Estimação

Tópicos de Inferência Estatística

José Passos

ISEG-ULisboa

16 de Outubro de 2019

Tabela de conteúdos

Conceitos

- Vamos assumir que a população X tem distribuição na família $\mathcal{F} = \{f(.|\theta) : \theta \in \Theta\}$.
- Com base numa amostra casual $X_1, \ldots, X_n \overset{iid}{\sim} X$ procura-se fazer inferência sobre (estimar pontualmente) o parâmetro desconhecido θ mediante o uso de uma determinada estatística $T = T(X_1, \ldots, X_n)$.
- Estimar pontualmente θ corresponde a escolher uma aplicação T que a cada ponto do espaço amostral $x \in \mathcal{X}$ faça corresponder um ponto do espaço parâmetro, $T(x) \in \Theta$.
- A aplicação T é uma estatística que se designa por estimador de θ
- Cada valor de T, $T(x)$, designa-se por estimativa.

Conceitos

- Qualquer estatística T opera uma redução nos dados.
- Estamos interessados em métodos de redução de dados que não despreze informação relevante sobre θ.
- O princípio da suficiência constitui um método de redução dos dados sem perda de informação sobre θ.

Conceitos

- Em termos intuitivos, uma estatística diz-se suficiente para θ se retira da amostra casual toda a informação relevante que esta tem sobre o parâmetro.
- Definição: Seja $(X_1, \ldots, X_n)$ uma amostra casual de uma população $\mathcal{F} = \{f(.|\theta) : \theta \in \Theta\}$. Uma estatística T diz-se suficiente para θ se a distribuição de $(X_1, \ldots, X_n)$ condicionada por $T = t$ não depende de θ, $\forall \theta \in \Theta$ e $\forall t$.

Critério da factorização

- A definição de estatística suficiente é pouco prática na pesquisa de estatísticas suficientes.
- Teorema (critério da factorização): A estatística T diz-se suficiente para θ sse existem funções não negativas $g(.)$ e $h(.)$ tais que para todo o $x_1, \ldots, x_n$,

$$f(x_1, \ldots, x_n) = g[T(x_1, \ldots, x_n), \theta]h(x_1, \ldots, x_n)$$

Note-se que $g(.)$ depende de θ e da amostra unicamente através de T enquanto que $h(.)$ depende unicamente da amostra.

Critério da factorização: exemplo

- Exemplo: seja $(X_1, \ldots, X_n)$ uma amostra casual proveniente de uma população de Poisson de parâmetro θ. Tem-se,

$$f(x_1, \ldots, x_n) = \prod_{i=1}^{n} e^{-\theta} \frac{\theta^{x_i}}{x_i!} = e^{-n\theta} \theta^{\sum_{i=1}^{n} x_i} \left(\prod_{i=1}^{n} x_i! \right)^{-1}$$

Pelo critério da factorização $T = \sum_{i=1}^{n} x_i$ é suficiente para θ.

Suficiência mínima

- Existem muitas estatísticas suficientes para θ. Qualquer estatística que seja função de uma estatística suficiente é ainda suficiente. De entre todas as estatísticas suficientes estamos interessados naquelas que operam a máxima redução dos dados, sem perda de informação.
- Definição (suficiência mínima): A estatística T diz-se suficiente mínima para θ sse for suficiente e função de toda a estatística suficiente para θ.

Suficiência mínima

- Teorema: Seja $f(x_1, \ldots, x_n \mid \theta)$ a função densidade (ou probabilidade) da amostra casual $(X_1, \ldots, X_n)$. Se a estatística T é tal que, para duas amostras observadas $(x_1, \ldots, x_n)$ e $(y_1, \ldots, y_n)$ o quociente,

$$\frac{f(x_1, \ldots, x_n \mid \theta)}{f(y_1, \ldots, y_n \mid \theta)}$$

não depende de θ sse $T(x_1, \ldots, x_n) = T(y_1, \ldots, y_n)$, então T é suficiente mínima para θ.

Verosimilhança

- No que se segue vamos considerar modelos uniparamétricos, $\mathcal{F} = \{f(.|\theta) : \theta \in \Theta\}$, onde $\Theta \subset \Re$.
- Definição: Seja $(x_1, \ldots, x_n)$ uma amostra observada. A função de verosimilhança de θ é dada por,

$$\begin{aligned} L(\theta \mid x_1, \ldots, x_n) &= f(x_1, \ldots, x_n \mid \theta) \\ &= \prod_{i=1}^{n} f(x_i \mid \theta) \end{aligned}$$

Nota: como a amostra observada é fixa pode escrever-se a verosimilhança como $L(\theta)$.

