Ime in priimek: ______________________ Vpisna št: | | | | | | | | |

FAKULTETA ZA MATEMATIKO IN FIZIKO

Oddelek za matematiko

Statistika

Pisni izpit

2. julij 2020

Navodila

Pazljivo preberite besedilo naloge, preden se lotite reševanja.

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

1. (25) Naj bo dvorazsežni slučajni vektor $(X_0, Y_0) \sim \mathrm{N}(0, \mathbf{\Sigma})$ neodvisen od dvorazsežnih slučajnih vektorjev $\boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2, \ldots, \boldsymbol{\epsilon}_n$, ti pa naj bodo neodvisni med sabo in naj velja $\boldsymbol{\epsilon}_k \sim \mathrm{N}(0, \mathbf{\Lambda})$. Definirajte rekurzivno

$$\begin{pmatrix} X_{k+1} \\ Y_{k+1} \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} X_k \\ Y_k \end{pmatrix} + \boldsymbol{\epsilon}_{k+1}$$

za $k = 1, 2, \ldots, n-1$. Označite

$$\mathbf{Z}_k = \begin{pmatrix} X_k \\ Y_k \end{pmatrix}$$

za $k = 0, 1, \ldots, n$. Kot znano privzemite, da za večrazsežen normalni vektor $(X, Y, \mathbf{Z})^T$ velja

$$E\left(Y|X, \mathbf{Z}\right) = E(Y|\mathbf{Z}) + \mathrm{cov}(X, Y|\mathbf{Z})\mathrm{var}(X|\mathbf{Z})^{-1}\left(X - E(X|\mathbf{Z})\right) .$$

a. (5) Izračunajte $E(Y_{n+1}|\mathbf{Z}_0, \ldots, \mathbf{Z}_n)$.

Rešitev: Iz enačbe

$$\mathbf{Z}_{n+1} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \mathbf{Z}_n + \boldsymbol{\epsilon}_{n+1} \, ,$$

linearnosti in predpostavk o neodvisnosti takoj sledi

$$E(Y_{n+1}|\mathbf{Z}_0, \ldots, \mathbf{Z}_n) = X_n + Y_n \, .$$

b. (5) Izračunajte

$$\mathrm{cov}(X_{n+1}, Y_{n+1}|\mathbf{Z}_0, \ldots, \mathbf{Z}_n) \, .$$

Rešitev: Iz predpostavk o neodvisnosti sledi

$$\mathrm{cov}(X_{n+1}, Y_{n+1}|\mathbf{Z}_0, \ldots, \mathbf{Z}_n) = \lambda_{12} \, .$$

c. (5) Izračunajte

$$\mathrm{var}(X_{n+1}|\mathbf{Z}_0, \ldots, \mathbf{Z}_n) \, .$$

Rešitev: Iz predpostavk o neodvisnosti sledi

$$\mathrm{var}(X_{n+1}|\mathbf{Z}_0, \ldots, \mathbf{Z}_n) = \lambda_{11}^2 \, .$$

d. (10) Izračunajte

$$E\left(Y_{n+1}|X_{n+1}, \mathbf{Z}_0, \ldots, \mathbf{Z}_n\right) \, .$$

Rešitev: Vektor $(X_{n+1}, Y_{n+1}, \mathbf{Z}_0^T, \ldots, \mathbf{Z}_n^T)$ *je večrazsežen normalen. Uporabimo lahko formulo iz besedila naloge in dobimo*

$$E\left(Y_{n+1}|X_{n+1}, \mathbf{Z}_0, \ldots, \mathbf{Z}_n\right) = X_n + Y_n + \frac{\lambda_{12}}{\lambda_{11}}\left(X_{n+1} - X_n - Y_n\right) \, .$$

2. (25) Dana je populacija velikosti N, na njej pa statistična spremenljivka. Označimo z μ in σ^2 povprečje in varianco te spremenljivke na celotni populaciji.

a. (5) Recimo, da iz dane populacije izberemo enostavni slučajni vzorec velikosti n. Najdite nepristransko cenilko za kvadrat populacijskega povprečja μ^2 in utemeljite nepristranskost.

