Evgenija Burger
Statistika 2, 2022/23
Posodobljeno: 4. december 2022

Vaje pri Statistiki 2

1. prve vaje (6. 10. 2022)

Naloga 1. Zanima nas povprečni volumen kave esspreso, ki jo dobimo v domačem lokalu. Naročimo kavo, izmerimo volumen, označimo ga z X.

$$X \sim N(\mu, \sigma^2)$$

Eksperiment ponavljamo in dobimo slučajni vzorec $X_1, \dots X_n$. Želimo oceniti μ in σ^2 z metodo momentov in metodo največjega verjetja.

Rešitev 1. Kolikor imamo parametrov toliko momentov moramo poračunati.

$$m_i = E(X_i) \dots \ i\text{-ti moment}$$

Če želimo oceniti karakteristiko q jo najprej izrazimo kot funkcijo momentov

$$q = q(m_1, \dots, m_r)$$

Njena cenilka po metodi momentov je ista funkciji vzorčnih momentov $\hat{q} = q(\hat{m}_1, \dots, \hat{m}_n)$, kjer je $\hat{m}_k = \sum_{i=1}^n X_i^k$

$$E(X) = \mu$$
$$E(X^2) = \sigma^2 + \mu^2$$
$$var(X) = E(X^2) - E(X)^2$$

Metoda momentov (MM):

$$m_1 = \mu$$
$$m_2 = \sigma^2 + \mu^2$$
$$\hat{\mu} = \hat{m}_1 = \bar{(}X)$$
$$\hat{\sigma^2} = \hat{(}m_2) - \hat{m}_1 = \bar{X^2} - \bar{X}^2$$

Dobimo torej:

$$\hat{\mu} = \bar{X}$$
$$\hat{\sigma^2} = \bar{X^2} - \bar{X}^2$$

Izračunajmo verjetje.

$$L(X_i, \mu, \sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}} \ / \ ln$$
$$L(X_i, \mu, \sigma^2) = \sum_{i=1}^n \ln(\frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}})$$
$$l(X_i, \mu, \sigma^2) = n\ln(\frac{1}{\sqrt{2\pi\sigma^2}}) - \frac{1}{2\sigma^2}\sum_{i=1}^n (X_i - \mu)^2$$

$$\frac{\partial l}{\partial \mu} = \frac{1}{2\sigma^2}\sum_{i=1}^{n} 2(X_i - \mu)$$
$$\frac{\partial l}{\partial \sigma^2} = -\frac{n}{2}\frac{1}{2\pi\sigma^2}2\pi + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(X_i - \mu)^2$$
$$\text{i) } \frac{1}{2\hat{\sigma^2}}\sum_{i=1}^{n} 2(X_i - \hat{\mu})$$
$$\text{ii)} \frac{\mu}{2}\sigma^2 + \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2 = 0$$
$$\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2 = -\frac{\mu}{2}\sigma^2$$
$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Cenilka ni nepristranska, da dobimo nepristransko moramo deliti z $n - 1$.

Naloga 2. Predvidevamo, da je vzorec realnih števil realizacija neodvisnih ponovitev s.s. z gostoto

$$f(x, a, \lambda) = \frac{\lambda^a}{\Gamma(a)} x^{a-1} e^{-\lambda x}$$

Poišči kompletno zadostno statistiko za $\nu = (a, \lambda) \in (0, \infty) \times (0, \infty)$.

Rešitev 2.

$$f(\vec{x}, \nu) = C(\nu) e^{<T(\vec{x}, Q\nu)>} h(\vec{x})$$

Zgornja enačba je eksponentna družina. Tisti T bo zadostna statistika.

$$\begin{aligned}
f(\vec{x}, a, \lambda) &= \prod_{i=1}^{n} \frac{\lambda^a}{\Gamma(a)} x_i^{a-1} e^{-\lambda x_i} \\
&= \prod_{i=1}^{n} \frac{\lambda^a}{\Gamma(a)} e^{\ln(x_i^{a-1}) - \lambda x_i} \\
&= (\frac{\lambda^a}{\Gamma(a)})^n \prod_{i=1}^{n} e^{(a-1)\ln x_i} e^{-\lambda x_i} \\
&= (\frac{\lambda^a}{\Gamma(a)})^n e^{(a-1)\sum_{i=1}^{n} \ln x_i} e^{-\lambda \sum_{i=1}^{n} x_i} \\
&= (\frac{\lambda^a}{\Gamma(a)})^n e^{<(a-1, -\lambda), (\sum_{i=1}^{n} \ln x_i, \sum_{i=1}^{n} x_i)>}
\end{aligned}$$

