Lecture 5

Point estimators.

1 Estimators. Properties of estimators.

An estimator is a function of the data. If we have a parametric family with parameter θ, then an estimator of θ is usually denoted by $\hat{\theta}$.

Example For example, if $X_1, ..., X_n$ is a random sample from some distribution with mean μ and variance σ^2, then sample average $\hat{\mu} = \overline{X}_n$ is an estimator of the population mean, and sample variance $\hat{\sigma}^2 = s^2 = \sum_{i=1}^n (X_i - \overline{X}_n)^2/(n-1)$ is an estimator of the population variance.

1.1 Finite sample properties.

1.1.1 Unbiasness

Let X be our data. Let $\hat{\theta} = T(X)$ be an estimator where T is some function.

We say that $\hat{\theta}$ is *unbiased* for θ if $E_\theta[T(X)] = \theta$ for all possible values of θ where E_θ denotes the expectation when θ is the true parameter value.

Thus, the concept of unbiasness means that we are on average right. The *bias* of $\hat{\theta}$ is defined by $\text{Bias}(\hat{\theta}) = E_\theta[\hat{\theta}] - \theta$. Thus, $\hat{\theta}$ is unbiased if and only if its bias equals 0. For example, if X is a random sample $X_1, ..., X_n$ from some distribution with mean μ and variance σ^2, then, as we have already seen, $E[\hat{\mu}] = \mu$ and $E[s^2] = \sigma^2$. Thus, sample average and sample variance are unbiased estimators of population mean and population variance correspondingly.

There are some cases when unbiased estimators do not exist. As an example, let $X_1, ..., X_n$ be a random sample from a Bernoulli(p) distribution. Suppose that our parameter of interest $\theta = 1/p$. Let $\hat{\theta} = T(X)$ be some estimator. Then $E[\hat{\theta}] = \sum_{(x_1,...,x_n)\in\{0,1\}^n} T(x_1, ..., x_n) P\{(X_1, ..., X_n) = (x_1, ...x_n)\}$. We know that for any $(x_1, ..., x_n) \in \{0,1\}^n$, $P\{(X_1, ..., X_n) = (x_1, ...x_n)\} = p^{\sum x_i}(1-p)^{\sum (1-x_i)}$ which is a polynomial of degree n in p. Therefore, $E[\hat{\theta}]$ is a polynomial of degree at most n in p. However, $1/p$ is not a polynomial at all. Hence, there are no unbiased estimators in this case.

In some other cases, unbiased estimators may be quite weird. As an example, suppose there is an infinite number of independent trials, $X_1, ..., X_n, ...$ from Bernoulli(p) distribution. Suppose that $X_i = 0$ means failure in the i-th trial while $X_i = 1$ means success. Suppose that we do not observe $X_1, ..., X_n, ...$ Instead, we observe only the number of failures before the first success. Denote the number of failures by X. Then

Cite as: Anna Mikusheva, course materials for 14.381 Statistical Methods in Economics, Fall 2011. MIT OpenCourseWare (http://ocw.mit.edu), Massachusetts Institute of Technology. Downloaded on [DD Month YYYY].

$P\{X=0\}=p$, $P\{X=1\}=(1-p)p$, $P\{X=2\}=(1-p)^2p$, etc. Suppose that our parameter of interest $\theta=p$. Let $\hat{\theta}=T(X)$ be some estimator. Then $E[\hat{\theta}]=\sum_{i=0}^{\infty}T(i)(1-p)^ip$. If $\hat{\theta}$ is unbiased for θ, then $\sum_{i=0}^{\infty}T(i)(1-p)^ip=p$. Equivalently, $\sum_{i=0}^{\infty}T(i)(1-p)^i=1$. Thus, the only unbiased estimator is $T(0)=1$ and $T(i)=0$ for all $i\geq 1$. Does it seem reasonable?

