Unbiased Estimation and Risk Inequalities

MIT 18.655

Dr. Kempthorne

Spring 2016

Outline

Unbiased Estimation

Comments on Unbiased Estimation

- Estimation decision problem:
  - $X \sim P_\theta, \theta \in \Theta$
  - $\theta(P) = E[X \mid P_\theta]$
  - Estimation: $\mathcal{A} = \times$
  - Loss function: $L : \Theta \times \mathcal{A} \to R$.
  - Decision procedures: $\mathcal{D} = \{\delta : \mathcal{X} \to \mathcal{A}\}$
- Restrict estimation procedures to the subclass:
$$\mathcal{D}_0 = \{\delta \in \mathcal{D} : E[\delta(X) \mid \theta] = \theta, \text{ for all } \theta \in \Theta\}.$$
- Apply decision-theoretic principles to identify optimal procedures in $\mathcal{D}_0$.

Choice of $\mathcal{D}_0$ equivalent to choice of constraints:

- Unbiasedness
- Linearity (in X)
- Computational algorithms (e.g., orthogonal polynomials in X, Fourier series, generalized-basis series)

Unbiased Estimation

Comments on Unbiased estimation (continued)

- Significant role of *unbiasedness* in survey sampling.
- Bayes estimates are necessarily biased (Problem 3.4.20).
- Unbiasedness not preserved under non-linear re-parametrization (not equivariant).
- Asymptotic unbiasedness:
$$\frac{Bias^2(\hat{\theta}_n)}{Var[\hat{\theta}_n \mid \theta]} \to 0.$$

Outline

Information Inequality: Preliminaries

Definition: Regular Problem A statistical inference problem with $X \sim P_\theta, \theta \in \Theta$ which satisfies the following regularity conditions:

- $\mathcal{X} = \{x : p(x \mid \theta) > 0\}$ does not depend on θ.
- $\dfrac{\partial logp(x \mid \theta)}{\partial \theta}$ exists and is finite for all $x \in \mathcal{X}$ and $\theta \in \Theta$.
- For any statistic T such that $E[|T(X)| \mid \theta] < \infty$
$$\frac{\partial}{\partial \theta}\left[\int T(x)p(x \mid \theta)dx\right] = \int T(x)\frac{\partial}{\partial \theta}[p(x \mid \theta)]dx.$$

Definition: Efficient Score Function. For a fixed $\theta_0 \in \Theta$, the *efficient score* for X is
$$u(X;\theta_0) = \frac{\partial \log p(x \mid \theta)}{\partial \theta}|_{\theta=\theta_0}$$
Note: The magnitude of $u(X;\theta_0)$ scales how far θ_0 is from $\hat{\theta}_{MLE}$.

Proposition The Efficient Score Function has the following properties:

$$\begin{aligned} E[u(X;\theta_0) \mid \theta = \theta_0] &= 0. \\ Var[u(X;\theta_0) \mid \theta = \theta_0] &= E([u(X;\theta_0)]^2 \mid \theta = \theta_0) = I(\theta_0). \end{aligned}$$

$I(\theta)$ is the *Fisher information* about θ contained in X which satisfies the following identity

$$I(\theta_0) = Var[(u(X;\theta_0) \mid \theta_0] = E\left[-\frac{\partial^2 \log p(X \mid \theta_0)}{\partial\theta^2} \mid \theta_0\right]$$

