CE085 - Estatística Inferencial
Vetor de parâmetros

Prof. Wagner Hugo Bonat

Laboratório de Estatística e Geoinformação - LEG
Curso de Bacharelado em Estatatística
Universidade Federal do Paraná - UFPR

5 de outubro de 2018

Conteúdo

Conteúdo

- Vetor escore e matrix de informação de Fisher.
- Desigualdade de Cramér-Rao (generalizada).
- Propriedades do MLE.
- Ortogonalidade.
- Modelos exponenciais de dispersão.
- Regressão Linear.
- Exemplos.

Configuração geral

- Sejam $Y_1, \ldots, Y_n$ v.a com fdp ou fp $f(y, \boldsymbol{\theta})$.
- Vetor de parâmetros: $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^\top \in \Omega$, onde Ω é uma região no $\Re^p$.
- Exemplo 1: $Y_i \sim G(\mu, \sigma^2)$

$$f(y, \boldsymbol{\theta}) = \frac{1}{1\sqrt{2\pi\sigma}} \exp\left\{ -\frac{1}{2\sigma^2}(y-\mu)^2 \right\}, \quad y \in \Re \quad \text{e} \quad \boldsymbol{\theta} = (\mu, \sigma)^\top$$

- Exemplo 2: $Y_i \sim Ga(\theta, \lambda)$

$$f(y, \boldsymbol{\theta}) = \frac{\theta^\lambda}{\Gamma(\lambda)} y^{\lambda - 1} \exp^{-\theta y}, \quad y > 0 \quad \text{e} \quad \boldsymbol{\theta} = (\theta, \lambda)^\top \in \Re^2_+.$$

Verossimilhança e log-verossimilhança

- Função de verossimilhança $L : \Omega \rightarrow [0, \infty]$ é uma função aleatória de um **vetor** de argumentos definida por

$$L(\theta|\mathbf{y}) = \prod_{i=1}^{n} f(y_i, \theta), \quad \text{para} \quad \theta \in \Omega.$$

- Função de log-verossimilhança $l : \Omega \rightarrow \Re$ é uma função aleatória de um **vetor** de argumentos definida por

$$l(\theta|\mathbf{y}) = \sum_{i=1}^{n} \log f(y_i, \theta).$$

- Equivalentemente,

$$l(\theta|\mathbf{y}) = \log L(\theta|\mathbf{y}).$$

Vetor escore

- O vetor escore $\boldsymbol{U}(\theta|\boldsymbol{y}) : \Omega \rightarrow \Re^p$ é um vetor aleatório $p \times 1$ definido por

$$\boldsymbol{U}(\theta|\boldsymbol{y}) = \frac{\partial l(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \boldsymbol{\theta}} = \begin{pmatrix} \frac{\partial l(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \theta_1} \\ \vdots \\ \frac{\partial l(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \theta_p} \end{pmatrix}.$$

- Notação popular em termos de gradiente

$$\boldsymbol{U}(\theta|\boldsymbol{y}) = \nabla_{\boldsymbol{\theta}} l(\theta|\boldsymbol{y}).$$

- Notação j-ésimo componente de $\boldsymbol{U}(\theta|\boldsymbol{y})$ por $\boldsymbol{U}_j(\theta|\boldsymbol{y})$.

Esperança do vetor escore

- O vetor escore satisfaz as igualdades de Bartlett, ou seja,

$$E(\boldsymbol{U}(\theta|\boldsymbol{Y})) = \boldsymbol{0},$$

isso significa que

$$E(\boldsymbol{U}_j(\theta|\boldsymbol{Y})) = 0, \quad \text{para} \quad j = 1, \ldots, p.$$

Matriz de informação esperada

- A matriz $p \times p$ definida por

$$\begin{aligned} \mathbf{I}(\boldsymbol{\theta}) &= \mathbf{Var}(\boldsymbol{U}(\boldsymbol{\theta}|\boldsymbol{Y})) \\ &= \mathrm{E}(\boldsymbol{U}(\boldsymbol{\theta}|\boldsymbol{Y})\boldsymbol{U}^\top(\boldsymbol{\theta}|\boldsymbol{Y})). \end{aligned}$$