1.1.2 Efficiency: MSE

Another of the concepts that evaluates performance of estimators is the MSE (Mean Squared Error). By definition, $\text{MSE}(\hat{\theta})=E_\theta[(\hat{\theta}-\theta)^2]$. The theorem below gives a useful decomposition for MSE:

Theorem 1. *$MSE(\hat{\theta})=Bias^2(\hat{\theta})+V(\hat{\theta})$.*

Proof.

$$\begin{aligned}
E[(\hat{\theta}-\theta)^2] &= E[(\hat{\theta}-E[\hat{\theta}]+E[\hat{\theta}]-\theta)^2] \\
&= E[(\hat{\theta}-E[\hat{\theta}])^2+(E[\hat{\theta}]-\theta)^2+2(\hat{\theta}-E[\hat{\theta}])(E[\hat{\theta}]-\theta)] \\
&= V(\hat{\theta})+\text{Bias}(\hat{\theta})+2E[\hat{\theta}-E[\hat{\theta}]](E[\hat{\theta}]-\theta) \\
&= V(\hat{\theta})+\text{Bias}(\hat{\theta}).
\end{aligned}$$

□

Estimators with smaller MSE are considered to be better, or more *efficient*. Quite often there is a trade-off between bias of the estimator and its variance. Thus, we may prefer a slightly biased estimator to an unbiased one if the former has much smaller variance in comparison to the latter one.

Example Let $X_1,...,X_n$ be a random sample from $N(\mu,\sigma^2)$. Let $\hat{\sigma}_1^2=s^2=\sum_{i=1}^n(X_i-\overline{X}_n)^2/(n-1)$ and $\hat{\sigma}_2^2=\sum_{i=1}^n(X_i-\overline{X}_n)^2/n$ be two estimators of σ^2. We know that $E[\hat{\sigma}_1^2]=\sigma^2$. So $E[\hat{\sigma}_2^2]=((n-1)/n)E[\hat{\sigma}_1^2]=((n-1)/n)\sigma^2$, and $\text{Bias}(\hat{\sigma}_2^2)=\sigma^2/n$. We also know that $(n-1)\hat{\sigma}_1^2/\sigma^2\sim\chi^2(n-1)$. What is $V(\chi^2(n-1))$? Let $\xi_1,...,\xi_{n-1}$ be a random sample from $N(0,1)$. Then $\xi=\xi_1^2+...+\xi_{n-1}^2\sim\chi^2(n-1)$. By linearity of expectation, $E[\xi]=(n-1)$. By independence,

$$\begin{aligned}
E[\xi^2] &= E[(\xi_1^2+...+\xi_{n-1}^2)^2] \\
&= \sum_{i=1}^{n-1}E[\xi_i^4]+2\sum_{1\leq i<j\leq n-1}E[\xi_i^2\xi_j^2] \\
&= 3(n-1)+2\sum_{1\leq i<j\leq n-1}E[\xi_i^2]E[\xi_j^2] \\
&= 3(n-1)+(n-1)(n-2) \\
&= (n-1)(n+1),
\end{aligned}$$

since $E[\xi_i^4]=3$. So

$$V(\xi)=E[\xi^2]-(E[\xi])^2=(n-1)(n+1)-(n-1)^2=2(n-1).$$

Thus, $V(\hat{\sigma}_1^2) = V(\sigma^2\xi/(n-1)) = 2\sigma^4/(n-1)$ and $V(\hat{\sigma}_2^2) = ((n-1)/n)^2 V(\hat{\sigma}_1^2) = 2\sigma^4(n-1)/n^2$. Finally, $\text{MSE}(\hat{\sigma}_1^2) = 2\sigma^4/(n-1)$ and

$$\text{MSE}(\hat{\sigma}_2^2) = \sigma^4/n^2 + 2\sigma^4(n-1)/n^2 = (2n-1)\sigma^2/n^2.$$

So, $\text{MSE}(\hat{\sigma}_1^2) < \text{MSE}(\hat{\sigma}_2^2)$ if and only if $2/(n-1) < (2n-1)/n^2$, which is equivalent to $3n < 1$. So, for any $n \geq 1$, $\text{MSE}(\hat{\sigma}_1^2) > \text{MSE}(\hat{\sigma}_2^2)$ in spite of the fact that $\hat{\sigma}_1^2$ is unbiased.

