Stat 4620: Day 21

(4/9)

1 Sec. 6.4: Multiparameter estimation

Let $x_1, \ldots, x_n$ be a random sample from a normal distribution with mean θ_1 and standard deviation θ_2, i.e.

$$f(x; \theta_1, \theta_2) = \frac{1}{\sqrt{2\pi\theta_2^2}} e^{-\frac{(x-\theta_1)^2}{2\theta_2^2}}, \quad -\infty < x < \infty$$

Definition Let $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_p]'$.

- Let $L(\boldsymbol{\theta}) = \prod f(x_i; \boldsymbol{\theta}),\ \boldsymbol{\theta} \in \Omega$ where $\Omega \subset R_p$ be the *likelihood function.*
- The *maximum likelihood estimator* is $\hat{\boldsymbol{\theta}}$ such that

$$L(\hat{\boldsymbol{\theta}}) \geq L(\boldsymbol{\theta}), \quad \text{for all } \boldsymbol{\theta} \in \Omega$$

- Similarly to the one-parameter case, it can be shown that $g(\hat{\boldsymbol{\theta}})$ is the maximum likelihood estimator of $g(\boldsymbol{\theta})$

Example 6.4.1
Let $x_1, \ldots, x_n$ be a random sample from $N(\mu, \sigma^2)$. Find the mle of $\boldsymbol{\theta} = [\mu, \sigma]$.

Solution.

$$f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$L(\mu, \sigma) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} e^{-\frac{\sum(x_i-\mu)^2}{2\sigma^2}}$$

$$l(\mu, \sigma) = -\frac{n}{2}\ln(2\pi) - n\ln\sigma - \frac{\sum(x_i-\mu)^2}{2\sigma^2}$$

The *normal equations* are

1. $\frac{\partial}{\partial\mu} l(\mu, \sigma) = \frac{1}{\sigma^2}\sum(x_i - \mu)^2 = 0$
2. $\frac{\partial}{\partial\sigma} l(\mu, \sigma) = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum(x_i - \mu)^2 = 0$

which is a set of two equations with two unknowns. Equation 1 implies

$$0 = \sum x_i - n\hat{\mu} \ \Rightarrow\ \hat{\mu} = \overline{x}$$

Substituting into equation 2,

$$0 = -\frac{n}{\hat{\sigma}} + \frac{1}{\hat{\sigma}^3}\sum(x_i - \overline{x})^2 \ \Rightarrow\ \hat{\sigma} = \sqrt{\frac{\sum(x_i - \overline{x})^2}{n}}$$

so that

$$\hat{\boldsymbol{\theta}}_{mle} = \begin{bmatrix} \hat{\mu} \\ \hat{\sigma} \end{bmatrix} = \begin{bmatrix} \overline{x} \\ \sqrt{\frac{\sum(x_i-\overline{x})^2}{n}} \end{bmatrix}$$

Suppose we used the parameterization $\boldsymbol{\theta} = [\mu, \sigma^2]$, i.e. θ_2 represents the variance instead of the standard deviation. Then we rewrite the log-likelihood as

$$l(\mu, \sigma^2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{\sum(x_i-\mu)^2}{2\sigma^2}$$

and the *normal equations* are

1. $\frac{\partial}{\partial\mu} l(\mu, \sigma^2) = \frac{1}{\sigma^2}\sum(x_i-\mu)^2 = 0$
2. $\frac{\partial}{\partial\sigma^2} l(\mu, \sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{(\sigma^2)^2}\frac{\sum(x_i-\mu)^2}{2} = 0$

Solving simultaneously for μ and σ^2, we get

$$\hat{\boldsymbol{\theta}}_{mle} = \begin{bmatrix} \hat{\mu} \\ \hat{\sigma^2} \end{bmatrix} = \begin{bmatrix} \overline{x} \\ \frac{\sum(x_i-\overline{x})^2}{n} \end{bmatrix}$$

■

6.4.1 Fisher Information matrix

The gradient of $\ln f(x;\boldsymbol{\theta})$ is

$$\nabla \ln f(x;\boldsymbol{\theta}) = \begin{bmatrix} \frac{\partial}{\partial\theta_1}\ln f(x;\boldsymbol{\theta}) \\ \vdots \\ \frac{\partial}{\partial\theta_p}\ln f(x;\boldsymbol{\theta}) \end{bmatrix}$$

