Tilastollinen päättely II, syksy 2015 – kevät 2016
Harjoitus 4 (24. ja 26. 11. 2015) Esimerkkiratkaisut

Tehtävät 1–3 liittyvät su-estimointiin ja informaation käsitteeseen. Tehtävässä 1 tarvitaan myös säännöllisen mallin erityispiirteitä. Luentomonisteen jaksot 2.2–2.5.

1. (Monisteen teht. 2.14.) Olkoon $f_{\mathbf{Y}}(\mathbf{y};\theta)$ tilastollinen malli, jonka parametri θ on yksiulotteinen. Olkoon $\phi = \phi(\theta)$ kääntäen yksikäsitteinen parametrimuunnos, jonka käänteismuunnos on $\theta = \theta(\phi)$. Tarkastellaan uudelleenparametroitua mallia $f^*_{\mathbf{Y}}(\mathbf{y};\phi) = f_{\mathbf{Y}}(\mathbf{y};\theta(\phi))$. Näytä, että sen havaittu informaatio ja Fisherin informaatio saadaan alkuperäisen mallin informaatioista kaavoilla

$$j^*(\hat{\phi};\mathbf{y}) = j(\hat{\theta};\mathbf{y})\,\theta'(\hat{\phi})^2, \qquad i^*(\phi) = i(\theta(\phi))\,\theta'(\phi)^2.$$

Oletetaan, että malli täyttää kaikki tarpeelliset säännöllisyysehdot ja että parametrimuunnos on riittävän monta kertaa derivoituva. [*Apu.* $l'(\hat{\theta};\mathbf{y}) = 0$ ja $E[l'(\theta;\mathbf{Y})] = 0$.]

Ratkaisu Selvitetään ensin uudelleenparametroidun mallin log-uskottavuusfunktion toinen derivaatta. Uudelleenparametroitu log-uskottavuusfunktio on

$$\ell^*(\phi;\mathbf{y}) = \ell(\theta(\phi);\mathbf{y}),$$

joten sen ensimmäisen kertaluvun derivaataksi saadaan

$$(\ell^*)'(\phi;\mathbf{y}) = \ell'(\theta(\phi);\mathbf{y})\theta'(\phi).$$

Toinen derivaatta saadaan ensimmäisestä tulon derivaatan laskusäännön avulla:

$$\begin{aligned}(\ell^*)''(\phi;\mathbf{y}) &= \frac{d}{d\phi}\left(\ell'(\theta(\phi);\mathbf{y})\theta'(\phi)\right)\\ &= \ell''(\theta(\phi);\mathbf{y})\theta'(\phi)^2 + \ell'(\theta(\phi);\mathbf{y})\theta''(\phi).\end{aligned}$$

SU-estimaattorin invarianssiominaisuuden nojalla $\hat{\theta} = \theta(\hat{\phi})$, joten pisteessä $\phi = \hat{\phi}$ pätee

$$\begin{aligned}j^*(\hat{\phi};\mathbf{y}) &= -\ell''(\theta(\hat{\phi});\mathbf{y})\theta'(\hat{\phi})^2 - \ell'(\theta(\hat{\phi});\mathbf{y})\theta''(\hat{\phi})\\ &= -\ell''(\theta(\hat{\phi});\mathbf{y})\theta'(\hat{\phi})^2\\ &= j(\hat{\theta};\mathbf{y})\theta'(\hat{\phi})^2,\end{aligned}$$

missä toinen yhtäsuuruus seuraa vihjeenä annetusta tiedosta $\ell'(\theta(\hat{\phi});\mathbf{y}) = 0$.

Mallin säännöllisyydestä seuraa, että $E[l'(\theta;\mathbf{Y})] = 0$ (monisteessa kohta 2.5.3) ja tämän avulla saadaan johdettua Fisherin informaatio kysyttyyn muotoon, sillä

$$\begin{aligned}i^*(\phi) &= E\left(j^*(\phi;\mathbf{Y})\right)\\ &= E\left(-\ell''(\theta(\phi);\mathbf{Y})\theta'(\phi)^2 - \ell'(\theta(\phi);\mathbf{Y})\theta''(\phi)\right)\\ &= E\left(-\ell''(\theta(\phi);\mathbf{Y})\right)\theta'(\phi)^2 - E\left(\ell'(\theta(\phi);\mathbf{Y})\right)\theta''(\phi)\\ &= i(\theta(\phi))\theta'(\phi)^2.\end{aligned}$$

2. "Sensuroidut" eksponenttihavainnot. Kurssilla on esitelty kaksi vaihtoehtoista tapaa tutkia elinaikojen jakaumaa: kaikkien otosyksiköiden (ihmisten, sähkölaitteiden, radioaktiivisten atomien, jne.) elinaikojen mittaus (ks. monisteen esim. 1.2.2) tai tietyllä

ajanhetkellä tapahtuva elävien ja kuolleiden lukumäärien laskeminen, joka voitiin tulkita toistokokeena (ks. harjoituksen 1 tehtävä 3).

