STAT 517:Maximum Likelihood Methods

Lecture 9: Multiparameter case: Testing

Prof. Michael Levine

February 9, 2016

Framework

- $X_1, \ldots, X_n \sim f(x; \boldsymbol{\theta})$ for $\boldsymbol{\theta} \subset \Omega \in \mathbb{R}^p$
- The likelihood function is

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i; \boldsymbol{\theta})$$

- The log-likelihood function is

$$l(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(x_i; \boldsymbol{\theta})$$

- As before, with probability going to 1, $L(\boldsymbol{\theta})$ is maximized at the true value of $\boldsymbol{\theta}$
- An MLE $\hat{\boldsymbol{\theta}}$ is a maximizer of $L(\boldsymbol{\theta})$ or a solution of the normal equation

$$\frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 0$$

- Again, as before, the MLE of $\boldsymbol{\eta} = g(\boldsymbol{\theta})$ is $\hat{\boldsymbol{\theta}} = g(\hat{\boldsymbol{\theta}})$

Example

- Take $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ with $\boldsymbol{\theta} = (\mu, \sigma^2)'$ and $\Omega = (-\infty, \infty) \times (0, \infty)$
- The *system* of estimating equations is now

$$\frac{\partial l}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) = 0$$

and

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

- Solutions are $\hat{\mu} = \bar{X}$ and $\hat{\sigma} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2}$; Moreover, the MLE of σ^2 is $\frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2$
- These estimators are both consistent with μ being unbiased and $\hat{\sigma}^2$ being asymptotically unbiased

Example

- Laplace family $F_X(x) = \frac{1}{2b} e^{\{-|x-a|/b\}}$ for $-\infty < x < \infty$
- The parameter space is $\Omega = \{(a, b) : -\infty < a < \infty, b > 0\}$
- The estimating equation for a is

$$\frac{\partial l(a, b)}{\partial a} = \frac{1}{b} sgn\{x_i - a\}$$

  and the MLE of a is $Q_2 = \text{med}\{X_1, \ldots, X_n\}$ *regardless* of the value of b
- Simultaneous solution of *both* equations produces also the MLE of b as $\hat{b} = \frac{1}{n} \sum_{i=1}^{n} |x_i - a|$

Fisher information

- The score function is the gradient $\nabla \log f(x;\boldsymbol{\theta})$ while the Fisher information is now a $p \times p$ matrix

$$I(\boldsymbol{\theta}) = Cov\left(\nabla \log f(X;\boldsymbol{\theta})\right)$$

- Using regularity conditions, it is easy to verify that

$$I_{jk} = \mathbb{E}\left[\frac{\partial}{\partial\theta_j}\log f(X;\boldsymbol{\theta})\right] = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta_j\partial\theta_k}\log f(X;\boldsymbol{\theta})\right]$$

Rao-Cramer lower bound

- For a sample $X_1, \ldots, X_n$ the gradient is $\nabla \log L(\boldsymbol{\theta}; \mathbf{X}) = \sum_{i=1}^{n} \nabla \log f(X_i; \boldsymbol{\theta})$ and the $Cov\ (\nabla \log L(\boldsymbol{\theta}; \mathbf{X})) = nI(\boldsymbol{\theta})$
- The diagonal entries of $I(\boldsymbol{\theta})$ are

$$I_{ii}(\boldsymbol{\theta}) = Var\left[\frac{\partial \log f(\mathbf{X}; \boldsymbol{\theta})}{\partial \theta_i}\right] = -\mathbb{E}\left[\frac{\partial^2}{\partial \theta_i^2} \log f(X; \boldsymbol{\theta})\right]$$

- If $Y_j = u_j(X_1, \ldots, X_n)$ is an unbiased estimator of θ_j,

$$Var\ Y_j \geq \frac{1}{n}[I^{-1}(\boldsymbol{\theta})]_{jj}$$

- The estimator is **efficient** if its variance attains this lower bound

Example

- $X \sim N(\mu, \sigma^2)$ with both μ and σ^2 unknown
- Second partial derivatives produce $I_{11} = \frac{1}{\sigma^2}$
- On the other hand, the MLE of μ is $\bar{X}$ and so $\bar{X}$ is an efficient estimator of μ for finite samples
- Note that the Fisher's information for μ does not depend on μ!

