE should be. However, it turns out given the information inequality, the MLE's asymptotic distribution can be further simplified.

¹In this step, note the importance of assumption A3 in CB, pg. 516. If x has support depending on θ , then it will not integrate to 1 for all θ .

Theorem 10.1.12: Asymptotic efficiency of MLE

Proof: (following Amemiya, pp. 143–144)

- $\hat{\theta}_n^{MLE}$ satisfies the FOC of the MLE problem:

$$0 = \frac{\partial \log L(\theta | \vec{X}_n)}{\partial \theta} \Big|_{\theta = \hat{\theta}_n^{MLE}}.$$

- Using the mean value theorem:

$$\begin{aligned} 0 &= \frac{\partial \log L(\theta | \vec{X}_n)}{\partial \theta} \Big|_{\theta = \theta_0} + \frac{\partial^2 \log L(\theta | \vec{X}_n)}{\partial \theta^2} \Big|_{\theta = \theta_n^*} \left(\hat{\theta}_n^{MLE} - \theta_0 \right) \\ \implies \sqrt{n} \left(\hat{\theta}_n - \theta_0 \right) &= \sqrt{n} \frac{-\frac{\partial \log L(\theta | \vec{X}_n)}{\partial \theta} \Big|_{\theta = \theta_0}}{\frac{\partial^2 \log L(\theta | \vec{X}_n)}{\partial \theta^2} \Big|_{\theta = \theta_n^*}} = \sqrt{n} \frac{-\frac{1}{n} \sum_i \frac{\partial \log f(x_i | \theta)}{\partial \theta} \Big|_{\theta = \theta_0}}{\frac{1}{n} \sum_i \frac{\partial^2 \log f(x_i | \theta)}{\partial \theta^2} \Big|_{\theta = \theta_n^*}} \quad (**) \end{aligned}$$

- Note that, by the LLN,

$$\frac{1}{n} \sum_i \frac{\partial \log f(x_i | \theta)}{\partial \theta} \Big|_{\theta = \theta_0} \xrightarrow{p} E_{\theta_0} \frac{\partial \log f(X | \theta)}{\partial \theta} \Big|_{\theta = \theta_0} = \int \frac{\partial f(x_i | \theta)}{\partial \theta} \Big|_{\theta = \theta_0} dx.$$

Using same argument as in the information inequality result above, the last term is:

$$\int \frac{\partial f}{\partial \theta} dx = \frac{\partial}{\partial \theta} \int f dx = 0.$$

- Hence, the CLT can be applied to the numerator of (**):

$$\text{numerator of (**)} \xrightarrow{d} N \left(0, E_{\theta_0} \left(\frac{\partial \log f(x_i | \theta)}{\partial \theta} \Big|_{\theta = \theta_0} \right)^2 \right).$$

- By LLN, and uniform convergence of Hessian term:

$$\text{denominator of (**)} \xrightarrow{p} E_{\theta_0} \frac{\partial^2 \log f(X | \theta)}{\partial \theta^2} \Big|_{\theta = \theta_0}.$$

- Hence, by Slutsky theorem:

$$\sqrt{n} \left(\hat{\theta}_n - \theta_0 \right) \xrightarrow{d} N \left(0, \frac{E_{\theta_0} \left(\frac{\partial \log f(x_i | \theta)}{\partial \theta} \Big|_{\theta = \theta_0} \right)^2}{\left[E_{\theta_0} \frac{\partial^2 \log f(X | \theta)}{\partial \theta^2} \Big|_{\theta = \theta_0} \right]^2} \right).$$

- By the information inequality:

$$E_{\theta_0} \left(\frac{\partial \log f(x_i|\theta)}{\partial \theta} \Big|_{\theta=\theta_0} \right)^2 = -E_{\theta_0} \frac{\partial^2 \log f(X|\theta)}{\partial \theta^2} \Big|_{\theta=\theta_0}$$

so that

$$\sqrt{n} \left(\hat{\theta}_n - \theta_0 \right) \xrightarrow{d} N \left(0, \frac{1}{E_{\theta_0} \left(\frac{\partial \log f(x_i|\theta)}{\partial \theta} \Big|_{\theta=\theta_0} \right)^2} \right)$$

so that the asymptotic variance is the CRLB.

Hence, the asymptotic approximation for the finite-sample distribution is

$$\hat{\theta}_n^{MLE} \stackrel{a}{\sim} N \left(\theta_0, \frac{1}{n} \frac{1}{E_{\theta_0} \left(\frac{\partial \log f(x_i|\theta)}{\partial \theta} \Big|_{\theta=\theta_0} \right)^2} \right).$$