Namig: ali lahko nepristransko ocenite količino

$$\gamma = \frac{1}{N}\sum_{k=1}^{N} x_k^2 \text{ ?}$$

Rešitev: Količino γ lahko ocenimo nepristransko z vzorčnim povprečjem kvadratov vzorčnih vrednosti. Velja tudi

$$\sigma^2 = \frac{1}{N}\sum_{k=1}^{N} x_k^2 - \mu^2 = \gamma - \mu^2 .$$

Vemo, da je

$$\hat{\sigma}^2 = \frac{N-1}{N(n-1)}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

nepristranska cenilka za σ^2, kjer z $X_1, \ldots, X_n$ označimo vzorčne vrednosti. V zgornji enačbi torej znamo nepristransko oceniti dve od treh količin, zato znamo nepristransko oceniti tudi tretjo, torej μ^2.

b. (10) Zdaj pa populacijo razdelimo na K skupinic velikosti M, tako da je $N = KM$, ter vzorčimo tako, da izberemo enostavni slučajni vzorec velikosti k skupinic in vključimo v vzorec vse enote iz izbranih skupinic. Označimo z μ_r in σ_r^2 povprečje in varianco dane spremenljivke na r-ti skupinici, $r = 1, 2, \ldots, K$. Za cenilko populacijskega povprečja μ vzemimo povprečje na tako dobljenem vzorcu. Utemeljite, da je ta cenilka nepristranska, in navedite njeno standardno napako.

Rešitev: Ker so vse skupinice enako velike, je $\mu = \frac{1}{K}\sum_{k=1}^{K}\mu_k$. Če si mislimo, da vzorčimo iz populacije skupnic, potem z enostavnim slučajnim vzorčenjem ocenjujemo populacijsko povprečje. Cenilka je zato nepristranska, varianca pa je enaka

$$\mathrm{var}(\bar{X}) = \frac{\tau^2}{k}\cdot\frac{K-k}{K-1},$$

kjer je

$$\tau^2 = \frac{1}{K}\sum_{r=1}^{K}(\mu_r - \mu)^2 .$$

c. (10) Recimo, da želimo z vzorcem, opisanem v prejšnji točki, oceniti populacijsko varianco σ^2. Predlagajte nepristransko cenilko in utemeljite nepristranskost.

Namig: oglejte si točko a.

Rešitev:

Prvi način. *Če razmišljamo o skupinicah kot primarnih enotah, lahko na podlagi prvega dela naloge sklepamo, da znamo nepristransko oceniti* μ^2. *Če se potem vrnemo k opisanemu vzorčenju, pa vemo, da lahko nepristransko ocenimo vsoto*

$$\frac{1}{N}\sum_{k=1}^{N} x_k^2 \,.$$

Ker je

$$\sigma^2 = \frac{1}{N}\sum_{k=1}^{N} x_k^2 - \mu^2$$

in znamo nepristransko oceniti obe količini na desni, znamo nepristransko oceniti tudi σ^2.

Za eksplicitno izražavo cenilke označimo z X_{ij} *vrednost spremenljivke na* j*-ti enoti v* i*-ti skupinici, izbrani v vzorec, z* A_i *povprečje na tej skupinici, z* $\bar{A}$ *pa povprečje teh povprečij po izbranih skupinicah (to je torej iskana cenilka v točki b.). Tedaj velja*

$$\begin{aligned}
\hat{\sigma}^2 &= \frac{1}{kM}\sum_{i=1}^{k}\sum_{j=1}^{M} X_{ij}^2 - \frac{1}{k}\sum_{i=1}^{k} A_i^2 + \frac{K-1}{K(k-1)}\sum_{i=1}^{k}(A_i-\bar{A})^2 \\
&= \frac{1}{kM}\sum_{i=1}^{k}\sum_{j=1}^{M}(X_{ij}-A_i)^2 + \frac{K-1}{K(k-1)}\sum_{i=1}^{k}(A_i-\bar{A})^2 \,.
\end{aligned}$$

Drugi način. *Upoštevamo, da populacijska varianca razpade na varianco znotraj skupinic in varianco med skupinicami: velja* $\sigma^2 = \sigma_{\mathrm{B}}^2 + \sigma_{\mathrm{W}}^2$, *kjer je*

$$\sigma_{\mathrm{B}}^2 = \frac{1}{K}\sum_{r=1}^{k}(\mu_r-\mu)^2 \quad \textit{in} \quad \sigma_{\mathrm{W}}^2 = \frac{1}{KM}\sum_{r=1}^{k}\sum_{s=1}^{M}(x_{rs}-\mu_r)^2 \,,$$

x_{rs} *pa je vrednost spremenljivke na* s*-ti enoti v* r*-ti skupinici. Vemo, da je*

$$\hat{\sigma}_{\mathrm{B}}^2 = \frac{K-1}{K(k-1)}\sum_{i=1}^{k}(A_i-\bar{A})^2$$

nepristranska cenilka za σ_{B}^2. *Zdaj pa opazimo, da je* σ_{W}^2 *populacijsko povprečje statistične spremenljivke, ki je na* s*-ti enoti v* r*-ti skupinici enaka* $y_{rs} := (x_{rs} - \mu_r)^2$. *Točka b., uporabljena na tej spremenljivki, pove, da je*

$$\frac{1}{kM}\sum_{i=1}^{k}\sum_{j=1}^{M}(X_{ij}-A_i)^2$$

nepristranska cenilka za $\hat{\sigma}_{\mathrm{B}}^2$. *Ko seštejemo, dobimo isto cenilko kot pri prvem načinu.*

