Topic 14
Unbiased Estimation
Compensating for Bias and the Information Inequaltiy

Outline

Compensating for Bias

Let's return to the method of moments estimator for the Pareto distribution with parameter β.

- The transformation
$$\beta = g(\mu) = \frac{\mu}{\mu - 1}$$
is a convex function of the distribution mean μ.
- For a convex function, the tangent lines lie *below* the function

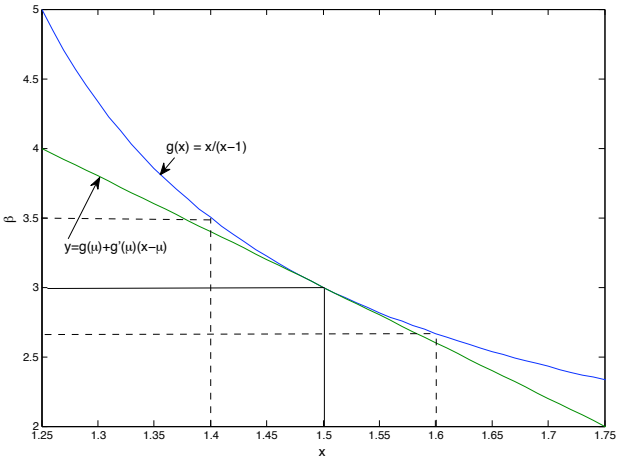
Let examine the case having $\mu = 1.5$ and so $\beta = 3$.

Compensating for Bias

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.



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.

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

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

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

$$E\bar{X}_n = \mu,$$

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 is continuous at μ, the method of moments estimators $\hat{\theta}_n$ satisfy

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

and so we have that $g(\bar{X}_n)$ is a consistent sequence of estimators for θ.

Information Inequality

The smallest possible variance of a unbiased estimator can be derived from the property that the correlation ρ of two random variables Y and Z satisfies

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

or

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

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

$$\theta = E_\theta d(X) = \int_{\mathbb{R}^n} d(\mathbf{x})\mathbf{f}(\mathbf{x}|\theta)\ d\mathbf{x}.$$

Information Inequality

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

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

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

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

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

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 $N(\mu, \sigma)$, we estimate μ with $\bar{X}$. Then

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

- Show that $I_1(\mu) = 1/\sigma^2$ 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.

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.