  é chamada de matriz de informação esperada.
- As entradas j e k são expressadas por

$$\begin{aligned} \mathbf{I}_{jk}(\boldsymbol{\theta}) &= \mathrm{Cov}(\boldsymbol{U}_j(\boldsymbol{\theta}|\boldsymbol{Y}), \boldsymbol{U}_k(\boldsymbol{\theta}|\boldsymbol{Y})) \quad (1) \\ &= \mathrm{E}(\boldsymbol{U}_j(\boldsymbol{\theta}|\boldsymbol{Y}), \boldsymbol{U}_k(\boldsymbol{\theta}|\boldsymbol{Y}))). \quad (2) \end{aligned}$$

Reparametrização

- Seja $\theta = g(\psi), g : 1-1$ diferenciável o vetor escore fica dada por

$$\tilde{\mathbf{U}}(\theta|\mathbf{Y}) = \frac{\partial \boldsymbol{\theta}^\top}{\partial \boldsymbol{\psi}} \mathbf{U}(\theta|\mathbf{Y}).$$

- Matriz de informação esperada

$$\tilde{\mathbf{I}}(\psi) = \frac{\partial \boldsymbol{\theta}^\top}{\partial \boldsymbol{\psi}} \mathbf{I}(\boldsymbol{\theta}) \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\psi}^\top}$$

Matriz de informação observada

- A matriz $p \times p$ definida por

$$\mathrm{J}(\boldsymbol{\theta}) = -\frac{\partial^2 l(\boldsymbol{\theta}|\mathbf{y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}.$$

- As entradas j e k da matriz de informação observada é dada por

$$\mathrm{J}_{jk}(\boldsymbol{\theta}) = -\frac{\partial^2 l(\boldsymbol{\theta}|\mathbf{y})}{\partial \theta_j \partial \theta_k}.$$

- Segunda igualdade de Bartlett

$$\mathrm{I}(\boldsymbol{\theta}) = \mathrm{E}(\mathrm{J}(\boldsymbol{\theta})).$$

Desigualdade de Cramér-Rao generalizada

- Define $\mathrm{I}^{jk}(\boldsymbol{\theta}) = \{\mathrm{I}^{-1}(\boldsymbol{\theta})\}_{jk}$. Se $\tilde{\theta} = \tilde{\theta}(Y_1, \ldots, Y_n)$ é um estimador não-viciado para θ_1, ou seja,

$$\mathrm{E}(\tilde{\theta}) = \theta_1,$$

  então

$$\mathrm{Var}(\tilde{\theta}) \geq \mathrm{I}^{11}(\boldsymbol{\theta}).$$

- Demonstração análoga ao caso univariado usando a desigualdade de Cauchy-Schwarz generalizada.

Normalidade assintótica do vetor escore

- Vetor escore

$$\boldsymbol{U}(\boldsymbol{\theta}|\boldsymbol{Y}) = \sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}} \log f(\boldsymbol{\theta}|y_i),$$

é a soma de v.a. iid com média zero, ou seja,

$$\mathrm{E}\left\{\sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}} \log f(\boldsymbol{\theta}|y_i)\right\} = \boldsymbol{0}$$

e variância igual a matriz de informação de Fisher.

$$\mathrm{Var}\left\{\sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}} \log f(\boldsymbol{\theta}|y_i)\right\} = \mathrm{I}(\boldsymbol{\theta}).$$

Normalidade assintótica do vetor escore

- Usando a versão multivariada do Teorema Central do Limite, temos
$$\mathrm{I}^{-\frac{1}{2}}(\theta)\boldsymbol{U}(\theta|\boldsymbol{Y}) \xrightarrow{D} N_p(\mathbf{0}, \mathbf{I}),$$
onde $\mathbf{I}$ denota a $p \times p$ matriz identidade.
- De forma equivalente,
$$\boldsymbol{U}(\theta|\boldsymbol{Y}) \xrightarrow{D} N_p(\mathbf{0}, \mathrm{I}(\theta)).$$
- Note que pela lei dos grandes números
$$\mathrm{J}(\theta) \xrightarrow{P} \mathrm{I}(\theta),$$
assim podemos usar $\mathrm{J}(\theta)$ quando calcular $\mathrm{I}(\theta)$ for difícil/impossível.