3. (25) Opazovane vrednosti naj bodo pari $(x_1, y_1), \ldots, (x_n, y_n)$, $n \geq 2$, za katere predpostavljamo, da so vzorec iz dvorazsežne normalne porazdelitve

$$\mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}\right).$$

a. (10) Poiščite cenilke za parametre po metodi največjega verjetja. Kot znano privzemite, da za dano pozitivno definitno matriko $\mathbf{A}$ med vsemi pozitivno definitnimi matrikami $\mathbf{X}$ izraz

$$\frac{1}{(\det \mathbf{X})^{n/2}} \exp\left(-\frac{1}{2}\mathrm{Sl}\left(\mathbf{X}^{-1}\mathbf{A}\right)\right)$$

maksimizira matrika $\mathbf{X} = \frac{1}{n}\mathbf{A}$.

Rešitev: Označimo $\mathbf{x}_k = (x_k, y_k)$. Funkcijo verjetja tako lahko zapišemo kot

$$L(\mathbf{\Sigma}|\mathbf{x}_1, \ldots, \mathbf{x}_n) = \frac{1}{(2\pi)^{n/2}(\det \mathbf{\Sigma})^{n/2}} \exp\left(-\frac{1}{2}\sum_{k=1}^{n} \mathbf{x}_k^T \mathbf{\Sigma}^{-1} \mathbf{x}_k\right).$$

Če želimo upoštevati namig, pišemo:

$$\mathbf{x}_k^T \mathbf{\Sigma}^{-1} \mathbf{x}_k = \mathrm{Sl}\big(\mathbf{x}_k^T \mathbf{\Sigma}^{-1} \mathbf{x}_k\big) = \mathrm{Sl}\big(\mathbf{\Sigma}^{-1} \mathbf{x}_k \mathbf{x}_k^T\big).$$

Iz namiga zdaj razberemo, da maksimum dobimo pri matriki

$$\hat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{k=1}^{n} \mathbf{x}_k \mathbf{x}_k^T = \frac{1}{n}\begin{bmatrix} \sum_{k=1}^n x_k^2 & \sum_{k=1}^n x_k y_k \\ \sum_{k=1}^n x_k y_k & \sum_{k=1}^n y_k^2 \end{bmatrix}.$$

b. (15) Želimo preizkusiti domnevo

$$H_0\colon \sigma_{12} = 0 \qquad \text{proti} \qquad H_1\colon \sigma_{12} \neq 0.$$

Poiščite testno statistiko po metodi kvocienta verjetij in navedite njeno aproksimativno porazdelitev pri velikem vzorcu.

Rešitev: Najprej moramo oceniti parametre pri domnevi H_0. Če je $\sigma_{12} = 0$, funkcija verjetja razpade na faktor, ki je odvisen samo od x-ov in faktor, ki je odvisen samo od y-ov. Maksimum dobimo, če maksimiziramo vsak faktor posebej, kar pa je iskanje ocen po metodi največjega verjetja za normalno porazdelitev. Maksimum pri stranskem pogoju torej dobimo za

$$\hat{\sigma}_{11} = \frac{1}{n}\sum_{k=1}^{n} x_k^2 \qquad \textit{in} \qquad \hat{\sigma}_{22} = \frac{1}{n}\sum_{k=1}^{n} y_k^2.$$

Iz prvega dela naloge sledi, da je maksimum brez stranskih pogojev enak

$$L(\hat{\mathbf{\Sigma}}|\mathbf{x}_1, \ldots, \mathbf{x}_n) = \frac{1}{(2\pi)^{n/2}(\det \hat{\mathbf{\Sigma}})^{n/2}} e^{-n},$$

maksimum s stranskimi pogoji pa

$$\frac{1}{(2\pi)^{n/2}(\hat{\sigma}_{11}\hat{\sigma}_{22})^{n/2}}\,e^{-n}\,.$$

Sledi

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

Statistika $\lambda = 2\log\Lambda$ *ima približno porazdelitev* $\chi^2(1)$, *t. j. hi kvadrat z eno prostostno stopnjo.*

4. (25) Privzemite regresijski model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}$$

kjer je $\mathbf{Z}$ matrika velikosti $r \times n$ ranga r in velja $n \geq r$, za $\mathbf{u}$ pa velja $E(\mathbf{u}) = 0$ in $\mathrm{var}(\mathbf{u}) = \sigma^2\mathbf{I}$.

a. (10) Pokažite, da je

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T(\mathbf{Z}\mathbf{Z}^T)^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\left(\mathbf{Z}\mathbf{Z}^T\right)^{-1}\mathbf{Y}$$

najboljša nepristranska linearna cenilka za $\boldsymbol{\beta}$.

