Ime in priimek: _____ Vpisna št: | | | | | | | | |

FAKULTETA ZA MATEMATIKO IN FIZIKO

Oddelek za matematiko

Statistika

Pisni izpit

16. junij 2017

Navodila

Pazljivo preberite besedilo naloge, preden se lotite reševanja. Nalog je 6, 5 rešenih nalog pa je že 100%. Na razpolago imate 2 uri.

Naloga	a.	b.	c.	d.	
1.			•	•	
2.				•	
3.					
4.					
5.			•	•	
6.				•	
Skupaj	•	•	•	•	

1. (20) Naj bosta (X, Y, Z) in (U, V, W) slučajna vektorja. Privzemite, da so X,Y in Z neodvisne, enako porazdeljene strogo pozitivne celoštevilske omejene slučajne spremenljivke. Privzemite, da za vsako omejeno zvezno funkcijo $h\colon \mathbb{R}^3 \to \mathbb{R}$ velja

$$E\big[h(U,V,W)\big] = E\left[\frac{3X}{X+Y+Z}\, h\,(X,Y,Z)\right].$$

a. (10) Utemeljite, da za vsako omejeno funkcijo $g(Y,Z)$ velja

$$E\big[g(Y,Z)|X, Y+Z\big] = E\big[g(Y,Z)|Y+Z\big].$$

Namig: vsako omejeno funkcijo $f(x,s)$ lahko na zalogi vrednosti vektorja $(X, Y+Z)$ napišemo kot končno vsoto

$$f(x,s) = \sum_{k=1}^{n} f_{1,k}(x) f_{2,k}(s)\,,$$

kjer so $f_{1,k}$ in $f_{2,k}$ omejene funkcije.

Rešitev: Prvi način: *upoštevamo namig in preverimo po definiciji za $f(x,s) = f_1(x)f_2(s)$, pri čemer upoštevamo neodvisnost slučajnih spremenljivk X in $Y+Z$:*

$$\begin{aligned}
&E\big[E\left[g(Y,Z)|Y+Z\right] f_1(X)\, f_2(Y+Z)\big] \\
&\quad = E\left[E\big[g(Y,Z)|Y+Z\big]\, f_2(Y+Z)\right] E\left[f_1(X)\right] \\
&\quad = E\left[g(Y,Z)\, f_2(Y+Z)\right] E\left[f_1(X)\right] \\
&\quad = E\left[g(Y,Z)\, f_1(X)\, f_2(Y+Z)\right].
\end{aligned}$$

S tem je trditev dokazana.

Drugi način: *neposredno. Velja:*

$$\begin{aligned}
&E\big[g(Y,Z)|X=x, Y+Z=s\big] \\
&\quad = \frac{E\big[g(Y,Z)\,\mathbf{1}(X=x, Y+Z=s)\big]}{P(X=x, Y+Z=s)} \\
&\quad = \frac{E\big[\mathbf{1}(X=x)\big] E\big[g(Y,Z)\,\mathbf{1}(Y+Z=s)\big]}{P(X=x)\, P(Y+Z=s)} \\
&\quad = \frac{E\big[g(Y,Z)\,\mathbf{1}(Y+Z=s)\big]}{P(Y+Z=s)} \\
&\quad = E\big[g(Y,Z)|Y+Z=s\big].
\end{aligned}$$

b. (10) Naj za omejeno zvezno funkcijo g velja

$$E\left[g(Y,Z)|Y+Z\right] = \psi(Y+Z)\,.$$

Pokažite, da je tudi

$$E\left[g(V,W)|V+W\right] = \psi(V+W)\,.$$

Namig: v pravem trenutku pogojujte na $(X, Y+Z)$.

Rešitev: Računamo po definiciji za omejeno funkcijo f *in upoštevamo prvi del naloge:*

$$\begin{aligned}
E\big[g(V,W)\,f(V+W)\big] &= E\left[\frac{3X}{X+Y+Z}\,g(Y,Z)\,f(Y+Z)\right] \\
&= E\left[E\left[\frac{3X}{X+Y+Z}\,g(Y,Z)\,f(Y+Z)\;\middle|\;X,Y+Z\right]\right] \\
&= E\left[\frac{3X}{X+Y+Z}\,E\big[g(Y,Z)|X,Y+Z\big]f(Y+Z)\right] \\
&= E\left[\frac{3X}{X+Y+Z}\,E\big[g(Y,Z)|Y+Z\big]f(Y+Z)\right] \\
&= E\left[\frac{3X}{X+Y+Z}\,\psi(Y+Z)\,f(Y+Z)\right] \\
&= E\big[\psi(V+W)\,f(V+W)\big]\,.
\end{aligned}$$

Trditev je s tem dokazana.

