

Prof. H. Bensmail

Stat-664: Adv Stat Infer

Spring Semester 2006

Chapter 2:

Likelihood inference

February 1, 2006

Likelihood function:

y is a random process

$p.d.f$ of y is $p(y|\theta)$ is a function of y

Likelihood is a function of θ that is proportional to $p(y|\theta)$

$$L(\theta) = L(\theta|y) \propto p(y|\theta)$$

- $p.d.f$ $p(y|\theta)$ is varying with y at fixed values of θ
- likelihood $L(\theta|y)$ is a function of the parameter θ for fixed data.

$$\begin{aligned} L(\theta|y) &= k(y)p(y|\theta) \\ &\propto p(y|\theta) \end{aligned}$$

Maximum likelihood estimator:

condition for θ to be a M.L.E

$$\left\{ \begin{array}{l} \frac{l^{(1)}(\theta|y)}{\partial\theta} = 0 \\ \frac{\partial l^{(2)}(\theta|y)}{\partial\theta\partial\theta'} < 0 \end{array} \right\}$$

Gaussian model:

$$y = 10 \sim N(m, 3)$$

$$L(m|y = 10) = \underbrace{\frac{1}{\sqrt{6\pi}}}_{k(y)} \underbrace{\exp\left(-\frac{(10-m)^2}{6}\right)}_{f(m)}$$

$$L(m) = \exp\left(-\frac{(10-m)^2}{6}\right)$$

if $y = (y_1 = 10, y_2 = 8, y_3 = 12, y_4 = 9, y_5 = 11)$

$$\begin{aligned} & L(m|\sigma^2 = 3, y) \\ &= \frac{1}{(\sqrt{6\pi})^5} \exp\left(-\frac{(10-m)^2 + \dots + (11-m)^2}{6}\right) \\ &= k(y) \times \exp\left(\frac{-5(\bar{y} - m)^2}{6}\right); \hat{m} = \bar{y} \end{aligned}$$

Fischer (1922) proposed to rescale likelihood values relative to $\hat{\theta}$:

$$r(\theta) = \frac{L(\theta|y)}{L(\hat{\theta}|y)}$$

so that rescaled values are between 0 and 1.

Fisher's information measure:

$$I(\theta) = E_y \left[\left(\frac{dl}{d\theta} \right)^2 \right] > 0$$

$$l = c + \log p(y|\theta)$$

$$= c + \sum_{i=1}^n \log p(y_i|\theta)$$

$$I(\theta) = \sum_{i=1}^n I_i(\theta)$$

$$I_i(\theta) = E \left[\left(\frac{d \log p(y_i|\theta)}{d\theta} \right)^2 \right]$$

because

$$E \left(\frac{d}{d\theta} \log p(y_i|\theta) \right) = 0; \quad i = 1, \dots, n$$

and also

$$E \left(\frac{d}{d\theta} \log p(y_i|\theta) \times \frac{d}{d\theta} \log p(y_j|\theta) \right) \\ E \left(\frac{d}{d\theta} \log p(y_i|\theta) \right) \times E \left(\frac{d}{d\theta} \log p(y_j|\theta) \right) \\ \text{independency}$$

and

$$E \left(\frac{d}{d\theta} \log p(y_i|\theta) \right) = 0$$

Here we show that:

$$I(\theta) = -E_y(l'')$$

Alternative representation of the Information

$$l = \log p(y|\theta)$$

$$l' = \frac{dl}{d\theta}; l'' = \frac{d^2l}{d\theta^2}$$

$$I(\theta) = -E_y(l'') = -\int l'' p(y|\theta) dy$$

$$= E_y(l')^2 = E_y(-l'')$$

$$E_y(l') = 0 \text{ and } \int l'' p(y|\theta) dy + \int (l')^2 p(y|\theta) dy = 0$$

means:

$$\begin{aligned} I(\theta) &= E_y[(l')^2] \\ &= -E_y(l'') \end{aligned}$$

Curvature as a measure of information:

