Lecture 2: Point Estimation

Econ 205A: Econometric Methods I

Ruoyao Shi

Fall 2018

Contents

1 Methods of Finding Estimators

Estimators

- Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ be a random sample with a density $f(\mathbf{x}|\theta)$, which depends on a finite vector of unknown parameters $\theta \in \mathbb{R}^k$.
- Note that $\mathbf{X}_i$ can be a random variable or a random vector.
- **Definition 1** *An **estimator** $\hat{\theta}$ of θ is a measurable real function of the sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$; that is $\hat{\theta} = \hat{\theta}(\mathbf{X}_1, \ldots, \mathbf{X}_n)$.*
- Note that any statistic is an estimator.
- An *estimator* is a function of the sample; while an *estimate* is a realized value (a number) of the estimator.
- **Example 1** *Suppose the random sample is drawn from $\mathcal{N}(0, \sigma^2)$. An estimator of σ^2 could be $\bar{X}_n$. (Although not a good one!)*

Method of Moments (MM) Estimators

- **Definition 2** ***(Method of moments)*** *estimators are found by equating the first k sample moments to the corresponding k population moments, and solving the resulting system of simultaneous equations. More precisely, define*

$$\begin{aligned} m_1 &= \frac{1}{n}\sum_{i=1}^{n} X_i,\ \mu_1 = \mathbb{E}(X); \\ m_2 &= \frac{1}{n}\sum_{i=1}^{n} X_i^2,\ \mu_2 = \mathbb{E}(X^2); \\ &\vdots \\ m_k &= \frac{1}{n}\sum_{i=1}^{n} X_i^k,\ \mu_k = \mathbb{E}(X^k); \end{aligned}$$

  Suppose that the population moments μ_j are functions of $\theta_1, \ldots, \theta_k$, say $\mu_j(\theta_1, \ldots, \theta_k)$.[1] *The method of moments estimator $(\hat{\theta}_1, \ldots, \hat{\theta}_k)$ of $(\theta_1, \ldots, \theta_k)$ is obtained by solving the following system of equations for $(\theta_1, \ldots, \theta_k)$:*

$$\begin{aligned} m_1 &= \mu_1(\theta_1, \ldots, \theta_k), \\ m_2 &= \mu_2(\theta_1, \ldots, \theta_k), \\ &\vdots \\ m_k &= \mu_k(\theta_1, \ldots, \theta_k). \end{aligned}$$

- Because of LLN, MM estimators usually will be consistent. However, since the functions μ_j are not necessarily linear, MM estimators won't be unbiased in general.

[1] This condition ensures that the parameters $\theta_1, \ldots, \theta_k$ are **identified**, which might not always be the case. In general, identification is an important question in econometrics, and we will discuss in details later.

- **Example 2** *Suppose $X_1, \dots, X_n$ are i.i.d $\mathcal{N}(\mu, \sigma^2)$. Then the method of moments gives rise to the following system of equations:*

$$\begin{aligned} m_1 &= \bar{X}_n = \mu = \theta_1, \\ m_2 &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 = \mu^2 + \sigma^2 = \theta_2. \end{aligned}$$

Solving for (μ, σ^2) yields the method of moments estimators

$$\begin{aligned} \hat{\mu} &= \bar{X}_n, \\ \hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2. \end{aligned}$$

Note that $\hat{\sigma}^2$ is different from the sample variance S_n^2, the estimator of the population variance we have seen before. $\hat{\sigma}^2$ is consistent but biased, S_n^2 is consistent and unbiased.

- The moments used for MM estimator don't have to be the first k moments of X, it could be any moments whose link with the unknown parameters θ is some known functions.