Rešitev:
Prvi način: *neposredno. Nepristranskost lahko takoj preverimo. Naj bo* $\tilde{\boldsymbol{\beta}} = \mathbf{L}\mathbf{Y}$ *alternativna linearna nepristranska cenilka. Iz*

$$E\big(\tilde{\boldsymbol{\beta}}\big) = \mathbf{L}\mathbf{X}\boldsymbol{\beta}$$

za vse $\boldsymbol{\beta}$ *sledi* $\mathbf{L}\mathbf{X} = \mathbf{I}$. *Računamo*

$$\begin{aligned}
\mathrm{var}\big(\tilde{\boldsymbol{\beta}}\big) &= \mathrm{var}\big(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\beta}}\big) \\
&= \mathrm{var}\big(\hat{\boldsymbol{\beta}}\big) + \mathrm{var}\big(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}\big) + 2\,\mathrm{cov}\big(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\beta}}\big)\,.
\end{aligned}$$

Cenilka $\tilde{\boldsymbol{\beta}}$ *bo najboljša, brž ko bo kovarianca na desni enaka* 0. *Upoštevajoč*

$$\mathrm{cov}\,(\mathbf{Y}, \mathbf{Y}) = \sigma^2\mathbf{Z}\mathbf{Z}^T\,.$$

izračunamo, da je to res:

$$\begin{aligned}
&\mathrm{cov}\big(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\beta}}\big) = \\
&= \Big(\mathbf{L} - \big(\mathbf{X}^T(\mathbf{Z}\mathbf{Z}^T)^{-1}\mathbf{X}\big)^{-1}\mathbf{X}^T(\mathbf{Z}\mathbf{Z}^T)^{-1}\Big)\,\mathrm{cov}(\mathbf{Y}, \mathbf{Y})\cdot \\
&\qquad \cdot\Big(\big(\mathbf{X}^T(\mathbf{Z}\mathbf{Z}^T)^{-1}\mathbf{X}\big)^{-1}\mathbf{X}^T(\mathbf{Z}\mathbf{Z}^T)^{-1}\Big)^T \\
&= \sigma^2\Big(\mathbf{L} - \big(\mathbf{X}^T(\mathbf{Z}\mathbf{Z}^T)^{-1}\mathbf{X}\big)^{-1}\mathbf{X}^T(\mathbf{Z}\mathbf{Z}^T)^{-1}\Big)\,\mathbf{Z}\mathbf{Z}^T \\
&\qquad \cdot(\mathbf{Z}\mathbf{Z}^T)^{-1}\mathbf{X}\,\big(\mathbf{X}^T(\mathbf{Z}\mathbf{Z}^T)^{-1}\mathbf{X}\big)^{-1} \\
&= \sigma^2\Big(\mathbf{L} - \big(\mathbf{X}^T(\mathbf{Z}\mathbf{Z}^T)^{-1}\mathbf{X}\big)^{-1}\mathbf{X}^T(\mathbf{Z}\mathbf{Z}^T)^{-1}\Big)\,\mathbf{X}\,\big(\mathbf{X}^T(\mathbf{Z}\mathbf{Z}^T)^{-1}\mathbf{X}\big)^{-1} \\
&= \sigma^2\Big(\mathbf{L}\mathbf{X} - \big(\mathbf{X}^T(\mathbf{Z}\mathbf{Z}^T)^{-1}\mathbf{X}\big)^{-1}\mathbf{X}^T(\mathbf{Z}\mathbf{Z}^T)^{-1}\mathbf{X}\Big)\,\big(\mathbf{X}^T(\mathbf{Z}\mathbf{Z}^T)^{-1}\mathbf{X}\big)^{-1} \\
&= \sigma^2(\mathbf{I} - \mathbf{I})\,\big(\mathbf{X}^T(\mathbf{Z}\mathbf{Z}^T)^{-1}\mathbf{X}\big)^{-1} \\
&= 0\,.
\end{aligned}$$

