ITAM
Maestría en Ciencia de Datos
Estadística Computacional

Ariana Judith López Coronel
Denisse Aneth Martínez Mejorado
Eduardo David Martínez Neri

Corrección del Sesgo mediante Bootstrap

La estimación del sesgo en *bootstrap* se plantea como

$$\sum_{i=1}^{B}\left(\widehat{\theta}_i^{*}-\theta\right)/B \;=\; \overline{\theta}^{\,*}-\widehat{\theta}$$

El bootstrap puede ser utilizado para reducir el sesgo de una muestra finita de un estimador.

Para ejemplificar, sea X un vector aleatorio y fijemos $\mu = E(X)$. Considerando que el valor verdadero de θ es $\theta_o = g(u)$, donde g es una función continua conocida. Suponga que los datos consisten en una muestra aleatoria $\{X_i : i = 1,...,n\}$ de X. Definiendo el vector $\overline{X} = n^{-1}\sum_{i=1}^{n} X_i$. Entonces θ es estimado consistentemente por

$$\theta_n = g\left(\overline{X}\right) \tag{1}$$

Si θ_n tienen una media finita, entonces $E(\theta_n) = E\left[g\left(\overline{X}\right)\right]$. Sin embargo, $E\left[g\left(\overline{X}\right)\right] \neq g(u)$ en general, a menos de que g sea una función lineal. Por lo tanto, $E(\theta_n) \neq \theta_o$ y θ_n es un estimador sesgado de θ. En particular, $E(\theta_n) \neq \theta_o$ si θ_n es cualquier variedad de máxima verosimilitud o estimadores por método generalizado de momentos.

Para observar cómo un *bootstrap* puede reducir el sesgo de θ_n, suponga que g es cuatro veces continuamente diferenciable en el vecindario de μ y que los componentes de X tienen 4 momentos absolutos finitos. Dejemos que G_1 muestre el vector de primeras derivadas de g y G_2 muestre la matriz de segundas derivadas. Una expansión de series de Taylor agregada a la ecuación (1) sobre $\overline{X} = \mu$ da

$$\theta_n - \theta_o = G_1(u)'\left(\overline{X} - \mu\right) + \tfrac{1}{2}\left(\overline{X} - \mu\right)'G_2(u)\left(\overline{X} - \mu\right) + R_n \tag{2}$$

donde R_n es el término remanente que satisface $E(R_n) = O\left(n^{-2}\right)$. Por lo tanto, sacando la esperanza en ambos lados de la ecuación (2) obtenemos

$$E(\theta_n - \theta_o) = \tfrac{1}{2}E\left[\left(\overline{X} - \mu\right)'G_2(u)\left(\overline{X} - \mu\right)\right] + O\left(n^{-2}\right) \tag{3}$$

El primer elemento del lado derecho de la ecuación (3) tiene tamaño $O\left(n^{-1}\right)$[1]. Por lo que a través de $O\left(n^{-1}\right)$ el sesgo de θ_n es

$$B_n = \tfrac{1}{2}E\left[\left(\overline{X}-\mu\right)'G_2(u)\left(\overline{X}-\mu\right)\right] \tag{4}$$

Ahora consideremos el bootstrap. El bootstrap muestrea la distribución empírica de los datos. Sea $\{X_i^* : i = 1,...,n\}$ una muestra bootstrap que es obtenida de esta manera. Definiendo $\overline{X}^* = n^{-1}\sum_{i=1}^{n} X_i^*$ como el vector de medias de la muestra bootstrap. El estimador bootstrap de θ es $\theta_n^* = g\left(\overline{X}^*\right)$. Condicional en los datos, la verdadera media de una distribución muestreada por bootstrap es $\overline{X}$. Por lo tanto, $\overline{X}$ es el bootstrap análogo de μ, y $\theta_n = g\left(\overline{X}\right)$ es el bootstrap análogo de θ_o. El bootstrap análogo de la ecuación (2) es