Proof:

$$\begin{aligned} \int p(x \mid \theta)dx &= 1 \\ \implies \int \frac{\partial p(x \mid \theta)}{\partial\theta}dx &= \frac{\partial}{\partial\theta}(1) = 0 \\ \implies \int [\frac{\partial p(x \mid \theta)}{\partial\theta}/p(x \mid \theta)]p(x \mid \theta)dx &= 0 \\ \implies \int [\frac{\partial log[p(x \mid \theta)]}{\partial\theta}p(x \mid \theta)dx &= 0 \\ \implies E[u(X;\theta) \mid \theta] &= 0 \end{aligned}$$

$$
\begin{aligned}
E[u(X;\theta) \mid \theta] &= 0 \\
\iff \quad \int [\frac{\partial log[p(x \mid \theta)]}{\partial \theta} p(x \mid \theta) dx &= 0 \\
\frac{\partial}{\partial \theta}\left(\int [\frac{\partial log[p(x \mid \theta)]}{\partial \theta} p(x \mid \theta) dx \right) &= \frac{\partial}{\partial \theta}(0) \\
\int \left(\frac{\partial^2 log[p(x \mid \theta)]}{\partial \theta^2} p(x \mid \theta) + \frac{\partial log[p(x \mid \theta)]}{\partial \theta} (\frac{\partial p(x \mid \theta)}{\partial \theta}) \right) dx &= 0
\end{aligned}
$$

The last line can be written as:

$$
\int [\frac{\partial^2 log[p(x \mid \theta)]}{\partial \theta^2} p(x \mid \theta) dx] + \int [\frac{\partial log[p(x \mid \theta)]}{\partial \theta}]^2 p(x \mid \theta) dx = 0
$$

I.e.,

$$
E\left[\frac{\partial^2 log[p(x \mid \theta)]}{\partial \theta^2} \mid \theta \right] + E\left[\left(\frac{\partial log[p(x \mid \theta)]}{\partial \theta} \right)^2 \mid \theta \right] = 0
$$

So we have

$$
\begin{aligned}
I(\theta) &= E[(u(X;\theta))^2 \mid \theta] = -E\left[\frac{\partial^2 log[p(x \mid \theta)]}{\partial \theta^2} \mid \theta \right]. \\
&= Var[u(X;\theta) \mid \theta]
\end{aligned}
$$

Proposition 3.4.1 Suppose P_θ is a one-parameter exponential family with density/pmf function:

$$p(x \mid \theta) = h(x) exp\{\eta(\theta)T(x) - B(\theta)\}$$

which has non-vanishing continuous derivative on Θ. Then the statistical inference problem for θ given X is a regular problem.

Theorem 3.4.1. Information Inequality

For a regular problem, let $T(X)$ be any statistic such that

$E[T(X) \mid \theta] = \psi(\theta)$.

$Var[T(X) \mid \theta] < \infty$, for all θ.

Then for all θ:

- $Var[T(X) \mid \theta] \geq \dfrac{[\psi'(\theta)]^2}{I(\theta)}$,

  ($\psi(\theta)$ is differentiable and $I(\theta) =$ Fisher Information of P_θ).

Proof: By the conditions of a regular problem:

$$\begin{aligned}
\psi'(\theta) &= \frac{\partial}{\partial\theta}\left(\int T(x)p(x \mid \theta)dx\right) \\
&= \int \left(T(x)\frac{\partial}{\partial\theta}[p(x \mid \theta)]\right) dx \\
&= \int \left(T(x)\frac{\partial}{\partial\theta}[\log p(x \mid \theta)]p(x \mid \theta)\right) dx \\
&= E[T(X)U(X;\theta) \mid \theta] = Cov[T(X), U(X;\theta) \mid \theta]
\end{aligned}$$

(the last equation follows since $E[U(X;\theta) \mid \theta] = 0$.)

The theorem follows from the Cauchy-Schwarz Inequality for two random variables:
$(Cov[T(X), U(X;\theta) \mid \theta])^2 \leq Var[T(X) \mid \theta] \times Var[U(X;\theta) \mid \theta]$
i.e.,

$$[\psi'(\theta)]^2 \leq Var[T(X) \mid \theta] \times I(\theta)$$

Corollary 3.4.1 Suppose $T(X)$ is unbiased estimate of θ in a regular problem, then

$$Var(T(X) \mid \theta) \geq \frac{1}{I(\theta)} \quad \text{(Cramer-Rao Lower Bound)}$$

