Lecture 6

Efficient estimators. Rao-Cramer bound.

1 Common methods for constructing an estimator

1.1 Method of Analogy (plug-in)

A method of analogy is another name for the plug-in estimator we have seen before. If we are interested in estimating $\theta = \theta(F)$ where F denotes the population distribution, we can estimate θ by $\hat{\theta} = \theta(\hat{F})$ where $\hat{F}$ is some estimator of F. We have seen a number of plug-in estimators. For example, $\mu = EX_i = \int x dF(x)$ is a functional of the cdf. An analog estimator is $\hat{\mu} = \int x d\hat{F}(x) = \bar{X}$. Another example: if we wish to estimate $\theta = P\{X_i \in A\}$ we may use $\hat{\theta} = \frac{1}{n}\sum_{i=1}^n \mathbb{I}\{X_i \in A\}$

1.2 Method of Moments

Let $X_1, ..., X_n$ be a random sample from some distribution. Suppose that the k-dimensional parameter of interest θ satisfies the system of equations $E[X_i] = m_1(\theta)$, $E[X_i^2] = m_2(\theta)$,..., $E[X_i^k] = m_k(\theta)$ where $m_1, ..., m_k$ are some known functions. Then the method-of-moments estimator $\hat{\theta}_{MM}$ of θ is the solution of the above system of equations when we substitute $\sum_{i=1}^n X_i/n$, $\sum_{i=1}^n X_i^2/n$,..., $\sum_{i=1}^n X_i^k/n$ for $E[X_i]$, $E[X_i^2]$,..., $E[X_i^k]$ correspondingly. In other words $\hat{\theta}_{MM}$ solves the following system of equations: $\sum_{i=1}^n X_i/n = m_1(\hat{\theta})$, $\sum_{i=1}^n X_i^2/n = m_2(\hat{\theta})$,..., $\sum_{i=1}^n X_i^k/n = m_k(\hat{\theta})$. We implicitly assume here that the solution exists and is unique.

Example Let $X_1, ..., X_n$ be a random sample from $N(\mu, \sigma^2)$. Then $E[X_i] = \mu$ and $E[X_i^2] = \mu^2 + \sigma^2$. Thus, $\hat{\mu}_{MM} = \sum_{i=1}^n X_i/n$ and $\hat{\mu}_{MM}^2 + \hat{\sigma}_{MM}^2 = \sum_{i=1}^n X_i^2/n$. So $\hat{\sigma}_{MM}^2 = \sum_{i=1}^n X_i^2/n - (\sum_{i=1}^n X_i/n)^2$.

Example Let $X_1, ..., X_n$ be a random sample from an exponential distribution:

$$f(x; \lambda) = \lambda e^{-\lambda x} \mathbb{I}\{x > 0\}.$$

One can calculate that $EX_i = \frac{1}{\lambda}$. So one suggestion for an estimator is a solution to

$$\bar{X} = \frac{1}{\hat{\lambda}}$$

or

$$\hat{\lambda} = \frac{1}{\bar{X}}.$$

We may use higher moments as well. For example

$$EX_i^2 = \frac{2}{\lambda^2}.$$

As such, another method-of-moments estimator for λ is:

$$\hat{\lambda} = \sqrt{\frac{2}{\frac{1}{n}\sum_{i=1}^{n} X_i^2}}$$

So, the method of moments estimator is not unique and depends on the moments chosen. One can easily prove that if function $m(\theta)$ has a unique inverse which is continuous, then the method-of-moments estimator is consistent (do this for your own practice! Hint: use the continuous mapping theorem). And if the inverse is continuously differentiable, we can use the delta method and prove asymptotic gaussianity (try to do this as well).

The idea of the method-of-moments is a very old one. There is a generalization of it which allows for more moments than the dimensionality of the parameter and also allows for the data and parameter to be mixed up within the moment condition. It is called a GMM (Generalized Method of Moments) and will be studied extensively later on, as the main workhorse of Econometrics.

