Chapter 7
Point Estimation
Bias

Outline

Introduction

In creating a parameter estimator, a fundamental question is whether or not the estimator differs from the parameter in a *systematic* manner.

Definition. For observations $X = (X_1, X_2, \ldots, X_n)$ based on a distribution having parameter value θ, and for $d(X)$ an estimator for $h(\theta)$, the bias is the mean of the difference $d(X) - h(\theta)$, i.e.,

$$b_d(\theta) = E_\theta d(X) - h(\theta).$$

If $b_d(\theta) = 0$ for *all* values of the parameter, then $d(X)$ is called an unbiased estimator. Any estimator that is not unbiased is called biased.

Exercise. If $X_1, \ldots, X_n$ form a simple random sample with unknown finite mean μ, then $\bar{X}$ is an unbiased estimator of μ. If the X_i have variance σ^2, then

$$S^2 = \frac{1}{n-1} \sum_{n=1}^{n} (X_i - \bar{X})^2$$

is an unbiased estimator of σ^2.

Mean Square Error

We can assess the quality of an estimator by computing its mean square error, defined by

$$E_\theta[(d(X) - h(\theta))^2].$$

To derive a simple relationship between mean square error and variance, we begin by substituting the equation for bias into the question above, rearranging terms, and expanding the square.

$$\begin{aligned} E_\theta[(d(X) - h(\theta))^2] &= E_\theta[(d(X) - (E_\theta d(X) - b_d(\theta)))^2] \\ &= E_\theta[((d(X) - E_\theta d(X)) + b_d(\theta))^2] \\ &= E_\theta[(d(X) - E_\theta d(X))^2] + 2b_d(\theta)E_\theta[d(X) - E_\theta d(X)] + b_d(\theta)^2 \\ &= \mathrm{Var}_\theta(d(X)) + b_d(\theta)^2 \end{aligned}$$

NB. $E_\theta[d(X) - E_\theta d(X)] = 0$. So, bias *increases* mean square error.

Compensating for Bias

To estimate the size of the bias, we look at a quadratic approximation for g centered at the value μ

$$g(x) - g(\mu) \approx g'(\mu)(x - \mu) + \frac{1}{2}g''(\mu)(x - \mu)^2.$$

Replace x with the random variable $\bar{X}$ and then take expectations. Then, the bias

$$\begin{aligned} b_g(\mu) &= E_\mu[g(\bar{X})] - g(\mu) \approx E_\mu[g'(\mu)(\bar{X} - \mu)] + \frac{1}{2}E_\mu[g''(\mu)(\bar{X} - \mu)^2] \\ &= \frac{1}{2}g''(\mu)\mathrm{Var}(\bar{X}) = \frac{1}{2}g''(\mu)\frac{\sigma^2}{n}. \end{aligned}$$

Thus, the bias has the intuitive properties of being

- large for strongly convex functions,
- large for observations having high variance σ^2, and
- small when the number of observations n is large.

Compensating for Bias

Exercise. For $g(\mu) = \mu/(\mu - 1)$, show that $g''(\mu) = 2(\mu - 1)^{-3}$.

Because $\mu > 1$, g is a convex function. To estimate bias,

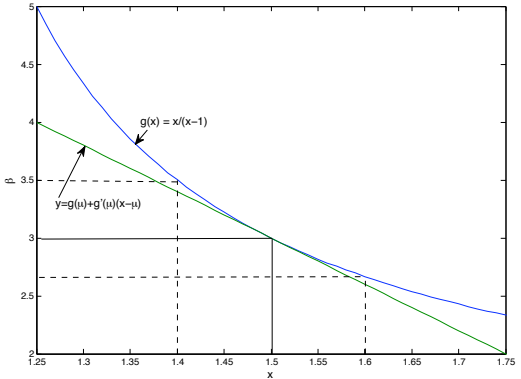
$$g''\left(\frac{\beta}{\beta - 1}\right) = \frac{2}{\left(\frac{\beta}{\beta-1} - 1\right)^3} = 2(\beta - 1)^3.$$

Thus, the bias

$$b_g(\beta) \approx \frac{1}{2} g''(\mu) \frac{\sigma^2}{n} = \frac{1}{2} 2(\beta - 1)^3 \frac{\beta}{n(\beta - 1)^2(\beta - 2)} = \frac{\beta(\beta - 1)}{n(\beta - 2)}.$$

So, for $\beta = 4$ and $n = 225$, the bias is approximately 0.027. Compare this to the estimated value of 0.035 from the simulation.