2. (20) Pri revizijah zahtevkov za izplačila iz evropskih strukturnih skladov se uporablja vzorčenje na naslednji način: recimo, da je bilo v koledarskem letu N zahtevkov v višini $v_1, v_2, \ldots, v_N$. Označimo vsoto vseh izplačil z $v = v_1 + \cdots + v_N$. Vzorčimo tako, da izmed vsemi izplačanimi euri izberemo enostavni slučajni vzorec $w < v$ eurov, nato pa izberemo zahtevke, na podlagi katerih so bili izbrani euri izplačani. Velikosti vzorca zahtevkov tako ne moremo določiti vnaprej. Namen vzorčenja je oceniti delež na podlagi zahtevkov z nepravilnostmi izplačanih eurov.

a. (5) Označite z S slučajno število zahtevkov, ki bodo izbrani v vzorec. Pokažite, da je pričakovana velikost vzorca zahtevkov enaka

$$E(S) = \sum_{k=1}^{N} \left(1 - \frac{\binom{v-v_k}{w}}{\binom{v}{w}}\right).$$

Namig: definirajte ustrezne indikatorje. Laže je določiti verjetnost, da določen zahtevek ne bo izbran, kot da bo.

Rešitev: Naj bo I_k indikator dogodka, da bo v vzorec izbran k-ti zahtevek. Tedaj je $S = \sum_{k=1}^{N} I_k$, torej je dovolj izračunati $E(I_k) = P(I_k = 1)$.

Če želimo, da zahtevek ne bo *izbran, moramo vse eure izbrati izmed $v - v_k$ eurov, ki ne pripadajo zahtevku; takih izbir je $\binom{v-v_k}{w}$. Torej je*

$$E(I_k) = P(I_k = 1) = 1 - \frac{\binom{v-v_k}{w}}{\binom{v}{w}}.$$

Pričakovana velikost vzorca zahtevkov je $E(S) = \sum_{k=1}^{N} E(I_k)$, kar res pride tako kot zahtevano.

b. (10) Naj bo I_k indikator dogodka, da bo izbran k-ti zahtevek. Za $k \neq l$ izračunajte $\operatorname{cov}(I_k, I_l)$.

Namig:

$$P(I_k = 1, I_l = 1) = 1 - P(I_k = 0) - P(I_l = 0) + P(I_k = 0, I_l = 0).$$

Rešitev:

Prvi način. *Potrebujemo $P(I_k = 1, I_l = 1)$, to pa skladno z namigom izračunamo s pomočjo $P(I_k = 0, I_l = 0)$, kar je verjetnost dogodka, da vse eure izberemo izmed $v - v_k - v_l$ eurov, ki ne pripadajo niti k-temu niti l-temu zahtevku; takih izbir je $\binom{v-v_k-v_l}{w}$. Sledi*

$$P(I_k = 1, I_l = 1) = 1 - \frac{\binom{v-v_k}{w}}{\binom{v}{w}} - \frac{\binom{v-v_l}{w}}{\binom{v}{w}} + \frac{\binom{v-v_k-v_l}{w}}{\binom{v}{w}}.$$

in

$$\begin{aligned}\operatorname{cov}(I_k, I_l) &= P(I_k = 1, I_l = 1) - P(I_k = 1)P(I_l = 1) \\ &= \frac{\binom{v-v_k-v_l}{w}}{\binom{v}{w}} - \frac{\binom{v-v_k}{w}\binom{v-v_l}{w}}{\binom{v}{w}^2}\,.\end{aligned}$$

Drugi način. *Upoštevamo, da je*

$$\operatorname{cov}(I_k, I_l) = \operatorname{cov}(1 - I_k, 1 - I_l) = P(I_k = 0, I_l = 0) - P(I_k = 0)\,P(I_l = 0)\,.$$

Vstavimo prej izračunane verjetnosti in vidimo, da pride isto kot prej

c. (5) Izrazite varianco vsote izplačil T na podlagi zahtevkov, izbranih v vzorec.

Rešitev: Pišimo $T = \sum_{k=1}^{N} v_k I_k$. Varianco lahko izračunamo po običajni formuli za varianco linearnih kombinacij slučajnih spremenljivk, še elegantneje pa gre, če zapišemo

$$T = \sum_k v_k - \sum_k v_k(1 - I_k)$$

in izračunamo

$$\begin{aligned}\operatorname{var}(T) &= \operatorname{var}\left(\sum_k v_k(1 - I_k)\right) \\ &= E\left[\left(\sum_k v_k(1 - I_k)\right)^2\right] - \left[E\left(\sum_k v_k(1 - I_k)\right)\right]^2 \\ &= \sum_k v_k^2\,P(I_k = 0) + \sum_{k,l} v_k v_l\,P(I_k = 0, I_l = 0) - \left[\sum_k v_k\,P(I_k = 0)\right]^2 \\ &= \sum_k v_k^2\,\frac{\binom{v-v_k}{w}}{\binom{v}{w}} + \sum_{k\neq l} v_k v_l\,\frac{\binom{v-v_k-v_l}{w}}{\binom{v}{w}} - \left(\sum_k v_k\,\frac{\binom{v-v_k}{w}}{\binom{v}{w}}\right)^2\,.\end{aligned}$$