◦ Curvature

$$c(\theta) = \frac{l''(\theta)}{[1 + l'(\theta)^2]^{3/2}}$$

$$l'(\hat{\theta}) = 0 \Rightarrow c(\hat{\theta}) = l''(\hat{\theta})$$

◦ *Example* : Quadratic approximation to log-likelihood:

Taylor:

$$\begin{aligned} l(\theta) &\simeq l(\hat{\theta}) + l'(\hat{\theta})(\theta - \hat{\theta}) + \frac{1}{2}l''(\hat{\theta})(\theta - \hat{\theta})^2 \\ &= l(\hat{\theta}) - \frac{1}{2}J(\hat{\theta})(\theta - \hat{\theta})^2 \end{aligned}$$

$$\left\{ \begin{array}{l} \frac{L(\theta|y)}{L(\hat{\theta}|y)} \approx \exp\left[-\frac{1}{2}J(\hat{\theta})(\theta - \hat{\theta})^2\right] \\ \hat{\theta} \sim N(\theta, J(\hat{\theta})^{-1}) \end{array} \right\}$$

$$p(\hat{\theta}) \simeq (2\pi)^{-1/2} |J(\hat{\theta})|^{1/2} \frac{L(\theta|y)}{L(\hat{\theta}|y)}$$

Score function:

$$S(\theta|y) = l'$$

S is a function of y so it is a random variable with unknown distribution which has a mean =0

We show that

$$l' = S(\theta|y) \sim [0, I(\theta)]$$

$$\begin{aligned} \text{proof: } \text{Var}(l') &= E[(l')^2] - [E(l')]^2 \\ &= E[(l')^2] = I(\theta) \end{aligned}$$

Example: Normal distribution

$$y_1, \dots, y_n \sim N(m, \sigma^2)$$

$$S = \frac{dl}{d\mu}(\mu|\sigma^2, y) = \frac{n(\bar{y} - \mu)}{\sigma^2}$$

$$\begin{aligned} E(S) &= E_y\left(\frac{n(\bar{y} - \mu)}{\sigma^2}\right) \\ &= \frac{n}{\sigma^2} E_y(\bar{y} - \mu) = 0 \end{aligned}$$

$$\begin{aligned} \text{Var}(S) &= \text{Var}\left(\frac{n(\bar{y} - \mu)}{\sigma^2}\right) = \frac{n}{\sigma^2} \\ S &\sim \left(0, \frac{n}{\sigma^2}\right) \end{aligned}$$

Multivariate:

$$\begin{aligned} I(\theta) &= E_y \left[\left(\frac{\partial l}{\partial \theta} \right) \left(\frac{\partial l}{\partial \theta} \right)' \right] \\ &= E_y \left[\frac{\partial^2 l}{\partial \theta \partial \theta'} \right] \\ S &\sim [0, \mathbf{I}(\theta)] \end{aligned}$$

Example:

$$y \sim N(X\beta, V)$$

β : unknown

X : Matrix of explanatory variable

V : known variance matrix

(gave a review of the matricial and vectorial derivation: Thursday 12: Anderson 1984)

$$L(\beta|V, y) \propto \exp \left[-\frac{1}{2} (y - X\beta)' V^{-1} (y - X\beta) \right]$$

$$\begin{aligned} S(\beta|y) &= \frac{\partial}{\partial \beta} \left[-\frac{1}{2} (y - X\beta)' V^{-1} (y - X\beta) \right] \\ &= X' V^{-1} (y - X\beta) \end{aligned}$$

$$\begin{aligned} \mathbf{I}(\beta) &= E \left(\frac{\partial^2 l}{\partial \beta \partial \beta'} \right) = E \left(\frac{\partial S(\beta|y)}{\partial \beta'} \right) \\ &= E(X' V^{-1} X) = X' V^{-1} X \end{aligned}$$

Cramer-Rao Lower Bound:

if $E(\hat{\theta}) = m(\theta)$, then Cramer-Rao Bound is

$$\text{Var}(\hat{\theta}) \geq \frac{[m'(\theta)]^2}{I(\theta)}$$

if the $\text{Var}(\hat{\theta}) \geq \frac{[m'(\theta)]^2}{I(\theta)}$, then θ is said to have a minimum variance.