Location-Scale Family

- Let $X_1, \ldots, X_n$ with pdf $f_X(x) = \frac{1}{b} f\left(\frac{x-a}{b}\right)$ where $(a, b) \in \Omega$ with $\Omega = \{(a, b) : -\infty < a < \infty, b > 0\}$
- It is not hard to realize that

$$X_i = a + be_i$$

  where e_i are iid with pdf $f(z)$
- Verify that

$$I_{11} = \frac{1}{b^2} \int_{-\infty}^{\infty} \left[\frac{f'(z)}{f(z)}\right]^2 f(z)\, dz$$

  and so the information on the location parameter *never* depends on the value of the parameter a itself
- Verify that the off-diagonal entries of the information matrix are 0 if $f(z)$ is symmetric around 0

Multi(tri)nomial case example

- Snake's head flower is a very nice garden flower grown in many European countries...it comes in three color morphs: violet, white, pink
- Let the total number of flowers observed be n, with n_i in each of the three categories; respective probabilities are p_i, $i = 1, 2, 3$
- The trinomial likelihood is

$$l(\boldsymbol{\theta}) = n_1 \log p_1 + n_2 \log p_2 + (n - n_1 - n_2) \log(1 - (p_1 + p_2))$$

- The score vector is

$$\left(\frac{n_1}{p_1} - \frac{n - n_1 - n_2}{1 - p_1 - p_2}, \frac{n_2}{p_2} - \frac{n - n_1 - n_2}{1 - p_1 - p_2} \right)'$$

- Since each marginal $N_j \sim b(n, p_j)$ it is easy to verify that the expectation of the score vector is $(0, 0)'$

Multi(tri)nomial case example

- Verify that

$$I(\boldsymbol{\theta})_{jj} = \frac{np_j}{p_j^2} + \frac{n(1-p_1-p_2)}{(1-p_1-p_2)^2}$$

  and

$$I(\boldsymbol{\theta})_{12} = \frac{n(1-p_1-p_2)}{(1-p_1-p_2)^2}$$

- The information matrix is

$$\begin{pmatrix} \frac{1}{p_1}+\frac{1}{p_3} & \frac{1}{p_3} \\ \frac{1}{p_3} & \frac{1}{p_1}+\frac{1}{p_3} \end{pmatrix}$$

Multi(tri)nomial case example

- Let x_{ij} be the record of ith observation, $i = 1, \ldots, n$
- Verify that MLE $\hat{p}_h = \frac{\sum_{i=1}^{n} x_{ij}}{n}$ with variances $np_h(1 - p_h)$, $h = 1, \ldots, k - 1$ - MLE's are efficient in this case

Consistency and asymptotic normality of MLE

- Let $X_1, \ldots, X_n$ be iid with pdf $f(x; \boldsymbol{\theta})$ for $\theta \in \Omega$. All of the regularity conditions hold Then,
  1. The likelihood equation
  $$\frac{\partial}{\partial \boldsymbol{\theta}} l(\boldsymbol{\theta}) = 0$$
  has a solution $\hat{\theta}_n$ s.t. $\hat{\boldsymbol{\theta}}_n \xrightarrow{P} \boldsymbol{\theta}$
  2. For any sequence that satisfies (1),
  $$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{D} N_p(\mathbf{0}, I^{-1}(\boldsymbol{\theta}))$$

Practical corollary

- $\hat{\boldsymbol{\theta}}_n$ are asymptotically efficient estimators in the sense that for $j = 1, \ldots, p$

$$\sqrt{n}(\hat{\theta}_{n,j} - \theta_j) \xrightarrow{D} N(0, [I^{-1}(\boldsymbol{\theta})]_{jj})$$