

Tilastollinen päättely, syksy 2013 - kevät 2014

Harjoitus 5

Esimerkkiratkaisut

1. Tarkastellaan Poisson-mallia $Y_1, \dots, Y_n \sim P(\mu)$ $\perp\!\!\!\perp$. Varmista, että su-estimaattori $\hat{\mu} = \bar{Y} = (Y_1 + \dots + Y_n)/n$ on harhaton, ja laske sen varianssi. Onko $\hat{\mu}$ täystehokas?

Ratk. Lasketaan su-estimaattorin $\hat{\mu}$ odotusarvo:

$$E(\hat{\mu}) = E\left[\frac{1}{n} \sum_{i=1}^n Y_i\right] = \frac{1}{n} \sum_{i=1}^n E(Y_i) = \frac{1}{n} \sum_{i=1}^n \mu = \mu \quad \forall \mu > 0,$$

joten $\hat{\mu}$ on parametrin μ harhaton estimaattori.

Estimaattorin varianssiksi saadaan

$$\text{var}(\hat{\mu}) = \text{var}\left[\frac{1}{n} \sum_{i=1}^n Y_i\right] \stackrel{\perp\!\!\!\perp}{=} \frac{1}{n^2} \sum_{i=1}^n \text{var}(Y_i) = \frac{1}{n^2} \sum_{i=1}^n \mu = \frac{\mu}{n}.$$

Harhaton parametrin μ estimaattori $\hat{\mu}$ on täystehokas, jos sen varianssi yhtyy informaatioepäyhtälön antamaan alarajaan $1/i(\mu)$. Lasketaan seuraavaksi Fisherin informaatio $i(\mu)$.

Poisson-mallin ytf on

$$f_{\mathbf{Y}}(\mathbf{y}; \mu) = \prod_{i=1}^n \frac{\mu^{y_i}}{y_i!} e^{-\mu} = \frac{\mu^{\sum_{i=1}^n y_i}}{\prod_{i=1}^n y_i!} e^{-n\mu} = \frac{\mu^{n\bar{y}}}{\prod_{i=1}^n y_i!} e^{-n\mu}.$$

Asettamalla $c(\mathbf{y}) = \prod_{i=1}^n y_i!$, saadaan uskottavuusfunktioiksi

$$L(\mu; \mathbf{y}) = c(\mathbf{y}) f_{\mathbf{Y}}(\mathbf{y}; \mu) = \mu^{n\bar{y}} e^{-n\mu}$$

ja log-uskottavuusfunktioiksi

$$l(\mu; \mathbf{y}) = \log L(\mu; \mathbf{y}) = n\bar{y} \log(\mu) - n\mu.$$

Log-uskottavuusfunktion toinen derivaatta on

$$l''(\mu; \mathbf{y}) = \frac{d^2}{d\mu^2} l(\mu; \mathbf{y}) = \frac{d}{d\mu} \left(\frac{n\bar{y}}{\mu} - n \right) = -\frac{n\bar{y}}{\mu^2},$$

joten Fisherin informaatioksi saadaan

$$i(\mu) = E[-l''(\mu; \mathbf{Y})] = \frac{nE[\bar{Y}]}{\mu^2} = \frac{n\mu}{\mu^2} = \frac{n}{\mu}.$$

Informaatioepäyhtälön antama alaraja on siis $1/i(\mu) = \mu/n = \text{var}(\hat{\mu})$, joten $\hat{\mu}$ on täystehokas.

2. Tarkastellaan ns. Poisson-regressiomallia: $Y_1, \dots, Y_n \perp\!\!\!\perp$ ja $Y_i \sim P(\lambda x_i)$, jossa x_1, \dots, x_n ovat tunnettuja positiivisia lukuja (selittävän muuttujan arvoja). Konkreettisenä esimerkkinä voidaan ajatella, että Y_i on johonkin tautiin kuolneiden lukumäärä populaatiossa, jonka koko on x_i .

a) Muodosta tämän mallin log-uskottavuusfunktio ja johda parametrin λ suurimman uskottavuuden estimaattorille lauseke

$$\hat{\lambda} = \frac{\sum_{i=1}^n Y_i}{\sum_{i=1}^n x_i}.$$

b) Näytä, että $\hat{\lambda}$ on harhaton.

c) Laske estimaattorin $\hat{\lambda}$ varianssi ja osoita, että se yhtyy informaatioepäyhtälön antamaan alarajaan.