1.2 Asymptotic properties.

1.2.1 Consistency

Imagine a thought experiment in which the number of observations n increases without bound, i.e. $n \to \infty$. Suppose that for each n, we have an estimator $\hat{\theta}_n$.

We say that $\hat{\theta}_n$ is *consistent* for θ if $\hat{\theta}_n \to_p \theta$.

Example Let $X_1, ..., X_n$ be a random sample from some distribution with mean μ and variance σ^2. Let $\hat{\mu} = \hat{\mu}_n = \overline{X}_n$ be our estimator of μ and $s^2 = s_n^2 = \sum_{i=1}^n (X_i - \overline{X}_n)^2/(n-1)$ be our estimator of σ^2. By the Law of large numbers, we know that $\hat{\mu} \to_p \mu$ as $n \to \infty$. In addition,

$$\begin{aligned} s^2 &= \sum_{i=1}^n (X_i - \overline{X}_n)^2/(n-1) \\ &= \sum_{i=1}^n (X_i - \mu)^2/(n-1) - (n/(n-1))(\overline{X}_n - \mu)^2 \\ &= (n/(n-1))(\sum_{i=1}^n (X_i - \mu)^2/n) - (n/(n-1))(\overline{X}_n - \mu)^2 \end{aligned}$$

By the Law of large numbers, $\sum_{i=1}^n (X_i - \mu)^2/n \to_p E[(X_i - \mu)^2 = \sigma^2$ and $\overline{X}_n - \mu = \sum_{i=1}^n (X_i - \mu)/n \to_p E[X_i - \mu] = 0$. By the Continuous mapping theorem, $(\overline{X}_n - \mu)^2 \to_p 0$. In addition, $n/(n-1) \to_p 1$. So, by the Slutsky theorem, $s^2 \to_p \sigma^2$. So $\hat{\mu}$ and s^2 are consistent for μ and σ^2 correspondingly.

1.2.2 Asymptotic Normality

We say that $\hat{\theta}$ is *asymptotically normal* if there are sequences $\{a_n\}_{n=1}^\infty$ and $\{r_n\}_{n=1}^\infty$ and constant σ^2 such that $r_n(\hat{\theta} - a_n) \Rightarrow N(0, \sigma^2)$. Then r_n is called the *rate of convergence*, a_n - the *asymptotic mean*, and σ^2 - the *asymptotic variance*. In many cases, one can choose $a_n = \theta$ and $r_n = \sqrt{n}$. We will use the concept of asymptotic normality for confidence set construction later on. For now, let us consider an example.

Example Let $X_1, ..., X_n$ be a random sample from some distribution with mean μ and variance σ^2. Let $\hat{\mu}$ and s^2 be the sample mean and the sample variance correspondingly. Then, by the Central limit theorem,

$\sqrt{n}(\hat{\mu}-\mu) \Rightarrow N(0,\sigma^2)$. As for s^2,

$$\sqrt{n}(s^2-\sigma^2) = (n/(n-1))[\sum_{i=1}^{n}((X_i-\mu)^2-\sigma^2)/\sqrt{n} - (\sqrt{n}(\overline{X}_n-\mu)/n^{1/4})^2] + (\sqrt{n}/(n-1))\sigma^2$$

By the Central limit theorem, $\sum_{i=1}^{n}((X_i-\mu)^2-\sigma^2)/\sqrt{n} \Rightarrow N(0,\tau^2)$ with $\tau^2 = E[((X_i-\mu)^2-\sigma^2)^2]$. Note that $\tau^2 = \mu_4 - 2\sigma^2 E[(X_i-\mu)^2] + \sigma^4 = \mu_4 - \sigma^4$ with $\mu_4 = E[(X_i-\mu)^4]$. By Slutsky theorem, $\sqrt{n}(\overline{X}_n-\mu)/n^{1/4} \to_p 0$. In addition, $(\sqrt{n}/(n-1))\sigma^2 \to_p 0$. So, by the Slutsky theorem again, $\sqrt{n}(s^2-\sigma^2) \Rightarrow N(0,\tau^2)$.

2 Common methods for constructing an estimator

2.1 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.