$$\theta_n^* - \theta_n = G_1\left(\overline{X}\right)'\left(\overline{X}^* - \overline{X}\right) + \tfrac{1}{2}\left(\overline{X}^* - \overline{X}\right)'G_2\left(\overline{X}\right)\left(\overline{X}^* - \overline{X}\right) + R^*_n \tag{5}$$

donde R^*_n es el término remanente del bootstrap. Sea E^* la esperanza bajo la muestra bootstrap, esto es, la esperanza relativa a la distribución empírica de los datos estimación. Sea $B_n^* \equiv E^*(\theta_n^* - \theta_n)$ el sesgo de θ_n^* como un estimador de θ_n. Calculando la esperanza E^* en los dos lados de la ecuación (5) se muestra que

$$B_n = \tfrac{1}{2}E^*\left[\left(\overline{X}^* - \overline{X}\right)'G_2\left(\overline{X}\right)\left(\overline{X}^* - \overline{X}\right)\right] + O\left(n^{-2}\right) \tag{6}$$

Debido a que la distribución muestreada por bootstrap es conocida, B_n^* puede ser calculada con precisión arbitraria por una simulación Monte Carlo. Así, B_n^* es un estimador factible del sesgo de θ_n
.

Comparando las ecuaciones (4) y (6), se puede observar que las únicas diferencias entre B_n y el término principal de B_n^* es que $\overline{X}$ reemplaza a μ en B_n^* y la esperanza empírica de E^*, reemplaza la esperanza de la población, E. Más aún, $E(B_n^*) = B_n + O\left(n^{-2}\right)$. Por lo tanto, a través de $O\left(n^{-1}\right)$, el uso de la estimación de sesgo bootstrap B_n^* provee la misma reducción del sesgo en la población que sería obtenido si utilizaramos B_n. Esta es la fuente de la habilidad del bootstrap para reducir el sesgo de θ_n. El estimador con sesgo corregido resultante de θ es $\theta_n - B_n^*$. Este satisface $E(\theta_n - \theta_o - B_n^*) = O\left(n^{-2}\right)$. Por tanto, el sesgo del estimador con sesgo corregido es $O\left(n^{-2}\right)$, mientras que el sesgo del estimador no corregido θ_n es $O\left(n^{-1}\right)$.

Se procedió a generar el siguiente ejercicio tomado de referencia: “Chapter 52: The Bootstrap” página 3,174 (http://www.unc.edu/~saraswat/teaching/econ870/fall11/JH_01.pdf):

Suponga $X \sim N(0,6)$ y $n = 1{,}000$. Suponga $g(\mu) = exp(\mu)$. Entonces $\theta_o = 1$ y $\theta_n = exp(\overline{X})$. B_n y el sesgo de θ_n - B_n^* se estima mediante el siguiente procedimiento Monte Carlo:

---

[1] el error en la aproximación del bootstrap a una distribución simétrica. Las aproximaciones asintóticas de primer orden a las distribuciones muestrales finitas de estadísticas chi cuadrada asintótica, típicamente genera errores de tamaño On^-1.

1) Se genera un set de datos para estimación de tamaño n muestreando de la distribución $N(0,6)$. Se usan estos datos para calcular θ_n.
2) Para estimar B_n^* mediante Bootstrap Monte Carlo:
    a) Se estima θ_n.
    b) Se generan muestras bootstrap de tamaño n muestreando con reemplazo del set de datos. Se estima $\theta_n^* = g(\overline{X}^*)$.
    c) Se estima $E^*\theta_n^*$ promediando los resultados de varias repeticiones del paso b. Obtenemos $B_n^* = E^*\theta_n^* - \theta_n$.
3) Se estima $E(\theta_n - \theta_o)$ y $E(\theta_n - B_n^* - \theta_o)$ promediando los resultados de varias repeticiones de los pasos 1 y 2. Se estiman los errores cuadráticos medios de θ_n y $\theta_n - B_n^*$ a través de promediar los valores de $(\theta_n - \theta_o)^2$ y $(\theta_n - B_n^* - \theta_o)^2$.

Ver ejercicio: https://eduardomtz.shinyapps.io/Bootstrap

Referencia:

http://www.unc.edu/~saraswat/teaching/econ870/fall11/JH_01.pdf