Proposition 3.4.2 For a random sample $\mathbf{X} = (X_1, \ldots, X_n)$ from a distribution P_θ with density $p(x \mid \theta)$ satisfying the conditions of a regular problem. If $I_1(\theta) = E\left[\left(\left(\frac{\partial}{\partial\theta}[\log p(x_1 \mid \theta)] \right) \right)^2 \mid \theta \right]$ then

$$I(\theta) = nI_1(\theta) \quad \text{and}$$
$$Var[T(\mathbf{X}) \mid \theta] \geq \frac{[\psi'(\theta)]^2}{nI_1(\theta)}$$

Proof: This follows directlly from the results above upon noting that

$$\begin{aligned} U(\mathbf{X};\theta) &= \frac{\partial}{\partial\theta}[\log p(\mathbf{x} \mid \theta)] \\ &= \frac{\partial}{\partial\theta}[\sum_1^n \log p(x_i \mid \theta)] \\ &= \textstyle\sum_1^n \frac{\partial}{\partial\theta}[\log p(x_i \mid \theta)] \\ &= \textstyle\sum_1^n U(X_i;\theta) \end{aligned}$$

By the independence of the terms,

$$Var[U(\mathbf{X};\theta) \mid \theta] = \textstyle\sum_1^n Var[U(X_i;\theta)] = nI_1(\theta) = I(\theta).$$

Theorem 3.4.2 Consider a *regular problem* with $X \sim P_\theta, \theta \in \Theta$, and $T^*(X)$ is an estimator of $\psi(\theta)$ which is

- Unbiased: $E[T^*(X) \mid \theta] = \psi(\theta)$, for all $\theta \in \Theta$.
- Achieves the Cramer-Rao Lower Bound:

$$Var(T^*(X) \mid \theta) = \frac{|\psi'(\theta)|^2}{I(\theta)}, \text{ for all } \theta \in \Theta.$$

Then $\{P_\theta\}$ is a one-parameter exponential family with density/pmf:

$$p(x \mid \theta) = h(x) exp\{\eta(\theta) T^*(x) - B(\theta)\}$$

Proof: From the proof of Theorem 3.4.1 for any unbiased estimator of $\psi(\theta)$,

$$\begin{aligned}
\psi(\theta) &= E[T(x) \mid \theta] = \int T(x) p(x \mid \theta) dx \\
\Longrightarrow \psi'(\theta) &= \int T(x) U(x;\theta) p(x \mid \theta) dx \\
& \quad \text{where } U(x;\theta) = \partial \log p(x \mid \theta)/\partial\theta \\
&= Cov(T(X), U(X;\theta) \mid \theta) \\
\Longrightarrow |\psi'(\theta)| &\leq \sqrt{Var(T(X) \mid \theta) \times Var(U(X;\theta) \mid \theta)}
\end{aligned}$$

with equality if and only if $U(X;\theta) = a_1(\theta) + a_2(\theta) T(X)$ for some functions $a_1(\theta)$ and $a_2(\theta)$.

Technical Details of Proof:

- For each $\theta_0 \in \Theta$, define
  $$A_{\theta_0} = \{x : U(x;\theta_0) = a_1(\theta_0)T^*(x) + a_2(\theta_0)\}$$
  Note: $P_{\theta_0}(A_{\theta_0}) = 1$
  (otherwise the absolute correlation would be less than 1)
- Define $\{\theta_i, i = 1, 2, \ldots\}$ to be a denumerable dense subset of Θ.
- Define $A^{**} = \cap_i A_{\theta_i}$. Then
  $$P_{\theta_i}(A^{**}) = 1, \text{ for all } \theta_i.$$
- Fix any two values $x_1, x_2 \in A^{**}$, for which $T^*(x_1) \neq T^*(x_2)$. Solve the equations:
  $$\begin{aligned} U(x_1;\theta) &= a_1(\theta)T^*(x_1) + a_2(\theta) \\ U(x_2;\theta) &= a_1(\theta)T^*(x_2) + a_2(\theta) \end{aligned}$$
  to obtain equations for $a_1(\theta), a_2(\theta)$ as linear combinations of
  $$U(x_1;\theta) \text{ and } U(x_2;\theta).$$
  Since $U(x;\theta)$ is continuous in θ, so are $a_1(\theta)$ and $a_2(\theta)$.