Drugi način: *prevedemo na standardno linearno regresijo. Razlika je namreč le v tem, da šum* $\mathbf{Z}\mathbf{u}$ *nima kovariančne matrike* $\sigma^2\mathbf{I}$, *temveč* $\sigma^2\mathbf{Z}\mathbf{Z}^T$. *Če definiramo*

$\mathbf{Y}' = (\mathbf{Z}\mathbf{Z}^T)^{-1/2}\mathbf{Y}$, $\mathbf{X}' = (\mathbf{Z}\mathbf{Z}^T)^{-1/2}\mathbf{X}$ *in* $\boldsymbol{\varepsilon} = (\mathbf{Z}\mathbf{Z}^T)^{-1/2}\mathbf{Z}\mathbf{u}$, *velja* $\mathbf{Y}' = \mathbf{X}'\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, *kar je standardni linearni regresijski model, saj je:*

$$\begin{aligned} \operatorname{var}(\boldsymbol{\varepsilon}) &= (\mathbf{Z}\mathbf{Z}^T)^{-1/2}\mathbf{Z}\operatorname{var}(\mathbf{u})\,\mathbf{Z}^T(\mathbf{Z}\mathbf{Z}^T)^{-1/2} \\ &= \sigma^2(\mathbf{Z}\mathbf{Z}^T)^{-1/2}\mathbf{Z}\mathbf{I}\mathbf{Z}^T(\mathbf{Z}\mathbf{Z}^T)^{-1/2} \\ &= \sigma^2\mathbf{I}\,. \end{aligned}$$

Seveda se linearne cenilke v podanem nestandardnem modelu ujemajo z linearnimi cenilkami v prirejenem standardnem modelu, nepristranskost in standardna napaka pa sta tako ali tako univerzalna pojma. Zato je iskana cenilka tudi najboljša nepristranska linearna cenilka v prirejenem standardnem modelu, to pa je:

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \big(\mathbf{X}'^T\mathbf{X}'\big)^{-1}\mathbf{X}'^T\mathbf{Y}' \\ &= \big(\mathbf{X}^T(\mathbf{Z}\mathbf{Z}^T)^{-1/2}(\mathbf{Z}\mathbf{Z}^T)^{-1/2}\mathbf{X}\big)^{-1}\mathbf{X}^T(\mathbf{Z}\mathbf{Z}^T)^{-1/2}(\mathbf{Z}\mathbf{Z}^T)^{-1/2}\mathbf{Y} \\ &= \big(\mathbf{X}^T(\mathbf{Z}\mathbf{Z}^T)^{-1}\mathbf{X}\big)^{-1}\mathbf{X}^T(\mathbf{Z}\mathbf{Z}^T)^{-1}\mathbf{Y}\,. \end{aligned}$$

b. (15) Naj bo

$$\hat{\mathbf{u}} = \mathbf{Z}^T(\mathbf{Z}\mathbf{Z}^T)^{-1}\big(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\big)$$

cenilka za $\mathbf{u}$. Izračunajte

$$\operatorname{cov}(\hat{\mathbf{u}}, \hat{\boldsymbol{\beta}})\,.$$

Rešitev: Računamo

$$\begin{aligned} &\operatorname{cov}(\hat{\mathbf{u}}, \hat{\boldsymbol{\beta}}) = \\ &\quad = \mathbf{Z}^T(\mathbf{Z}\mathbf{Z}^T)^{-1}\Big(\mathbf{I} - \mathbf{X}\left(\mathbf{X}^T(\mathbf{Z}\mathbf{Z}^T)^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T(\mathbf{Z}\mathbf{Z}^T)^{-1}\Big)\operatorname{cov}(\mathbf{Y}, \mathbf{Y})\cdot \\ &\qquad\qquad \cdot (\mathbf{Z}\mathbf{Z}^T)^{-1}\mathbf{X}\left(\mathbf{X}^T(\mathbf{Z}\mathbf{Z}^T)^{-1}\mathbf{X}\right)^{-1} \\ &\quad = \mathbf{Z}^T(\mathbf{Z}\mathbf{Z}^T)^{-1}\Big(\mathbf{I} - \mathbf{X}\left(\mathbf{X}^T(\mathbf{Z}\mathbf{Z}^T)^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T(\mathbf{Z}\mathbf{Z}^T)^{-1}\Big)\cdot \\ &\qquad\qquad \cdot \mathbf{Z}\mathbf{Z}^T(\mathbf{Z}\mathbf{Z}^T)^{-1}\mathbf{X}\left(\mathbf{X}^T(\mathbf{Z}\mathbf{Z}^T)^{-1}\mathbf{X}\right)^{-1} \\ &\quad = \mathbf{0}\,. \end{aligned}$$