3. (20) Privzemite, da so podatki $x_1, x_2, \ldots, x_n$ nastali kot med sabo neodvisne, enako porazdeljene slučajne spremenljivke s porazdelitvijo

$$P(X_1 = x) = \binom{2x}{x} \frac{\beta^x}{4^x \, (1+\beta)^{x+\frac{1}{2}}}$$

za $x = 0, 1, \ldots$ in $\beta > 0$.

a. (5) Poiščite oceno za β po metodi največjega verjetja.

Rešitev: Logaritemska funkcija verjetja je

$$\ell\,(\beta|\mathbf{x}) = \sum_{k=1}^{n} \log \binom{2x_k}{x_k} + \log\beta \sum_{k=1}^{n} x_k - \log 4 \sum_{k=1}^{n} x_k - \log(1+\beta) \sum_{k=1}^{n} \left(x_k + \frac{1}{2}\right) .$$

Odvajamo po β in izenačimo z 0. Dobimo enačbo

$$\frac{1}{\beta} \sum_{k=1}^{n} x_k - \frac{1}{1+\beta} \sum_{k=1}^{n} \left(x_k + \frac{1}{2}\right) = 0\,.$$

Sledi

$$\hat{\beta} = \frac{2\sum_{k=1}^{n} x_k}{n}\,.$$

b. (5) Prepričajte se, da je

$$\begin{aligned} E(X_1) &= \sum_{k=0}^{\infty} k\, P(X_1 = k) \\ &= \frac{2\beta}{4(1+\beta)} \sum_{k=1}^{\infty} \big[2(k-1)+1\big] P(X_1 = k-1) \\ &= \frac{\beta}{1+\beta}\, E(X_1) + \frac{\beta}{2(1+\beta)} \end{aligned}$$

in pokažite, da je cenilka $\hat{\beta}$ po metodi največjega verjetja nepristranska.

Rešitev: Najprej opazimo, da lahko v formuli za pričakovano vrednost izpustimo vrednost 0 – velja $E(X_1) = \sum_{k=1}^{\infty} k\, P(X_1 = k)$. Nadalje izračunamo

$$\begin{aligned} k\, P(X_1 = k) &= k\, \frac{(2k)!}{(k!)^2}\, \frac{\beta^k}{4^k (1+\beta)^{k+1/2}} \\ &= k\, \frac{2k(2k-1)\,(2k-2)!}{k^2 \big[(k-1)!\big]^2}\, \frac{\beta^k}{4^k (1+\beta)^{k+1/2}} \\ &= \frac{(2k-1)\,(2k-2)!}{\big[(k-1)!\big]^2}\, \frac{2\beta}{4(1+\beta)}\, \frac{\beta^{k-1}}{4^{k-1} (1+\beta)^{k-1/2}} \\ &= \frac{2\beta}{4(1+\beta)} \big[2(k-1)+1\big] P(X_1 = k-1)\,. \end{aligned}$$

Sledi

$$\begin{aligned} E(X_1) &= \frac{2\beta}{4(1+\beta)} \sum_{k=1}^{\infty} \big[2(k-1)+1\big] P(X_1 = k-1) \\ &= \frac{\beta}{2(1+\beta)} \sum_{l=0}^{\infty} (2l+1) P(X_1 = l) \\ &= \frac{\beta}{1+\beta}\, E(X_1) + \frac{\beta}{2(1+\beta)}\,, \end{aligned}$$

od koder izračunamo

$$E(X_1) = \frac{\beta}{2}\,.$$

Sledi

$$E\big(\hat{\beta}\big) = E\left(\frac{2\sum_{k=1}^{n} X_k}{n}\right) = \beta\,,$$

torej je cenilka res nepristranska.

c. (5) Izračunajte Fisherjevo informacijo in navedite aproksimativno standardno napako za $\hat{\beta}$.