The Fisher information is then defined by the $p \times p$ matrix

$$\mathbf{I}(\boldsymbol{\theta}) = COV\left[\nabla \ln f(x;\boldsymbol{\theta})\right]$$

$$= \begin{bmatrix} var\left(\frac{\partial}{\partial\theta_1}\ln f(x;\boldsymbol{\theta})\right) & cov\left(\frac{\partial}{\partial\theta_1}\ln f(x;\boldsymbol{\theta}), \frac{\partial}{\partial\theta_2}\ln f(x;\boldsymbol{\theta})\right) & \dots & cov\left(\frac{\partial}{\partial\theta_1}\ln f(x;\boldsymbol{\theta}), \frac{\partial}{\partial\theta_p}\ln f(x; \text{[illegible]} \\ cov\left(\frac{\partial}{\partial\theta_1}\ln f(x;\boldsymbol{\theta}), \frac{\partial}{\partial\theta_2}\ln f(x;\boldsymbol{\theta})\right) & var\left(\frac{\partial}{\partial\theta_2} f(x;\boldsymbol{\theta})\right) & \dots & cov\left(\frac{\partial}{\partial\theta_2}\ln f(x;\boldsymbol{\theta}), \frac{\partial}{\partial\theta_p}\ln f(x; \text{[illegible]} \\ \vdots & & & \vdots \\ cov\left(\frac{\partial}{\partial\theta_1}\ln f(x;\boldsymbol{\theta}), \frac{\partial}{\partial\theta_p}\ln f(x;\boldsymbol{\theta})\right) & cov\left(\frac{\partial}{\partial\theta_2}\ln f(x;\boldsymbol{\theta}), \frac{\partial}{\partial\theta_p}\ln f(x;\boldsymbol{\theta})\right) & \dots & var\left(\frac{\partial}{\partial\theta_p}\ln f(x;\boldsymbol{\theta})\right) \end{bmatrix}$$

$$\vdots$$

$$= \begin{bmatrix} -E\frac{\partial^2}{\partial\theta_1^2}\ln f(x;\boldsymbol{\theta}) & -E\frac{\partial^2}{\partial\theta_1\partial\theta_2}\ln f(x;\boldsymbol{\theta}) & \dots & -E\frac{\partial^2}{\partial\theta_1\partial\theta_p}\ln f(x;\boldsymbol{\theta}) \\ -E\frac{\partial^2}{\partial\theta_1\partial\theta_2}\ln f(x;\boldsymbol{\theta}) & -E\frac{\partial^2}{\partial\theta_2^2}\ln f(x;\boldsymbol{\theta}) & \dots & -E\frac{\partial^2}{\partial\theta_2\partial\theta_p}\ln f(x;\boldsymbol{\theta}) \\ \vdots & & & \vdots \\ -E\frac{\partial^2}{\partial\theta_1\partial\theta_p}\ln f(x;\boldsymbol{\theta}) & -E\frac{\partial^2}{\partial\theta_2\partial\theta_p}\ln f(x;\boldsymbol{\theta}) & \dots & -E\frac{\partial^2}{\partial\theta_p^2}\ln f(x;\boldsymbol{\theta}) \end{bmatrix}$$

Comments:

1. Since $\ln f(x_1,\dots,x_n;\boldsymbol{\theta}) = \sum \ln f(x_i;\boldsymbol{\theta})$, then the information for a random sample is

$$\mathbf{I}_n(\boldsymbol{\theta}) = n\mathbf{I}(\boldsymbol{\theta})$$

2. (Rao-Cramer lower bound) If $Y = u(x_1, \dots, x_n)$ is an unbiased estimator of θ_j then

$$V(Y) \geq \frac{1}{n}\left[\mathbf{I}^{-1}(\boldsymbol{\theta})\right]_{jj}$$

3. We call an estimator *efficient* if it attains the lower bound

Example 6.4.3 (Information matrix of $N(\mu, \sigma^2)$)

$$f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$\ln f(x;\mu,\sigma) = -\frac{1}{2}\ln(2\pi) - \ln\sigma - \frac{(x-\mu)^2}{2\sigma^2}$$

The first and second partial derivatives are

$$\frac{\partial \ln f}{\partial \mu} = \frac{x-\mu}{\sigma^2}$$

$$\frac{\partial^2 \ln f}{\partial \mu^2} = -\frac{1}{\sigma^2}$$

$$\frac{\partial \ln f}{\partial \sigma} = -\frac{1}{\sigma} + \frac{1}{\sigma^3}(x-\mu)^2$$

$$\frac{\partial^2 \ln f}{\partial \sigma^2} = \frac{1}{\sigma^2} - \frac{3}{\sigma^4}(x-\mu)^2$$

$$\frac{\partial^2 \ln f}{\partial \mu \partial \sigma} = \frac{-2}{\sigma^3}(x-\mu)$$