Verosimilhança

- Nota: a verosimilhança não é uma probabilidade e portanto não tem uma escala natural associada. Está definida a menos de uma constante multiplicativa,

$$L(\theta \mid x_1, \ldots, x_n) \propto f(x_1, \ldots, x_n \mid \theta)$$

A constante multiplicativa pode depender da amostra mas não do parâmetro.

Condições de regularidade

Alguns dos resultados seguintes dependem da verificação de um conjunto de condições, conhecidas como condições de regularidade. Essas condições de regularidade são as seguintes:

- C1 - Θ é um intervalo de $\Re$
- C2 - o conjunto $\{x : f(x \mid \theta) > 0\}$ não depende de θ
- C3 - $f(x \mid \theta)$ é diferenciável em ordem a θ para todo o x
- C4 - $0 < E[\partial \ln f(X \mid \theta)/\partial\theta]^2 < +\infty$ para todo o θ
- C5 - $\partial/\partial\theta$ pode permutar-se com $\int dx$

Score

- Definição (Função Score): dada a amostra observada, $x_1, \ldots, x_n$, a função score define-se como,

$$S(\theta \mid x_1, \ldots, x_n) = \frac{\partial \ln L(\theta \mid x_1, \ldots, x_n)}{\partial \theta}$$

  Se a amostra é casual, $S(\theta \mid x_1, \ldots, x_n) = \sum_{i=1}^{n} S(\theta \mid x_i)$

- Teorema: verificando-se as condições de regularidade, tem-se $\forall \theta \in \Theta$,

$$E\left[S(\theta \mid X_1, \ldots, X_n)\right] = 0$$

Hessiana

- Definição (Função Hessiana): dada a amostra observada, $x_1, \ldots, x_n$, a função hessiana define-se como,

$$H(\theta \mid x_1, \ldots, x_n) = \frac{\partial^2 \ln L(\theta \mid x_1, \ldots, x_n)}{\partial \theta^2} = \frac{\partial S(\theta \mid x_1, \ldots, x_n)}{\partial \theta}$$

Se a amostra é casual, $H(\theta \mid x_1, \ldots, x_n) = \sum_{i=1}^{n} H(\theta \mid x_i)$

Informação

- Definição (Informação de Fisher): a quantidade de informação de Fisher sobre θ contida na observação de $X_1, \ldots, X_n$ é definida por,

$$I_{(X_1,\ldots,X_n)}(\theta) = E\left[S(\theta \mid X_1, \ldots, X_n)^2\right] = Var\left[S(\theta \mid X_1, \ldots, X_n)\right]$$

- Teorema: Se $T(X_1, \ldots, X_n)$ é uma estatística, tem-se,

$$I_{(X_1,\ldots,X_n)}(\theta) \geq I_T(\theta)$$

A igualdade verifica-se sse T é suficiente para θ.

Informação (cont.)

Propriedades:

- Se a amostra é casual, $I_{(X_1,\ldots,X_n)}(\theta) = nI_{X_i}(\theta)$
- Verificando-se as condições de regularidade,

$$\begin{aligned} I_{(X_1,\ldots,X_n)}(\theta) &= -E\left[\frac{\partial^2 \ln f(X_1,\ldots,X_n \mid \theta)}{\partial\theta^2}\right] \\ &= -E\left[H(\theta \mid X_1,\ldots,X_n)\right] \end{aligned}$$

ou,

$$E\left[S(\theta \mid X_1,\ldots,X_n)^2\right] = -E\left[H(\theta \mid X_1,\ldots,X_n)\right]$$

Ideia base

- A avaliação da qualidade de um estimador é feito através da sua distribuição por amostragem.
- Um estimador é sempre classificado em relação a um conjunto de critérios e é em relação a estes critérios que avaliamos a sua qualidade.
- Em geral, não existe estimador que seja o melhor em todos os critérios.
- No que se segue, os estimadores vão ser avaliados à luz dos seguintes critérios: não enviesamento, eficiência, erro quadrático médio e consistência.

Não enviesamento

- Definição: um estimador T diz-se não-enviesado ou centrado de $\tau(\theta)$ sse,

$$E(T) = \tau(\theta)$$

- Um estimador não centrado diz-se enviesado. O seu enviesamento é dado por $b_\theta = E(T) - \tau(\theta)$
- Se a população tem média, a média amostral é sempre estimador não enviesado.
- Se a população tem variância, a variância amostral corrigida é sempre estimador não enviesado.

Eficiência

- Definição: sejam T e T^* estimadores centrados de $\tau(\theta)$. Diz-se que T é mais eficiente que T^* na estimação de $\tau(\theta)$ se, $\forall\theta \in \Theta$

$$Var(T) \leq Var(T^*)$$

Eficiência (cont.)