Torej:

$$T(x) = (\sum_{i=1}^{n} \ln x_i, \sum_{i=1}^{n} x_i)$$

Gostoto vektorja $(x_1, \ldots, x_n)$ smo zapisali kot eksponentno družino. Kako preverimo, da je kompletna zadostna statistika? Če ima naša ekponentna družina poln rang je ZS tudi KZS. Q je bijekcija, zato ima eksponentna družina poln rang in zato je T KZS.

2. DRUGE VAJE (13. 10. 2022)

Naloga 3. Obravnavamo splošni linearni model

$$X = Z\beta + \epsilon$$

kjer je $X \in \mathbb{R}$ vektor opazovanj, $\beta \in \mathbb{R}^d$ vektor parametrov, $Z \in \mathbb{R}^{n\times d}$ matrika konstant, $\epsilon \in \mathbb{R}^n$ vektor slučajnih odstopanj, za katerega vedno privzamemo $E(\epsilon) = 0$. Zapišemo $Z = SP$ za matriko $S \in \mathbb{R}^{n\times d}$ s paroma pravokotnimi stolpci, od katerih so nekateri lahko 0 (na stolpcih matrike Z izvajamo Gramm-Schmidtovo ortogonalizacijo). P je zgornje trikotna matrika s pozitivnimi elementi na glavni diagonali. Tedaj je

$$Z^T Z = P^T JP$$

kjer je $J = diag(j_1, \ldots, j_d), j_i = ||S^i||$. Za reševanje enačbe $(Z^T Z)\hat{\beta} = Z^T x$ konstruiramo konkretno rešitev

$$\hat{\beta} = (Z^T Z)' Z^T x$$

kjer je $(Z^T Z)' = P^{-1} JP^{-T}$.

Za dano matriko Z in vektor X skonstruiraj posplošeni inverz $(Z^T Z)'$ ter poišči oceno $\hat{\beta}$.

Imamo linearni model: $X = Z\beta + \epsilon$, v splošnem iščemo cenilko za β.

$$\hat{\beta} = (Z^T Z)' Z^T x$$

kjer $(Z^T Z)$ predstavlja klasičen inverz (Z mora imeti poln rang = imeti mora linearno neodvisne stolpce).

Matriko Z razcepimo na $Z = SP$ pri čemer uporabimo Gram-Schmidtovo ortogonalizacijo.

Torej želimo $Z = SP$:

$$Z = [Z^1, Z^2, \ldots Z^d] \text{ in } S = [S^1, S^2, \ldots S^d]$$

<u>Gram-Schmidt</u>

Spomnimo se pravokotne projekcije v splošnem:

$$\text{proj}_u v = \frac{< u, v >}{< u, u >} u, \text{ pri nas je } < u, u >= 1$$

$$S_1 = \frac{Z^1}{||Z^1||}$$
$$S_2 = \frac{Z^2 - < Z^2, S^1 > S^1}{||Z^2 - < Z^2, S^1 > S^1||}$$
$$S_3 = \frac{Z^3 - < Z^3, S^1 > S^1 - < Z^3, S^2 > S^2}{||Z^3 - < Z^3, S^1 > S^1 - < Z^3, S^2 > S^2||}$$

in tako naprej

Če imamo na primer dva linearno odvisna stolpca $Z^2 = \lambda Z^1$ potem:

$$Z^2 - < Z^2, S^1 > S^1 = 0 \Rightarrow S^2 = 0$$

Kkao bi izgledala matrika P?

$$Z = SP = [Z^1, Z^2, \ldots Z^d] = [S^1, S^2, \ldots S^d]P$$

P je zgornje trikotna matrika s pozitivnimi elementi na diagonali:

$$P = \begin{bmatrix} poz & * & \cdots & * \\ & poz & \cdots & * \\ & & \ddots & * \\ & & & poz \end{bmatrix}$$