Compensating for Bias



Graph of a convex function. Note that the tangent line is below the graph of g. Here we show the case in which $\mu = 1.5$ and $\beta = g(\mu) = 3$. Notice that the interval from $x = 1.4$ to $x = 1.5$ has a longer range than the interval from $x = 1.5$ to $x = 1.6$. Because g spreads the values of $\bar{X}$ above $\beta = 3$ more than below, the estimator $\hat{\beta}$ for β is biased upward. We can use a second order Taylor series expansion to correct most of this bias.

Consistency

Definition. Given data $X_1, X_2, \ldots$ and a real valued function h of the parameter space, a sequence of estimators d_n, based on the first n observations, is called consistent if for every choice of θ

$$\lim_{n\to\infty} d_n(X_1, X_2, \ldots, X_n) = h(\theta)$$

in probability whenever θ is the true state of nature.

For circumstances in which a bias estimator is not available, we, instead, look for circumstances

- in which the bias disappears in the limit of a large number of observations and
- the distribution of the estimators $d_n(X_1, X_2, \ldots, X_n)$ become more and more concentrated near $h(\theta)$.

Consistency

For a method of moments estimator of a single parameter, we have independent observations, $X_1, X_2, \ldots,$ having mean $\mu = k(\theta)$, we have that

$$E_\theta \bar{X}_n = \mu,$$

i. e. $\bar{X}_n$, the sample mean for the first n observations, is an unbiased estimator for $\mu = k(\theta)$. Also, by the law of large numbers, we have that

$$\lim_{n\to\infty} \bar{X}_n = \mu.$$

If $g = k^{-1}$ is continuous at μ, the method of moments estimators $\hat{\theta}_n$ satisfy

$$\lim_{n\to\infty} \hat{\theta}_n(X_1, X_2, \ldots, X_n) = \lim_{n\to\infty} g(\bar{X}_n) = g(\lim_{n\to\infty} \bar{X}_n) = g(\mu) = \theta$$

with probability one and $g(\bar{X}_n)$ is a consistent sequence of estimators for θ.

Cauchy-Schwarz Inequality

Theorem. For random variables Y and Z having finite variance,

$$\mathrm{Cov}(Y,Z)^2 \leq \mathrm{Var}(Y)\mathrm{Var}(Z).$$

Proof. For the random variable $Y - \beta Z$,

$$0 \leq \mathrm{Var}(Y - \beta Z) = \mathrm{Var}(Y) - 2\beta\mathrm{Cov}(Y,Z) + \beta^2\mathrm{Var}(Z),$$

a non-negative quadratic expression of β. Thus, its discriminant

$$4\mathrm{Cov}(Y,Z)^2 - 4\mathrm{Var}(Y)\mathrm{Var}(Z)$$

is non-positive. Thus, $\mathrm{Cov}(Y,Z)^2 \leq \mathrm{Var}(Y)\mathrm{Var}(Z)$.

Information Inequality

Consequently, the square of the correlation

$$\rho(Y,Z)^2 = \frac{\mathrm{Cov}(Y,Z)^2}{\mathrm{Var}(Y)\cdot\mathrm{Var}(Z)} \le 1$$

with equality if and only if $0 = \mathrm{Var}(Y - \beta Z)$. i.e., for some α,

$$Y = \alpha + \beta Z$$

with probability 1.

We begin with independent observations $X = (X_1, \ldots, X_n)$ drawn from an unknown probability P_θ from a 1-dimensional parameter space Θ. Denote the joint density of these random variables

$$\mathbf{f}(\mathbf{x}|\theta) = f(x_1|\theta)f(x_2|\theta)\cdots f(x_n|\theta), \quad \text{where } \mathbf{x} = (x_1 \ldots, x_n).$$

For d be an unbiased estimator of θ, then

$$\theta = E_\theta d(X) = \int_{\mathcal{X}} d(\mathbf{x})\mathbf{f}(\mathbf{x}|\theta)\ \nu(d\mathbf{x}).$$

Information Inequality

Using one of the two basic properties of the density, we take the derivative with respect to θ to see that

$$
\begin{aligned}
1 &= \int_{\mathcal{X}} \mathbf{f}(\mathbf{x}|\theta)\ \nu(d\mathbf{x}) \\
0 &= \int_{\mathcal{X}} \frac{\partial \mathbf{f}(\mathbf{x}|\theta)/\partial\theta}{\mathbf{f}(\mathbf{x}|\theta)}\ \mathbf{f}(\mathbf{x}|\theta)\ \nu(d\mathbf{x}) \\
0 &= \int_{\mathcal{X}} \left(\frac{\partial}{\partial\theta} \ln \mathbf{f}(\mathbf{x}|\theta)\right) \mathbf{f}(\mathbf{x}|\theta)\ \nu(d\mathbf{x}) = E_\theta\left[\frac{\partial}{\partial\theta} \ln \mathbf{f}(X|\theta)\right]
\end{aligned}
$$

So the random variable $Y = \frac{\partial}{\partial\theta} \ln \mathbf{f}(X|\theta)$ has mean 0.