Rešitev: Za $n = 1$ *izračunamo:*

$$\ell''(\beta|k) = -\frac{k}{\beta^2} + \frac{k+\frac{1}{2}}{(1+\beta)^2}\,,$$

torej

$$E\big[-\ell''(\beta|X_1)\big] = \frac{\frac{\beta}{2}}{\beta^2} - \frac{\frac{\beta}{2}+\frac{1}{2}}{(1+\beta)^2} = \frac{1}{2\beta(\beta+1)}\,.$$

Sledi

$$\hat{\mathrm{se}}(\hat{\beta}) = \frac{\sqrt{2\beta(1+\beta)}}{\sqrt{n}}\,.$$

d. (5) Prepričajte se, da je

$$\begin{aligned} E(X_1^2) &= \sum_{k=0}^{\infty} k^2 P(X_1 = k) \\ &= \frac{\beta}{2(1+\beta)} \sum_{k=1}^{\infty} \big[2(k-1)^2 + 3(k-1) + 1\big] P(X_1 = k-1) \\ &= \frac{\beta}{2(1+\beta)} \big(2E(X_1^2) + 3E(X_1) + 1\big)\,. \end{aligned}$$

Uporabite to za izračun eksaktne standardne napake za $\hat{\beta}$.

Rešitev: V zgornjo enakost se prepričamo podobno kot pri točki b. V formuli za pričakovano vrednost spet izpustimo vrednost 0 *– velja* $E(X_1^2) = \sum_{k=1}^{\infty} k^2 P(X_1 = k)$. *Nadalje uporabimo račun iz točke b.:*

$$\begin{aligned} k^2 P(X_1 = k) &= \frac{\beta}{2(1+\beta)}\, k(2k-1) P(X_1 = k-1) \\ &= \frac{\beta}{2(1+\beta)} \big[2(k-1)^2 + 3(k-1) + 1\big] P(X_1 = k-1)\,. \end{aligned}$$

Nato seštejemo na enak način kot pri točki b. in dobimo zahtevano izražavo, iz katere izračunamo

$$E(X_1^2) = \frac{\beta(2+3\beta)}{4}$$

in posledično

$$\operatorname{var}(X_1) = \frac{\beta(1+\beta)}{2}\,.$$

Eksaktna standardna napaka cenilke $\hat{\beta}$ *bo torej*

$$\operatorname{se}(\hat{\beta}) = \sqrt{\operatorname{var}(\hat{\beta})} = \sqrt{\frac{4\operatorname{var}(X_1)}{n}} = \frac{\sqrt{2\beta(1+\beta)}}{\sqrt{n}}\,,$$

kar se ujema z oceno na podlagi Fisherjeve informacije.

4. (20) Privzemite, da so podatki $(x_1, y_1), \ldots, (x_n, y_n)$ nastali kot med sabo neodvisni enako porazdeljeni slučajni pari $(X_1, Y_1), \ldots, (X_n, Y_n)$ z gostoto

$$f_{X,Y}(x, y) = e^{-x} \cdot \frac{1}{\sigma\sqrt{2\pi x}}\, e^{-\frac{(y-\theta x)^2}{2\sigma^2 x}}$$

za $\sigma > 0$, $x > 0$, $-\infty < y < \infty$. Želimo preizkusiti domnevo

$$H_0 \colon \theta = 0 \quad \text{proti} \quad H_1 \colon \theta \neq 0\,.$$

a. (5) Poiščite Wilksovo λ testno statistiko.

Rešitev: Logaritemska funkcija verjetja je

$$\ell\,(\theta, \sigma|\mathbf{x}, \mathbf{y}) = -\frac{n}{2}\log 2\pi - n\log\sigma - \frac{1}{2}\sum_{k=1}^{n}\left[\log x_k + \frac{(y_k - \theta x_k)^2}{\sigma^2 x_k}\right].$$

Najprej poiščimo maksimum v širšem modelu. Parcialno odvajamo in dobimo

$$\frac{\partial \ell}{\partial \theta} = \sum_{k=1}^{n} \frac{(y_k - \theta x_k)}{\sigma^2}$$

$$\frac{\partial \ell}{\partial \sigma} = -\frac{n}{\sigma} + \sum_{k=1}^{n} \frac{(y_k - \theta x_k)^2}{\sigma^3 x_k}$$

Izenačimo parcialna odvoda z 0*. Iz prve enačbe sledi*

$$\hat{\theta} = \frac{\sum_{k=1}^{n} y_k}{\sum_{k=1}^{n} x_k}\,,$$

iz druge pa

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{k=1}^{n}\frac{(y_k - \hat{\theta}x_k)^2}{x_k} = \frac{1}{n}\left[\sum_{k=1}^{n}\frac{y_k^2}{x_k} - \frac{\left(\sum_{k=1}^{n} y_k\right)^2}{\sum_{k=1}^{n} x_k}\right].$$

V ožjem modelu, ko je $\theta = 0$*, maksimiziramo samo po* σ*. Z odvajanjem dobimo*

$$\frac{\partial \ell}{\partial \sigma} = -\frac{n}{\sigma} + \sum_{k=1}^{n}\frac{y_k^2}{\sigma^3 x_k}\,.$$