Ratk.

a) Poisson-regressioon liittyvä tilastollinen malli on

$$f_{\mathbf{Y}}(\mathbf{y}; \lambda) = \prod_{i=1}^n e^{-\lambda x_i} \frac{(\lambda x_i)^{y_i}}{y_i!} = e^{-n\lambda \bar{x}} \lambda^{n\bar{y}} \frac{\prod_{i=1}^n x_i^{y_i}}{\prod_{i=1}^n y_i!}.$$

Valitsemalla

$$c(\mathbf{y}) = \frac{\prod_{i=1}^n y_i!}{\prod_{i=1}^n x_i^{y_i}},$$

saadaan uskottavuusfunktioksi

$$L(\lambda; \mathbf{y}) = c(\mathbf{y}) f_{\mathbf{Y}}(\mathbf{y}; \lambda) = e^{-n\lambda \bar{x}} \lambda^{n\bar{y}}$$

ja logaritmiseksi uskottavuusfunktioksi

$$l(\lambda; \mathbf{y}) = -n\lambda \bar{x} + n\bar{y} \log(\lambda).$$

Derivoidaan log-uskottavuusfunktio parametrin λ suhteen:

$$l'(\lambda; \mathbf{y}) = \frac{d}{d\lambda} l(\lambda; \mathbf{y}) = \frac{n\bar{y}}{\lambda} - n\bar{x}.$$

Kun $\bar{y} > 0$, niin derivaatalla on alueessa $(0, \infty)$ ainoastaan yksi nollakohta, $\hat{\lambda} = \bar{y}/\bar{x}$. Se on myös suurimman uskottavuuden estimaatti, koska

$$l''(\lambda; \mathbf{y}) = -\frac{n\bar{y}}{\lambda^2} < 0, \quad \lambda \in (0, \infty).$$

Kun $\bar{y} = 0$, niin logaritminen uskottavuusfunktio yksinkertaistuu muotoon

$$l(\lambda; \mathbf{y}) = -n\lambda\bar{x},$$

joka maksimoituu pisteessä $\hat{\lambda} = 0 = 0/\bar{x} = \bar{y}/\bar{x}$. Siten suurimman uskottavuuden estimaattori on

$$\hat{\lambda} = \frac{\bar{Y}}{\bar{x}} = \frac{\sum_{i=1}^n Y_i}{\sum_{i=1}^n x_i}.$$

b) Odotusarvoksi saadaan

$$E(\hat{\lambda}) = \frac{E(\bar{Y})}{\bar{x}} = \frac{\frac{1}{n} \sum_{i=1}^n E(Y_i)}{\bar{x}} = \frac{\frac{1}{n} \sum_{i=1}^n \lambda x_i}{\bar{x}} = \frac{\bar{x}}{\bar{x}} \lambda = \lambda \quad \forall \lambda \geq 0,$$

joten $\hat{\lambda}$ on parametrin λ harhaton estimaattori.

c) Lasketaan suurimman uskottavuuden estimaattorin $\hat{\lambda}$ varianssi:

$$\begin{aligned} \text{var}(\hat{\lambda}) &\stackrel{\text{H}}{=} \frac{1}{(\sum_{i=1}^n x_i)^2} \sum_{i=1}^n \text{var}(Y_i) = \frac{1}{(\sum_{i=1}^n x_i)^2} \sum_{i=1}^n \lambda x_i = \frac{\sum_{i=1}^n x_i}{(\sum_{i=1}^n x_i)^2} \lambda \\ &= \frac{\lambda}{\sum_{i=1}^n x_i} = \frac{\lambda}{n\bar{x}}. \end{aligned}$$

