Técnicas computacionais em probabilidade e estatística II

Márcia D'Elia Branco

Universidade de São Paulo
Instituto de Matemática e Estatística
http:www.ime.usp.br/ mbranco

AULA 2: *Bootstrap* **e** *Jackknife*.

Introdução

- O método de *bootstrap* foi introduzido por Efron (1979). Para mais informação sobre a metodologia ver o livro de Efron and Tibshirani(1998). Chapman and Hall/CRC.
- O termo provem de uma frase do romance "'As aventuras do Barão de Munchausen"' escrito por Rudolph Raspe, "' *to pull oneself up by one´s bootstrap*"'.
- Na língua portuguesa poderia ser interpretado como "'pular sobre suas próprias botas"' ou "'subir com seus próprios esforços"'.
- O método de *Jackknife* é mais simples e anterior ao *bootstrap*.
- O objetivo dos métodos é estimar a variabilidade de um estimador re-amostrando da própria amostra observada.

O método de *Jackknife*

Considere $x_1, x_2, \ldots, x_n$ a amostra observada.
As amostras de *Jack* serão construídas retirando-se um elemento da amostra original,

$$x_{(i)} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) \ , \ i = 1, \ldots, n-1.$$

Obtemos assim n amostras de *Jack*. Para cada uma destas amostras calcula-se o valor do estimador de interesse $\hat{\theta}(i)$.
A estimativa de *Jackknife* para o Erro Quadrático Médio (EQM) do estimador $\hat{\theta}$ é dada por

$$EQM_{Jack}(\hat{\theta}) = \frac{n-1}{n} \sum_{i=1}^{n} [\hat{\theta(i)} - \bar{\theta}]^2$$

com $\bar{\theta} = \frac{1}{n} \sum_{i=1}^{n} \hat{\theta(i)}$.

Exemplo 1: Estimando a média populacional

Parâmetro de interesse $\mu = E[X]$. Estimador $\bar{X}$.
Considere a amostra $x = (10, 27, 31, 40, 46, 50, 52, 104, 146)$.
Amostras de *Jack* de tamanho (n-1):

$x_{(i)}$	$\bar{x}$	mediana
(27,31,40,46,50,52,104,146)	62.00	48
(10,31,40,46,50,52,104,146)	59.87	48
(10,27,40,46,50,52,104,146)	59.37	48
(10,27,31,46,50,52,104,146)	58.25	48
(10,27,31,40,50,52,104,146)	57.50	45
(10,27,31,40,46,52,104,146)	57.00	43
(10,27,31,40,46,50,104,146)	56.75	43
(10,27,31,40,46,50,52,146)	50.25	43
(10,27,31,40,46,50,52,104)	45.00	43

$\bar{\theta} = 56.22$, o qual é igual a média amostra é $\bar{x}$.

$$EQM_{Jack}(\bar{X}) = \frac{8}{9}\sum_{i=1}^{9}[\bar{x}(i) - \bar{\theta}]^2 = 199.9$$

O real valor de erro quadrático médio para $\bar{X}$ é dado por

$$\frac{Var(X)}{n}$$

Usando a variância amostral $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ como estimador de $Var(X)$, temos que $E\hat{Q}M = 1799.19/9 = 199.9$
É possível mostrar que

$$EQM_{Jack}(\bar{X}) = \frac{s^2}{n}$$

Problema com o *Jackknife*: uso de funções não suaves.
Considere $\hat{\theta} = med(x)$, então

$$EQM_{Jack}(md) = 47.11$$

Se utilizarmos a metodologia de *bootstrap* obtemos

$$EQM_{Boot} = 91.78$$

com base em 1000 amostras simuladas.

O método de *bootstrap* não paramétrico

Um amostra de *bootstrap* é obtida simulando, com reposição, n valores segundo a distribuição empírica F_e. Denotada por $x^* = (x_1^*, x_2^*, \ldots, x_n^*)$.
Simula-se um número B de amostras e para cada uma avalia-se o estimador. Obtendo-se $\hat{\theta}_1, \ldots, \hat{\theta}_B$.

No exemplo, possíveis amostras (ou réplicas) são:

x^*	$\bar{x}$	mediana
(10,27,31,40,46,50,52,104,146)	56.22	46
(10,10,27,27,40,40,50,50,104)	39.78	40
(10,10,27,40,40,50,50,104,104)	48.33	40
(10,27,27,27,46,50,52,104,104)	49.67	46
(27,27,31,46,50,140,140,146,146)	83.67	50
(40,50,52,52,104,104,104,104,146)	84.00	104

A função de distribuição empírica é

$$F_e(x_i) = \frac{\text{numero de valores menores ou iguais a } x_i}{n}$$

Como simular de uma v.a. discreta?