- **Example 3** ***(Linear regression)*** *Consider the simple linear regression model*

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i,$$

and the classical assumptions that

$$\begin{aligned} \mathbb{E}(\epsilon_i) &= 0, \\ \mathbb{E}(\epsilon_i X_i) &= \mathbb{E}[\mathbb{E}(\epsilon_i | X_i) X_i] = 0. \end{aligned}$$

Since we know that ϵ_i can be written as $Y_i - \beta_0 - \beta X_i$, a function of the parameters and the data only, we can match the population version of these two moment conditions with their sample analog:

$$\begin{aligned} \mathbb{E}(\epsilon_i) = \mathbb{E}(Y_i - \beta_0 - \beta_1 X_i) &= n^{-1}\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i), \\ \mathbb{E}(\epsilon_i X_i) = \mathbb{E}[(Y_i - \beta_0 - \beta_1 X_i) X_i] &= n^{-1}\sum_{i=1}^{n}[(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) X_i]. \end{aligned}$$

By solving this system of two equations, we can get the MM estimators (also the familiar OLS estimators) for β:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X}_n)(Y_i - \bar{Y}_n)}{\sum_{i=1}^n (X_i - \bar{X}_n)^2},$$
$$\hat{\beta}_0 = \bar{Y}_n - \hat{\beta}_1 \bar{X}_n.$$

Maximum Likelihood Estimators (MLE)

- **Definition 3** *Let $X_1, \dots, X_n$ be a random sample from a population with pdf or pmf $f(x|\theta)$. Then the **likelihood function** is*

$$L(\theta|\mathbf{x}) = L(\theta_1, \dots, \theta_k|x_1, \dots, x_n) \equiv \Pi_{i=1}^n f(x_i|\theta_1, \dots, \theta_k).$$

- The likelihood function looks like a joint pdf or joint pmf, but they are different. $L(\theta|\mathbf{x})$ is a function of θ given the values of a sample $\mathbf{x}$, while the joint pdf $f(\mathbf{x}|\theta)$ is a function of $\mathbf{x}$ given particular parameter value θ.

- The likelihood is the probability that the observed sample would have been generated if the true parameter value were to be θ.

- **Example 4** *Suppose you have two cents, one is fair ($p = 1/2$) and the other is not ($p = 1/5$). You don't know which is which, but you decide to figure out by choosing one of them and flipping it* 1000 *times. Suppose after a while, you get* 201 *heads and* 799 *tails. If the coin was the fair one, then the probability of having obtained such a sample is $(\frac{1}{2})^{201} \cdot (\frac{1}{2})^{799} = 9.33e - 302$; if the coin was the unfair one, then the probability of having obtained such a sample is $(\frac{1}{5})^{201} \cdot (\frac{4}{5})^{799} = 1.19e - 218$. So **ex post**, the likelihood of having chosen an unfair coin to start with is* 1.28e83 *times higher than the likelihood of having chosen a fair coin in the first place. Therefore, given the sample, we are confident that the coin you chose was the unfair one.*

- **Definition 4** *For any $\mathbf{x} \equiv (x_1, \dots, x_n)'$, let $\hat{\theta}(\mathbf{x})$ denote the parameter value at which $L(\theta|\mathbf{x})$ attains its maximum; that is*

$$L(\hat{\theta}(\mathbf{x})|\mathbf{x}) = \sup_{\theta \in \Theta} L(\theta|\mathbf{x}).$$

  *$\hat{\theta}(\mathbf{x})$ is a function of θ with $\mathbf{x}$ held fixed. Then $\hat{\theta}(\mathbf{X})$ is called the **maximum likelihood estimator (MLE)** of the parameter θ.*

- **Example 5** *Let* $X_1, \ldots, X_n$ *be i.i.d. from Uniform*$[0, \theta]$, *where* $0 < \theta < \infty$ *is the unknown parameter. The likelihood function is*

$$L(\theta|x) = \begin{cases} (1/\theta)^n & \textit{if } \theta \geq \max_i X_i, \\ 0 & \textit{otherwise.} \end{cases}$$

Note that $(1/\theta)^n$ *is a positive, monotonically decreasing function of* θ, *so the MLE of* θ *is* $\max_i X_i$.

- More often than not, using the **log likelihood function** defined as:

$$l(\theta|\mathbf{x}) \equiv \log[L(\theta|\mathbf{x})] = \log\left[\Pi_{i=1}^n f(X_i|\theta)\right] = \sum_{i=1}^n \log f(X_i|\theta),$$

instead of the likelihood function is more convenient.

