Bootstrap

Patricia Kisbye

FaMAF

13 de mayo, 2010

Técnica de bootstrap

La técnica de Bootstrap fue introducida por B. Efron, 1982.

- Consiste en aproximar la precisión de un estimador a partir de una muestra de datos u observaciones.
- La precisión se mide como la inversa de la varianza de un estimador.
- Conocer la precisión o ECM de estimadores de la varianza, coef. de asimetría, curtosis, etc., suele ser algebraicamente complicado, o bien no se conoce la distribución de los datos.
- El método Bootstrap permite obtener una buena aproximación del ECM y otras estimaciones a partir de la muestra, aún sin conocer la distribución de donde provienen los datos.
- Bootstrap: Levantarse tirando de las propias correas de las botas. Método autosuficiente.

Algunas consideraciones

Consideramos X una v. a., y F una distribución:

$$F(x) = P(X \leq x)$$

Notación: $\theta = E[X]$, $\sigma^2 = \mathrm{Var}(X)$

- θ y σ^2 dependen de la distribución F.
- Distribución empírica: Sean $X_1, X_2, \ldots, X_n$ v. a. independientes con la misma distribución que X (observaciones), y supongamos que se observan los valores

$$X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n.$$

Sea F_e la distribución empírica dada por

$$P_{F_e}(X = x) = \frac{\#\{i \mid x_i = x,\ 1 \leq i \leq n\}}{n}.$$

Distribución empírica

Según la distribución empírica F_e, la media y la varianza de X está dada por:

$$\begin{aligned} \theta_{F_e} &= E_{F_e}[X] = \sum_{i=1}^{n} \frac{x_i}{n} = \frac{\sum_{i=1}^{n} x_i}{n} \\ \sigma^2_{F_e} &= \mathrm{Var}_{F_e}(X) = \sum_{i=1}^{n} \frac{(x_i - \theta_{F_e})^2}{n} = \frac{\sum_{i=1}^{n} (x_i - \theta_{F_e})^2}{n}. \end{aligned}$$

- También se definen otras medidas y parámetros: mediana, curtosis, coeficiente de asimetría, etc., todos ellos dependientes de la distribución.

Estimaciones

Si $X_1, X_2, \ldots, X_n$ son v.a. independientes, con distribución común F, tal que

$$E[X] = \theta, \qquad \text{y } \mathrm{Var}(X) = \sigma^2,$$

se tiene que para esta distribución F:

- $\overline{X}(n)$ es un estimador insesgado de $\theta = E[X]$.

$$E[\overline{X}(n)] = \theta.$$

- $S^2(n)$ es un estimador insesgado de $\sigma^2 = \mathrm{Var}[X] = E[(X-\theta)^2]$.

$$E[S^2(n)] = \sigma^2.$$

- La varianza del estimador $\overline{X}(n)$ está dada por

$$\mathrm{Var}(\overline{X}(n)) = \sigma^2/n.$$

Notar que en el cálculo de E y Var subyace la distribución F

Estimaciones

Distribución F

$$E[\overline{X}(n)] = \theta, \qquad E[S^2(n)] = \sigma^2, \qquad \text{Var}(\overline{X}(n)) = \sigma^2/n.$$

Distribución F_e

$$E_{F_e}[\overline{X}(n)] = \theta_{F_e}, \qquad E_{F_e}[S^2(n)] = \sigma^2_{F_e}, \qquad \text{Var}_{F_e}(\overline{X}(n)) = \sigma^2_{F_e}/n.$$

$$\theta_{F_e} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \sigma^2_{F_e} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \theta_{F_e})^2$$

Técnica de bootstrap

- Así como $\overline{X}(n)$ y $S^2(n)$, pueden definirse otros estimadores para un determinado parámetro θ.
- Si $\hat{\theta} = g(X_1, X_2, \ldots, X_n)$ es un estimador para un parámetro θ, interesa conocer

$$\text{Var}(\hat{\theta}) \qquad \text{y} \qquad ECM(\hat{\theta}, \theta) = E[(\hat{\theta} - \theta)^2].$$

- Estos valores suelen ser difíciles de calcular algebraicamente o numéricamente, más aún si no se conoce la distribución. Por ejemplo

$$ECM(\hat{\theta}, \theta) = \int \ldots \int (g(x_1, \ldots, x_n) - \theta)^2 \, f(x_1) \ldots f(x_n) dx_1 \ldots dx_n.$$

- La técnica de Bootstrap propone aproximar esta estimación utilizando la distribución empírica.

