Ocenjevanje parametrov

Nataša Kejžar

Povzetek

Metoda momentov

Pri tej metodi parametre vedno lahko ocenimo, vendar njihova varianca ni vedno najmanjša možna.

Momenti:

$$\mu_k = E(X^k)$$

Na vzorcu n enot:

$$\hat{\mu}_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k$$

Cenilka po metodi momentov:

$$\hat{\boldsymbol{\theta}} = f(\boldsymbol{\mu}),$$

kjer sta $\boldsymbol{\theta}$ vektor parametrov, ki jih ocenjujemo, in $\boldsymbol{\mu}$ vektor najmanjših možnih momentov.

Metoda največjega verjetja

Za reprezentativen vzorec velikosti n predpostavljamo, da so opazovanja x_i porazdeljena iid. Funkcijo verjetja lahko zato zapišemo kot produkt gostot f (ali verjetnosti v diskretnem primeru)

$$L(x,\theta) = \prod_{i=1}^{n} f(X_i = x_i, \theta).$$

Poiskati želimo vrednost θ, kjer bo verjetje največje.

Postopek:

1. zapišemo logaritmiramo verjetje

$$l(x,\theta) = \sum_{i=1}^{n} \log f(X_i = x_i, \theta)$$

2. odvajamo l po parametru(ih) θ in ga (jih) izenačimo z 0 (tj. iščemo ekstreme)

$$\frac{\partial l(x,\theta)}{\partial \theta} = 0$$

3. izrazimo cenilko $\widehat{\theta}$ kot funkcijo opazovanj x_i. To je cenilka MLE.

Varianca cenilke ima asimptotično varianco

$$var(\hat{\theta}) = \frac{1}{n} \cdot I(\theta_0)^{-1},$$

kjer je $I(\theta_0)$ Fisherjeva informacija.

$$\begin{aligned} I(\theta_0) =& E\left\{\left[(\log f(X,\theta_0))'\right]^2\right\} \\ =& -E\left[(\log f(X,\theta_0))''\right], \end{aligned}$$

kjer je θ_0 prava vrednost parametra. Te v praksi ne poznamo, zato jo ocenimo z $\widehat{\theta}$.

Srednja kvadratna napaka

Ena izmed mer za presojanje kakovosti cenilke. Angleško: mean squared error (MSE). Pove nam, kako daleč od prave vrednosti parametra pričakujemo, da bo njegova ocenjena vrednost iz podatkov:

$$E\Big[(\hat{\theta}-\theta_0)^2\Big] = var(\hat{\theta}) + \Big[E(\hat{\theta})-\theta_0\Big]^2$$

oziroma

$$MSE_\theta = var(\hat{\theta}) + pristranost^2$$

Metoda delta - metoda za aproksimativni izračun variance cenilke

Naj bo $\widehat{\theta}_n$ zaporedje cenilk za θ, za katerega velja

$$\sqrt{n}[\widehat{\theta}_n - \theta] \to N(0,\sigma^2)$$

v porazdelitvi. Potem za poljubno funkcijo g in dano vrednost θ (predpostavimo, da $g'(\theta)$ obstaja in da ni enako 0) velja

$$\sqrt{n}[g(\widehat{\theta}_n) - g(\theta)] \to N(0,[g'(\theta)]^2\sigma^2)$$

v porazdelitvi.

Varianco cenilke $g(\widehat{\theta})$ lahko torej ocenimo z $[g'(\theta)]^2\sigma^2$, kjer je σ^2 varianca $\widehat{\theta}$, torej $var(\widehat{\theta})$, ki je funkcija parametra θ.