- **Example 6** *Let* $X_1, \ldots, X_n$ *be a random sample from a population with pdf*

$$f(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!},\ x = 0, 1, 2, \ldots$$

That is, X_i *is a Poisson random variable with mean and variance* λ. *The log likelihood function is*

$$L(\lambda|\mathbf{x}) = \log\lambda \sum_{i=1}^n x_i - n\lambda - \sum_{i=1}^n \log(x_i!),$$

which implies that

$$\frac{\partial l(\lambda|\mathbf{x})}{\partial\lambda} = \sum_{i=1}^n x_i/\lambda - n.$$

Setting this derivative to zero and solving gives $\hat{\lambda} = n^{-1}\sum_{i=1}^n x_i$. *And the SOC is*

$$\frac{\partial^2 l(\lambda|\mathbf{x})}{\partial\lambda^2} = -\sum_{i=1}^n x_i/\lambda^2 < 0.$$

So the MLE of λ *is* $\hat{\lambda} = n^{-1}\sum_{i=1}^n x_i$.

- In the above example, the SOC is to ensure that the solution to the FOC is the *glocal maximum.* Finding global maximum is crucial to MLE.

- **Example 7** *Let $X_1, \ldots, X_n$ be i.i.d. Bernoulli(p). Then the likelihood function is*

$$L(p|x) = \Pi_{i=1}^n p^{x_i}(1-p)^{1-x_i} = p^y(1-p)^{n-y},$$

where $y = \sum x_i$. And the log likelihood function is

$$l(p|x) = y \log p + (n-y)\log(1-p).$$

To find the MLE, first consider the case $0 < y < n$. The FOC gives rise to

$$\frac{dl(p|x)}{dp} = \frac{y}{\hat{p}} - \frac{n-y}{1-\hat{p}} = 0 \Rightarrow \hat{p} = \frac{y}{n},$$

and the SOC is

$$\frac{d^2 l(p|x)}{dp^2} = -\frac{y}{\hat{p}^2} - \frac{n-y}{(1-\hat{p})^2} < 0.$$

If $y = 0$, then $l(p|x) = n\log(1-p)$; if $y = n$, then $l(p|x) = n\log p$. In either case, $\hat{p} = y/n$. To summarize, the MLE of p is $\hat{p} = \sum_{i=1}^n X_i/n$.

- **Example 8** ***(Normal MLE)*** *Let $X_1, \ldots, X_n$ be a random sample from $\mathcal{N}(\mu, \sigma^2)$, with both μ and σ^2 unknown. Then the likelihood function is*

$$L(\mu, \sigma^2|\mathbf{x}) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\sum_{i=1}^n (x_i-\mu)^2/2\sigma^2},$$

and the log likelihood function is

$$l(\mu, \sigma^2|\mathbf{x}) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2}\sum_{i=1}^n (x_i-\mu)^2/\sigma^2.$$

The partial derivatives, with respect to μ and σ^2, are

$$\frac{\partial}{\partial\mu} l(\mu, \sigma^2|\mathbf{x}) = \frac{1}{\sigma^2}\sum_{i=1}^n (x_i-\mu),$$

$$\frac{\partial}{\partial\sigma^2} l(\mu, \sigma^2|\mathbf{x}) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^n (x_i-\mu)^2.$$

Setting these partial derivatives to zero and solving for μ and σ^2 yields the solution

$$\hat{\mu} = \bar{x}_n,$$
$$\hat{\sigma}^2 = n^{-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2.$$

Given these FOC and the fact that $\sum_{i=1}^{n}(x_i-\mu)^2$ is a convex function in μ, we can conclude that for any $\mu \neq \bar{x}_n$, $\sum_{i=1}^{n}(x_i-\mu)^2 > \sum_{i=1}^{n}(x_i-\bar{x}_n)^2$. As a result, for any value of $\sigma^2 > 0$, we have

$$\frac{1}{(2\pi\sigma^2)^{n/2}}e^{-\sum_{i=1}^{n}(x_i-\bar{x}_n)^2/2\sigma^2} \geq \frac{1}{(2\pi\sigma^2)^{n/2}}e^{-\sum_{i=1}^{n}(x_i-\mu)^2/2\sigma^2}.$$

Thus we have shown that $\bar{x}_n$ is indeed a global maximum. Plugging it into the FOC with respect to σ^2, we get

$$\frac{\partial}{\partial\sigma^2}l(\sigma^2|\mathbf{x}, \mu = \bar{x}_n) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2,$$

which is non-negative if $\sigma^2 < \hat{\sigma}^2$ and negative if $\sigma^2 > \hat{\sigma}^2$. Thus we show that $\hat{\sigma}^2$ is the global maximum.