Fisherin informaatio on

$$i(\lambda) = E[-l''(\lambda; \mathbf{Y})] = E\left[\frac{n\bar{Y}}{\lambda^2}\right] = \frac{\sum_{i=1}^n E(Y_i)}{\lambda^2} = \frac{\sum_{i=1}^n \lambda x_i}{\lambda^2} = \frac{n\bar{x}}{\lambda},$$

joten informaatioepäyhtälö antaa parametrin λ harhattoman estimaattorin varianssin alarajaksi

$$\frac{1}{i(\lambda)} = \frac{\lambda}{n\bar{x}},$$

joka on sama kuin suurimman uskottavuuden estimaattorin $\hat{\lambda}$ varianssi.

3. Erään elektronisen komponentin kestoikä noudattaa eksponenttijakaumaa, jonka odotusarvo on θ/t , jossa $t > 0$ on komponentin käyttölämpötila ja $\theta > 0$ on tuntematon parametri. Parametrin θ estimoimiseksi testataan n komponenttia toisistaan riippumattomasti lämpötiloissa t_1, \dots, t_n ja mitataan niiden kestoajat Y_1, \dots, Y_n . Osoita, että

$$T = \sum_{i=1}^n Y_i / \sum_{i=1}^n \frac{1}{t_i}$$

on θ :n harhaton mutta ei täystehokas estimaattori.

Vihje: Voit käyttää hyväksi seuraavia aputuloksia:

$$\begin{aligned} \left(\sum_{i=1}^n a_i \right)^2 &= n^2 \bar{a}^2 \\ \sum_{i=1}^n a_i^2 &= \sum_{i=1}^n (a_i - \bar{a})^2 + n\bar{a}^2 \end{aligned}$$

Ratk. Lasketaan estimaattorin T odotusarvo:

$$E(T) = E \left[\frac{\sum_{i=1}^n Y_i}{\sum_{i=1}^n t_i^{-1}} \right] = \frac{\sum_{i=1}^n E(Y_i)}{\sum_{i=1}^n t_i^{-1}} = \frac{\sum_{i=1}^n \theta/t_i}{\sum_{i=1}^n t_i^{-1}} = \theta \quad \forall \theta > 0,$$

joten T on parametrin θ harhaton estimaattori. Estimaattorin varianssiksi saadaan

$$\text{var}(T) = \text{var} \left[\frac{\sum_{i=1}^n Y_i}{\sum_{i=1}^n t_i^{-1}} \right] \stackrel{\text{||}}{=} \frac{\sum_{i=1}^n \text{var}(Y_i)}{[\sum_{i=1}^n t_i^{-1}]^2} = \frac{\sum_{i=1}^n \theta^2/t_i^2}{[\sum_{i=1}^n t_i^{-1}]^2} = \frac{\sum_{i=1}^n t_i^{-2}}{[\sum_{i=1}^n t_i^{-1}]^2} \theta^2.$$

Mallin ytf on

$$f_{\mathbf{Y}}(\mathbf{y}; \theta) = \prod_{i=1}^n (t_i/\theta) e^{-(t_i/\theta)y_i} = \frac{\prod_{i=1}^n t_i}{\theta^n} \exp \left[-\frac{1}{\theta} \sum_{i=1}^n t_i y_i \right].$$

Asettamalla $c(\mathbf{y}) = (\prod_{i=1}^n t_i)^{-1}$ saadaan uskottavuusfunktiksi

$$L(\theta; \mathbf{y}) = c(\mathbf{y}) f_{\mathbf{Y}}(\mathbf{y}; \theta) = \frac{1}{\theta^n} \exp \left[-\frac{1}{\theta} \sum_{i=1}^n t_i y_i \right]$$

ja log-uskottavuusfunktiksi

$$l(\theta; \mathbf{y}) = -n \log(\theta) - \frac{1}{\theta} \sum_{i=1}^n t_i y_i.$$

