Intervalos de sesgo corregido y acelerado

Carlos González Munguía

5 de octubre de 2015

El método bootstrap acelerado y corrector de sesgo (Efron yTibshirani, 1993) o por sus siglas en ingles $\boldsymbol{BC_a}$ (bias-corrected and accelerated), realiza una corrección por el sesgo encontrado en el caso de Bootstrap por Percentiles.

Sólo para entender un poco lo que se está haciendo, en el caso de *bootstrap* por percentiles se denota a G como la función de distribución acumulada de $\widehat{\theta}^*$ que es el estadístico calculado de las n muestras *bootstrap*, si se quiere un intervalo para el estadístico, se define por los percentiles α y $1-\alpha$ de G como:

$$(\theta^*_{inf}, \theta^*_{sup}) = (G^{-1}(\alpha), G^{-1}(1-\alpha))$$

Por definifción, $G^{-1}(\alpha) = \widehat{\theta}^*(\alpha)$, esto es el percentil 100α de la distribución bootstrap, por lo que podemos escribir el intervalo bootstrap como

$$(\theta^*_{inf}, \theta^*_{sup}) = (\widehat{\theta}^*(\alpha), \widehat{\theta}^*(1-\alpha))$$

En la práctica, usualmente la distribución empírica de $\widehat{\theta}^*$ es asimétrica, por lo que no coinciden los intervalos, lo cual se pueden corregir aplicando una transformación, después calcular los intervalos usando la normal y aplicar la transformación inversa para volver a la escala original, el problema con este método es que requiere que se conozca la transformación adecuada para cada parámetro. Por otra parte, podemos pensar en el método del percentil como un algoritmo que incorpora la transformación de manera automática, es decir:

Supongamos que la transformación $\hat{\phi} = m(\hat{\theta})$ normaliza la distribución de $\hat{\theta}$ de manera perfecta,

$$\hat{\phi} \approx N(\phi, c^2)$$

para alguna desviación estándar c. Entonces el intervalo de percentil basado en $\hat{\theta}$ es igual a

$$(m^{-1}(\hat{\phi} - z^{(1-\alpha)}c), m^{-1}(\hat{\phi} - z^{(\alpha)}c))$$

Sin embargo, no se tienen cubiertos todos los casos ya que existen ocasiones en que $\hat{\theta}$ tiene una distribución no sesgada, es decir,

$$\hat{\theta} \approx N(\theta + sesgo, se^2)$$

y en este caso, no se puede encontrar una transformación $m\theta$ que pueda volver simétrico el intervalo, por lo que se recurre al método de bootstrap acelerado $\boldsymbol{BC_a}$.

El método $\boldsymbol{BC_a}$ corrige el sesgo de manera automática, lo cuál es una de sus principales ventajas comparado con el método del percentil. La idea de la construcción de dichos intervalos es la siguiente.

Supongamos que existe una función ϕ monótonamente creciente con a y b constantes tales que

$$U = \frac{\phi(\hat{\theta}) - \phi(\theta)}{1 + a\phi(\theta)} + b$$

dando origen una distribución $N(0,1)$ con $1 + a\phi(\theta) > 0$. Notar que si a=b=0, se tiene el método de percentil.

Por el principio de bootstrap se tiene que

$$U^* = \frac{\phi(\hat{\theta}^*) - \phi(\theta)}{1 + a\phi(\theta)} + b$$

se distribuye aproximadamente normal estándar, para cualquier cuantil de una distribución normal estándar, decimos que z_a,

$$\alpha \approx P^*[U^* \leq z_\alpha] = P^*[\hat{\theta}^* \leq \phi^{-1}(\phi(\theta) + (z_\alpha - b)[1 + a\phi(\theta)])]$$

Sin embargo, el cuantil α de la distribución empírica de la distribución de $\hat{\theta}^*$, es observable de la distribución *bootstrap*. Entonces, el principio de bootstrap puede ser aplicado para concluir que θ aproximará al $1 - \alpha$ límite de confianza superior.