Important Properties of Maximum Likelihood Estimators

- Let $X_1, \ldots, X_n$ be a random sample from a population with a density $f(x|\theta)$. The log likelihood function is

  $$l(\theta|x) = \log L(\theta|x) = \log \Pi_{i=1}^{n} f(x_i|\theta) = \sum_{i=1}^{n}\log f(x_i|\theta).$$

  It implies that for any θ,

  $$\int_{\mathbb{R}^n} L(\theta|\mathbf{x})d\mathbf{x} = 1.$$

  If integral and derivative can be exchanged, this further implies that

  $$\frac{\partial}{\partial\theta}\int_{\mathbb{R}^n} L(\theta|\mathbf{x})d\mathbf{x} = \int_{\mathbb{R}^n}\frac{\partial L(\theta|\mathbf{x})}{\partial\theta}d\mathbf{x} = \int_{\mathbb{R}^n}\frac{\partial L(\theta|\mathbf{x})}{\partial\theta}\frac{L(\theta|\mathbf{x})}{L(\theta|\mathbf{x})}d\mathbf{x}$$
  $$= \int_{\mathbb{R}^n}\frac{\partial l(\theta|\mathbf{x})}{\partial\theta}L(\theta|\mathbf{x})d\mathbf{x} = 0.$$

Define the **score function**

$$S(\theta|\mathbf{X}) \equiv \frac{\partial l(\theta|\mathbf{x})}{\partial\theta} = \frac{\partial}{\partial\theta}\left[\sum_{i=1}^{n}\log f(x_i|\theta)\right] = \sum_{i=1}^{n}\left(\frac{\partial \log f(x_i|\theta)}{\partial\theta}\right).$$

Then the above equation states that the expectation of the score function is zero. That is,

$$\mathbb{E}[S(\theta|\mathbf{X})] = \int_{\mathbb{R}^n}\frac{\partial l(\theta|\mathbf{x})}{\partial\theta}L(\theta|\mathbf{x})d\mathbf{x} = 0.$$

- Differentiate both sides of this equation with respect to θ, we get

$$\begin{aligned}
\frac{\partial}{\partial\theta}\int_{\mathbb{R}^n}\frac{\partial l(\theta|\mathbf{x})}{\partial\theta}L(\theta|\mathbf{x})d\mathbf{x} &= \int_{\mathbb{R}^n}\frac{\partial^2 l(\theta|\mathbf{x})}{\partial\theta^2}L(\theta|\mathbf{x})d\mathbf{x} + \int_{\mathbb{R}^n}\frac{\partial l(\theta|\mathbf{x})}{\partial\theta}\frac{\partial L(\theta|\mathbf{x})}{\partial\theta}d\mathbf{x} \\
&= \int_{\mathbb{R}^n}\frac{\partial^2 l(\theta|\mathbf{x})}{\partial\theta^2}L(\theta|\mathbf{x})d\mathbf{x} + \int_{\mathbb{R}^n}\left(\frac{\partial l(\theta|\mathbf{x})}{\partial\theta}\right)^2 L(\theta|\mathbf{x})d\mathbf{x} \\
&= \mathbb{E}\left[\frac{\partial^2 l(\theta|\mathbf{x})}{\partial\theta^2}\right] + \mathbb{E}\left[\left(\frac{\partial l(\theta|\mathbf{x})}{\partial\theta}\right)^2\right] \\
&= 0.
\end{aligned}$$

This equation is called **information matrix equality**:

$$\mathbb{E}\left[\frac{\partial^2 l(\theta|\mathbf{x})}{\partial\theta^2}\right] + \mathbb{E}\left[\left(\frac{\partial l(\theta|\mathbf{x})}{\partial\theta}\right)^2\right] = 0$$

Note that $\mathbb{E}\left[\left(\frac{\partial l(\theta|\mathbf{x})}{\partial\theta}\right)^2\right] = \mathbb{E}[(S(\theta|\mathbf{x}))^2] = var[S(\theta|\mathbf{x}))]$.