Lasketaan log-uskottavuusfunktion ensimmäinen ja toinen derivaatta:

$$l'(\theta; \mathbf{y}) = \frac{d}{d\theta} l(\theta; \mathbf{y}) = -\frac{n}{\theta} + \frac{1}{\theta^2} \sum_{i=1}^n t_i y_i.$$

$$l''(\theta; \mathbf{y}) = \frac{d^2}{d\theta^2} l(\theta; \mathbf{y}) = \frac{n}{\theta^2} - \frac{2}{\theta^3} \sum_{i=1}^n t_i y_i.$$

Fisherin informaatioksi saadaan nyt

$$\begin{aligned} i(\theta) &= \mathbb{E}[-l''(\theta; \mathbf{Y})] = -\frac{n}{\theta^2} + \frac{2}{\theta^3} \sum_{i=1}^n t_i \mathbb{E}(Y_i) = -\frac{n}{\theta^2} + \frac{2}{\theta^3} \sum_{i=1}^n t_i (\theta/t_i) \\ &= -\frac{n}{\theta^2} + \frac{2n\theta}{\theta^3} = \frac{n}{\theta^2}. \end{aligned}$$

Informaatioepäyhtälön antama alaraja on tällöin $1/i(\theta) = \theta^2/n$. Verrataan seuraavaksi alarajaa estimaattorin varianssiin:

$$\text{var}(T) - i(\theta)^{-1} = \frac{\sum_{i=1}^n t_i^{-2}}{[\sum_{i=1}^n t_i^{-1}]^2} \theta^2 - \frac{1}{n} \theta^2 = \frac{\sum_{i=1}^n t_i^{-2} - \frac{1}{n} [\sum_{i=1}^n t_i^{-1}]^2}{[\sum_{i=1}^n t_i^{-1}]^2} \theta^2.$$

Olkoon $a_i = 1/t_i$. Tällöin saadaan

$$\begin{aligned} \text{var}(T) - i(\theta)^{-1} &= \frac{\sum_{i=1}^n a_i^2 - \frac{1}{n} [\sum_{i=1}^n a_i]^2}{[\sum_{i=1}^n a_i]^2} \theta^2 = \frac{\sum_{i=1}^n a_i^2 - \frac{1}{n} [n\bar{a}]^2}{[n\bar{a}]^2} \theta^2 \\ &= \frac{\sum_{i=1}^n a_i^2 - n\bar{a}^2}{n^2\bar{a}^2} \theta^2 = \frac{\sum_{i=1}^n (a_i - \bar{a})^2}{n^2\bar{a}^2} \theta^2 > 0, \end{aligned}$$

Nähdään, että $\text{var}(T) > i(\theta)^{-1}$, joten estimaattori T ei ole täystehokas.

4. Olkoot $Y_1, \dots, Y_n \sim N(\theta, 1) \perp\!\!\!\perp$.

- a) Osoita, että $U(\mathbf{Y}) = \bar{Y}^2 - 1/n$ on funktion $g(\theta) = \theta^2$ harhaton estimaattori. Laske estimaattorin $U(\mathbf{Y})$ varianssi ja näytä, että se on suurempi kuin informaatioepäyhtälön antama alaraja.
- b) Funktion $g(\theta) = \theta^2$ suurimman uskottavuuden estimaattori $V(\mathbf{Y}) = \bar{Y}^2$ on a)-kohdan perusteella harhainen. Laske estimaattorin $V(\mathbf{Y})$ keskineliövirhe ja vertaa sitä estimaattorin $U(\mathbf{Y})$ keskineliövirheeseen.