Técnica bootstrap

- Si n es suficientemente grande, suele ser cierto que:
  - (Glivenko-Cantelli): F_e converge uniformemente en x a F, con probabilidad 1.
  - Puede suponerse que los parámetros $\theta(F_e)$ de F_e se aproximan a los parámetros θ de F de manera continua.
- Entonces, por ejemplo: el error cuadrático medio del estimador

$$\hat{\theta} = g(X_1, X_2, \ldots, X_n)$$

podría aproximarse por:

$$ECM_e(\hat{\theta}, \theta) = E_{F_e}[(g(X_1, X_2, \ldots, X_n) - \theta(F_e))^2],$$

- $ECM_e(\hat{\theta}, \theta)$: aproximación bootstrap al error cuadrático medio.

Ejemplo

A partir de las 2 observaciones

$$X_1 = 1, \qquad X_2 = 3,$$

calcular la aproximación bootstrap de $ECM(\overline{X}, \theta)$ y $ECM(S^2, \sigma^2)$, siendo $\overline{X} = \frac{1}{2}(X_1 + X_2)$ y $S^2 = \frac{1}{2-1}\sum_{i=1}^{2}(X_i - \overline{X})^2$.

- Dado que $\overline{X}$ y S^2 son estimadores insesgados de la media y de la varianza respectivamente, se tiene que el error cuadrático medio con respecto a estos parámetros es igual a la varianza.
- $\text{Var}(\overline{X}) = E[(\overline{X} - E[\overline{X}])^2]$.
- $\text{Var}(S^2) = E[(S^2 - E[S^2])^2]$.
- Para la aproximación bootstrap utilizamos la distribución empírica.
- La distribución empírica da peso $p_1 = p_2 = \frac{1}{2}$.

Aproximación bootstrap

Varianza de la media muestral: $\mathrm{Var}(\overline{X})$

$$\mathrm{Var}_{F_e}(\overline{X}) = E_{F_e}[(\overline{X} - E_{F_e}[\overline{X}])^2]$$

- $E_{F_e}[\overline{X}] = \theta_{F_e} = \frac{1+3}{2} = 2.$

Muestras x_1	Muestras x_2	$\overline{X}$	$(\overline{X}-2)^2$
1	1	1	1
1	3	2	0
3	1	2	0
3	3	3	1

$$\mathrm{Var}_{F_e}(\overline{X}) = \frac{1}{4}\cdot 1 + \frac{1}{4}\cdot 0 + \frac{1}{4}\cdot 0 + \frac{1}{4}\cdot 1 = \frac{1}{2}.$$

Aproximación bootstrap

Varianza de la varianza muestral: $\mathrm{Var}(S^2)$

$$\mathrm{Var}_{F_e}(S^2) = E_{F_e}[(S^2 - E_{F_e}[S^2])^2]$$

- $E_{F_e}[S^2] = \sigma^2_{F_e} = \frac{(1-2)^2+(3-2)^2}{2} = 1.$

Muestras x_1	x_2	$\overline{X}$	S^2	$(S^2-1)^2$
1	1	1	0	1
1	3	2	2	1
3	1	2	2	1
3	3	3	0	1

$$\mathrm{Var}_{F_e}(S^2) = \frac{1}{4}\cdot 1 + \frac{1}{4}\cdot 1 + \frac{1}{4}\cdot 1 + \frac{1}{4}\cdot 1 = 1.$$

Aproximación bootstrap para $ECM(\overline{X}(n), \theta)$

En general, para calcular $ECM(\overline{X}(n))$ con respecto a la media θ:

- Obtener una muestra $x_1, x_2, \ldots, x_n$ (datos observados).
- F_e le asigna probabilidad $1/n$ a cada uno de estos datos (sumando pesos si no son todos distintos).
- Bootstrap: estimar el ECM por

$$ECM_{F_e}(\overline{X}(n), \theta_{F_e}) = E_{F_e}[(\overline{X}(n) - \theta_{F_e})^2].$$

- $ECM(\overline{X}(n), \theta) = \text{Var}(\overline{X}(n))$, ya que $\overline{X}(n)$ es insesgado al estimar la media.
- $\text{Var}(\overline{X}(n)) = \sigma^2/n = E[S^2/n]$.
- Para estimar $\text{Var}(\overline{X}(n))$ podemos plantearlo como:
  - $E_{F_e}[S^2/n]$.
  - $E_{F_e}[(\overline{X}(n) - \theta_{F_e})^2]$.

