Intervalos de confianza bootstrap
métodos percentil

Grau en Estadística

Mètodes no paramètrics i de remostreig

Jordi Ocaña Rebull

Departament d'Estadística

El método percentil. Definición

- Situación de partida: θ parámetro de interés, $\hat{\theta}$ su estimador "plug-in", $\hat{G}$ estimador bootstrap de la distribución de $\hat{\theta}$
- IC *percentil bootstrap* de θ, con recubrimiento nominal $1-\alpha$: $\left[\hat{G}^{-1}\left(\frac{\alpha}{2}\right),\quad \hat{G}^{-1}\left(1-\frac{\alpha}{2}\right)\right]$
- En la práctica se suele aproximar mediante $\left[\hat{\theta}^*_{(\alpha/2)}, \hat{\theta}^*_{(1-\alpha/2)}\right]$ donde $\hat{\theta}^*_{(p)}$ corresponde al percentil muestral p obtenido a partir de B réplicas bootstrap: $\hat{\theta}^*_1, \hat{\theta}^*_2, \ldots, \hat{\theta}^*_B$

Motivación método percentil (i)

- Sea $\hat{\sigma}_{\hat{\theta}}$ un estimador de la desviación estándar de $\hat{\theta}$
- Recordemos: si $\dfrac{\hat{\theta}-\theta}{\hat{\sigma}_{\hat{\theta}}} \approx N(0,1)$

  entonces $\hat{\theta} \pm z_{\alpha/2}\hat{\sigma}_{\hat{\theta}}$ es IC “estándar” con recubrimiento $1-\alpha$, aproximadamente
- Extremos del IC estándar equivalentes a percentiles $\alpha/2$ y $1-\alpha/2$ de la distribución de $\hat{\theta}^{*} \sim N\left(\hat{\theta}, \hat{\sigma}_{\hat{\theta}}\right)$

Motivación método percentil (ii)

- Si cierto $\hat{\theta} \approx N\left(\theta, \sigma_{\hat{\theta}}\right)$, en general también $\hat{\theta}^* \approx N\left(\hat{\theta}, \hat{\sigma}_{\hat{\theta}}\right)$, a su vez bien emulada por $\hat{G}$, la estima bootstrap de la distribución de $\hat{\theta}$
- Es decir: $$\hat{\theta} - z_{\alpha/2}\hat{\sigma}_{\hat{\theta}} \approx \hat{G}^{-1}\left(\alpha / 2\right)$$ $$\hat{\theta} + z_{\alpha/2}\hat{\sigma}_{\hat{\theta}} \approx \hat{G}^{-1}\left(1 - \alpha / 2\right)$$
- Pero ¿y si $\hat{\theta}$ no normal? (p.e. $\theta = \rho$, coef. de correlación): ¿existe transformación normalizadora y estabilizadora de varianza?

Motivación método percentil (y iii)

- Si existe h, monótona creciente, tal que $\phi = h\left(\theta\right)$, $\hat{\phi} = h\left(\hat{\theta}\right)$, $\hat{\phi} \approx N\left(\phi, \sigma_{\hat{\phi}}\right)$, $\sigma_{\hat{\phi}}$ cte.

  p.e. $\phi = \dfrac{1}{2}\log\dfrac{1+\rho}{1-\rho} = \tanh^{-1}(\rho)$, $\quad \hat{\phi} \approx N\left(\phi, \frac{1}{\sqrt{n-3}}\right)$

- En escala ϕ, intervalo percentil $\approx$ estándar
- Monotonicidad de $h \Rightarrow$ en escala $\theta = h^{-1}(\phi)$, IC percentil (¡obtenido directamente, sin conocer h!) aproximadamente correcto

En resumen:

- Esquemáticamente:

$$\hat{\theta}^*_{(p)} = \hat{G}^{-1}(p), \qquad \phi^*_{(p)} = \underbrace{\hat{Q}^{-1}(p) \cong \hat{\phi} + \Phi^{-1}(p)\sigma_{\hat{\theta}}}_{\text{interpretable como percentil}}$$

h^{-1}

donde $\hat{Q}$ es la distribución bootstrap de $\hat{\phi}$ y Φ es la función de distribución $N(0,1)$

- Entonces justificada la validez del IC

$$\left[\hat{G}^{-1}\left(\frac{\alpha}{2}\right), \hat{G}^{-1}\left(1-\frac{\alpha}{2}\right)\right] = \hat{G}^{-1}\left(\Phi\left(-z_{\frac{\alpha}{2}}\right), \Phi\left(z_{\frac{\alpha}{2}}\right)\right)$$

Modelo para el sesgo y la heteroscedasticidad

- Supongamos que existe una transformación h, normalizadora, pero que no corrige el sesgo ni estabiliza la varianza, en concreto sea

$$\phi = h(\theta), \quad \hat{\phi} = h(\hat{\theta})$$

$$\frac{\hat{\phi} - \phi}{\sigma_{\hat{\phi}}(\phi)} + z_0 \approx N(0,1), \quad \text{con } \sigma_{\hat{\phi}}(\phi) = 1 + a\phi$$

y por lo tanto, también $\dfrac{\hat{\phi}^* - \hat{\phi}}{1 + a\hat{\phi}} + z_0 \approx N(0,1)$