NB. If $EY = 0$, then $\mathrm{Cov}(Y, Z) = EYZ$.

Information Inequality

If $d(X)$ is an unbiased estimator of θ, then again we take the derivative with respect to θ to see that

$$\begin{aligned}
\theta &= E_\theta[d(X)] = \int_{\mathcal{X}} d(x)\mathbf{f}(\mathbf{x}|\theta)\ \nu(d\mathbf{x}) \\
1 &= \int_{\mathcal{X}} d(x)\frac{\partial \mathbf{f}(\mathbf{x}|\theta)/\partial\theta}{\mathbf{f}(\mathbf{x}|\theta)}\ \mathbf{f}(\mathbf{x}|\theta)\ \nu(d\mathbf{x}) \\
1 &= \int_{\mathcal{X}} d(x)\left(\frac{\partial}{\partial\theta}\ln \mathbf{f}(\mathbf{x}|\theta)\right)\mathbf{f}(\mathbf{x}|\theta)\ \nu(d\mathbf{x}) = E_\theta\left[d(X)\frac{\partial}{\partial\theta}\ln \mathbf{f}(X|\theta)\right] \\
1^2 &= \mathrm{Cov}_\theta\left(d(X), \frac{\partial}{\partial\theta}\ln \mathbf{f}(X|\theta)\right)^2 \leq \mathrm{Var}_\theta(d(X)) \cdot \mathrm{Var}_\theta(\frac{\partial}{\partial\theta}\ln \mathbf{f}(X|\theta))
\end{aligned}$$

$$\frac{1}{\mathrm{Var}_\theta\left(\frac{\partial}{\partial\theta}\ln \mathbf{f}(X|\theta)\right)} \leq \mathrm{Var}_\theta(d(X))$$

Information Inequality

$$\mathrm{Var}_\theta(d(X)) \geq \frac{1}{I_n(\theta)}$$

where the Fisher information is the variance of the score function, $\frac{\partial}{\partial\theta} \ln \mathbf{f}(X|\theta)$.

$$\begin{aligned}
I_n(\theta) &= \mathrm{Var}_\theta\left(\frac{\partial}{\partial\theta} \ln \mathbf{f}(X|\theta)\right) = \mathrm{Var}_\theta\left(\frac{\partial}{\partial\theta} \ln f(X_1|\theta) f(X_2|\theta) \cdots f(X_n|\theta)\right) \\
&= \mathrm{Var}_\theta\left(\frac{\partial}{\partial\theta} \left(\ln f(X_1|\theta) + \ln f(X_2|\theta) + \cdots + \ln f(X_n|\theta)\right)\right) \\
&= \mathrm{Var}_\theta\left(\frac{\partial}{\partial\theta} \ln f(X_1|\theta) + \frac{\partial}{\partial\theta} \ln f(X_2|\theta) + \cdots + \frac{\partial}{\partial\theta} \ln f(X_n|\theta)\right) \\
&= \mathrm{Var}_\theta\left(\frac{\partial}{\partial\theta} \ln f(X_1|\theta)\right) + \mathrm{Var}_\theta\left(\frac{\partial}{\partial\theta} \ln f(X_2|\theta)\right) + \cdots + \mathrm{Var}_\theta\left(\frac{\partial}{\partial\theta} \ln f(X_n|\theta)\right) \\
&= n I_1(\theta)
\end{aligned}$$

Thus, the information increases linearly with the number of observations.

Information Inequality

An alternate expression for Fisher information is

$$I(\theta) = -E_\theta \left[\frac{\partial^2}{\partial \theta^2} \ln \mathbf{f}(X|\theta) \right]$$

Exercise. For $X_1, \ldots, X_n$ independent $N(\mu, \sigma_0^2)$, σ_0 known, we estimate μ with $\bar{X}$. Then

$$E_\mu[\bar{X}] = \mu \quad \text{and} \quad \mathrm{Var}_\mu(\bar{X}) = \sigma_0^2/n.$$

- Show that $I_1(\mu) = 1/\sigma_0^2$ and so information is the *inverse* of the variance.
- Show that

$$\mathrm{Var}_\mu(\bar{X}) = \frac{1}{I_n(\mu)}$$

and so $\bar{X}$ has the minimum possible variance for an unbiased estimator.