Estimador de máxima verossimilhança

- O EMV $\hat{\boldsymbol{\theta}} \in \Omega$ é definido por $L(\hat{\boldsymbol{\theta}}) \geq L(\boldsymbol{\theta})$ para qualquer $\boldsymbol{\theta} \in \Omega$. Em geral, $\hat{\boldsymbol{\theta}}$ satisfaz a equação de verossimilhança

$$\boldsymbol{U}(\boldsymbol{\theta}|\boldsymbol{Y}) = \boldsymbol{0}.$$

- Um sistema com p equações e p incógnitas

$$\begin{pmatrix} U_1(\boldsymbol{\theta}|\boldsymbol{y}) = 0 \\ \vdots \\ U_p(\boldsymbol{\theta}|\boldsymbol{y}) = 0 \end{pmatrix}$$

Propriedades do Estimador de máxima verossimilhança

- Sendo $\boldsymbol{\theta}_0$ o verdadeiro valor do vetor de parâmetros $\boldsymbol{\theta}$. Então

1. Consistência: $\hat{\boldsymbol{\theta}} \xrightarrow{P} \boldsymbol{\theta}_0$, ou seja,

$$P\left(\left\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right\| > \epsilon\right) \to 0, \quad \text{quando} \quad n \to \infty.$$

2. Normalidade assintótica:

$$\hat{\boldsymbol{\theta}} \xrightarrow{D} N_p(\boldsymbol{\theta}, \mathrm{I}^{-1}(\boldsymbol{\theta})), \quad \text{quando} \quad n \to \infty.$$

- Exemplos: Distribuição Normal, distribuição Gamma.

Parâmetros ortogonais

- Considere um modelo estatístico parametrizado por $\boldsymbol{\theta} = (\theta_1, \theta_2)^\top$. No caso da matriz de informação de Fisher ser diagonal

$$\mathrm{I}(\boldsymbol{\theta}) = \begin{pmatrix} \mathrm{I}_{11}(\boldsymbol{\theta}) & 0 \\ 0 & \mathrm{I}_{22}(\boldsymbol{\theta}) \end{pmatrix},$$

  os parâmetros θ_1 e θ_2 são ditos **ortogonais**.
- O inverso da informação de Fisher é também diagonal

$$\mathrm{I}^{-1}(\boldsymbol{\theta}) = \begin{pmatrix} 1/\mathrm{I}_{11}(\boldsymbol{\theta}) & 0 \\ 0 & 1/\mathrm{I}_{22}(\boldsymbol{\theta}) \end{pmatrix}.$$

- Assim, $\hat{\theta}_1$ e $\hat{\theta}_2$ são assintóticamente independentes com distribuição

$$\hat{\theta}_j \overset{a}{\sim} N(\theta_j, 1/\mathrm{I}_{jj}(\boldsymbol{\theta})).$$

Parâmetros ortogonais: Generalização

- Considere um modelo estatístico parametrizado por $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)^\top$. No caso da matriz de informação de Fisher ser bloco diagonal

$$\mathrm{I}(\boldsymbol{\theta}) = \begin{pmatrix} \mathrm{I}_1(\boldsymbol{\theta}) & 0 \\ 0 & \mathrm{I}_2(\boldsymbol{\theta}) \end{pmatrix},$$

  os vetores de parâmetros $\boldsymbol{\theta}_1$ e $\boldsymbol{\theta}_2$ são ditos **ortogonais**.
- A distribuição assintótica de $\boldsymbol{\theta}_1$ é a mesma se $\boldsymbol{\theta}_2$ é considerado conhecido ou desconhecido.
- Definição similar pode ser feita usando a matriz de informação observada.
- Exemplos: Modelos exponenciais de dispersão, regressão linear.