Ratk.

a) Koska Y_1, \dots, Y_n ovat riippumattomia $N(\theta, 1)$ -jakautuneita satunnaismuuttujia, niin

$$\bar{Y} \sim N(\theta, 1/n) \text{ ja } Z = \frac{\bar{Y} - \theta}{1/\sqrt{n}} \sim N(0, 1).$$

Standardoidusta normaalijakaumasta tiedetään, että $E(Z) = 0$, $E(Z^2) = 1$, $E(Z^3) = 0$ ja $E(Z^4) = 3$. Helposti nähdään, että $\bar{Y} = \theta + Z/\sqrt{n}$. Lasketaan keskiarvon \bar{Y} toinen ja neljäs momentti.

$$\begin{aligned} E[(\bar{Y})^2] &= E[(\theta + Z/\sqrt{n})^2] = E\left[\theta^2 + \frac{2\theta}{\sqrt{n}}Z + \frac{1}{n}Z^2\right] \\ &= \theta^2 + \frac{2\theta}{\sqrt{n}} \cdot 0 + \frac{1}{n} \cdot 1 = \theta^2 + \frac{1}{n} \end{aligned}$$

$$\begin{aligned} E[(\bar{Y})^4] &= E[(\theta + Z/\sqrt{n})^4] = E\left[\theta^4 + \frac{4\theta^3}{\sqrt{n}}Z + \frac{6\theta^2}{n}Z^2 + \frac{4\theta}{n\sqrt{n}}Z^3 + \frac{1}{n^2}Z^4\right] \\ &= \theta^4 + \frac{4\theta^3}{\sqrt{n}} \cdot 0 + \frac{6\theta^2}{n} \cdot 1 + \frac{4\theta}{n\sqrt{n}} \cdot 0 + \frac{1}{n^2} \cdot 3 \\ &= \theta^4 + \frac{6\theta^2}{n} + \frac{3}{n^2} \end{aligned}$$

Nyt voidaan laskea estimaattorin $U(\mathbf{Y}) = (\bar{Y})^2 - 1/n$ odotusarvo ja varianssi.

$$E[U(\mathbf{Y})] = E[(\bar{Y})^2] - \frac{1}{n} = \theta^2 + \frac{1}{n} - \frac{1}{n} = \theta^2 \quad \forall \theta \in \mathbb{R},$$

joten $U(\mathbf{Y})$ on harhaton funktion $g(\theta) = \theta^2$ estimaattori.

$$\begin{aligned} \text{Var}[U(\mathbf{Y})] &= \text{Var}[(\bar{Y})^2] = \text{E}[(\bar{Y})^4] - (\text{E}[(\bar{Y})^2])^2 \\ &= \theta^4 + \frac{6\theta^2}{n} + \frac{3}{n^2} - \left(\theta^2 + \frac{1}{n}\right)^2 \\ &= \theta^4 + \frac{6\theta^2}{n} + \frac{3}{n^2} - \theta^4 - \frac{2\theta^2}{n} - \frac{1}{n^2} \\ &= \frac{4\theta^2}{n} + \frac{2}{n^2} \end{aligned}$$

Johdetaan seuraavaksi informaatioepäyhtälön antama alaraja. Luentomonisteen esimerkin 2.4.3 perusteella mallin Fisherin informaatio on $i(\theta) = n/1 = n$. Luentomonisteen kappaleen 3.4.3 perusteella funktion $g(\theta)$ harhattomalle estimaattorille pätee

$$\text{Var}(T) \geq \frac{g'(\theta)^2}{i(\theta)} = \frac{(2\theta)^2}{n} = \frac{4\theta^2}{n}.$$

Välittömästi nähdään, että

$$\text{Var}[T(\mathbf{Y})] = \frac{4\theta^2}{n} + \frac{2}{n^2} > \frac{4\theta^2}{n}.$$

b) Lasketaan suurimman uskottavuuden estimaattorin $V(\mathbf{Y}) = \bar{Y}^2$ keskineliövirhe. Kappaleen 3.4.1 perusteella keskineliövirheelle pätee hajotelma

$$\text{E}_\theta[(V(\mathbf{Y}) - \theta^2)^2] = \text{var}_\theta(V(\mathbf{Y})) + b(\theta)^2,$$

jossa

$$\text{var}_\theta(V(\mathbf{Y})) = \text{var}_\theta(U(\mathbf{Y}))$$

ja

$$b(\theta) = \text{E}_\theta(V(\mathbf{Y})) - \theta^2 = \text{E}_\theta(U(\mathbf{Y}) + 1/n) - \theta^2 = \theta^2 + \frac{1}{n} - \theta^2 = \frac{1}{n}.$$