1.3 Maximum Likelihood Estimator

In this section we consider parametric estimation. We have a parametric estimation problem when we know the distribution of the data up to a finite-dimensional parameter θ (the only unknown part). We denote the joint pdf of $X = (X_1, ..., X_n)$ as $f(x|\theta) = f(x_1, ..., x_n|\theta)$ (for i.i.d. sample we will have$f(x|\theta) = \prod_{i=1}^{n} f_1(x_i|\theta)$, where $f_1(x_i|\theta)$ is the pdf of one observation). That is, if we knew θ we have known the exact distribution of the data.

Let $x = (x_1, ..., x_n)$ denote the realization of $X = (X_1, ..., X_n)$. By definition, the maximum likelihood estimator $\hat{\theta}_{ML}$ of θ is the value that maximizes $f(x|\theta)$, i.e.

$$\hat{\theta}_{ML} = \arg\max_{\theta \in \Theta} f(x_1, ..., x_n|\theta).$$

The function $f(x|\theta)$, when considered as a function of θ for fixed values $x = (x_1, ..., x_n)$, is called the likelihood function. It is usually denoted by $\mathcal{L}(\theta|x)$. Thus, the maximum likelihood estimator maximizes the likelihood function, which explains the name of this estimator. Since $\log(x)$ is increasing in x, it is easy to see that $\hat{\theta}_{ML}$ also maximizes $\ell(\theta|x_1, ..., x_n) = \log \mathcal{L}(\theta|x_1, ..., x_n)$. Function $\ell(\theta|x_1, ..., x_n)$ is called the log-likelihood. If $\ell(\theta|x_1, ..., x_n)$ is differentiable in θ, then $\hat{\theta}_{ML}$ satisfies first order condition (FOC): $\frac{d\ell}{d\theta}(\hat{\theta}_{ML}|x_1, ..., x_n) = 0$. If the data comes from an i.i.d. sample then it is equivalent to $\sum_{i=1}^{n} \partial \log f_1(x_i|\hat{\theta}_{ML})/\partial\theta = 0$. The reason we took the log of the likelihood function now can be seen: it is easier to take the derivative of the sum than the derivative of the product. Function $S(\theta|x) = \partial \log f(x|\theta)/\partial\theta$ is called the score. Thus, $\hat{\theta}_{ML}$ solves the

equation $S(\theta|x) = 0$.

Example Let $X_1, ...X_n$ be a random sample from $N(\mu, \sigma^2)$. Then

$$\log f_1(\theta|x_i) = -\log\sqrt{2\pi} - (1/2)\log\sigma^2 - (x_i - \mu)^2/(2\sigma^2),$$

where $\theta = (\mu, \sigma^2)$. So

$$\ell(\theta|x_1, ..., x_n) = -n\log\sqrt{2\pi} - (n/2)\log\sigma^2 - \sum_{i=1}^{n}(x_i - \mu)^2/(2\sigma^2)$$

FOCs are

$$\partial\ell/\partial\mu = \sum_{i=1}^{n}(x_i - \mu)/\sigma^2 = 0,$$

$$\partial\ell/\partial\sigma^2 = -n/(2\sigma^2) + \sum_{i=1}^{n}(x_i - \mu)^2/(2\sigma^4) = 0.$$

So $\hat{\mu}_{ML} = \overline{X}_n$ and $\hat{\sigma}^2_{ML} = \sum_{i=1}^{n}(X_i - \overline{X}_n)^2/n$.

Example As another example, let $X_1, ..., X_n$ be a random sample from $U[0, \theta]$. Then $f_1(x_i|\theta) = 1/\theta$ if $x \in [0, \theta]$ and 0 otherwise. So $f(x_1, ...x_n|\theta) = 1/\theta^n$ if $0 \leq x_{(1)} \leq x_{(n)} \leq \theta$ and 0 otherwise. Thus, $\mathcal{L}(\theta|x) = (1/\theta^n)I\{\theta \geq x_{(n)}\}I\{x_{(1)} \geq 0\}$. We conclude that $\hat{\theta}_{ML} = X_{(n)}$.