Izenačimo z 0 *in sledi*

$$\tilde{\sigma}^2 = \frac{1}{n}\sum_{k=1}^{n}\frac{y_k^2}{x_k}\,.$$

Z nekaj računanja sledi

$$\begin{aligned}\lambda &= 2\big[\ell(\hat{\theta}, \hat{\sigma}|\mathbf{x}, \mathbf{y}) - \ell(0, \tilde{\sigma}|\mathbf{x}, \mathbf{y})\big] \\ &= -2n\log\hat{\sigma} + 2n\log\tilde{\sigma} \\ &= n\log\frac{\sum_{k=1}^{n} x_k \sum_{k=1}^{n}\frac{y_k^2}{x_k}}{\sum_{k=1}^{n} x_k \sum_{k=1}^{n}\frac{y_k^2}{x_k} - \left(\sum_{k=1}^{n} y_k\right)^2}\,.\end{aligned}$$

b. (5) Predpostavite, da je $\theta = 0$. Definirajte

$$Z_k = \frac{Y_k}{\sqrt{X_k}} \quad \text{in} \quad W_k = \sqrt{X_k}$$

za $k = 1, 2, \ldots, n$. Pokažite, da so $Z_1, \ldots, Z_n, W_1, \ldots, W_n$ neodvisne z $Z_k \sim \mathrm{N}(0, \sigma^2)$.

Rešitev: Ker so pari neodvisni, je treba pokazati samo neodvisnost med Z_k in W_k. Splača se najprej poiskati porazdelitev slučajnih spremenljivk X_k. Ko integriramo po y, opazimo, da po izpostavitvi faktorja e^{-x} preostane natančno gostota normalne porazdelitve $\mathrm{N}(0, \sigma^2 x)$. Sledi, da je $X_k \sim \exp(1)$ in $Y_k \mid X_k = x_k \sim \mathrm{N}(0, \sigma^2 x_k)$, torej $Z_k \mid X_k = x_k \sim \mathrm{N}(0, \sigma^2)$. Ker je pogojna porazdelitev neodvisna od x_k, sledita obe trditvi.

c. (5) Privzemite, da H_0 zavrnemo, če je $\lambda > \lambda_\alpha$ za tako izbran λ_α, da bo stopnja tveganja dani $\alpha \in (0, 1)$. Utemeljite, da je kritično območje za tak test enako

$$\left\{ \frac{\left(\sum_{k=1}^n y_k\right)^2 \Big/ \sum_{k=1}^n x_k}{\sum_{k=1}^n y_k^2 / x_k} > c_\alpha \right\}$$

za ustrezen c_α.

Rešitev: Wilksovo statistiko prepišemo v obliki

$$\lambda = -n \log \left(1 - \frac{\left(\sum_{k=1}^n y_k\right)^2}{\sum_{k=1}^n x_k \sum_{k=1}^n \frac{y_k^2}{x_k}} \right) .$$

Neenačba $\lambda > \lambda_\alpha$ je torej ekvivalentna neenačbi

$$\frac{\left(\sum_{k=1}^n y_k\right)^2}{\sum_{k=1}^n x_k \sum_{k=1}^n y_k^2 / x_k} > 1 - e^{-\lambda_\alpha / n} .$$

d. (5) Kot znano predpostavite, da v primeru, ko so $Z_1, \ldots, Z_n, W_1, \ldots, W_n$ neodvisne, $W_k > 0$ za vse k in $Z_k \sim \mathrm{N}(0, \sigma^2)$ za vse k, velja

$$\frac{\left(\sum_{k=1}^n Z_k W_k\right)^2 \Big/ \sum_{k=1}^n W_k^2}{\sum_{k=1}^n Z_k^2} \sim \mathrm{Beta}\left(\frac{1}{2}, \frac{n-1}{2} \right) .$$

Navedite eksakten test za zgornjo domnevo z uporabo kvantilov porazdelitve $\mathrm{Beta}\left(\frac{1}{2}, \frac{n-1}{2}\right)$.

Rešitev: Če definiramo

$$z_k = \frac{y_k}{\sqrt{x_k}} \quad \textit{in} \quad w_k = \sqrt{x_k} ,$$

dobi kritično območje iz prejšnje točke obliko

$$\frac{\left(\sum_{k=1}^{n} z_k w_k\right)^2 \Big/ \sum_{k=1}^{n} w_k^2}{\sum_{k=1}^{n} z_k^2} > c_\alpha \,.$$

Ustrezni test dobimo, če za c_α namesto funkcije kvantila porazdelitve hi kvadrat postavimo kvantil porazdelitve $\mathrm{Beta}\left(\frac{1}{2}, \frac{n-1}{2}\right)$ *za verjetnost* $1-\alpha$.