- Teorema (Desigualdade de Fréchet-Cramer-Rao): Considere uma população X. Suponha que se verificam as condições de regularidade e seja $\tau(\theta)$ uma função real e diferenciável de θ. Considere também que T é estimador centrado de $\tau(\theta)$. Se $\forall\theta \in \Theta$

$$E\left[\left|T(X_1,\ldots,X_n)\frac{\partial f(X_1,\ldots,X_n)}{\partial\theta}\right|\right] < \infty$$

então,

$$Var(T) \geq \frac{[\tau'(\theta)]^2}{I_{(X_1,\ldots,X_n)}(\theta)}$$

Eficiência (cont.)

Observações:

- Se $\tau(\theta) = \theta$ e $X_1, \ldots, X_n$ é amostra casual de X então qualquer estimador centrado de θ, T, verifica

$$Var(T) \geq \frac{1}{I_X(\theta)}$$

- O limite inferior de FCR só é válido se as condições de regularidade forem verificadas
- Mesmo que se verifiquem as condições de regularidade não há garantias que exista estimador centrado de $\tau(\theta)$ cuja variância atinja o limite inferior de FCR.

Eficiência (cont.)

Observações (cont.):

- Ao quociente entre o limite inferior de FCR e a variância de um estimador centrado de $\tau(\theta)$ dá-se o nome de eficiência,

$$e_\theta(T) = \frac{\frac{[\tau'(\theta)]^2}{I_{(X_1,\ldots,X_n)}(\theta)}}{Var(T)}$$

- Se as condições de regularidade forem satisfeitas, $0 \leq e_\theta(T) \leq 1$.

Eficiência (cont.)

O seguinte teorema diz-nos em que condições existem estimadores mais eficientes para $\tau(\theta)$.
Teorema: Seja $T = T(X_1, \ldots, X_n)$ um estimador regular não enviesado de $\tau(\theta)$. T é mais eficiente sse existir $\beta(\theta)$ tal que

$$S(\theta \mid x_1, \ldots, x_n) = \beta(\theta)\,[T(x_1, \ldots, x_n) - \tau(\theta)]$$

Eficiência (cont.)

Observações:

- se existir estimador mais eficiente este será necessariamente suficiente
- eficiência relativa: se T e T^* são estimadores centrados de $\tau(\theta)$, a eficiência relativa de T^* relativamente a T define-se por,

$$e_\theta(T^*, T) = \frac{Var(T)}{Var(T^*)}$$

Erro Quadrático Médio

Definição (EQM): o EQM de um estimador T de $\tau(\theta)$ é definido como,

$$EQM_\theta(T) = E\left[(T - \tau(\theta))^2\right]$$

Observações:

- T é preferível a T^* na estimação de $\tau(\theta)$ se $\forall\theta \in \Theta$, $EQM_\theta(T) \leq EQM_\theta(T^*)$
- $EQM_\theta(T) = Vart(T) + [b_\theta(T)]^2$

Consistência

A consistência refere-se ao comportamento probabilístico de um estimador $T_n = T(X_1, \ldots, X_n)$ à medida que $n \to \infty$
Definição (Consistência): O estimador T_n diz-se (fracamente) consistente para $\tau(\theta)$ se $\forall \theta \in \Theta$ e $\forall \epsilon > 0$,

$$lim_{n\to\infty} P(|T_n - \tau(\theta)| > \epsilon) = 0$$

e diz-se sumariamente, $T_n \xrightarrow{p} \tau(\theta)$
Definição (Consistência em média quadrática): O estimador T_n diz-se consistente em média quadrática para $\tau(\theta)$ se $\forall \theta \in \Theta$,

$$lim_{n\to\infty} E\left[(T_n - \tau(\theta))^2\right] = 0$$

e diz-se sumariamente, $T_n \xrightarrow{mq} \tau(\theta)$

Consistência (cont.)

Observações:

- a consistência em média quadrática implica a consistência (fraca).
- condição suficiente para que T_n seja consistente para $\tau(\theta)$

$$lim_{n\to\infty} E(T_n) = \tau(\theta)$$
$$lim_{n\to\infty} Var(T_n) = 0$$

Nota: estas condições são necessárias e suficientes para a consistência em média quadrática.

Método dos momentos

- O método dos momentos é uma técnica que permite obter o estimador dos parâmetros da população igualando os momentos amostrais aos momentos populacionais.
- Seja $\theta = (\theta_1, \ldots, \theta_k)$ o vector de parâmetros desconhecidos da população.
- Vamos admitir que os momentos populacionais $\mu'_r = E(X^r)$ são uma função conhecida de θ: $\mu'_r = \psi_r(\theta)$
- Os correspondentes momentos amostrais são dados por $M'_r = \sum_{i=1}^{n} X_i^r / n$
- A solução do sistema de equações $M'_r = \psi_r(\theta)$, $r = 1, \ldots, k$ fornece os estimadores obtidos pelo método dos momentos.