Če iz S^i izrazimo Z^i dobimo:

$$\begin{aligned} Z_1 &= ||Z_1|| S_1 \\ Z_2 &= ||Z^2 - <Z^2, S^1> S^1|| S^2 + <Z^2, S^1> S^1 \\ &\cdots \\ Z^j &= \text{poz. št. } S^j + \text{ lin. komb. } (S^1, \cdots S^{j-1}) \end{aligned}$$

Torej mora P izgledat takole:

$$P = \begin{bmatrix} ||Z_1|| & <Z^2, S^1> & <Z^3, S^1> & \cdots \\ & ||Z^2 - <Z^2, S^1> S^1|| & <Z^3, S^2> & \cdots \\ & & ||Z^3 - <Z^3, S^2> S^2 - <Z^3, S^1> S^1|| & \\ & & & \ddots \end{bmatrix}$$

Opomba 1. Na diagonali P je lahko 0, lahko pa na tisto mesto napišemo karkoli, ker se zmnoži s S^d, ki je itak ničeln.

Implementacija v R:

```r
# norma vektorja
norma = function(x){
  sqrt(sum(x^2))
}

# GRAM-SCHMIDT
GS = function(Z){
  n <- nrow(Z)
  d <- ncol(Z)
  S <- matrix(0, nrow = n, ncol = d)
  P <- matrix(0, nrow = d, ncol = d)
  # prvi stolpec S samo normiramo
  # drop = F uporabljamo zato, da se ohrani oblika (rezultat
  # bo se vedno matrika in ne vektor)
  S[, 1] = Z[, 1, drop = F] / norma(Z[, 1])
  P[1, 1] = norma(Z[, 1])
  for (i in 2:d){
    S[, i] = Z[, i, drop = F]
    # odstevamo pravokotne projekcije
    for (j in 1:(i-1)){
      koef = sum(Z[, i] * S[, j]) # skalarni produkt
      S[, i] = S[, i] - koef * S[, j]
      P[j, i] = koef
    }
    normaS = norma(S[,i])
    # ce je pripadajoc stolpec enak 0, bo tudi norma enaka 0
    if (normaS < 1e-10){
      # na diagonalo P lahko postavimo katerokoli pozitivno stevilo, ker se mnozi z 0 (preko S)
      P[i, i] = 1
      }else{
        S[, i] = S[, i, drop = F] / normaS
```

```
        P[i, i] = normaS
      }
    }
    # vrnemo seznam z S in P
    list(S, P)
}

podatki2 = read.csv("primer2.csv")
Z = as.matrix(podatki2[,3:4])
X = as.matrix(podatki2[,2])
dim(Z)
dim(X)

# Z = S * P
S = GS(Z)[[1]]
P = GS(Z)[[2]]

# Preverimo, ce res dobimo Z, ko zmnozimo S in P
z = round(S %*% P, 10)
Z == z # vse OK :)

# se konstrukcija cenilke za beta
# lahko pa bi J dobili tudi preko norm stolpcev od S
J = t(S) %*% S
J = round(J, digits = 10)
# posploseni inverz
inv_splosni = solve(P) %*% J %*% solve(t(P))

# cenilka
beta_hat = inv_splosni %*% t(Z) %*% X
beta_hat
```

3. tretje vaje (20. 10. 2022)

Naloga 4. Linearna regresija

Predpostavimo linearni model

$$x_i = \beta_0 + \beta_1 z_{i1} + \beta_2 z_{i2} + \epsilon_i,\ \ i = 1, \dots, n$$

ϵ_i so neodvisno slušajne spremenljivke, vse porazdeljene po zakonu $N(0, \sigma^2)$. Vrednosti vektorjev X, Z_1, Z_2 so v datoteki *primer2.csv*. Poiščite oceno za $\hat{\beta}$ na tri načine:

1) $\hat{\beta} = (Z^T Z)' Z^T x$
2) $\hat{\beta} = (Z^T Z)^{-1} Z^T x$
3) z uporabo funkcije lm v R