$\underline{E_{F_e}[S^2/n]}$

$$E_{F_e}[S^2/n] = \frac{\text{Var}_{F_e}(X)}{n} = \frac{1}{n^2}\sum_{i=1}^{n}(x_i - \theta_{F_e})^2$$

siendo

$$\theta_{F_e} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$\underline{E_{F_e}[(\overline{X}(n) - \theta_{F_e})^2]}$

$$\begin{aligned} E_{F_e}[(\overline{X}(n) - \theta_{F_e})^2] &= \text{Var}_{F_e}\left(\overline{X}(n)\right) \\ &= \frac{\text{Var}_{F_e}(X)}{n} = \frac{1}{n^2}\sum_{i=1}^{n}(x_i - \theta_{F_e})^2 \end{aligned}$$

Valor observado de S^2/n: $\dfrac{1}{n(n-1)}\displaystyle\sum_{i=1}^{n}(x_i - \theta_{F_e})^2$.

Aproximación bootstrap para Var(S^2)

- Obtener una muestra $x_1, x_2, \ldots, x_n$ (datos observados).
- F_e le asigna probabilidad $1/n$ a cada uno de estos datos (sumando pesos si no son todos distintos).
- Utilizar la distribución empírica para calcular:

$$E_{F_e}[(S^2(n) - E_{F_e}[S^2(n)])^2] = E_{F_e}[(S^2(n) - \text{Var}\, F_e(X))^2]$$

siendo

$$\begin{aligned} \theta_{F_e} &= \frac{1}{n}(x_1 + x_2 + \cdots + x_n) \\ \text{Var}_{F_e}(X) &= \frac{1}{n}\left((x_1 - \theta_{F_e})^2 + (x_2 - \theta_{F_e})^2 + \cdots + (x_n - \theta_{F_e})^2\right) \end{aligned}$$

Bootstrap y Montecarlo

- Una aproximación bootstrap requiere una suma de n^n términos, si la muestra es de n observaciones.
- Cualquiera sea el estimador $\hat{\theta} = g(X_1, \ldots, X_n)$ de un parámetro θ, la estimación bootstrap de $ECM(\hat{\theta})$:

$$\sum_{1 \leq i_1 \leq n} \cdots \sum_{1 \leq i_n \leq n} \frac{(g(x_{i_1}, x_{i_2}, \ldots, x_{i_n}) - \theta(F_e))^2}{n^n}$$

- Por ejemplo, para la aproximación de $ECM(S^2(n)$, se debe calcular:
  - θ_{F_e}: una vez.
  - Var_{F_e}: una vez.
  - Por cada una de las n^n muestras calcular el promedio $\overline{x}(n)$ y la varianza muestral $s^2(n)$ y hacer

$$(s^2(n) - \mathrm{Var}_{F_e})^2.$$

Bootstrap y Montecarlo

- Montecarlo: Este promedio puede aproximarse con un promedio de r términos, tomando r muestras aleatorias $(X_1^j, X_2^j, \ldots, X_n^j)$, $1 \leq j \leq r$:

$$\begin{aligned} Y_1 &= \left(g(X_1^1, X_2^1, \ldots, X_n^1) - \theta(F_e)\right)^2 \\ Y_2 &= \left(g(X_1^2, X_2^2, \ldots, X_n^2) - \theta(F_e)\right)^2 \\ &\vdots \\ Y_r &= \left(g(X_1^r, X_2^r, \ldots, X_n^r) - \theta(F_e)\right)^2 \end{aligned}$$

$$ECM_e(\hat{\theta}) \approx \frac{\sum_{j=1}^{r} Y_j}{r}.$$

Ejemplo

A partir de las 15 observaciones

$$5, 4, 9, 6, 21, 17, 11, 20, 7, 10, 21, 15, 13, 16, 8,$$

calcular la aproximación bootstrap de $\text{Var}(S^2) = \text{Var}(S^2(15))$.