- Under i.i.d., the above equation implies

$$\begin{aligned}
\mathbb{E}\left[\left(\frac{\partial l(\theta|\mathbf{x})}{\partial\theta}\right)^2\right] &= \mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log f(\mathbf{X}|\theta)\right)^2\right] \\
&= \mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log\sum_{i=1}^{n} f(X_i|\theta)\right)^2\right] \\
&= \sum_{i=1}^{n}\mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log f(X_i|\theta)\right)^2\right] \\
&= n\mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log f(X_i|\theta)\right)^2\right],
\end{aligned}$$

  where $I(\theta) \equiv \mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log f(X_i|\theta)\right)^2\right]$ is called **information matrix**.

- $\left[\frac{\partial^2 \log f(X_i|\theta)}{\partial\theta^2}\right]$ is called **Hessian matrix**. And define $H(\theta) \equiv \mathbb{E}\left[\frac{\partial^2 \log f(X_i|\theta)}{\partial\theta^2}\right]$.

Invariance Property of MLE

- **Theorem 1** *If $\hat{\theta}$ is the MLE of θ, and $g(\theta)$ is a one-to-one function over Θ. Then $g(\hat{\theta})$ is an MLE of $g(\theta)$.*

2 Methods of Evaluating Estimators

Mean Squared Error (MSE)

- **Definition 5** *The* ***bias*** *of an estimator $\hat{\theta}_n$ of parameter θ is $b(\hat{\theta}_n) \equiv \mathbb{E}(\hat{\theta}_n) - \theta$.*

- **Example 9** *In Example 1, the bias of $\bar{X}_n$ as an estimator of σ^2 is $\mathbb{E}(\bar{X}_n) - \sigma^2 = \mu - \sigma^2 = 0 - \sigma^2 = -\sigma^2$. Its bias as an estimator of μ is zero.*

- **Definition 6** *An estimator $\hat{\theta}_n$ of a parameter θ is* ***unbiased*** *if $\mathbb{E}(\hat{\theta}_n) = \theta$.*

- **Definition 7** *The **mean squared error (MSE)** of an estimator* $\hat{\theta}_n$ *of parameter* θ *is* $MSE(\hat{\theta}_n) \equiv \mathbb{E}[(\hat{\theta}_n - \theta)^2]$.

- To better understand MSE, consider

$$\begin{aligned}
MSE(\hat{\theta}_n) &= \mathbb{E}[(\hat{\theta}_n - \theta)^2] \\
&= \mathbb{E}[(\hat{\theta}_n - \mathbb{E}(\hat{\theta}_n) + \mathbb{E}(\hat{\theta}_n) - \theta)^2] \\
&= \mathbb{E}[(\hat{\theta}_n - \mathbb{E}(\hat{\theta}_n))^2] + \mathbb{E}[(\mathbb{E}(\hat{\theta}_n) - \theta)^2] + 2\mathbb{E}[(\hat{\theta}_n - \mathbb{E}(\hat{\theta}_n))(\mathbb{E}(\hat{\theta}_n) - \theta)] \\
&= var(\hat{\theta}_n) + (\mathbb{E}(\hat{\theta}_n) - \theta)^2 + 2(\mathbb{E}(\hat{\theta}_n) - \theta)\mathbb{E}[\hat{\theta}_n - \mathbb{E}(\hat{\theta}_n)] \\
&= var(\hat{\theta}_n) + (b(\hat{\theta}_n))^2 + 0 \\
&= var(\hat{\theta}_n) + (b(\hat{\theta}_n))^2;
\end{aligned}$$

  this is the bias-variance decomposition of MSE.