Con esto, los extremos en los intervalos $\boldsymbol{BC_a}$ están dados por percentiles de la distribución bootstrap, por lo que hay que notar que los percentiles usados dependen de dos números $\hat{a}$ y $b = \hat{z_0}$, que se denominan la aceleración y la corrección del sesgo.

$$\boldsymbol{BC_a} : (\theta_{inf}, \theta_{sup}) = (\widehat{\theta}^*(\alpha_1), \widehat{\theta}^*(\alpha_2))$$

donde

$$\alpha_1 = \Phi(\hat{z_0} + \frac{\hat{z_0} + z^{(\alpha)}}{1 - \hat{\alpha}(\hat{z_0} + z^{(\alpha)})})$$

$$\alpha_2 = \Phi(\hat{z_0} + \frac{\hat{z_0} + z^{(1-\alpha)}}{1 - \hat{\alpha}(\hat{z_0} + z^{(1-\alpha)})})$$

donde Φ es la función de distribución acumulada de la distribución normal estándar y z^α es el percentil 100α de una distribución normal estándar.

El valor de la corrección por sesgo $\hat{z_0}$ se obtiene de la propoción de replicaciones bootstrap menores a la estimación original $\hat{\theta}$,

$$\hat{a} = \frac{\sum_{i=1}^{n}(\hat{\theta(.)} - \hat{\theta(i)})^3)}{6(\sum_{i=1}^{n}(\hat{\theta(.)} - \hat{\theta(i)})^2)^{3/2}}$$

donde $\hat{\theta(.)}$ es la *i-ésima* estimación de θ y $\hat{\theta(i)}$ es el promedio de todos los $\hat{\theta(.)}$; por otro lado, $\hat{z_0}$ mide la mediana del sesgo de $\hat{\theta^*}$, esto es, la discrepancia entre la mediana de $\hat{\theta^*}$ y $\hat{\theta}$ en unidades normales. La aceleración $\hat{a}$ se refiere a la tasa de cambio del error estándar de $\hat{\theta}$ respecto al verdadero valor del parámetro θ. La aproximación.

Una vez calculado el primer factor de corrección $\hat{z_0}$ y el segundo factor $\hat{a}$, se calculan las cotas L y U del intervalo con niveles de confianza $\alpha 100\%$ de la siguiente manera:

$$L = B * \Phi(\hat{z_0} + \frac{\hat{z_0} + z^{(\alpha)}}{1 - \hat{\alpha}(\hat{z_0} + z^{(\alpha)})})$$

$$U = B * \Phi(\hat{z_0} + \frac{\hat{z_0} + z^{(1-\alpha)}}{1 - \hat{\alpha}(\hat{z_0} + z^{(1-\alpha)})})$$

Los intervalos de confianza $\boldsymbol{BC_a}$ son: $L \leq \theta \leq U$

Para que el intervalo $\boldsymbol{BC_a}$ sea suficientemente confiable es recomendable que B sea de al menos 1000, de acuerdo a Effron y Tibshirani.

Ventajas:

- Respetan transformaciones, esto nos dice que los extremos del intervalo se transforman de manera adecuada si cambiamos el parámetro de interés.
- los intervalos $\boldsymbol{BC_a}$ tienen precisión de segundo orden.

Ejemplo:

Para ejemplificar los intervalos $\boldsymbol{BC_a}$ se tomarán los datos de *spatial*, el cuál es un ejemplo de Effron y Tibshirani, los datos constan de los resultados de 2 pruebas espaciales aplicadas a 26 niños con algún problema neurológico.

```
library(bootstrap)
library(boot)
data(spatial)
str(spatial)
```

```
## 'data.frame':    26 obs. of  2 variables:
##  $ A: num  48 36 20 29 42 42 20 42 22 41 ...
##  $ B: num  42 33 16 39 38 36 15 33 20 43 ...
```

El ejemplo se encuentra en un shiny con código en github https://github.com/montactuaria/CompuStat/tree/master/Tarea4

Referencias

- Computational Statistics, Geof H. Givens and Jennifer A. Hoeting
- An introduction to the bootstrap, B. Effron, R. J. Tibshirani.
- Bootstrap Methods and their applications, A. C. Davison, D.V. Hinkley.
- Efron, B. and Tibshirani, R. (1993) An Introduction to the Bootstrap. Chapman and Hall, New York, London.