Ali se vrednosti ocen ujemajo?

```
data = read.csv("primer2.csv")

X = as.vector(data[,2])
Z = as.matrix(data[,3:4])
# treba dodat se stolpec enic
z = cbind(1, Z)

beta_hat = solve(t(z) %*% z) %*% t(z) %*% X

model1 = lm(x ~ z1 + z2, data=data[,2:4])
model1$coefficients
```

Naloga 5. ANOVA z enojno klasifikacijo

Obravnavamo m neodvisnih skupin, v vsaki od njih ($1 \leq j \leq m$) imamo n_j neodvisnih meritev, za katere privzamemo, da so porazdeljene kot $N(\mu + \alpha_j, \sigma_j^2)$. Mislimo si, da so α_j majhna števila. Pišimo $n = sum_{j=1}^{m} n_j$ in naj bo

$$X_i = \mu + \alpha_j + \epsilon_i,\ \ \epsilon_i \sim N(o, \sigma_j^2) \text{ za } \sum_{k=1}^{j-1} n_k + 1 \leq i \leq \sum_{k=1}^{j} n_k$$

S primerno izbiro matrike Z lahko zgornji model zapišemo kot splošni linearni model.

Naj bo $m = 4, \mu = 2, (\alpha_1, \alpha_2, \alpha_3, \alpha_4) = (0.2, -0.5, 1, 2), (\sigma_1, \sigma_2, \sigma_3, \sigma_4) = (0.05, 1, 0.03, 0.08)$ in $(n_1, n_2, n_3, n_4) = (25, 30, 25, 27)$.

Simuliraj vrednosti X_i; konstruiraj primerno matriko Z ter preko splošnega inverza matrike Z poišči ocene za $\mu, \alpha_1, \alpha_2, \alpha_3, \alpha_4$.

```
  m = 4
mi = 2
alpha = c(0.2, -0.5, 1, 2)
sigma = c(0.05, 1, 0.03, 0.08)
n = c(25, 30, 25, 27)

# Z = matrix(0, sum(n), m)
# z = cbind(1, Z)
Z = matrix(0,0,m)

for (i in length(n):1){
  nov_n = n[i]
  matrika = matrix(0, nov_n, m)
  nov_stolpec = as.vector(matrix(1, nov_n, 1))
  matrika[,i] = nov_stolpec
  print(matrika)
  Z = rbind(matrika, Z)
}

Z = cbind(1, Z)

vektor = c(mi, alpha)

Z %*% vektor

X = Map(function(x,y,z) rnorm(n=x, mean=y, sd = z), n, mi + alpha, sigma)
# library(dplyr)
# X = X %>% unlist() %>% as.matrix()
```

4. Četrte vaje (27. 10. 2022)

Izrek 1. Gauss-Markov
Naj za $\vec{X} = Z\vec{\beta} + \vec{\epsilon}$ velja $E(\epsilon) = 0$ in $\mathrm{var}(\vec{\epsilon}) = \sigma^2 I$. Tedaj ima cenilka $UZ\vec{\beta}$ med vsemi linearnimi nepristranskimi cenilkami za $UZ\vec{\beta}$ enakomerno najmanjšo varianco.

Naloga 6. Privzemimo model linearne regresije $Y = X\beta + \epsilon$ kjer $E(\epsilon) = 0$ in $\mathrm{var}(\epsilon) = \sigma^2 V$
Velja

$$v_{ij} = \frac{p^{|i-j}}{1-p^2}$$

kjer je σ^2 neznana konstanta, $p \in (-1, 1)$ pa znana.
Naj bodo komponente Z dane z

$$Z_1 = \sqrt{1-p^2} y_1 \text{ in } Z_i = Y_i - p y_{i-1}$$

Izračunaj $\mathrm{var}(Z_i)$ in $\mathrm{cov}(Z_i, Z_j)$ za $i \neq j$ in poišči najboljšo linearno nepristransko cenilko za β.

5. pete vaje - odpadejo (3. 11. 2022)

6. Šeste vaje (10. 11. 2022)

7. sedme vaje (17. 11. 2022)

8. osme vaje (24. 11. 2022)

9. devete vaje (1. 12. 2022)