75570 TEORETISK STATISTIKK - VÅREN 1993

JACKKNIFING

Bo Lindqvist

Jackknife-estimatoren

La $X_1, \ldots, X_n$ være u.i.f. observasjoner av en stokastisk variabel X med en sannsynlighetsfordeling som avhenger av en ukjent parameter θ (der θ kan være en vektor). Anta at $\tau(\theta)$ skal estimeres.

La $T_n \equiv T_n(X_1, \ldots, X_n)$ være en estimator for $\tau(\theta)$. Vi definerer de skalte *pseudo-verdier* $T_n^{(i)}$ $(i = 1, \ldots, n)$ ved

$$T_n^{(i)} = T_{n-1}(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$$

dvs. at $T_n^{(i)}$ er det estimat vi ville få ved å ta bort observasjon nr. i $(i = 1, \ldots, n)$.

La $T_n^{(\cdot)}$ være gjennomsnittet av pseudo-verdiene, dvs.

$$T_n^{(\cdot)} = \frac{1}{n} \sum_{i=1}^{n} T_n^{(i)}$$

Da er *Jackknife-estimatoren* for $\tau(\theta)$ gitt ved

$$JK(T_n) = nT_n - (n-1)T_n^{(\cdot)}$$

Denne estimatoren ble foresltt av Quenoille (1956) som et middel for å redusere en eventuell forventningsskjevhet i estimatoren T_n. En begrunnelse er gitt i neste avsnitt.

Begrunnelse for Jackknife-estimatoren

Svært ofte vil estimatoren T_n ha en forventning av formen

$$E_\theta T_n = \tau(\theta) + \frac{a}{n} + \frac{b}{n^2} + O(\frac{1}{n^3})$$

der a og b godt kan avhenge av θ og der $O(\frac{1}{n^3})$ er en størrelse av *orden* $\frac{1}{n^3}$.

Da er

$$E_\theta T_n^{(\cdot)} = E_\theta T_n^{(i)} = \tau(\theta) + \frac{a}{n-1} + \frac{b}{(n-1)^2} + O(\frac{1}{(n-1)^3})$$

og dermed

$$E_\theta JK(T_n) = n\tau(\theta)+a+\frac{b}{n}+O(\frac{1}{n^2})-(n-1)\tau(\theta)-a-\frac{b}{n-1}+O(\frac{1}{n^2}) = \tau(\theta)+O(\frac{1}{n^2})$$

siden

$$\frac{b}{n} - \frac{b}{n-1} = \frac{b}{n(n-1)} = O(\frac{1}{n^2})$$

og

$$O(\frac{1}{(n-1)^2}) = O(\frac{1}{n^2})$$

Det følger at mens forventningsskjevheten i T_n er av orden $\frac{1}{n}$ vil skjevheten i $JK(T_n)$ være av orden $\frac{1}{n^2}$. Vi ser også at dersom estimatoren T_n er forventningsrett i utgangspunktet, vil Jackknifingen ikke ødelegge denne egenskapen (dvs. også $JK(T_n)$ vil være forventningsrett).

Jackknife-estimatoren for varians

Det finnes også en Jackknife estimator for *variansen* til estimatoren T_n. Denne er gitt ved

$$(\widehat{\mathrm{Var}T_n})_{\mathrm{JACK}} = \frac{n-1}{n}\sum_{i=1}^{n}(T_n^{(i)} - T_n^{(\cdot)})^2$$

Som en slags motivasjon for denne varians-estimatoren skal vi se hva den blir i et velkjent konkret tilfelle.

Eksempel: La $\tau(\theta) = E_\theta X$, og la vår estimator være $T_n = \overline{X}_n$, der som vanlig

$$\overline{X}_n = \sum_{i=1}^{n} X_i$$

Da er

$$T_n^{(i)} = \frac{1}{n-1}\sum_{j\neq i} X_j = \frac{n}{n-1}\overline{X}_n - \frac{1}{n-1}X_i$$

Det følger da at

$$T_n^{(\cdot)} = \sum_{i=1}^{n} T_n^{(i)} = \frac{n}{n-1}\overline{X}_n - \frac{1}{n-1}\overline{X}_n$$

slik at

$$T_n^{(i)} - T_n^{(\cdot)} = \frac{1}{n-1}(\overline{X}_n - X_i)$$

og dermed

$$(\widehat{\mathrm{Var}T_n})_{\mathrm{JACK}} = \frac{n-1}{n}\frac{1}{(n-1)^2}\sum_{i=1}^{n}(X_i-\overline{X}_n)^2 = \frac{1}{n(n-1)}\sum_{i=1}^{n}(X_i-\overline{X}_n)^2 \equiv \frac{S_n^2}{n}$$

der S_n^2 er den vanlige (forventningsrette) estimatoren for $\sigma^2 \equiv \mathrm{Var}_\theta X$.

Numerisk eksempel

Dette er egentlig Oppgave 2 fra øving 7, våren 1993.

La $X_1, \ldots, X_n$ være u.i.f. $\mathrm{N}(\theta, \theta^2)$, der θ er en ukjent parameter. Anta at vi har $n = 10$ observasjoner gitt ved:

$$\begin{array}{rrrrr} 22.2 & 10.0 & 8.3 & 25.3 & 15.1 \\ 8.8 & 16.6 & 18.0 & -15.9 & -4.5 \end{array}$$

SME for θ er gitt ved

$$T_n = \frac{-\bar{X}_n + \sqrt{(\bar{X}_n)^2 + 4Y_n}}{2} \tag{1}$$

der

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$Y_n = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

Med de gitte dataene får vi

$$\bar{X}_n = 10.39 \quad \text{og} \quad Y_n = 247.989$$

som gir

$$T_n = 11.38743$$

Følgende tabell gir de 10 pseudo-verdiene $T_n^{(i)}$, som vi får ved å bruke (1) når observasjon nr. i tas bort, $(i = 1, \ldots, n)$.

i	x_i	$T_n^{(i)}$
1	22.2	10.99768
2	10.0	11.86098
3	8.3	11.89635
4	25.3	10.58291
5	15.1	11.63612
6	8.80	11.88794
7	16.6	11.53438
8	18.0	11.42399
9	-15.9	10.42513
10	-4.5	11.57211

Dette gir

$$T_n^{(\cdot)} = 11.38176$$

og

$$JK(T_n) = 11.43850$$

mens variansestimatet blir

$$(\widehat{\mathrm{Var}T_n})_{\mathrm{JACK}} = 2.319747$$