The information matrix is

$$\mathbf{I}(\mu,\sigma) = \begin{bmatrix} -E\left(\frac{\partial^2 \ln f}{\partial \mu^2}\right) & -E\left(\frac{\partial^2 \ln f}{\partial \mu \partial \sigma}\right) \\ -E\left(\frac{\partial^2 \ln f}{\partial \mu \partial \sigma}\right) & -E\left(\frac{\partial^2 \ln f}{\partial \sigma^2}\right) \end{bmatrix} = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{bmatrix}$$

Note that

$$\frac{1}{n}\mathbf{I}^{-1}(\mu,\sigma) = \frac{1}{n}\begin{bmatrix} \sigma^2 & 0 \\ 0 & \frac{\sigma^2}{2} \end{bmatrix}$$

so if $Y_1 = u(x_1, \dots, x_n)$ is an unbiased estimator of μ, then

$$V(Y_1) \geq \frac{\sigma^2}{n}$$

and if $Y_1 = w(x_1, \dots, x_n)$ is an unbiased estimator of σ, then

$$V(Y_2) \geq \frac{\sigma^2}{2n}$$

Theorem 6.4.1. *(The mle is consistent, asymptotically normal and efficient) Let $x_1, \dots, x_n$ be a random sample from a distribution $f(x;\boldsymbol{\theta})$ for $\boldsymbol{\theta} \in \Omega$. Assume the regularity conditions hold, including existence of a unique solution $\hat{\boldsymbol{\theta}}$ to the normal equations. Then*

1. $\hat{\boldsymbol{\theta}} \xrightarrow{P} \boldsymbol{\theta}$

2. $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{D} N_p(0, \mathbf{I}^{-1}(\boldsymbol{\theta}))$

(See, e.g. Lehmann and Casella (1998) for proof.)

Theorem 6.4.2. *Let* $\mathbf{g}(\boldsymbol{\theta}) = [g_1(\boldsymbol{\theta}), \ldots, g_k(\boldsymbol{\theta})]'$ *be such that* $k \leq p$ *and the* $k \times p$ *matrix of partial derivatives*

$$\mathbf{B} = \left[\frac{\partial g_i}{\partial \theta_j}\right], \quad i = 1, \ldots, k, \quad j = 1, \ldots, p$$

has continuous elements and does not vanish in a neighborhood of $\boldsymbol{\theta}$*. Then*

1. $\mathbf{g}(\hat{\boldsymbol{\theta}})$ *is the mle of* $\mathbf{g}(\boldsymbol{\theta})$

2. $\sqrt{n}(\mathbf{g}(\hat{\boldsymbol{\theta}}) - \mathbf{g}(\boldsymbol{\theta}) \xrightarrow{D} N_p(0, B\mathbf{I}^{-1}(\boldsymbol{\theta})B')$

3. *The information matrix for* $\mathbf{g}(\boldsymbol{\theta})$ *is*

$$\left[B\mathbf{I}^{-1}(\boldsymbol{\theta})B')\right]^{-1}$$

provided the inverse exists.

Example 6.4.3 con't. Recall that the information matrix for $[\mu, \sigma]$ is

$$\mathbf{I}(\mu, \sigma) = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{bmatrix}$$

Find the information matrix for $[\mu, \sigma^2]$.

Solution. Note that $\mathbf{g}(\mu, \sigma) = (\mu, \sigma^2)$. The 2×2 matrix of partial derivatives is

$$B = \begin{bmatrix} \frac{\partial \mu}{\partial \mu} & \frac{\partial \mu}{\partial \sigma} \\ \frac{\partial \sigma^2}{\partial \mu} & \frac{\partial \sigma^2}{\partial \sigma} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 2\sigma \end{bmatrix}$$

and

$$B\mathbf{I}^{-1}(\mu, \sigma)B' = \begin{bmatrix} 1 & 0 \\ 0 & 2\sigma \end{bmatrix} \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{bmatrix}^{-1} \begin{bmatrix} 1 & 0 \\ 0 & 2\sigma \end{bmatrix} = \begin{bmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{bmatrix}$$

so the information matrix for $[\mu, \sigma^2]$ is

$$\mathbf{I}(\mu, \sigma^2) = \left[B\mathbf{I}^{-1}(\mu, \sigma)B'\right]^{-1} = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{bmatrix}$$

The Rao-Cramer lower bound for the variance of an estimator of σ^2 is $\frac{2\sigma^4}{n}$. The sample variance is unbiased for σ^2, but its variance is $(2\sigma^4)/(n-1)$. It is not efficient for finite sample size, but is asymptotically efficient. ■