- **Example 10** ***(Normal MSE)*** *Let* $X_1, \ldots, X_n$ *be a random sample from* $\mathcal{N}(\mu, \sigma^2)$. *We know that sample mean and sample variance are unbiased estimators since*

$$\mathbb{E}(\bar{X}_n) = \mu, \text{ and } \mathbb{E}(S_n^2) = \sigma^2 \text{ for all } \mu \text{ and } \sigma^2.$$

  So the MSE of these estimators are (you may want to prove them yourselves)

$$\begin{aligned}
\mathbb{E}[(\bar{X}_n - \mu)^2] &= var(\bar{X}_n) = \frac{\sigma^2}{n}, \\
\mathbb{E}[(S_n^2 - \sigma^2)^2] &= var(S_n^2) = \frac{2\sigma^4}{n-1}.
\end{aligned}$$

  On the other hand, the method of moments estimator (also MLE) of σ^2 *is biased:*

$$\mathbb{E}(\hat{\sigma}^2) = \mathbb{E}\left(\frac{n-1}{n} S_n^2\right) = \frac{n-1}{n}\sigma^2;$$

  but its variance is smaller than that of S_n^2:

$$var(\hat{\sigma}^2) = \left(\frac{n-1}{n}\right)^2 var(S_n^2) = \frac{2(n-1)\sigma^4}{n^2}.$$

  As a result, the MSE of $\hat{\sigma}^2$ *is*

$$MSE(\hat{\sigma}^2) = \frac{(2n-1)\sigma^4}{n^2} < \frac{2\sigma^4}{n-1} = MSE(S_n^2).$$

To summarize, the MLE estimator $\hat{\sigma}^2$ is biases, but has smaller MSE than sample variance S_n^2.

- The estimator that attains the Cramér-Rao lower bound is called **efficient** estimator. The Cramér-Rao lower bound is also referred to as (finite sample parametric)[2] **efficiency bound**.

Best Unbiased Estimators

- **Definition 8** *An estimator $\hat{\theta}_n$ of θ is called a* ***uniformly minimum variance unbiased estimator (UMVUE)*** *(also called* ***best unbiased estimator*** *or* ***efficient estimator****) if $\hat{\theta}_n$ is unbiased and for any other unbiased estimator $\tilde{\theta}_n$ and any true unknown parameter value θ, we have $var_\theta(\hat{\theta}_n) \leq var_\theta(\tilde{\theta}_n)$, where the subscript $_\theta$ emphasizes that the variance is taken under the true parameter value θ.*

- Find the estimator whose MSE is the smallest among all possible estimators is sometimes a difficult task. But find the estimator whose variance is the smallest among all unbiased estimators might be easier.

- How small the variance can be? The following theorem gives a lower bound.

- **Theorem 2** ***(Cramér-Rao Inequality)*** *Let $X_1, \ldots, X_n$ be a random sample from a population with pdf $f(\mathbf{x}|\theta)$, and let $l(\theta|\mathbf{x})$ denote the log likelihood function. If $\hat{\theta}_n$ is an unbiased estimator of θ, then under some regularity conditions*[3]

$$var(\hat{\theta}_n) \geq \frac{1}{var[S(\theta|\mathbf{X})]},$$

where $S(\theta|\mathbf{X})$ is the score function.

- In the previous section, we have shown that under i.i.d.,

$$var[S(\theta|\mathbf{X})] = nI(\theta) = \mathbb{E}\left[\left(\frac{\partial l(\theta|\mathbf{x})}{\partial \theta}\right)^2\right] = -\mathbb{E}\left[\frac{\partial^2 l(\theta|\mathbf{x})}{\partial \theta^2}\right],$$

[2] At this moment, don't let these qualifiers bother you.

[3] We ellaborate them in the proof of the theorem later.

so

$$(var[S(\theta|\mathbf{X})])^{-1} = (nI(\theta))^{-1} = \left(\mathbb{E}\left[\left(\frac{\partial l(\theta|\mathbf{x})}{\partial\theta}\right)^2\right]\right)^{-1} = \left(-\mathbb{E}\left[\frac{\partial^2 l(\theta|\mathbf{x})}{\partial\theta^2}\right]\right)^{-1}$$

is called the **Cramér-Rao (Lower) Bound (CRLB)** for the unbiased estimator $\hat{\theta}_n$.

- To prove Theorem 2, we need to use the following corollary of the Cauchy-Schwarz inequality.