2 Fisher information

Let $f(x|\theta)$ with $\theta \in \Theta$ be some parametric family. For given $\theta \in \Theta$, let $Supp_\theta = \{x : f(x|\theta) > 0\}$. $Supp_\theta$ is usually called the support of distribution $f(x|\theta)$. Assume that $Supp_\theta$ does not depend on θ. As before, $\ell(\theta|x) = \log f(x|\theta)$ is called the log-likelihood function. Assume that $\ell(\theta|x)$ is twice continuously differentiable in θ for all $x \in Supp$ and $\frac{\partial^2\ell(\theta|x)}{\partial\theta^2}$ is bounded above by some function $g(x)$ such that $Eg(X) < \infty$ for random variable X with distribution $f(x|\theta)$. Then:

Definition 1. $I(\theta) = E_\theta[(\partial\ell(\theta|X)/\partial\theta)^2]$ is called Fisher information.

Fisher information plays an important role in maximum likelihood estimation. The theorem below gives two information equalities:

Theorem 2. *In the setting above,*

(1) $E_\theta[\partial\ell(\theta|X)/\partial\theta] = 0$

(2) $I(\theta) = -E_\theta[\partial^2\ell(\theta|X)/\partial\theta^2]$.

Proof. Since $\ell(\theta|x)$ is twice differentiable in θ, $f(x|\theta)$ is twice differentiable in θ as well. Let us differentiate the following identity with respect to θ:

$$\int f(x|\theta)dx \equiv 1.$$

The restrictions on dominance by $g(x)$ allow us to interchange differentiation and integration signs below:

$$\int \frac{\partial f(x|\theta)}{\partial \theta} dx = 0 \text{ for all } \theta \in \Theta.$$

The second differentiation with respect to θ yields

$$\int \frac{\partial^2 f(x|\theta)}{\partial \theta^2} dx = 0 \text{ for all } \theta \in \Theta. \tag{1}$$

Now notice that

$$\frac{\partial \ell(\theta|x)}{\partial \theta} = \frac{\partial \log f(x|\theta)}{\partial \theta} = \frac{1}{f(x|\theta)} \frac{\partial f(x|\theta)}{\partial \theta}$$

and

$$\frac{\partial^2 \ell(\theta|x)}{\partial \theta^2} = -\frac{1}{f^2(x|\theta)} \left(\frac{\partial f(x|\theta)}{\partial \theta} \right)^2 + \frac{1}{f(x|\theta)} \frac{\partial^2 f(x|\theta)}{\partial \theta^2}.$$

The former equality yields

$$E_\theta \left[\frac{\partial \ell(\theta|X)}{\partial \theta} \right] = E_\theta \left[\frac{1}{f(X|\theta)} \frac{\partial f(X|\theta)}{\partial \theta} \right] = \int \frac{1}{f(x|\theta)} \frac{\partial f(x|\theta)}{\partial \theta} f(x|\theta) dx = \int \frac{\partial f(x|\theta)}{\partial \theta} dx = 0,$$

which is our first result. The latter equality yields

$$E_\theta \left[\frac{\partial^2 \ell(X, \theta)}{\partial \theta^2} \right] = -\int \frac{1}{f(x|\theta)} \left(\frac{\partial f(x|\theta)}{\partial \theta} \right)^2 dx,$$

here the second term disappears due to equation (1). So,

$$\begin{aligned} I(\theta) = \quad & E_\theta \left[\left(\tfrac{\partial \ell(\theta|X)}{\partial \theta} \right)^2 \right] = \quad \int \left(\frac{1}{f(x|\theta)} \frac{\partial f(x|\theta)}{\partial \theta} \right)^2 f(x|\theta) dx \\ = \quad & \int \tfrac{1}{f(x|\theta)} \left(\tfrac{\partial f(x|\theta)}{\partial \theta} \right)^2 dx = \quad -E_\theta \left[\frac{\partial^2 \ell(\theta|X)}{\partial \theta^2} \right]. \end{aligned}$$