Käytännössä tavallisempi koejärjestely on seuraava: Ryhdytään seuraamaan n otosyksikköä ajanhetkellä 0 ja lopetetaan seuranta ennalta päätetyllä ajanhetkellä a. Niiden otosyksiköiden osalta, jotka kuolivat aikavälillä $(0, a]$, saadaan selville tarkka elinaika. Lisäksi havaitaan hetkellä a yhä elossa olevien yksiköiden lukumäärä; näiden yksiköiden elinajat ovat siis $> a$ mutta niitä ei mitata tarkasti. Tällaista menettelyä kutsutaan "sensuroinniksi" (*censoring*).

Oletetaan, että tutkittavien otosyksiköiden elinajat satunnaismuuttujina ovat riippumattomia ja Exp(λ)-jakautuneita, ts. $Y_1, \ldots, Y_n \sim \text{Exp}(\lambda)$ $\perp\!\!\!\perp$. Oletetaan, että havaituista elinajoista k kpl on välillä $(0, a]$; olkoot ne $y_1 \ldots, y_k$. Ajanhetkellä a on vielä elossa $n - k$ otosyksikköä. Koska $P(Y_i > a) = e^{-\lambda a}$, voidaan perustella, että aineistoa vastaava uskottavuusfunktio on

$$L(\lambda) = \lambda e^{-\lambda y_1} \cdots \lambda e^{-\lambda y_k} \cdot (e^{-\lambda a})^{n-k} = \lambda^k e^{-\lambda s},$$

jossa $s = \sum_{i=1}^{k} y_i + (n - k)a$.

a) Muodosta log-uskottavuusfunktio ja johda kaava λ:n su-estimaatille $\hat{\lambda}$ sekä keskimääräisen (eli odotettavissa olevan) elinajan $\mu = 1/\lambda$ su-estimaatille $\hat{\mu}$. Totea, että tapauksessa $k = n$ ne yhtyvät eksponenttimallista saataviin estimaatteihin (harjoituksen 2 tehtävä 2).

b) Oletetaan, että $n = 10$ ja otosyksiköiden elinajat suuruusjärjestyksessä (päivinä) ovat

$$4 \quad 5 \quad 8 \quad 11 \quad 20 \quad 29 \quad 35 \quad 40 \quad 66 \quad 70.$$

Mikä on $\hat{\mu}$ eksponenttimalliin perustuen (kaikkien elinaikojen tarkka mittaus)? Entä jos päätettiin suorittaa "sensurointi" ajanhetkellä $a = 50$ (päivää)? Entä jos $a = 25$?

Ratkaisu

a) Tehtävänannossa on annettu uskottavuusfunktio, josta logaritmoimalla saadaan

$$\ell(\lambda; \boldsymbol{y}) = k \log(\lambda) - \lambda s.$$

Derivoimalla log-uskottavuusfunktiota saadaan

$$\ell'(\lambda; \boldsymbol{y}) = \frac{k}{\lambda} - s = 0 \Leftrightarrow \frac{k}{\lambda} = s \Leftrightarrow \lambda = \frac{k}{s}.$$

Lisäksi $\ell''(\lambda; \boldsymbol{y}) = -k/\lambda^2 < 0$, joten $\hat{\lambda} = k/s$ ja invarianssiominaisuuden nojalla

$$\hat{\mu} = s/k = \frac{\sum_{i=1}^{k} y_i + (n - k)a}{k}. \tag{1}$$

Sijoittamalla $k = n$ yhtälöön (1) saadaan

$$\hat{\mu} = \frac{s}{n} = \frac{\sum_{i=1}^{n} y_i + (n - n)a}{n} = n^{-1} \sum_{i=1}^{n} y_i = \bar{y}, \tag{2}$$

joka on sama kuin eksponenttimallin tapauksessa. Jälleen invarianssiominaisuuden perusteella vastaava pätee myös estimaattorille $\hat{\lambda}$.

b) Havaintojen summa on

$$\sum_{i=1}^{10} y_i = 288,$$

joten sijoittamalla tulos yhtälöön (2) saadaan eksponenttimallin tapauksessa tulokseksi $\hat{\mu} = 28{,}8$.

Tapauksessa $a = 50$ saadaan $k = 8$ ja sijoittamalla arvot yhtälöön (1)

$$\hat{\mu} = \frac{\sum_{i=1}^{8} y_i + 100}{8} = \frac{4+5+8+11+20+29+35+40+100}{8} = \frac{252}{8} = 31{,}5.$$

Vastaavasti tapauksessa $a = 25$ saadaan $k = 5$ ja

$$\hat{\mu} = \frac{\sum_{i=1}^{5} y_i + 125}{5} = \frac{4+5+8+11+20+125}{5} = \frac{173}{5} = 34{,}6.$$

3. Jatkoa edellisen tehtävän a-kohtaan. Laske Fisherin informaatio $i_a(\lambda)$ "sensuroitujen" eksponenttihavaintojen mallissa (a = sensurointihetki eli seuranta-ajan pituus). Huomaa, että k on tulkittava erään satunnaismuuttujan K toteutuneeksi arvoksi: se on niiden otosyksiköiden lukumäärä, jotka ovat kuolleet viimeistään ajanhetkellä a, joten se noudattaa erästä binomijakaumaa (mitä?).