- La distribución empírica da peso $p(21) = \frac{2}{15}$ y $p(x) = \frac{1}{15}$ a los restantes 13 valores.
- $\theta_{F_e} = 12.2$
- $\text{Var}_{F_e}(X) = 32.03$.
- Para cada una de las 15^{15} muestras $y_1, \ldots, y_{15}$ calcular
  - $\overline{y} = \frac{1}{15} \sum_{i=1}^{15} y_i$,
  - $s^2(n) = \frac{1}{14} \sum_{i=1}^{15} (y_i - \overline{y})^2$,
  - $(s^2(n) - 32.03)^2$,
- y promediar.

Consideraciones

X con distribución F_e
Generar $U \sim \mathcal{U}(0,1)$; $I \leftarrow \lfloor nU \rfloor + 1$; $X \leftarrow x[I]$

- Heurística: con 100 simulaciones se obtiene una buena aproximación de ECM_{F_e}.
- Esta aproximación bootstrap es a su vez una aproximación de ECM.

Ejemplo

Si se quiere estimar el tiempo promedio que un cliente pasa en un sistema debido a:

- Tiempo de espera en cola.
- Tiempo(s) de servicio.
- $W_i \leftarrow$ tiempo que permanece el i-ésimo cliente en el sistema.
- Se quiere calcular

$$\theta = \lim_{n \to \infty} \frac{W_1 + \cdots + W_n}{n}.$$

Ejemplo

- Notar: los tiempos W_i no son independientes ni idénticamente distribuidos.

En un caso simple de un solo servidor, en el que los clientes son atendidos por orden de llegada:

A_i: tiempo de arribo del cliente i.
S_i: tiempo de servicio del cliente i.
D_i: tiempo de salida del cliente i.

$$D_i = \max\{A_i, D_{i-1}\} + S_i, \qquad D_0 = 0$$

W_i: tiempo que pasa el cliente i en el sistema,

$$W_i = D_i - A_i = \max\{A_i, D_{i-1}\} + S_i - A_i.$$

$N_i \leftarrow$ número de clientes el día i:
$D_i \leftarrow$ suma de tiempos que permanecen los clientes en el sistema el día i:

$$\begin{aligned} D_1 &= W_1 + \cdots + W_{N_1} \\ D_2 &= W_{N_1+1} + \cdots + W_{N_1+N_2} \\ &\vdots \\ D_i &= W_{N_1+\cdots+N_{i-1}+1} + \cdots + W_{N_1+\cdots+N_i} \end{aligned}$$

$$\begin{aligned} \theta &= \lim_{m\to\infty} \frac{D_1 + \cdots + D_m}{N_1 + \cdots + N_m} = \lim_{m\to\infty} \frac{(D_1 + \cdots + D_m)/m}{(N_1 + \cdots + N_m)/m} \\ &= \frac{E[D]}{E[N]} \end{aligned}$$

Estimación de θ

- Simular el sistema k días.
- $\overline{D} = \dfrac{D_1 + \cdots + D_k}{k}$.
- $\overline{N} = \dfrac{N_1 + \cdots + N_k}{k}$.
- $\hat{\theta} = \dfrac{\overline{D}}{\overline{N}}$.

$$ECM\left(\frac{\overline{D}}{\overline{N}}\right) = E\left[\left(\frac{\sum_i D_i}{\sum_i N_i} - \theta\right)^2\right].$$

Aproximación bootstrap

- Observar valores d_i, n_i, $1 \leq i \leq k$.
- Distribución empírica:

$$P_{F_e}(D = d_i,\ N = n_i) = \frac{1}{k}$$

- $E_{F_e}(D) = \overline{d} = \sum_{i=1}^{k} d_i/k$.
- $E_{F_e}(N) = \overline{n} = \sum_{i=1}^{k} n_i/k$.
- $\theta_{F_e} = \dfrac{\overline{d}}{\overline{n}}$.

$$ECM_{F_e}\left(\frac{\overline{D}}{\overline{N}}\right) = \frac{1}{k^k} \sum_{(i_1,\ldots,i_k)} \left(\frac{d_{i_1} + \cdots + d_{i_k}}{n_{i_1} + \cdots + n_{i_k}} - \frac{\overline{d}}{\overline{n}}\right)^2.$$