□

Example Let us calculate Fisher information for one random draw from the $N(\mu, \sigma^2)$ distribution where σ^2 is known. Thus, our parameter $\theta = \mu$. The density of a normal distribution is $f(x|\mu) = \exp(-(x-\mu)^2/(2\sigma^2))/\sqrt{2\pi\sigma^2}$. The log-likelihood is $\ell(\mu|x) = -\log(2\pi\sigma^2)/2 - (x-\mu)^2/(2\sigma^2)$. So $\partial \ell(\mu|x)/\partial \mu = (x-\mu)/\sigma^2$ and $\partial^2 \ell(\mu|x)/\partial \mu^2 = -1/\sigma^2$. So $-E_\theta[\partial^2 \ell(\mu|X)/\partial \mu^2] = 1/\sigma^2$. At the same time,

$$I(\theta) = E_\mu[(\partial \ell(\mu|X)/\partial \mu)^2] = E_\mu[(X-\mu)^2/\sigma^4] = 1/\sigma^2.$$

So, as was expected in view of the theorem above, $I(\theta) = -E_\mu[\partial^2 \ell(\mu|X)/\partial \mu^2]$ in this example.

Example Let us calculate Fisher information for one random draw from a Bernoulli(θ) distribution. Note that a Bernoulli distribution is discrete. So we use a probability mass function (pms) instead of a pdf. The

pms of Bernoulli(θ) is $f(x|\theta) = \theta^x(1-\theta)^{1-x}$ for $x \in \{0,1\}$. The log-likelihood is $\ell(\theta|x) = x\log\theta + (1-x)\log(1-\theta)$. So $\partial\ell(\theta|x)/\partial\theta = x/\theta - (1-x)/(1-\theta)$ and $\partial^2\ell(\theta|x)/\partial\theta^2 = -x/\theta^2 - (1-x)/(1-\theta)^2$. So

$$\begin{aligned} E_\theta[(\partial\ell(\theta|X)/\partial\theta)^2] &= E_\theta[(X/\theta - (1-X)/(1-\theta))^2] \\ &= E_\theta[X^2/\theta^2] - 2E_\theta[X(1-X)/(\theta(1-\theta))] + E_\theta[(1-X)^2/(1-\theta)^2] \\ &= E_\theta[X/\theta^2] + E_\theta[(1-X)/(1-\theta)^2] \\ &= 1/(\theta(1-\theta)), \end{aligned}$$

since $x = x^2$, $x(1-x) = 0$, and $(1-x) = (1-x)^2$ if $x \in \{0,1\}$. At the same time,

$$\begin{aligned} -E_\theta[\partial^2\ell(\theta|X)/\partial\theta^2] &= E_\theta[X/\theta^2 + (1-X)/(1-\theta)^2] \\ &= \theta/\theta^2 + (1-\theta)/(1-\theta)^2 \\ &= 1/\theta + 1/(1-\theta) \\ &= 1/(\theta(1-\theta)). \end{aligned}$$

So $I(\theta) = -E_\theta[\partial^2\ell(\theta|X)/\partial\theta^2]$, as it should be.

2.1 Information for a random sample

Let us now consider Fisher information for a random sample. Let $X = (X_1, ..., X_n)$ be an i.i.d. random sample from distribution $f_1(x_i|\theta)$. Then the joint pdf is $f(x) = \prod_{i=1}^n f_1(x_i|\theta)$ where $x = (x_1, ..., x_n)$. The joint log-likelihood is $l(x,\theta) = \sum_{i=1}^n l_1(x_i,\theta)$. So Fisher information for the sample X is

$$I(\theta) = -E_\theta\left[\frac{\partial^2\ell_n(\theta|X)}{\partial\theta^2}\right] = -E_\theta\sum_{i=1}^n\left[\frac{\partial^2\ell_1(\theta|X_i)}{\partial\theta^2}\right] = nI_1(\theta).$$

Here $I_1(\theta)$ denotes Fisher information for one random draw from the distribution $f_1(x_i|\theta)$.

3 Rao-Cramer bound

An important question in the theory of statistical estimation is whether there is a nontrivial bound such that no estimator can be more efficient than this bound. The theorem below is a result of this sort:

Theorem 3 (Rao-Cramer bound)**.** *Let $X = (X_1, ..., X_n)$ be a random sample from distribution $f(x|\theta)$ with information $I(\theta)$. Let $W(X)$ be an estimator of θ such that*

(1) $\frac{d}{d\theta}E_\theta[W(X)] = \int W(x)\frac{\partial f(x|\theta)}{\partial\theta}dx$, *where* $x = (x_1, ...x_n)$

(2) $Var(W) < \infty$.