Opomba. *Iz teorije velikih vzorcev torej dobimo, da, če so $\xi_1, \xi_2, \ldots$ slučajne spremenljivke s $\xi_n \sim \mathrm{Beta}\left(\frac{1}{2}, \frac{n-1}{2}\right)$, porazdelitve slučajnih spremenljivk $-n\log(1-\xi_n)$ konvergirajo proti $\chi^2(1)$. Še drugače, če je $\eta_n \sim \mathrm{Beta}\left(\frac{n-1}{2}, \frac{1}{2}\right)$, porazdelitve slučajnih spremenljivk $-n\log(\eta_n)$ konvergirajo proti $\chi^2(1)$.*

5. (20) Predpostavljamo model linearne regresije

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

z znano fiksno matriko $\mathbf{X}$ in $E(\boldsymbol{\epsilon}) = 0$ ter $\mathrm{var}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$.

Za dan n-dimenzionalni vektor $\mathbf{a} \neq 0$ bi radi našli najboljšo linearno cenilko oblike $\hat{v} = \mathbf{LY}$ z lastnostjo $E(\mathbf{LY}) = 0$ za količino $v = \mathbf{a}^T\boldsymbol{\epsilon}$ v smislu, da je kvadratična napaka

$$E\Big[\big(\hat{v} - v\big)^2\Big]$$

najmanjša možna.

a. (10) Pokažite, da je $\mathbf{LY} = \mathbf{a}^T\left(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)$ med vsemi cenilkami z navedenimi lastnostmi najboljša cenilka za v. Pri tem je $\hat{\boldsymbol{\beta}}$ cenilka parametra $\boldsymbol{\beta}$ po metodi najmanjših kvadratov.

Namig: uporabite idejo dokaza izreka Gaussa in Markova.

Rešitev: Spomnimo se, da je $\hat{\boldsymbol{\beta}} = \big(\mathbf{X}^T\mathbf{X}\big)^{-1}\mathbf{X}^T\mathbf{Y}$. *Od tod sledi, da je dana cenilka za v res oblike* $\mathbf{LY}$, *kjer je* $\mathbf{L} = \mathbf{a}^T(\mathbf{I} - \mathbf{H})$ *in* $\mathbf{H} = \mathbf{X}\big(\mathbf{X}^T\mathbf{X}\big)^{-1}\mathbf{X}^T$. *Ker je* $(\mathbf{I}-\mathbf{H})\mathbf{X} = 0$, *je tudi* $E\left(\mathbf{LY}\right) = 0$. *Naj bo* $\tilde{\mathbf{L}}\mathbf{Y}$ *konkurenčna cenilka z želenimi lastnostmi. Ker je* $E(\tilde{\mathbf{L}}\mathbf{Y}) = \tilde{\mathbf{L}}\mathbf{X}\boldsymbol{\beta} = 0$ *za vse* $\boldsymbol{\beta}$, *mora veljati* $\tilde{\mathbf{L}}\mathbf{X} = 0$. *Velja tudi* $\mathbf{LX} = 0$. *Računamo*

$$\begin{aligned} E\Big[\big(\tilde{\mathbf{L}}\mathbf{Y} - v\big)^2\Big] &= E\Big[\big(\tilde{\mathbf{L}}\mathbf{Y} - \mathbf{L}\mathbf{Y} + \mathbf{L}\mathbf{Y} - v\big)^2\Big] \\ &= E\Big[\big(\tilde{\mathbf{L}}\mathbf{Y} - \mathbf{L}\mathbf{Y}\big)^2\Big] + E\Big[\big(\mathbf{L}\mathbf{Y} - v\big)^2\Big] \\ &\quad + 2E\Big[\big(\tilde{\mathbf{L}}\mathbf{Y} - \mathbf{L}\mathbf{Y}\big)\big(\mathbf{L}\mathbf{Y} - v\big)\Big]. \end{aligned}$$

Pričakovana vrednost produkta je podobno kot v dokazu izreka Gaussa in Markova enaka nič. Da to preverimo opazimo, da je enaka kovarianci, saj je $E(\mathbf{LY}) = E(\tilde{\mathbf{L}}\mathbf{Y}) = E(v) = 0$. *Torej je*

$$E\Big[\big(\tilde{\mathbf{L}}\mathbf{Y} - \mathbf{L}\mathbf{Y}\big)\big(\mathbf{L}\mathbf{Y} - v\big)\Big] = \mathrm{cov}\big(\tilde{\mathbf{L}}\mathbf{Y} - \mathbf{L}\mathbf{Y}, \mathbf{L}\mathbf{Y} - v\big) = \mathrm{cov}\big(\tilde{\mathbf{L}}\mathbf{Y} - \mathbf{L}\mathbf{Y}, \mathbf{L}\mathbf{Y} - \mathbf{a}^T\boldsymbol{\epsilon}\big).$$