Technical Details of Proof (continued):

- Since
  $$U(x;\theta) = a_1(\theta)T^*(x) + a_2(\theta), \text{ for all } \theta_i \in \{\theta_i\}$$
  and both $U(x;\theta)$ and $a_1(\theta)$ and $a_2(\theta)$ are continuous, this equation must hold for all θ.
- So $A^{**} = \cap_i A_{\theta_i}$ must equal
  $$A^* = \{x : U(x;\theta) = a_1(\theta)T^*(x) + a_2(\theta), \text{ for all } \theta \in \Theta\}.$$
  and $P(A^*) = 1$.
- With
  $$U(x;\theta) = \frac{\partial \log p(x|\theta)}{\partial\theta} = a_1(\theta)T^*(x) + a_2(\theta)$$
  Define: $\eta(\theta) = \int_{\theta_0}^{\theta} a_1(t)dt$ and $B(\theta) = -\int_{\theta_0}^{\theta} a_2(t)dt$,
  Then
  $$\log\left[\frac{p(x|\theta)}{p(x|\theta_0)}\right] = \int_{\theta_0}^{\theta}\left[\frac{\partial \log p(x|\theta)}{\partial\theta}\right]d\theta = T^*(x)\eta(\theta) - B(\theta),$$
  and we have:
  $$p(x \mid \theta) = h(x)exp\{\eta(\theta)T^*(x) - B(\theta)\}, \ x \in A^*$$
  where $h(x) = p(x \mid \theta_0)$ (for a fixed value θ_0).

Multiparameter Case

Definition: Regular Problem A statistical inference problem with $X \sim P_\theta, \theta \in \Theta$ which satisfies the following regularity conditions:

- $\mathcal{X} = \{x : p(x \mid \theta) > 0\}$ does not depend on θ.
- $\dfrac{\partial logp(x \mid \theta)}{\partial \theta}$ exists and is finite for all $x \in \mathcal{X}$ and $\theta \in \Theta$.
- For any statistic T such that $E[|T(X)| \mid \theta] < \infty$
$$\frac{\partial}{\partial \theta}\left[\int T(x)p(x \mid \theta)dx\right] = \int T(x)\frac{\partial}{\partial \theta}[p(x \mid \theta)]dx.$$

Definition: Efficient Score Function. For a fixed $\theta_0 \in \Theta$, the *efficient score* for X is
$$u(X;\theta_0) = \frac{\partial \log p(x \mid \theta)}{\partial \theta}|_{\theta=\theta_0}$$
Note: The magnitude of $u(X;\theta_0)$ scales how far θ_0 is from $\hat{\theta}_{MLE}$.
The definitions extend to vector-valued θ immediately

Proposition (I). The Efficient Score Function has the following properties:

$$\begin{aligned} E[u(X;\theta_0) \mid \theta = \theta_0] &= 0. \\ Cov[u(X;\theta_0) \mid \theta = \theta_0] &= E([u(X;\theta_0)][u(X;\theta_0)]^T \mid \theta = \theta_0) \\ &= I(\theta_0). \end{aligned}$$

(II). $I(\theta)$ is the $(d \times d)$ *Fisher information* matrix whose elements satisfy the following identities

$$\begin{aligned} [I(\theta_0)]_{i,j} &= [Cov[u(X;\theta_0) \mid \theta_0]]_{i,j} \\ &= E[[u(X;\theta)]_i [u(X;\theta)]_j \mid \theta = \theta_0] \\ &= E[\frac{\partial \log p(X \mid \theta)}{\partial \theta_i} \frac{\partial \log p(X \mid \theta)}{\partial \theta_j} \mid \theta = \theta_0] \\ &= -E[\frac{\partial^2 \log p(X \mid \theta)}{\partial \theta_i \partial \theta_j} \mid \theta = \theta_0] \end{aligned}$$