- **Lemma 1** *Let X and Y be random variables, then*

$$var(Y)var(X) \geq [cov(X,Y)]^2.$$

- *Proof of Theorem 2.* Assume the following regularity conditions hold: (1) the log likelihood function is twice differentiable with respect to θ; and (2) the integral and the differential are exchangeable. Since $\hat{\theta}_n$ is unbiased, i.e.

$$\theta = \mathbb{E}(\hat{\theta}_n) = \int_{\mathbb{R}^n} \hat{\theta}_n L(\theta|\mathbf{x}) d\mathbf{x}.$$

This implies that

$$\begin{aligned} 1 &= \frac{\partial\theta}{\partial\theta} \\ &= \int_{\mathbb{R}^n} \hat{\theta}_n \frac{\partial L(\theta|\mathbf{x})}{\partial\theta} d\mathbf{x} \\ &= \int_{\mathbb{R}^n} \hat{\theta}_n \frac{\partial l(\theta|\mathbf{x})}{\partial\theta} L(\theta|\mathbf{x}) d\mathbf{x} \\ &= \int_{\mathbb{R}^n} \hat{\theta}_n S(\theta|\mathbf{x}) L(\theta|\mathbf{x}) d\mathbf{x} \\ &= \mathbb{E}[\hat{\theta}_n S(\theta|\mathbf{X})] \\ &= cov(\hat{\theta}_n, S(\theta|\mathbf{X})), \end{aligned}$$

where the last equality holds since $\mathbb{E}[S(\theta|\mathbf{x})] = 0$ and $\hat{\theta}_n$, which is a function of the data only, is unbiased for θ. By the Cauchy-Schwarz inequality, we have

$$1^2 = [cov(\hat{\theta}_n, S(\theta|\mathbf{x}))]^2 \leq var(\hat{\theta}_n) \cdot var(S(\theta|\mathbf{x})),$$

which implies the result of the theorem.

- **Example 11** ***(Bernoulli MLE)*** *Let $X_1, \ldots, X_n$ be a random sample from a population with pmf*

$$f(x|p) = p^x(1-p)^{1-x},$$

where $0 \leq p \leq 1$. The MLE (also the method of moments estimator) of p is $\bar{X}_n$. Note that it is unbiased, and

$$var(\bar{X}_n) = var(X_i)/n = p(1-p)/n.$$

Now consider the variance of the score function

$$\begin{aligned} var[S(p|\mathbf{X})] &= -\mathbb{E}\left[\frac{\partial^2}{\partial p^2}\sum_{i=1}^n \log f(x_i|p)\right] \\ &= \mathbb{E}\left[\frac{\sum_{i=1}^n X_i}{p^2} + \frac{n - \sum_{i=1}^n X_i}{(1-p)^2}\right] \\ &= \frac{np}{p^2} + \frac{n-np}{(1-p)^2} \\ &= \frac{n}{p} + \frac{n}{(1-p)} \\ &= \frac{n}{p(1-p)}. \end{aligned}$$

So the CRLB is $p(1-p)/n$, and the MLE is efficient.

- **Definition 9** *Let $\mathbf{X} \equiv (X_1, \ldots, X_n)'$ be a random sample. $\hat{\theta}$ is called a* ***linear estimator*** *of θ if $\hat{\theta} = \mathbf{a}'\mathbf{X}$ where $\mathbf{a} \in \mathbb{R}^n$ is a vector of* constants.

- **Example 12** *Sample mean $\bar{X}$ is a linear estimator, while sample variance S^2 is not.*

- Note that the *constants* in the above definition means that a does not depend on the parameter value. It could depend on the sample in more general cases. For example, $\hat{\beta} \equiv (X'X)^{-1}X'\mathbf{y}$, the ordinary least squared (OLS) estimator in the linear regression model is a linear estimator where $\mathbf{a} = (X'X)^{-1}X'$, and $\hat{\beta}$ is a linear combination of $\mathbf{y}$.

- **Definition 10** *If $\hat{\theta}$ is a linear and unbiased estimator and $var(\tilde{\theta}) \geq var(\hat{\theta})$ for any other linear and unbiased estimator $\tilde{\theta}$, then $\hat{\theta}$ is called the* ***best linear unbiased estimator (BLUE)*** *of θ.*

- **Example 13** *The OLS estimator is BLUE.*[4]

Loss Function Optimality

- Mean squared error is a special example of a **loss function**. In point estimation problems, if the estimator $\hat{\theta}$ is "close" to the unknown parameter θ, then the loss should be small; if $\hat{\theta}$ is "far" from θ, then the loss should be large. The "closeness" is measured by a loss function. In other words, a loss function is a non-negative function that increases as the distance between $\hat{\theta}$ and θ increases.