Naloge

1. Naj bo diskretna spremenljivka X porazdeljena po spodaj (v tabeli) zapisani porazdelitvi. Po metodi momentov poiščite cenilki za a in b. Ocenite ju za naslednji vzorec:

$$-1, 1, -1, 1, 2, 2, 0, 0, -1, 0, 1, 1, 1$$

X	-1	0	1	2
P(X)	1-2a-b	a	b	a

2. Ocenite parametra enakomerne porazdelitve po metodi momentov. V pomoč naj bo, da je $E(X) = \frac{a+b}{2}$ in $var(X) = \frac{(b-a)^2}{12}$. **Primer:** Poslušalci radijskega programa so uganjevali številko, ki si jo je zamislil radijski voditelj. Na začetku jim je povedal meje, med katerima leži prava številka, vendar pa smo se v poslušanje vključili prepozno. Zabeležili smo le, katere številke so poslušalci predlagali: 6.39, 4.33, -0.03, 3.79 in 1.98.

   a. Predlagajte obe meji z metodo momentov.
   b. Kateri meji boste predlagali, če veste, da so bile meje cela števila?

3. Imate vzorec treh opazovanj ($x_1 = 0.4; x_2 = 0.7; x_3 = 0.9$) iz zvezne porazdelitve z gostoto

$$f(x) = \theta x^{\theta-1}; \quad 0 < x < 1.$$

   a. Ocenite θ po metodi momentov.
   b. Ali je cenilka po metodi momentov nepristranska? Pokažite.
   c. (* gl. drugi semester predmeta) Kako bi generirali vzorec iz te porazdelitve (za $\theta = 10$)? Izvedite postopek in narišite histogram. Na histogram narišite še gostoto porazdelitve.

4. Vsak gen ima dva alela, možne so 3 kombinacije: AA, Aa, aa. Kadar so v ravnovesju (vzorčimo naključno iz neke populacije), so verjetnosti kombinacij enake: θ^2, $2\theta(1-\theta)$, $(1-\theta)^2$. Denimo, da imamo podatke za nek vzorec velikosti n:

   a. Parameter θ želimo oceniti po metodi momentov. V ta namen definiramo spremenljivko X:
   $$X_i = \begin{cases} -1 & \text{,če je } i \quad AA \\ 0 & \text{,če je } i \quad aA \\ 1 & \text{,če je } i \quad aa \end{cases}$$
   Zapišite pričakovano vrednost te slučajne spremenljivke.
   b. Zapišite cenilko po metodi momentov
   c. Ali je ta cenilka nepristranska?
   d. Izračunajte varianco slučajne spremenljivke X
   e. Zapišite varianco cenilke izpeljane po metodi momentov
   f. Izrazite oceno na vzorcu z n_1, n_2 in n_3.
   g. Kako bi s simulacijami primerjali varianco cenilke po metodi momentov in varianco cenilke $\theta = \sqrt{n_1/n}$?

	AA	Aa	aa
frekvenca	n_1	n_2	n_3

5. Imate podatke za 15 enot iz porazdelitve $Beta(\alpha, \beta)$. Zanjo veste (gl. Wikipedia), da velja:

$$E(X) = \frac{\alpha}{\alpha+\beta} \qquad var(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

   a. Izračunajte cenilki za oba parametra po metodi momentov.

b. Preverite s simulacijami za $\alpha = 2, \beta = 5$, da cenilki ocenjujeta prava parametra (simulirajte 95% intervala zaupanja za obe cenilki).
c. Komentirajte širino intervala zaupanja.
d. Kako se boste s simulacijami bolje prepričali, da sta vaši cenilki zares izračunani pravilno? Kateri teoretični rezultat boste pri tem uporabili?