(III). If $\mathbf{X} = (X_1, \ldots, X_n)$ is an iid sample from $X \sim P_\theta$ with Information $I_1(\theta)$, then

$$I(\mathbf{X}) = nI_1(\theta).$$

Theorem 3.4.3 For a regular problem with non-singular information matrix $I(\theta)$, consider a scalar-valued statistic $T(X)$ estimating the scalar $\psi(\theta)$, and suppose

$$E[T(X) \mid \theta] = \psi(\theta)$$

$$\dot{\psi}(\theta) = \nabla\psi(\theta) = \left(\frac{\partial\psi(\theta)}{\partial\theta_1}, \ldots, \frac{\partial\psi(\theta)}{\partial\theta_1}\right)^T$$

Then

$$Var[T(X) \mid \theta] \geq [\dot{\psi}(\theta)]^T [I(\theta)]^{-1} [\dot{\psi}(\theta)]$$

Proof. For a random variable Y, and a random d-vector Z, recall the minimum MSPE linear predictor $\mu_L(Z)$ of Y is given by:

$$\mu_L(Z) = \mu_Y + (Z - \mu_z)^T \Sigma_{Z,Z}^{-1} \Sigma_{Z,Y}$$

where $\mu_Y = E[Y]$, $\mu_Z = E[Z]$,

$$\Sigma_{Z,Z} = Cov(Z)\ (d \times d), \text{ and } \Sigma_{Z,Y} = Cov(Z,Y)\ (d \times 1).$$

The variance of $\mu_L(Z)$ satisfies

$$Var(\mu_L(Z)) = [\Sigma_{Z,Y}]^T \Sigma_{Z,Z}^{-1} \Sigma_{Z,Y} \leq Var(Y),$$

with equality only if $Y = \mu_L(Z)$.

The Theorem follows setting $Y = T(X)$ and $Z = u(X;\theta)$.

Theorem 3.4.4 For a regular problem as in Theorem 3.4.3 suppose:

$$T(X) = (T_1(X), \dots, T_d(X))^T \in R^d$$

$$E[T(X) \mid \theta] = \psi(\theta) \quad (d \times 1) \text{ vector}$$

$$\dot{\psi}(\theta) = \nabla\psi(\theta) = \left[\frac{\partial\psi(\theta)}{\partial\theta_1} | \cdots | \frac{\partial\psi(\theta)}{\partial\theta_d}\right] \quad (d \times d) \text{ matrix}$$

Then

$$Var[T(X) \mid \theta] \geq [\dot{\psi}(\theta)][I(\theta)]^{-1}[\dot{\psi}(\theta)]^T$$

where $A \geq B$ means $(A - B)$ is postive semi-definite:

$$a^T(A - B)a \geq 0, \text{ for all } a \in R^d.$$

Proof. Problem 3.4.21

Note: For $\hat{\theta} : E[\hat{\theta} \mid \theta] = \theta$,

$\psi(\theta) = \theta$, and $\dot{\psi}(\theta) = I_d$, the $(d \times d)$ identity matrix.

and

$$Var(\hat{\theta} \mid \theta) \geq [I(\theta)]^{-1}$$

Preview:

- When $\mathbf{X} = (X_1, \ldots, X_n)$ corresponds to a random sample from a population whose distribution has information $I_1(\theta)$ for a single observation, the information in a sample of size n is
$$I(\mathbf{X}) = nI_1(\theta)$$
- As the sample size grows large such samples, optimal estimators of parameters $q(\theta)$ are sought.
- The Cramer-Rao Lower Bound defines the golden standard of performance for estimators which are unbiased asymptotically.
- Such estimators will be called *asymptotically efficient*.

18.655 Mathematical Statistics
Spring 2016