if θ is unbiased, then CRB is $\text{Var}(\hat{\theta}) \geq I(\theta)^{-1}$

Using Cauchy-Schwarz inequality, we have

$$\text{cov}(\hat{\theta}, l') = [m'(\theta)]^2 \leq \text{Var}(\hat{\theta})I(\theta)$$

◦Consistency:

$\hat{\theta}_n$ is an estimator of θ , based on y_1, \dots, y_n

$\hat{\theta}_n$ is consistent if

$$P_r \{ |\hat{\theta}_n - \theta| < \epsilon \} \xrightarrow{n \rightarrow \infty} 1$$

(the sampling distribution of $\hat{\theta}_n$ becomes more and more concentrated around θ .)

◦Asymptotic normality and efficiency:

Under conditions:

1. θ lies inside Ω .
2. The support of p.d.f of data $A = \{y; p(y|\theta) > 0\}$ is independent of θ
3. $F(y|\theta_1) = F(y|\theta_2) \Rightarrow \theta_1 = \theta_2$ (the distribution of observations are distinct)
4. l', l'', l''' exist and l''' is continuous
5. $\int p(y|\theta) dy$ is three times differentiable under the integral sign
6. $0 < I(\theta) < \infty$
7. there exist $M(y)$, such that $|\frac{d^3}{(d\theta)^3}(\log p(y|\theta))| \leq M(y)$

then
when n is large

$$\sqrt{n} (\hat{\theta}_n - \theta_0) \rightarrow N\left(0, \frac{n}{I(\theta_0)}\right)$$

So the MLE is asymptotically unbiased, efficient and normal.

The same is true for multivariate models:

$$\hat{\theta}_n \sim N(\theta_0, \mathbf{I}^{-1}(\theta_0))$$

So the MLE is asymptotically unbiased, efficient and multivariate normal.

Functional invariance of MLE:

If $\hat{\theta}$ is MLE of θ then:
 $f(\hat{\theta})$ is MLE of $f(\theta)$.

Example:

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

$$\theta = (\beta_0, \beta_1)$$

$$\hat{\beta}_1 = \frac{\sum_i x_i y_i - \frac{1}{n} \sum_i x_i \sum_i y_i}{\sum_i x_i^2 - \frac{1}{n} (\sum_i x_i)^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$f(\theta) = \frac{\beta_0}{\beta_1} \quad \text{MLE of } f(\theta) \text{ is } \frac{\hat{\beta}_0}{\hat{\beta}_1}$$

In fact:

$$l(\theta, \sigma^2 | y) = -\frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_i (y_i - \beta_0 - \beta_1 x_i)^2$$

$$\eta = \frac{\beta_0}{\beta_1}$$

$$g : \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \rightarrow \begin{pmatrix} \eta \\ \beta_1 \end{pmatrix}$$

we have one to one transformation

$$l(\eta, \beta_1, \sigma^2 | y) = -\frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_i (y_i - \eta \beta_1 - \beta_1 x_i)^2$$

$$\begin{pmatrix} \frac{\partial l}{\partial \eta} \\ \frac{\partial l}{\partial \beta_1} \\ \frac{\partial l}{\partial \sigma^2} \end{pmatrix} = \mathbf{0}$$

$$\hat{\eta} = \frac{\bar{y} - \hat{\beta}_1 \bar{x}}{\hat{\beta}_1}$$

$$\hat{\beta}_1 = \frac{\sum_i x_i y_i - \frac{1}{n} \sum_i x_i \sum_i y_i}{\sum_i x_i^2 - \frac{1}{n} (\sum_i x_i)^2}$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_i (y_i - \hat{\eta} \hat{\beta}_1 - \hat{\beta}_1 \bar{x})^2$$

$$\hat{\beta}_0 = \hat{\eta} \hat{\beta}_1$$