Then

$$Var(W) \geq \left(\frac{d}{d\theta}E_\theta[W(X)]\right)^2 \frac{1}{I(\theta)}.$$

In particular, if W is unbiased for θ, then $Var(W) \geq \frac{1}{I(\theta)} = \frac{1}{nI_1(\theta)}$.

Proof. The first information equality gives $ES(\theta|X) = E_\theta\left[\frac{\partial \ell(\theta|X)}{\partial \theta}\right] = 0$. So,

$$\begin{aligned}
\text{cov}(W(X), S(\theta|X)) &= E\left[W(X)\frac{\partial \ell(\theta|X)}{\partial \theta}\right] \\
&= \int W(x)\frac{\partial \ell(\theta|x)}{\partial \theta} f(x|\theta)dx \\
&= \int W(x)\frac{\partial f(x|\theta)}{\partial \theta}\cdot\frac{1}{f(x|\theta)} f(x|\theta)dx \\
&= \int W(x)\frac{\partial f(x|\theta)}{\partial \theta}dx \\
&= \frac{d}{d\theta}E_\theta[W(X)].
\end{aligned}$$

By the Cauchy-Schwarz inequality,

$$(\text{cov}(W(X), S(\theta|X))^2 \leq Var(W(X))Var(S(\theta|X)) = Var(W(X))I(\theta).$$

Thus,

$$Var(W(X)) \geq \left(\frac{d}{d\theta}E_\theta[W(X)]\right)^2 / I(\theta).$$

If W is unbiased for θ, then $E_\theta[W(X)] = \theta$, $dE_\theta[W(X)]/d\theta = 1$, and $Var(W(X)) \geq 1/I(\theta)$. □

Example Let us calculate the Rao-Cramer bound for random sample $X_1, ..., X_n$ from a Bernoulli(θ) distribution. We have already seen that $I_1(\theta) = 1/(\theta(1-\theta))$ in this case. So Fisher information for the sample is $I(\theta) = n/(\theta(1-\theta))$. Thus, any unbiased estimator of θ, under some regularity conditions, has a variance no smaller than $\theta(1-\theta)/n$. On the other hand, let $\hat{\theta} = \overline{X}_n = \sum_{i=1}^n X_i/n$ be an estimator of θ. Then $E_\theta[\hat{\theta}] = \theta$, i.e. $\hat{\theta}$ is unbiased, and $V(\hat{\theta}) = \theta(1-\theta)/n$ which coincides with the Rao-Cramer bound. Thus, $\overline{X}_n$ is the uniformly minimum variance unbiased (UMVU) estimator of θ. The word "uniformly" in this situation means that $\overline{X}_n$ has the smallest variance among unbiased estimators for *all* $\theta \in \Theta$.

Example Let us now consider a counterexample to the Rao-Cramer theorem. Let $X_1, ..., X_n$ be a random sample from $U[0, \theta]$. Then $f(x_i|\theta) = 1/\theta$ if $x_i \in [0, \theta]$ and 0 otherwise. So $l(x_i, \theta) = -\log\theta$ if $x_i \in [0, \theta]$. Then $\partial l/\partial\theta = -1/\theta$ and $\partial^2 l/\partial\theta^2 = 1/\theta^2$. So $I(\theta) = 1/\theta^2$ while $-E_\theta[\partial^2 l(X_i, \theta)/\partial\theta^2] = -1/\theta^2 \neq I(\theta)$. Thus, the second information equality does not hold in this example. The reason is that support of the distribution depends on θ in this example. Moreover, consider an estimator $\hat{\theta} = ((n+1)/n)X_{(n)}$ of θ. Then $E_\theta[X_{(n)}] = \theta$ and

$$V(\hat{\theta}) = ((n+1)^2/n^2)V(X_{(n)}) = \theta^2/(n(n+2))$$

as we saw when we considered order statistics. So $\hat{\theta}$ is unbiased, but its variance is smaller than $1/I_n(\theta) = \theta^2/n^2$. Thus, the Rao-Cramer theorem does not work in this example either. Again, the reason is that the Rao-Cramer theorem assumes that support is independent of parameter.

14.381 Statistical Method in Economics

Fall 2018