Ker se $\mathbf{Y}$ *in* $\boldsymbol{\epsilon}$ *razlikujeta le za deterministično količino* $\mathbf{X}\boldsymbol{\beta}$, *smemo* $\mathbf{Y}$ *nadomestiti z* ϵ. *Sledi*

$$\begin{aligned} E\Big[\big(\tilde{\mathbf{L}}\mathbf{Y} - \mathbf{L}\mathbf{Y}\big)\big(\mathbf{L}\mathbf{Y} - v\big)\Big] &= \mathrm{cov}\Big(\big(\tilde{\mathbf{L}} - \mathbf{L}\big)\boldsymbol{\epsilon},\ \big(\mathbf{L} - \mathbf{a}^T\big)\boldsymbol{\epsilon}\Big) \\ &= \big(\tilde{\mathbf{L}} - \mathbf{L}\big) \cdot \sigma^2\mathbf{I} \cdot \big(\mathbf{L}^T - \mathbf{a}\big) \\ &= \sigma^2 \cdot \big(\tilde{\mathbf{L}} - \mathbf{L}\big)\big(-\mathbf{H}\big)\mathbf{a} \\ &= 0 \end{aligned}$$

(zadnja enakost sledi iz dejstva, da je $\mathbf{LX} = \tilde{\mathbf{L}}\mathbf{X} = 0$*). Končno dobimo*

$$E\Big[\big(\tilde{\mathbf{L}}\mathbf{Y} - v\big)^2\Big] = E\Big[\big(\tilde{\mathbf{L}}\mathbf{Y} - \mathbf{L}\mathbf{Y}\big)^2\Big] + E\Big[\big(\mathbf{L}\mathbf{Y} - v\big)^2\Big] \geq E\Big[\big(\mathbf{L}\mathbf{Y} - v\big)^2\Big]\,,$$

kar pomeni, da konkurenčna cenilka $\tilde{\mathbf{L}}\mathbf{Y}$ *ne more biti boljša od predlagane cenilke* $\mathbf{L}\mathbf{Y}$*.*

b. (10) Naj bo $\hat{\boldsymbol{\beta}}$ cenilka parametra $\boldsymbol{\beta}$ po metodi najmanjših kvadratov. Izračunajte

$$E\left[\big(\mathbf{a}^T\hat{\boldsymbol{\epsilon}} - \mathbf{a}^T\boldsymbol{\epsilon}\big)^2\right]\,,$$

kjer je $\hat{\boldsymbol{\epsilon}} = (\mathbf{I} - \mathbf{H})\,\mathbf{Y}$ in $\mathbf{H} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T$.

Rešitev: Z upoštevanjem, da je

$$\hat{\boldsymbol{\epsilon}} - \boldsymbol{\epsilon} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\big(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}\big) - \boldsymbol{\epsilon} = -\mathbf{H}\boldsymbol{\epsilon}\,,$$

izračunamo:

$$\begin{aligned} E\left[\big(\mathbf{a}^T\hat{\boldsymbol{\epsilon}} - \mathbf{a}^T\boldsymbol{\epsilon}\big)^2\right] &= E\left[\big(-\mathbf{a}^T\mathbf{H}\boldsymbol{\epsilon}\big)^2\right] \\ &= \operatorname{var}\big(\mathbf{a}^T\mathbf{H}\boldsymbol{\epsilon}\big) \\ &= \mathbf{a}^T\mathbf{H}\cdot\sigma^2\mathbf{I}\cdot\mathbf{H}\mathbf{a} \\ &= \sigma^2\mathbf{a}^T\mathbf{H}\mathbf{a}\,. \end{aligned}$$

6. (20) Na populaciji, v kateri je N enot, gledamo določeno lastnost. Vsaka enota ima to lastnost z verjetnostjo p, ki je ne poznamo. Privzamemo, da so enote pri tem med seboj neodvisne. Naj bo K število enot na celi populaciji, ki imajo dano lastnost.

Iz populacije vzamemo vzorec velikosti n in opazimo, da ima k enot dano lastnost. Privzamemo, da je vzorčenje neodvisno od zastopanosti dane lastnosti na populaciji.

a. (5) Postavimo se v situacijo, preden vzamemo vzorec, tako da je tudi k slučajna spremenljivka. Vzemimo konstanti $C, c \in \mathbb{R}$. Izračunajte $E(ck + CK)$ in $\mathrm{var}(ck + CK)$.

Namig: oglejte si slučajne spremenljivke $I_1, I_2, \ldots, I_N$, kjer je I_j indikator dogodka, da ima j-ta enota dano lastnost.