Information Inequality

The normal density,

$$
\begin{aligned}
f_X(x|\mu) &= \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma_0^2}\right) \\
\ln f_X(x|\mu) &= \frac{1}{2}\ln(2\pi\sigma_0^2) - \frac{(x-\mu)^2}{2\sigma_0^2} \\
\frac{\partial}{\partial\mu}\ln f_X(x|\mu) &= \frac{x-\mu}{\sigma_0^2} \\
\frac{\partial^2}{\partial\mu^2}\ln f_X(x|\mu) &= -\frac{1}{\sigma_0^2} \\
I(\mu) = -E_\mu\left[\frac{\partial^2}{\partial\mu^2}\ln f_X(x|\mu)\right] &= \frac{1}{\sigma_0^2}
\end{aligned}
$$

Information Inequality

Thus, $I_n(\mu) = nI(\mu) = n/\sigma_0^2$ and

$$\mathrm{Var}_\mu(\bar{X}) = \frac{\sigma_0^2}{n} = \frac{1}{I_n(\mu)}$$

NB. The information inequality is frequently called the Cramér-Rao lower bound in recognition of Harald Cramér in Sweden and C. R. Rao in India who were among the first to derive it.

Exponential Families

Recall that equality in the Cauchy-Schwarz inequality occurs precisely when the correlation is ± 1. This happens when the estimator $Z = d(X)$ and the score function $Y = \partial \ln f_X(X|\theta)/\partial\theta$ are linearly related with probability 1, i.e.,

$$\frac{\partial}{\partial\theta} \ln f_X(X|\theta) = \beta(\theta)d(X) + \alpha(\theta).$$

After integrating, we obtain,

$$\ln f_X(X|\theta) = \int \beta(\theta)d\theta d(X) + \int \alpha(\theta)d\theta \ + j(X) = \eta(\theta)d(X) - A(\theta) + j(X)$$

Note that the constant of integration is a function of X. Now exponentiate both sides of this equation

$$f_X(X|\theta) = h(X)\exp(\eta(\theta)d(X) - A(\theta))$$

Here $h(X) = \exp j(X)$. Thus, $f_X(X|\theta)$ is an exponential family.

Exponential Families

If we parameterize using the natural parameter η, then

$$f_X(X|\eta) = h(X)\exp(\eta d(X) - \tilde{A}(\eta))$$

with sufficient statistic $d(x)$. In this case,

$$\begin{aligned} \ln f_X(X|\theta) &= \ln h(X) + \eta d(X) - \tilde{A}(\eta) \\ \frac{d^2}{d\eta^2} \ln f_X(X|\theta) &= -\tilde{A}''(\eta) \\ I(\eta) &= \mathrm{Var}_\eta(d(X)) \end{aligned}$$

Thus, if we have independent random variables $X_1, X_2, \ldots X_n$, then the joint density is the product of the marginal densities, and

$$\overline{d(X)} = \frac{1}{n}(d(X_1) + \cdots + d(X_n))$$

is an unbiased estimator for η

Exponential Families

Example. For $X_1, \ldots, X_n$ independent $N(\mu, \sigma_0^2)$, σ_0 known, we write the density in the canonical expression for an exponential family

$$f_X(X|\mu) = \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-x^2/2\sigma_0^2} \exp\left(\mu \frac{x}{2\sigma_0^2} - \frac{\mu^2}{2\sigma_0^2}\right)$$

Thus,

$$\eta = \mu, \quad d(x) = \frac{x}{2\sigma_0^2}, \quad A(\mu) = -\frac{\mu^2}{2\sigma_0^2}$$

and

$$I(\mu) = -A''(\mu) = \frac{1}{\sigma_0^2}$$

Exponential Families

If we have that the parameter θ appears in the density as a function $\eta = \eta(\theta)$, then we have two forms for the Fisher information, I_θ and I_η for each parameterization. To connect the two expression

$$\begin{aligned} I_\theta(\theta) &= E_\theta\left[\left(\frac{\partial}{\partial\theta}\ln \mathbf{f}(X|\theta)\right)^2\right] = E_{\eta(\theta)}\left[\left(\frac{\partial}{\partial\theta}\ln \mathbf{f}(X|\eta(\theta))\right)^2\right] \\ &= E_{\eta(\theta)}\left[\left(\frac{\partial}{\partial\eta}\ln \mathbf{f}(X|\eta(\theta))\cdot\frac{d\eta(\theta)}{d\theta}\right)^2\right] = I_\eta(\eta(\theta))\left(\frac{d\eta(\theta)}{d\theta}\right)^2. \end{aligned}$$

The second equality uses the chain rule.