Bodite pozorni na pravo parametrizacijo, ko boste simulirali podatke iz porazdelitve v R! Na Wikipedii je gostota zapisana kot:

$$f(x, \alpha, \beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathrm{B}(\alpha, \beta)},$$

kjer je

$$\mathrm{B}(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

6. Za vzorec n i.i.d. spremenljivk, ki so porazdeljene po $N(\mu, \sigma^2)$, izračunajte cenilki za μ in za σ po metodi največjega verjetja.

   a. Komentirajte cenilki, ki ste ju dobili? Kako je z nepristranostjo, doslednostjo?

   b. Ocenite še standardno napako za cenilki in podajte oba IZ.

      i. Standardno napako dobite s pomočjo Fisherjeve matrike informacije.
      ii. Intervala zaupanja za μ in σ^2 dobite glede na teorijo (brez pomoči MLE).

   c. Kaj se zgodi s standardno napako, če ocenjujemo po metodi največjega verjetja varianco (σ^2) namesto standardnega odklona (σ)?

   d. Preverite z 1000 simulacijami vzorca $n = 100$, kakšna je standardna napaka ocene.

   e. Ocenite $\hat{\mu}$ in $\hat{\sigma}^2$ na vzorcu `set.seed(1); x = rnorm(100)`. Izračunajte 95% intervala zaupanja

      i. s pomočjo formul iz MLE
      ii. glede na teoretično izpeljavo (gl. nalogo b.ii.)

7. Z linearnim modelom bi radi ocenjevali, kako je prihodek podjetja odvisen od števila zaposlenih. Predpostavimo torej, da je prihodek podjetja normalno porazdeljen s povprečjem, ki je linarna funkcija logaritma števila zaposlenih.

   a. S pomočjo metode največjega verjetja dobite cenilki za regresijska koeficienta β_0 in β_1.
   b. Izpeljite tudi pripadajoče standardne napake.

8. Imamo 2 standardni normalni spremenljivki (Z_1 in Z_2). Zanima nas porazdelitev $\sqrt{Z_1 + Z_2}$, ki je znana kot Reyleighova porazdelitev. Uporablja se za modeliranje višine valov v oceanografiji in v komunikacijah za opis moči sprejetih radijskih valov. Gostota porazdelitvene funkcije je:

$$f(x, \theta) = \frac{x}{\theta^2} \cdot e^{-\frac{x^2}{2\theta^2}}, \qquad x > 0, \quad \theta > 0$$

   a. Izpeljite oceno MLE za parameter θ.
   b. Zapišite asimptotični 95% interval zaupanja za θ. Kot znano uporabite, da je

$$\int_0^\infty u e^{-au} du = \frac{1}{a^2}.$$

   c. Zapišite, kako bi s simulacijami preverili/ugotovili, ali je cenilka nepristranska. (Preverite s simulacijami, ali je cenilka nepristranska.)
   d. Pokažite teoretično, ali je vaša cenilka nepristranska. Ali znate najti nepristransko cenilko parametra θ^2?
   e. Izračunajte asimptotični interval zaupanja 95% interval zaupanja za θ za spodaj generirane podatke v R.

f. S simulacijami grafično predstavite vrednosti *verjetja* in *logaritmiranega verjetja* za Rayleighovo porazdelitev na vzorcu $n = 100$. Je logaritmirano verjetje lahko pozitivno?

```
set.seed(42)
#install.packages("VGAM")
library(VGAM)
n=100
vzorec=rrayleigh(n,3)
```

9. Računalnik nam generira n vrednosti $k = 1, 2, 3, 4$ za slučajno spremenljivko X. Predpostavimo, da so vrednosti med seboj neodvisne. Naj bodo $p_k = P(X = k)$ za $k = 1, 2, 3, 4$. Z metodo MLE določite cenilke za te verjetnosti.

10. *Potenčno normalna* porazdelitev ima gostoto

$$f(x, p) = p\phi(x)\Phi(-x)^{p-1}\,,$$

kjer je p pozitiven parameter, $\phi(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$ je gostota standardne normalne porazdelitve in $\Phi(x)$ porazdelitvena funkcija standardne normalne porazdelitve. Predpostavite, da so opazovane vrednosti neodvisne enako porazdeljene slučajne spremenljivke $X_1, \ldots, X_n$ s potenčno normalno porazdelitvijo.

   a. Privzemite, da so dane opazovane vrednosti $x_1, x_2, \ldots, x_n$. Poiščite oceno parametra p po metodi največjega verjetja.
   b. Zapišite interval zaupanja za parameter p s stopnjo tveganja α na podlagi opazovanih vrednosti $x_1, x_2, \ldots, x_n$.