- Two commonly used loss functions are:

$$\text{absolute error loss: } L(\theta, \hat{\theta}) \equiv |\hat{\theta} - \theta|;$$
$$\text{squared error loss: } L(\theta, \hat{\theta}) \equiv (\hat{\theta} - \theta)^2.$$

- The quality of an estimator is quantified by its **risk function**, the expectation of the loss function. That is,

$$R(\theta, \hat{\theta}) \equiv \mathbb{E}_\theta[L(\theta, \hat{\theta})].$$

- Note that the above expectation is taken under the true unknown parameter value θ, with respect to the randomness in $\hat{\theta}$, since $\hat{\theta}$ is a function of the sample.

- For the squared error loss function, the risk function is just the MSE.

- Risk functions generally depend on the value of θ. Since θ is unknown, the ideal case is that we can find an estimator that minimizes the risk regardless of the value of θ. An example is $\bar{X}$ for μ in normal samples. However, there not always exists such estimator.

- Minimizing loss functions can often give rise to estimators. For example, minimizing the absolute errors gives the **least absolute deviation (LAD)** estimator; minimizing the squared errors leads to the **least squares (LS)** estimator.[5]

[4]We will learn this later.
[5]We will discuss both in details later.

3 Exercises

1. Let $X_1, \ldots, X_n$ be i.i.d. binomial(k, p), and assume that both k and p are unknown. Find the method of moments estimator for them by matching the first two moments.

2. Let $X_1, \ldots, X_n$ be i.i.d. $\mathcal{N}(\theta, 1)$, find the MLE of θ. Remember to verify that you find the global maximum.

3. Let Y be a discrete (geometric) random variable with pdf

$$f(y) = p(1-p)^{y-1},\ y = 1, 2, 3, \ldots$$

   where p is an unknown parameter. Find the MLE of p if only one sample observation is available.

4. In the setting of Example 8, show that the Hessian evaluated at MLE equals

$$\left.\begin{bmatrix} \frac{\partial^2 l(\mu,\sigma^2|\mathbf{x})}{\partial \mu^2} & \frac{\partial^2 l(\mu,\sigma^2|\mathbf{x})}{\partial \mu \partial \sigma^2} \\ \frac{\partial^2 l(\mu,\sigma^2|\mathbf{x})}{\partial \mu \partial \sigma^2} & \frac{\partial^2 l(\mu,\sigma^2|\mathbf{x})}{\partial (\sigma^2)^2} \end{bmatrix}\right|_{\mu=\bar{x}_n, \sigma^2=\hat{\sigma}^2} = \begin{bmatrix} -\frac{n}{\hat{\sigma}^2} & 0 \\ 0 & -\frac{n}{2\hat{\sigma}^4} \end{bmatrix},$$

   and that it is negative definite.

5. Let $X_1, \ldots, X_n$ be a random sample from $\mathcal{N}(\mu, 1)$ population, where μ is unknown and non-negative. Find the MLE of μ under the constraint $\mu \geq 0$.

6. Let $X_1, \ldots, X_n$ be a random sample from a population with pmf

$$P_\theta(X = x) = \theta^x (1-\theta)^{1-x},\ x = 0 \text{ or } 1,\ 0 \leq \theta \leq \frac{1}{2}.$$

   (a) Find the method of moments estimator and MLE of θ;

   (b) Find the mean squared errors of each estimator;

   (c) Which estimator do you prefer? Justify your choice.

7. Let $X_1, \ldots, X_n$ be a random sample from $\mathcal{N}(\mu, \sigma^2)$ population. The normal pdf satisfies the regularity conditions for Theorem 2.

   (a) Find the Cramér-Rao lower bound for the unbiased estimators of σ^2;

(b) Show that sample variance S_n^2 does not attain the Cramér-Rao lower bound;

(c) Show that if μ is known, then the MLE of σ^2 attains the bound.