CE085 - Estatística Inferencial Função de Verossimilhança e suas derivadas

Prof. Wagner Hugo Bonat

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

5 de setembro de 2018

Conteúdo

Conteúdo

- Notação e definições.
- Verossimilhança e log-verossimilhança.
- Escore e informação de Fisher.
- Informação observada.
- Desigualdade de Cramér-Rao.

Notação e Definições

Notação

- O vetor $(n \times 1)$ de variáveis aleatórias (va) é denotado por $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$.
- O vetor $(n \times 1)$ de realizações de uma va é denotado por $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$.
- Denote $f(\boldsymbol{Y}; \boldsymbol{\theta})$ a função de probabilidade (fp) caso discreto ou função de densidade probabilidade (fdp) do vetor aleatório $\boldsymbol{Y}$.

Definições

- **Definição 1 (Parâmetro ou vetor de parâmetros)** Vetor de características numéricas de uma população. Denotaremos o vetor $(p \times 1)$ de parâmetros por $\boldsymbol{\theta}$. Em particular com $p = 1$(caso uniparamétrico) denotaremos θ.
- **Definição 2 (Espaço paramétrico) Espaço paramétrico** é o conjunto de todas as possíveis combinações entre todos os valores para todos os diferentes parâmetros envolvidos em uma fp ou fdp. Notação Ω.
- **Definição 3 (Suporte) Suporte** é conjunto de valores realizáveis de uma va.
- Exemplos: Binomial, Poisson e normal.

Verossimilhança - Caso uniparamétrico

- **Definição 4 (Verossimilhança)** Sejam dados **y** uma realização de um vetor aleatório $\boldsymbol{Y}$ com fp ou fdp $f(\boldsymbol{Y}, \theta)$. A **função de verossimilhança** $L(\theta) : \Omega \to [0, \infty]$ para θ é a função aleatória

$$L(\theta) \equiv f(\boldsymbol{Y}, \theta)$$

onde $f(y_1, \ldots, y_n|\theta)$ é a função de distribuição conjunta de $\boldsymbol{Y}$.

1. Caso discreto não há ambiguidade então

$$L(\theta) \equiv P_\theta[Y = y].$$

2. Caso contínuo em geral as observações são medidas com algum grau de precisão em um intervalo ($y_{iI} \leq y_i \leq y_{iS}$). Neste caso a verossimilhança é dada por

$$L(\theta) = P_\theta[y_{1I} \leq y_1 \leq y_{1S}, y_{2I} \leq y_2 \leq y_{2S}, \ldots, y_{nI} \leq y_n \leq y_{nS}]. \quad (1)$$

Verossimilhança - Caso uniparamétrico

- Suponha que as observações são independentes e medidas com o mesmo grau de precisão.
- Assim, cada dado é medido em um intervalo $(y_i - \delta/2 \leq Y_i \leq y_i + \delta/2)$.
- Com estas suposições a verossimilhança pode ser escrita como

$$
\begin{aligned}
L(\theta) &= \prod_{i=1}^{n} P_\theta[y_i - \delta/2 \leq Y_i \leq y_i + \delta/2] \\
&= \prod_{i=1}^{n} \int_{y_i-\delta/2}^{y_i+\delta/2} f(y_i, \theta) d(y_i).
\end{aligned}
$$

Verossimilhança - Caso uniparamétrico

- Se o grau de precisão é alto (δ é pequeno) em relação a variabilidade dos dados, a expressão se reduz a

$$L(\theta) \approx \left(\prod_{i=1}^{n} f(y_i, \theta) \right) \delta^n.$$

- Finalmente, se δ não depende de θ, temos

$$L(\theta) \approx \prod_{i=1}^{n} f(y_i, \theta), \tag{2}$$

- Para enfatizar que a verossimilhança é avaliada nas observações usamos a notação $L(\theta|\boldsymbol{y})$.