Rešitev: Naj bodo $1, 2, \ldots, n$ enote, zajete v vzorec, $n+1, n+2, \ldots, N$ pa enote izven vzorca. Tedaj je $k = \sum_{j=1}^{n} I_j$ in $K = \sum_{j=1}^{N} I_j$. Ker je vzorčenje neodvisno od zastopanosti lastnosti na populaciji, je $E(I_j) = p$, torej je

$$E(ck + CK) = (cn + CN)p\,.$$

Za izračun variance pa upoštevamo, da so indikatorji med seboj neodvisni (za to potrebujemo tako neodvisnost zastopanosti lastnosti po posameznih enotah kot tudi neodvisnost vzorčenja od zastopanosti). Pišimo:

$$ck + CK = (c + C)\sum_{j=1}^{n} I_j + C\sum_{i=n+1}^{N} I_j\,.$$

Ker je $\sum_{j=1}^{n} I_j \sim \mathrm{Bin}(n, p)$ in $\sum_{i=n+1}^{N} I_j \sim \mathrm{Bin}(N, p)$, je:

$$\mathrm{var}(ck + CK) = \big[(c + C)^2 n + C^2(N - n)\big]p(1 - p)\,.$$

b. (10) Znano je, da za dovolj velike vzorce velja centralni limitni izrek. Poleg tega je za velike vzorce znano, da imata slučajni spremenljivki:

$$\frac{ck + CK}{\sqrt{\frac{k}{n}\big(1 - \frac{k}{n}\big)}} \quad \text{in} \quad \frac{ck + CK}{\sqrt{p(1 - p)}}$$

za poljubni konstanti $C, c \in \mathbb{R}$ približno enako porazdelitev. Privzemimo, da je vzorec dovolj velik, da v okviru dogovorjene natančnosti velja oboje. Poiščite taki števili a in b (odvisni od N in n), da bo

$$ak - b\sqrt{k(n - k)} < K < ak + b\sqrt{k(n - k)}$$

pri zgornjih predpostavkah približen napovedni interval za K pri dani stopnji tveganja α, če p ni preblizu 0 ali 1.

Rešitev: Glede na to, da je $E(k) = np$ *in* $E(K) = Np$*, postavimo* $a = N/n$*. Napovedni interval prepišemo v obliki:*

$$-bn^2 < \frac{nK - Nk}{\sqrt{\frac{k}{n}\left(1 - \frac{k}{n}\right)}} < bn^2\,.$$

Po predpostavki ima slučajna spremenljivka na sredini približno enako porazdelitev kot $\frac{nK-Nk}{\sqrt{p(1-p)}}$*. Po centralnem limitnem izreku je torej slučajna spremenljivka* $nK - Nk$ *porazdeljena približno normalno. Pričakovano vrednost in varianco odčitamo iz prejšnje točke:*

$$\begin{aligned} E(nK - NK) &= 0\,, \\ \operatorname{var}(nK - Nk) &= (n-N)^2 np(1-p) + n^2(N-n)p(1-p) \\ &= (N-n)Nnp(1-p)\,. \end{aligned}$$

Slučajna spremenljivka $\frac{nK-Nk}{\sqrt{p(1-p)}}$ *in z njo tudi* $\frac{nK-Nk}{\sqrt{\frac{k}{n}\left(1-\frac{k}{n}\right)}}$ *pa je porazdeljena približno normalno s pričakovano vrednostjo nič in varianco* $(N-n)Nn$*. Napovedni inteval zdaj prepišemo v obliki:*

$$-b\sqrt{\frac{n^3}{(N-n)N}} < \frac{nK - Nk}{\sqrt{(N-n)Nn}} < b\sqrt{\frac{n^3}{(N-n)N}}\,.$$

Srednja slučajna spremenljivka je zdaj porazdeljena približno standardno normalno. Napovedni interval bo torej imel približno dano stopnjo tveganja, če bo $b = z_{1-\alpha/2}\sqrt{(N-n)N/n^3}$*, kjer je* $z_{1-\alpha/2}$ *kvantil standardne normalne porazdelitve za verjetnost* $1 - \alpha/2$*.*

c. (5) Za aktualnega predsednika ZDA Donalda Trumpa je bilo 9. 11. 2016 ob 7:30 po srednjeevropskem času znano, da je dobil 245 elektorskih glasov od 460. Vseh elektorskih glasov je 538. Kakšen bi bil po zgornjem modelu napovedni interval za skupno število dobljenih elektorskih glasov, če vzamemo stopnjo tveganja $\alpha = 0{,}05$?

Na koncu je Donald Trump dobil 304 elektorske glasove. Mislite, da je zgornji model ustrezen za predsedniške volitve v ZDA?

Rešitev: Vstavimo $k = 245$*,* $n = 460$ *in* $N = 538$*. Dobimo* $277{,}20 < K < 295{,}88$ *oziroma* $277 < K < 296$*. Dejansko število glasov je izven napovednega intervala, to pa zlahka razložimo z neustreznostjo zgornjega modela. Elektorski glasovi namreč niso neodvisni, saj gredo po državah – praviloma vsi elektorji iz določene države glasujejo enako.*