Keskineliövirheeksi saadaan siis

$$\text{E}_\theta[(V(\mathbf{Y}) - \theta^2)^2] = \text{var}_\theta(U(\mathbf{Y})) + 1/n^2 = \text{E}_\theta[(U(\mathbf{Y}) - \theta^2)^2] + 1/n^2.$$

Estimaattorin $V(\mathbf{Y})$ keskineliövirhe on siis estimaattorin $U(\mathbf{Y})$ keskineliövirhettä vakion $1/n^2$ verran suurempi.

5. Jatkoa edellisen harjoituksen tehtävään 5. Luentomonisteen perusteella säännöllisen mallin $f(\mathbf{y}; \theta)$ parametrin θ harhattoman estimaattorin T varianssille pätee, että

$$\text{var}_{\theta}(T) \geq \frac{1}{i(\theta)} = \frac{1}{E[\{l'(\theta; \mathbf{Y})\}^2]}.$$

Olkoot $Y_1, \dots, Y_n \sim \text{Tas}(0, \theta) \perp\!\!\!\perp$. Tarkastellaan jälleen parametrin θ harhatonta estimaattoria $\check{\theta} = [(n+1)/n]\hat{\theta}$. Vertaa estimaattorin $\check{\theta}$ varianssia informaatioepäyhtälön antamaan alarajaan

$$\frac{1}{E[\{l'(\theta; \mathbf{Y})\}^2]}$$

ja kommentoi tulosta.

Ratk. Mallia vastaava ytf on

$$\begin{aligned} f_{\mathbf{Y}}(\mathbf{y}; \theta) &= \prod_{i=1}^n \frac{1}{\theta} 1_{(0, \theta)}(y_i) = \frac{1}{\theta^n} \prod_{i=1}^n 1_{(0, \theta)}(y_i) \\ &= \begin{cases} \frac{1}{\theta^n}, & \text{kun } y_1 \in (0, \theta), \dots, y_n \in (0, \theta) \\ 0, & \text{muulloin.} \end{cases} \end{aligned}$$

Log-uskottavuusfunktiksi saadaan

$$l(\theta; \mathbf{y}) = -n \log(\theta), \quad \text{kun } y_1 \in (0, \theta), \dots, y_n \in (0, \theta).$$

Derivoimalla saadaan

$$l'(\theta; \mathbf{y}) = -\frac{n}{\theta}, \quad \text{kun } y_1 \in (0, \theta), \dots, y_n \in (0, \theta),$$

joten

$$E[\{l'(\theta; \mathbf{Y})\}^2] = E\left[\left\{-\frac{n}{\theta}\right\}^2\right] = \frac{n^2}{\theta^2}$$

ja siis

$$\frac{1}{E[\{l'(\theta; \mathbf{Y})\}^2]} = \frac{\theta^2}{n^2}.$$

Edellisen harjoituksen tehtävän 5 perusteella estimaattorin $\check{\theta}$ varianssi on

$$\text{var}(\check{\theta}) = \frac{\theta^2}{n(n+2)} = \frac{\theta^2}{n^2 + 2n}.$$

Välittömästi nähdään, että

$$\text{var}(\check{\theta}) = \frac{\theta^2}{n^2 + 2n} < \frac{\theta^2}{n^2} = \frac{1}{E[\{l'(\theta; \mathbf{Y})\}^2]},$$

joten informaatioepäyhtälö ei toimi tässä tapauksessa. Tämä johtuu siitä, että informaatioepäyhtälö pätee ainoastaan säännöllisillä malleilla. Tasainen jakauma $Tas(0, \theta)$ ei ole säännöllinen, koska sen alusta riippuu parametrusta θ . On helppo nähdä, että myöskään integroinnin ja derivoinnin järjestyksen vaihdannaisuus ei päde tasaisella jakaumalla (säännöllisen mallin ehdot c ja d luentomonisteen kappaleessa 2.5.2).