Método da máxima verosimilhança

- O método da máxima verosimilhança é uma técnica que propõe como estimativa do parâmetro desconhecido da população o valor que maximiza a função de verosimilhança.
- A estimativa da máxima verosimilhança, se existir, é o valor $\hat{\theta}$ tal que, $\forall \theta \in \Theta$,

$$L(\hat{\theta} \mid x_1, \ldots, x_n) \geq L(\theta \mid x_1, \ldots, x_n)$$

- Sob certas condições de regularidade, isto é, se a função de verosimilhança é diferenciável em θ, o EMV, $\hat{\theta}$, satisfaz a equação,

$$\frac{\partial \ln L(\hat{\theta} \mid x_1, \ldots, x_n)}{\partial \theta} = 0$$

Método da máxima verosimilhança (cont.)

Propriedades:

- Se $\hat{\theta}$ é EMV de θ e se $\tau(\theta)$ é função biunívoca de θ então $\tau(\hat{\theta})$ é EMV de $\tau(\theta)$.
- Se T é estatística suficiente e se existir EMV de θ então este é função de T.
- Se existe estimador mais eficiente, o EMV é único e coincide com esse estimador.
- Verificadas as condições de regularidade, o EMV de θ, $\hat{\theta}$, é consistente e assintoticamente mais eficiente, para além de ser assintoticamente normal,

$$\sqrt{I_{(X_1,\ldots,X_n)}(\theta)}(\hat{\theta}-\theta) \overset{a}{\sim} N(0,1)$$

Método da máxima verosimilhança (cont.)

Propriedades (cont.):

- Como $I_{(X_1,\ldots,X_n)}(\theta)$ depende em geral de θ desconhecido, é habitual substituir a informação de Fisher pelas seguintes expressões, assintoticamente equivalentes:
  - $I_{(X_1,\ldots,X_n)}(\hat{\theta})$
  - $-H(\hat{\theta}) = -\partial^2 \ln L(\hat{\theta} \mid x_1, \ldots, x_n)/\partial\theta^2$
  - $\sum_{i=1}^{n} S(\hat{\theta} \mid x_i)^2$

Ideia base

- Sob certas condições de regularidade o EMV, $\hat{\theta}$, obtém-se resolvendo a equação,

$$\frac{\partial \ln L(\theta \mid X_1, \ldots, X_n)}{\partial \theta} = 0$$

  em ordem a θ.
- Em muitas situações esta equação não é linear em θ, não havendo uma expressão analítica para o estimador. Neste caso a estimativa tem que ser obtida numericamente.

Algorítmo

Os algorítmos numéricos de estimação baseiam-se no seguinte:

- inicializar com um valor θ_0
- se na iteração $t = 0, \ldots, m$, θ_t não é um valor óptimo para θ, calcular uma direcção Δ_t e um passo δ_t, obtendo um novo valor,

$$\theta_{t+1} = \theta_t + \delta_t \Delta_t$$

- Repetir o procedimento anterior até que $|\theta_{t+1} - \theta_t| < \epsilon$, $\forall \epsilon > 0$, previamente fixado.

Algorítmo de Newton-Raphson

O algorítmo de Newton-Raphson baseia-se numa aproximação em série de Taylor de 1ª ordem da equação do score, $S(\theta \mid x) = 0$, avaliado em θ_{t+1}, em torno de θ_t,

$$S(\theta_{t+1} \mid x) \approx S(\theta_t \mid x) + \frac{dS(\theta_t \mid x)}{d\theta}(\theta_{t+1} - \theta_t) = 0$$

Resolvendo em ordem a θ_{t+1},

$$\begin{aligned}\theta_{t+1} &= \theta_t - \left(\frac{dS(\theta_t \mid x)}{d\theta}\right)^{-1} S(\theta_t \mid x) \\ &= \theta_t - H(\theta_t \mid x)^{-1} S(\theta_t \mid x)\end{aligned}$$

Variantes

O algorítmo de Newton-Raphson utiliza como direcção a informação de Fisher observada, $H(\theta \mid x)$. As variantes ao algorítmo de Newton envolvem diferentes formas para a direcção:

- Score eficiente: $H(\theta_t \mid x)$ é substituído pelo simétrico da informação de Fisher, $E\left[H(\theta_t \mid X)\right]$
- BHHH (Berndt, Hall, Hall e Hausman): $H(\theta_t \mid x)$ é substituído por $-\sum_{i=1}^{n} S(\theta_t \mid x_i)^2$

Bibliografia

- Casella, G. & Berger, R. (2002), *Statistical Inference*, 2nd Edition, Duxbury.
- Murteira, B. (1990), *Probabilidades e Estatística*, Volume 2, 2ª Edição, McGraw-Hill, Portugal.