11. Privzemite, da so vaši podatki med sabo neodvisne, enako porazdeljene slučajne spremenljivke $X_1, \ldots, X_n$ z gostoto

$$f(x, p) = pe^{-x} + 2(1 - p)\,e^{-2x}$$

za $x \geq 0$.

   a. Zapišite enačbo, ki ji mora ustrezati ocena parametra p po metodi največjega verjetja na osnovi podatkov $x_1, \ldots, x_n$.
   b. Izrazite aproksimativni interval zaupanja pri stopnji tveganja α za oceno parametra p po metodi največjega verjetja z oceno $\hat{p}$. Kot znano upoštevajte, da velja

$$\int_0^\infty \frac{(e^{-x} - 2e^{-2x})^2\,\mathrm{d}x}{pe^{-x} + 2(1-p)e^{-2x}} = \frac{1}{2(1-p)^3}\log\left(\frac{2-p}{p}\right) - \frac{1}{(1-p)^2}\,.$$

12. Na predavanjih (oz. v nalogi na začetku vaj iz ocenjevanja parametrov) ste si ogledali ravnotežje Hardy-Weinberg. Vsak gen ima dva alela, možne so 3 kombinacije: AA, Aa, aa. Kadar so v ravnovesju (vzorčimo naključno iz neke populacije), so verjetnosti kombinacij enake: θ^2, $2\theta(1-\theta)$, $(1-\theta)^2$.

   a. Generirajte vzorec iz populacije s $\theta = 0.3$. θ na vzorcu velikosti 100 ocenite z
      i. intuitivno metodo: z uporabo deleža genov AA
      ii. metodo največjega verjetja
   b. Simulirajte oceni s prejšnje točke in si oglejte nepristranskost. Ali sta obe cenilki nepristranski?
   c. Kakšni sta standardni napaki obeh cenilk? (ocenite ju iz simulacij)
   d. Simulirajte varianco cenilke MLE in jo primerjajte z izračunano.

13. Naj bo X spremenljivka, ki je porazdeljena po enakomerni porazdelitvi med 0 in a.

   a. Izračunajte cenilko za a po metodi momentov. Ali je nepristranska?
   b. Kako bi izračunali cenilko po metodi največjega verjetja? Intuitivno - kje ima pri tako porazdeljeni spremenljivki funkcija verjetja ekstrem; kje ima funkcija verjetja največjo vrednost?
   c. Kumulativna porazdelitvena funkcija slučajne spremenljivke $X_{(n)} = max(X_1, \ldots, X_n)$, če je porazdeljena $X \sim U(0, a)$ je

$$F_{X_{(n)}} = P(X_{(n)} < x) = \frac{x^n}{a^n}.$$

   Na tej podlagi izračunajte pristranskost cenilke $\hat{a}_{MLE}$ in povejte, ali precenjuje/podcenjuje a.

d. Za nepristranski različici cenilk $\hat{a}_{MM}$ in $\hat{a}_{MLE}$ in pristransko $\hat{a}_{MLE}$ izračunajte srednjo kvadratno napako. Povejte, katera izmed cenilk je **najučinkovitejša** s stališča MSE.
e. V R prikažite na istem grafu vrendosti MSE v odvisnosti od velikosti vzorca.

14. Vrnimo se nazaj na primer normalne porazdelitve, torej imamo vzorec n i.i.d. spremenljivk, ki so porazdeljene po $N(\mu, \sigma^2)$. Namesto izračuna cenilke za parameter σ po metodi MLE, bi radi izračunali cenilko za σ^2.

    a. Izračunajte cenilko po metodi največjega verjetja za parameter σ^2.
    b. Cenilka za σ je
    $$\hat{\sigma} = \sqrt{\frac{1}{n}\sum(x_i - \bar{x})^2}.$$
    Njena varianca pa
    $$var(\hat{\sigma}) = \frac{\sigma^2}{2n}.$$
    Katero funkcijo uporabite pri pretvorbi cenilke σ v σ^2?
    c. S pomočjo funkcije iz prejšnje točke s pomočjo metode delta izračunajte varianco cenilke za σ^2.
    d. Preverite izračunano varianco še s tem, da jo izračunate iz Fisherjeve matrike informacije.