Totea, että $i_a(\lambda) \to n/\lambda^2$, kun $a \to \infty$. Tässä n/λ^2 on eksponenttimallin Fisherin informaatio (harjoituksen 3 tehtävä 4). Kuinka suureksi a olisi pyrittävä valitsemaan (λ:sta riippuen), jotta sensuroidun mallin informaatio $i_a(\lambda)$ olisi ainakin puolet eksponenttimallin informaatiosta?

Ratkaisu

Eksponenttijakauman kertymäfunktiosta saadaan $\pi = P(Y_i \leq a) = 1 - e^{-\lambda a}$. Tätä voidaan pitää jossakin mielessä onnistumisena, jolloin $K \sim \text{Bin}(n, \pi)$. Edellisessä tehtävässä todettiin, että $\ell''(\lambda, \boldsymbol{y}) = -k/\lambda^2$, josta saadaan

$$i_a(\lambda) = \text{E}(-\ell''(\lambda)) = \text{E}(K)\lambda^{-2} = n(1 - e^{-\lambda a})\lambda^{-2}.$$

Koska $e^{-\lambda a} \to 0$, kun $a \to \infty$, pätee $n(1 - e^{-\lambda a})\lambda^{-2} \to n(1-0)\lambda^{-2} = n\lambda^{-2}$, kun $a \to \infty$.

Sensuroidun mallin Fisherin informaatio on vähintään puolet eksponenttimallin Fisherin informaatiosta silloin, kun $1 - e^{-\lambda a} \geq 1/2$. Logaritmifunktio on aidosti monotoninen, joten tämä on yhtäpitävästi

$$\begin{aligned} 1 - e^{-\lambda a} &\geq 1/2 \Leftrightarrow \\ e^{-\lambda a} &\leq 1/2 \Leftrightarrow \\ -\lambda a &\leq \log(1/2) \Leftrightarrow \\ a &\geq \frac{\log(2)}{\lambda} \approx \frac{0{,}69}{\lambda}. \end{aligned}$$

4. (Monisteen teht. 3.3.) Oletetaan, että havainnot $Y_1, \ldots, Y_n$ ovat riippumattomia ja noudattavat jakaumaa, jolla on odotusarvo μ ja varianssi σ^2; esimerkkinä $N(\mu, \sigma^2)$. Tarkastellaan parametrin σ^2 estimointia muotoa cV olevilla estimaattoreilla, kun $c > 0$ on vakio ja

$$V = \sum_{i=1}^{n} (Y_i - \overline{Y})^2.$$

Näytä, että cV on harhaton jos ja vain jos $c = 1/(n-1)$. [*Ehdotus.* Käytä harjoituksen 2 tehtävän 1 hajotelmaa valinnalla $a = \mu$.]

Ratkaisu Lähdetään liikkeelle tehtävänannon ehdotuksesta ja hyödynnetään harjoituksen 2 tehtävän 1 hajotelmaa sijoittamalla $a = \mu$:

$$\sum_{i=1}^{n}(Y_i-\mu)^2=\sum_{i=1}^{n}(Y_i-\overline{Y})^2+n(\overline{Y}-\mu)^2 \qquad \Leftrightarrow$$
$$\sum_{i=1}^{n}(Y_i-\overline{Y})^2=\sum_{i=1}^{n}(Y_i-\mu)^2-n(\overline{Y}-\mu)^2.$$

Satunnaismuuttujien Y_i, $i=1,2,\ldots,n$ riippumattomuuden ja odotusarvon lineaarisuuden nojalla

$$E(V)=E\left(\sum_{i=1}^{n}(Y_i-\overline{Y})^2\right)=\sum_{i=1}^{n}E\left[(Y_i-\mu)^2\right]-nE\left[(\overline{Y}-\mu)^2\right].$$

Tutkitaan oikealla puolella olevia termejä yksitellen. Varianssin määritelmän nojalla:

$$E\left[(Y_i-\mu)^2\right]=\mathrm{Var}Y_i=\sigma^2$$

ja

$$E\left[(\overline{Y}-\mu)^2\right]=\mathrm{Var}\overline{Y}=\frac{\sigma^2}{n}.$$

Siten

$$E\left(\sum_{i=1}^{n}(Y_i-\overline{Y})^2\right)=n\sigma^2-\sigma^2=(n-1)\sigma^2.$$

Siis $E(cV)=\sigma^2 \quad \Leftrightarrow \quad c=1/(n-1)$, jolloin $cV=S^2$.