Verossimilhança - Condições de regularidade

1. O parâmetro θ é **identificável**. Isso significa que se $f(\theta_1|\boldsymbol{y}) = f(\theta_2|\boldsymbol{y})$ para quase todos $y \in \Re$, então $\theta_1 = \theta_2$.
2. O suporte de $f(\theta|\boldsymbol{y})$ é o mesmo para todo $\theta \in \Re$.
3. O verdadeiro valor do parâmetro θ_0 pertence ao interior de Ω.
4. $f(\theta|\boldsymbol{y})$ é duas vezes continuamente diferenciável com relação θ para quase todo $\boldsymbol{y} \in \Re$.
5. $\frac{\partial}{\partial \theta}$ e $\int$ (caso contínuo), ou $\frac{\partial}{\partial \theta}$ e $\sum$ (caso discreto) podem ser intercambiada.

** Para quase todo significa que a condição não é verdadeira para um conjunto de $\boldsymbol{y}$ com probabilidade zero de ocorrência.

Log-Verossimilhança

- A função de log-verossimilhança é a função estocástica $l(\theta) : \Omega \to \Re$ definida por

$$l(\theta|\boldsymbol{y}) = \log\left(L(\theta|\boldsymbol{y})\right).$$

- No caso iid, tem-se

$$l(\theta|\boldsymbol{y}) = \sum_{i=1}^{n} \log\left(L(\theta|y_i)\right).$$

- $l(\theta|\boldsymbol{y}) = -\infty$ quando $L(\theta) = 0$, mas isso ocorre quando $f(y_1, \ldots, y_n|\theta) = 0$ que tem probabilidade de ocorrência igual a zero.

Exemplo

- Sejam $Y_1, \dots, Y_n$ iid ensaios Bernoulli com probabilidade de sucesso μ. Escreva a função de verossimilhança, log-verossimilhança e verifique se as condições de regularidade estão satisfeitas.

Transformação de parâmetros

- Suponha que o interesse seja trabalhar com ψ definido por $\theta = g(\psi)$ ao invés de θ. Assuma que g é $1-1$. Então, a log-verossimilhança para ψ é dada por

$$\tilde{l}(\psi) = \sum_{i=1}^{n} \log f(y_i, g(\psi)),$$

apenas inserimos $g(\psi)$ em $l(\cdot)$ para obter a nova log-verossimilhança.

Exemplo: Transformação de parâmetros

- Suponha que desejamos escrever a log-verossimilhança para $\psi = \log \frac{\mu}{1-\mu}$. Note que $\mu = \frac{\exp^{\psi}}{1+\exp^{\psi}}$.

Exemplo: Transformação de parâmetros

- Suponha que desejamos escrever a log-verossimilhança para $\psi = \log \frac{\mu}{1-\mu}$. Note que $\mu = \frac{\exp^{\psi}}{1+\exp^{\psi}}$.

$$\tilde{l}(\psi) = \sum_{i=1}^{n} y_i \psi - n \log(1 + \exp^{\psi}).$$

Transformação nos dados

- Considere uma transformação $1-1$ $Y_i = h(X_i)$ que será usada na log-verossimilhança ao invés de X_i. Considere o caso contínuo e que h é diferenciável. Então Y_i tem densidade dada por $f(h^{-1}(y), \theta)\frac{dx}{dy}(y)$, assim a nova verossimilhança será

$$
\begin{aligned}
\tilde{l}(\theta) &= \tilde{l}(\theta|Y_1, \ldots, Y_n) \\
&= \sum_{i=1}^{n} \log\left(f(h^{-1}(y_i), \theta)\frac{dx_i}{dy_i}Y_i\right) \\
&= \sum_{i=1}^{n} \log\left(f(h^{-1}(y_i), \theta)\right) + \log\left(\frac{dx_i}{dy_i}Y_i\right) \\
&= l(\theta|X_1 \ldots, X_n) + const.
\end{aligned}
$$

- Em geral estamos interessados nas derivadas de $l(\cdot)$ que não dependem da constante.
- Para fins de estimação a constante pode ser desconsiderada.
- Qual seria a melhor transformação para os dados?