2.2 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: $\sum_{i=1}^{n} X_i/n = m_1(\hat{\theta})$, $\sum_{i=1}^{n} X_i^2/n = m_2(\hat{\theta})$,..., $\sum_{i=1}^{n} X_i^k/n = m_k(\hat{\theta})$. It is implicitly assumed 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 Binomial(k, p). In other words, X_i is a discrete random variable and $P\{X_i = j\} = k!/(j!(k-j)!)p^j(1-p)^{k-j}$. Then $E[X_i] = kp$ and $E[X_i^2] = kp(1-p) + k^2p^2$. The method of moment estimator $(\hat{k}_{MM}, \hat{p}_{MM})$ solves $\sum_{i=1}^{n} X_i/n = \hat{k}_{MM}\hat{p}_{MM}$ and $\sum_{i=1}^{n} X_i^2/n = \hat{k}_{MM}\hat{p}_{MM}(1-\hat{p}_{MM}) + \hat{k}_{MM}^2\hat{p}_{MM}^2$. We can solve this system to find $\hat{k}_{MM}$ and $\hat{p}_{MM}$. Is $\hat{k}_{MM}$ always an integer?

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. It is called GMM (Generalized Method of Moments) and will be studied extensively later on, as the main work horse of Econometrics.

2.3 Maximum Likelihood Estimator

Let $f(x|\theta)$ with $\theta \in \Theta$ be some parametric family. Let $X_1, ..., X_n$ be a random sample from $f(x|\theta)$. The joint pdf of $X_1, ..., X_n$ is $f(x_1, ..., x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$. Let $x_1, ..., x_n$ denote the realization of $X_1, ..., X_n$.

By definition, the maximum likelihood estimator $\hat{\theta}_{ML}$ of θ is the value that maximizes $f(x_1, ..., x_n|\theta)$, i.e. $\hat{\theta}_{ML} = \arg\max_{\theta \in \Theta} f(x_1, ..., x_n|\theta)$. Thus, maximum likelihood estimator is a parameter value such that it gives the greatest probability of observing $x_1, ...x_n$. As we said before, $f(x_1, ..., x_n|\theta)$ as a function of θ for fixed values $x_1, ..., x_n$ is called the likelihood function. It is usually denoted by $\mathcal{L}(\theta|x_1, ..., x_n)$. 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 $l(\theta|x_1, ..., x_n) = \log \mathcal{L}(\theta|x_1, ..., x_n)$. Function $l(\theta|x_1, ..., x_n)$ is called the log-likelihood. If $l(\theta|x_1, ..., x_n)$ is differentiable in θ, then $\hat{\theta}_{ML}$ satisfies first order condition (FOC): $dl(\hat{\theta}_{ML}|x_1, ..., x_n)/d\theta = 0$ or, equivalently, $\sum_{i=1}^{n} \partial \log f(x_i|\hat{\theta}_{ML})/\partial\theta = 0$. The reason we took 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 $S(\theta|x_1, ...x_n) = \sum_{i=1}^{n} \partial \log f(x_i|\theta)/\partial\theta$ is called the score. Thus, $\hat{\theta}_{ML}$ solves the equation $S(\theta|x_1, ...x_n) = 0$.

Example Let $X_1, ...X_n$ be a random sample from $N(\mu, \sigma^2)$. Then $l(\theta|x_1) = -\log\sqrt{2\pi} - (1/2)\log\sigma^2 - (x_i - \mu)^2/(2\sigma^2)$ where $\theta = (\mu, \sigma^2)$. So

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

FOCs are

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

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

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 $f(x_1|\theta) = 1/\theta$ 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, $\mathcal{L}(\theta|x_1, ...x_n) = (1/\theta^n)I\{\theta \geq x_{(n)}\}I\{x_{(1)} \geq 0\}$. We conclude that $\hat{\theta}_{ML} = X_{(n)}$.

MIT OpenCourseWare
http://ocw.mit.edu

14.381 Statistical Method in Economics
Fall 2013

For information about citing these materials or our Terms of Use, visit: http://ocw.mit.edu/terms.