Construcción de los IC BCa. Intervalo en la escala normal

De $1-\alpha \cong \Pr\left\{-z_{\frac{\alpha}{2}} \le \dfrac{\hat{\phi}-\phi}{1+a\phi} + z_0 \le z_{\frac{\alpha}{2}}\right\}$

o, equivalentemente,

$$\alpha \cong \Pr\left\{-z_{\frac{\alpha}{2}} > \frac{\hat{\phi}-\phi}{1+a\phi} + z_0\right\} + \Pr\left\{z_{\frac{\alpha}{2}} > \frac{\hat{\phi}-\phi}{1+a\phi} + z_0\right\},$$

llegamos a la conclusión de que un IC de nivel, aproximado, $1-\alpha$ para ϕ viene dado por:

$$\Pr\left\{\hat{\phi} + \frac{z_0 - z_{\frac{\alpha}{2}}}{1-a\left(z_0 - z_{\frac{\alpha}{2}}\right)}\sigma_{\hat{\phi}}\left(\hat{\phi}\right) \le \phi \le \hat{\phi} + \frac{z_0 + z_{\frac{\alpha}{2}}}{1-a\left(z_0 + z_{\frac{\alpha}{2}}\right)}\sigma_{\hat{\phi}}\left(\hat{\phi}\right)\right\}$$

Construcción de los IC Bca. Intervalo en la escala normal

Los extremos del intervalo anterior pueden considerarse percentiles de una distribucion $N\left(\hat{\phi} - z_0\sigma_{\hat{\phi}}\left(\hat{\phi}\right), \sigma_{\hat{\phi}}\left(\hat{\phi}\right)\right)$ pero no los percentiles $\frac{\alpha}{2}$ y $1-\frac{\alpha}{2}$, sino percentiles

$$\alpha_1 = \Phi\left(z_0 + \frac{z_0 - z_{\frac{\alpha}{2}}}{1 - a\left(z_0 - z_{\frac{\alpha}{2}}\right)}\right) \text{ y}$$

$$1 - \alpha_2 = \Phi\left(z_0 + \frac{z_0 + z_{\frac{\alpha}{2}}}{1 - a\left(z_0 + z_{\frac{\alpha}{2}}\right)}\right)$$

respectivamente

Construcción de los IC Bca. Intervalo en la escala original

Por lo tanto, el intervalo en la escala θ:

$$\left[h^{-1}\left(\hat{\phi} + \frac{z_0 - z_{\frac{\alpha}{2}}}{1 - a\left(z_0 - z_{\frac{\alpha}{2}}\right)} \sigma_{\hat{\phi}}\left(\hat{\phi}\right) \right), h^{-1}\left(\hat{\phi} + \frac{z_0 + z_{\frac{\alpha}{2}}}{1 - a\left(z_0 + z_{\frac{\alpha}{2}}\right)} \sigma_{\hat{\phi}}\left(\hat{\phi}\right) \right) \right]$$

tendrá extremos de la forma:

$$\hat{\theta}^*_{(\alpha_1)} = \hat{G}^{-1}\left(\Phi\left(z_0 + \frac{z_0 - z_{\frac{\alpha}{2}}}{1 - a\left(z_0 - z_{\frac{\alpha}{2}}\right)} \right) \right) \quad \text{y}$$

$$\hat{\theta}^*_{(1-\alpha_2)} = \hat{G}^{-1}\left(\Phi\left(z_0 + \frac{z_0 + z_{\frac{\alpha}{2}}}{1 - a\left(z_0 + z_{\frac{\alpha}{2}}\right)} \right) \right)$$

Estima de la corrección del sesgo z_0

- Falta determinar el valor del parámetro de corrección del sesgo, z_0, y de la "constante de aceleración", a
- Estima de z_0: $\hat{z}_0 = \Phi^{-1}\left(\hat{G}\left(\hat{\theta}\right)\right).$
  - En efecto:

$$\hat{G}\left(\hat{\theta}\right) = P_*\left\{\hat{\theta}^* \leq \hat{\theta}\right\}$$
$$= P_*\left\{\hat{\phi}^* \leq \hat{\phi}\right\} = P_*\left\{\frac{\hat{\phi}^* - \hat{\phi}}{1 + a\hat{\phi}} + z_0 \leq z_0\right\} \cong \Phi\left(z_0\right)$$

Estima de la constante de aceleración

- Efron, para funcionales $\theta = t(F), \hat{\theta} = t\left(\hat{F}\right)$:

$$\hat{a} = \frac{\sum_{i=1}^{n} U_i^3}{6\left\{\sum_{i=1}^{n} U_i^2\right\}^{3/2}}$$

  – U_i es la función empírica de influencia asociada al dato i:

$$U_i = U\left(x_i, \hat{F}\right) = \lim_{\varepsilon \to 0} \frac{t\left((1-\varepsilon)\hat{F} + \varepsilon\delta_i\right) - t\left(\hat{F}\right)}{\varepsilon}$$

- Alternativamente, aproximación jackknife:

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

Resumen de intervalos percentil

- IC percentil:
  - Validez: ∃ transformación normalizante, centradora y estabilizadora de varianza (no necesario conocerla)
  - Definición: $\left[\hat{G}^{-1}\left(\frac{\alpha}{2}\right), \hat{G}^{-1}\left(1-\frac{\alpha}{2}\right)\right]$
- IC percentil corregido para el sesgo, acelerado (BCa):
  - Validez: ∃ transformación normalizante (no necesario conocerla)
  - Definición:

$$\left[\hat{G}^{-1}\left(\Phi\left(\hat{z}_0+\frac{\hat{z}_0-z_{\frac{\alpha}{2}}}{1-\hat{a}\left(\hat{z}_0-z_{\frac{\alpha}{2}}\right)}\right)\right), \hat{G}^{-1}\left(\Phi\left(\hat{z}_0+\frac{\hat{z}_0+z_{\frac{\alpha}{2}}}{1-\hat{a}\left(\hat{z}_0+z_{\frac{\alpha}{2}}\right)}\right)\right)\right]$$

- Si sesgo pero homoscedasticidad ⇒ $a = 0$: intervalo "corregido para el sesgo": BC