Desigualdade de Jensen e Máxima Verossimilhança

- Seja g uma função estritamente convexa e Y uma v.a com $E(|Y|) < \infty$ tal que a distribuição de Y é não degenerada. Então $g(E(Y)) < E(g(Y))$. Por outro lado, se g é estritamente concava, então $g(E(Y)) > E(g(Y))$.
- Demonstração: Exercício.
- Teorema 1: Seja θ_0 o verdadeiro valor do parâmetro. Então,

$$P_{\theta_0}(L(\theta_0|\mathbf{y}) > L(\theta|\mathbf{y})) \rightarrow 1, \quad \text{quando} \quad n \rightarrow \infty.$$

- Demonstração.
- Interpretação: $L(\theta_0) > L(\theta)$ com alta probabilidade para n grande. Assim, $L(\theta)$ vai tender a ter o seu máximo próximo a θ_0, o verdadeiro valor de θ.
- Motiva a ideia de estimação por máxima verossimilhança.

Definições

- Uma **estatística** é uma variável aleatória $T = t(\boldsymbol{Y})$, onde a função $t(\cdot)$ não depende de θ.
- Uma **estatística** T é um **estimador** para θ se o valor realizado $t = t(\boldsymbol{y})$ é usado como uma **estimativa** para o valor de θ.
- A distribuição de probabilidade de T é chamada de **distribuição amostral** do estimador $t(\boldsymbol{Y})$.
- O **viés** de um estimador T é a quantidade

  $$B(T) = E(T - \theta).$$

  O estimador T é dito não viciado para θ se $B(T) = 0$, tal que $E(T) = \theta$.
- O estimador T é assintóticamente não viciado para θ se $E(T) \to \theta$ quando $n \to \infty$.

Definições

- A **eficiência relativa** entre dois estimadores T_1 e T_2 é a razão $er = \frac{V(T_1)}{V(T_2)}$ em que $V(\cdot)$ denota a variância.
- O **erro quadrático médio** de um estimador T é a quantidade

$$EQM(T) = E((T - \theta)^2) = V(T) + B(T)^2.$$

- Um estimador T é **médio quadrático consistente** para θ se o $EQM(T) \to 0$ quando $n \to \infty$.
- O estimador T é **consistente em probabilidade** se $\forall \epsilon > 0$, $P(|T - \theta| > \epsilon) \to 0$, quando $n \to \infty$.

Função escore e Informação de Fisher

- Função escore para θ (efficient score)

$$
\begin{aligned}
U(\theta|\boldsymbol{Y}) &= U(\theta|Y_1, \ldots, Y_n) \\
&= \sum_{i=1}^{n} \frac{\partial}{\partial \theta} l(\theta, Y_i). \\
&= \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log f(\theta, Y_i).
\end{aligned}
$$

- Informação de Fisher ou Informação esperada

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

- Informação de Fisher também é chamada de intrinsic accuracy.

Igualdades de Bartlett

- Sob condições de regularidade (slide 10), tem-se
  1. Primeira igualdade $E(U(\theta|\boldsymbol{Y})) = 0$.
  2. Segunda igualdade $I_E(\theta) = -E(l''(\theta|\boldsymbol{Y})) = -E(U'(\theta|\boldsymbol{Y}))$.
- Implicação: $Var(U(\theta|\boldsymbol{Y})) = E(U(\theta|\boldsymbol{Y})^2)$.
- Exemplo: Bernoulli e Poisson.
- Demonstração.

Informação observada

- Informação observada para θ

$$I_O(\theta) = -l''(\theta|\boldsymbol{Y}).$$

- Note que

$$I_E(\theta) = E(I_O(\theta)).$$

- Além disso, pela lei dos grandes números

$$I_O(\theta) \stackrel{P}{\rightarrow} I_E(\theta) \quad \text{quando} \quad n \rightarrow \infty.$$

Desigualdade de Cramér-Rao

- Teorema: If $\tilde{\theta}(Y_1, \ldots, Y_n)$ é um estimador não viciado para θ, então
$$Var(\tilde{\theta}) \geq I_E(\theta)^{-1}.$$
- A quantidade $I_E(\theta)^{-1}$ é chamado de limite inferior de Cramér-Rao.
- Um estimador não viciado é chamado eficiente se $V(\tilde{\theta}) = I_E(\theta)^{-1}$.
- Exemplos: Geométrica e Poisson.
- Demonstração.
- Exemplo patológico: Distribuição uniforme.