(i) Gerar $u \sim U_{(0,1)}$
(ii) Se $F_e(x_{(i-1)}) < u \leq F_e(x_{(i)})$ fazer $x^* = x_{(i)}$

$i = 1, 2, \ldots, k$. $F_e(x_{(0)}) = -\infty$ e $x_{(1)} < x_{(1)} < \cdots < x_{(k)}$ são as estatísticas de ordem.

O erro quadrático médio de *bootstrap* é definido como

$$EQM_{Boot}(\hat{\theta}) = E_{F_e}[(\hat{\theta}(X) - \theta)^2]$$

e denominado estimativa ideal de *bootstrap* para o EQM.
No caso particular $\theta = \mu$ e $\hat{\theta} = \bar{X}$ obtemos

$$EQM_{Boot}(\hat{\theta}) = \frac{1}{n^2}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

a qual difere levemente da usual estimativa do $EQM(\bar{X})$ dada por

$$\frac{1}{n(n-1)}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Um medida de interesse em estatística é o erro padrão do estimador (desvio padrão do estimador).
O algoritmo proposto por Efron e Tibshirani para estimar o $ep(\hat{\theta})$ é dado por:
(i) Considere $x_1^*, x_2^*, \ldots, x_B^*$ réplicas de *Boot*.
(ii) Calcule o estimador em cada amostra obtida,
$\hat{\theta}(x_j^*)$, $j = 1, \ldots, B$.
(iii) Estime o erro padrão ($\hat{ep}_B$) por

$$\left\{ \frac{\sum_{j=1}^{B} [\hat{\theta}(x_j^*) - \bar{\theta}^*]^2}{B-1} \right\}^{1/2}$$

com $\bar{\theta}^* = \frac{1}{B} \sum_{j=1}^{B} \hat{\theta}(x_j^*)$.

Exemplo: Em uma amostra de 15 turmas de uma escola de direito duas medidas foram consideradas: LSAT, o escore médio da turma no exame nacional de admissão ao curso, e GPA, a nota média do curso de graduação.

LSTA	576	635	558	578	666	580	555	661
GPA	3.39	3.30	2.81	3.03	3.44	3.07	3.00	3.43
LSTA	651	605	653	575	545	572	594	
GPA	3.36	3.13	3.12	2.74	2.76	2.88	2.96	

O coeficiente de correlação amostral é $r_{xy} = 0.776$.

Qual o erro associado a esta estimativa?

A tabela a seguir apresenta a estimativa de *bootstrap* para o erro padrão do coeficiente de correlação amostral.

B	25	50	100	200	400	800	1600	3200
$\hat{ep}_B$	0.140	0.142	0.151	0.143	0.141	0.137	0.133	0.132

- Foi observada uma forte assimetria na distribuição de frequências dos valores obtidos para r_{xy}, indicando que o uso da aproximação normal para este estimador não é aconselhável.
- Assumindo que as observações tem distribuição normal é possível mostrar que o erro padrão de r_{xy} é 0.115, próximo dos valores obtidos

O método de *bootstrap* paramétrico

- A versão paramétrica do algoritmo de *bootstrap* assume parcialmente conhecida a distribuição de probabilidade F geradora dos dados observados, sendo necessário apenas definir os parâmetros dessa distribuição.
- O algoritmo para estimar o erro padrão de um estimador é igual ao estabelecido anteriormente, a única alteração é a maneira como simular as réplicas.
- A função empírica é substituída pela verdadeira F com os parâmetros estimados via amostra original.

O número de réplicas de *bootstrap*

- Efron mostra que o coeficiente de variação do $\hat{ep}_B$ é aproximado por

$$\left\{ cv(\hat{ep}_\infty)^2 + \frac{E(\Delta)+2}{4B} \right\}^{1/2}$$

- A partir dessa expressão conclui que para estimação do erro padrão $50 \leq B \leq 200$.

$cv(\hat{ep}_\infty)$	B=25	B=50	B=100	B=200
0.25	0.29	0.27	0.26	0.25
0.20	0.24	0.22	0.21	0.21
0.15	0.21	0.18	0.17	0.16
0.05	0.15	0.11	0.09	0.07

Outros usos para o *bootstrap*

- Estimar o viés de uma estimador, $Vies(\hat{\theta}) = E[\hat{\theta}] - \theta$.
- Construir intervalos de confiança.
- Teste de hipóteses.
- Técnicas de validação cruzada e estimação do erro de um preditor.
- .....