15. Obete za nek dogodek izračunamo kot
$$O = \frac{\pi}{1-\pi},$$
kjer je π verjetnost za dogodek. Kakšna je varianca ocenjenih obetov? Za izračun uporabite metodo delta.

16. Eksponentna porazdelitev ima naslednjo gostoto porazdelitve:
$$f(x|\theta) = \frac{1}{\theta}e^{-\frac{x}{\theta}}.$$

    a. Izračunajte cenilko MLE in njeno varianco za parameter θ.
    b. Reparametrizirajmo porazdelitev tako, da bo gostota porazdelitve zapisana kot
    $$f(x|\tau) = \tau e^{-x\tau}.$$
    Na podlagi funkcije, s katero smo reparametrizirali parameter θ, zapišite
        i. Cenilko po metodi največjega verjetja za parameter τ.
        ii. Varianco te cenilke.
        iii. Preverite izračunano varianco še s tem, da jo izračunate na podlagi metode največjega verjetja.

17. Oglejte si ponovno ravnotežje Hardy-Weinberg. Za cenilko vzamemo kar koren iz deleža genov AA. Izračunajte varianco take cenilke po metodi delta.

    a. Kaj veste o porazdelitvi deleža genov AA? Kakšna je njegova varianca?
    b. S simulacijami pokažite, da pri dovolj velikih vzorcih ($n \geq 1000$) metoda delta daje zanesljive približke pravi varianci (pomagate si lahko s simulacijami iz prejšnjih nalog o ravnotežju Hardy-Weinberg).

18. Zanimajo nas povprečja in variance hemoglobina različnih športnikov. Primerjati želimo meritve k športnikov, naj bodo vrednosti i-tega športnika ($i = 1, \ldots, k$) porazdeljene normalno, torej $X_{ij} \sim N(\mu_i, \sigma_i^2)$, kjer $j = 1, \ldots, n_i$ označujejo meritve pri posamezniku (n_i so lahko različni!). Predpostavimo, da so vse meritve med seboj neodvisne.

    a. Zapišite funkcijo verjetja iz katere bi ocenili povprečje in varianco za enega športnika.
    b. Zapišite funkcijo verjetja iz katere bi ocenili povprečja in variance za k športnikov. Komentirajte, v čem je razlika s prejšnjo funkcijo.
    c. Izračunajte in zapišite cenilke za povprečja vseh športnikov. Komentirajte korake, kjer je potrebno.
    d. Izračunajte in zapišite cenilke za variance vseh športnikov. Komentirajte korake, kjer je potrebno.

e. Kakšna je ocena variance po metodi največjega verjetja, če predpostavimo, da je pri vseh športnikih enaka?
f. Z besedami razložite, zakaj je dobljeni rezultat smiseln.
g. Simulirajte 10 športnikov s po 5 meritvami hemoglobina. Naj bodo povprečja porazdeljena po $N(148, 50)$, meritve znotraj vsakega posameznika pa po normalni porazdelitvi s standardnim odklonom 5. Ocenite povprečja in standardne odklone na oba načina.
h. Simulirajte 10 športnikov s po 5 meritvami hemoglobina. Naj bodo povprečja porazdeljena po $N(148, 50)$, meritve znotraj vsakega posameznika pa po normalni porazdelitvi, kjer standardni odklon za vsakega posameznika izračunate po lognormalni porazdelitvi s parametroma $\mu = 1,5, \sigma = 0,4$. Ocenite povprečja in standardne odklone spet na oba načina in komentirajte razlike.