Statistička analiza podataka – završni ispit

UNIZG FER, ak. god. 2021./2022.

27.1.2022.

Ispit traje 120 minuta i nosi 30 bodova. Svaki zadatak rješavajte na zasebnoj stranici. Pišite uredno i čitko – rješenja koja ispravljači ne mogu pročitati neće se bodovati.

1. (6 bodova) Dani su sljedeći podatci:

x	2	7	12	20
y	7	12	13	14

   (a) (3 boda) Prilagodite model linearne regresije $Y = \beta_0 + \beta_1 x + \beta_2 x^2$ danim podacima, tj. izračunajte regresijske koeficijente ako je dana matrica:

$$A^{-1} = (X^\top X)^{-1} = \begin{bmatrix} 2.05 & -0.4 & 0.02 \\ -0.4 & 0.1 & 0.0 \\ 0.02 & 0.0 & 0.0 \end{bmatrix}$$

   Pri tome jasno napišite matricu dizajna X i navedite kako pronalazimo regresijske koeficijente $\beta = [\beta_0, \beta_1, \beta_2]^\top$ u matričnom obliku. Koja je predikcija modela za $x = 2$?

   (b) (3 boda) Navedite pretpostavke modela linearne regresije i pojasnite kako grafički provjeravamo svaku od tih pretpostavki (jasno napišite što je na x i y osi te na što obraćamo pozornost u samom grafu).

2. (6 bodova) Banka želi napraviti model za predviđanje vjerojatnosti da će klijent kasniti s otplatom kredita. Koristeći podatke o klijentima procijenjeni su parametri modela logističke regresije. Dobiven je model:

$$P(Y = 1|x) = \frac{1}{1 + e^{1+7x_1-2x_2}},$$

   gdje je Y zavisna slučajna varijabla ($Y = 1$ ako klijent kasni s otplatom, a $Y = 0$ ako klijent redovito podmiruje obveze), a $x = [x_1, x_2]^\top$ je vektor nezavisnih varijabli signifikantnih na razini značajnosti od 5%.

   (a) (2 boda) Odredite $\frac{\partial P(Y=1|x)}{\partial x_1}$ te na temelju toga zaključite hoće li povećanje varijable x_1 smanjiti ili povećati vjerojatnost kašnjenja u otplati. Objasnite zaključak.

   (b) (2 boda) Ako se nezavisna varijabla x_1 poveća za 15%, možemo li odrediti koliko će se promijeniti zavisna varijabla? Ovisi li ta promjena zavisne varijable o razinama na kojima se nalaze x_1 i x_2 ili ne? Obrazložite.

   (c) (2 boda) Koja nezavisna varijabla ima veći utjecaj na zavisnu varijablu? Možemo li kvantificirati taj utjecaj i ovisi li on o razinama na kojima se nalaze x_1 i x_2? Obrazložite.

3. (6 bodova) Provedeno je istraživanje o povezanosti izmedu aktivnog bavljenja sportom i ocjena na četvrtoj godini studija. Razmatrane su tri gupe studenata: (a) oni koji se aktivno bave nogometom, (b) oni koji se aktivno bave nekim drugim sportom, a nije nogomet i (c) oni koji se ne bave aktivno sportom. Odabran je slučajan uzorak od po 10 studenata za svaku grupu te su dobiveni rezultati u tablici ispod.

	Nogomet	Ostali sportovi	Ne bave se sportom
Uzoračka sredina	4	4.1	3.9
Uzoračka standardna devijacija	0.12	0.15	0.17

   (a) (4 bodova) Odredite postoji li razlika u prosjeku ocjena za tri grupe na razini značajnosti $\alpha = 0.05$, uz pretpostavku da su ocjene normalno distribuirane te da su standardne devijacije jednake za sve tri grupe.

   (b) (2 boda) Navedite primjer dva ortogonalna kontrasta koje biste mogli primijeniti na dane podatke i odgovarajuće hipoteze koje biste njima testirali u ovom slučaju.

4. (6 bodova) Za uspješno polaganje jednog predmeta na fakultetu potrebno je riješiti dvije laboratorijske vježbe. U tablici su dani bodovi 5 studenata.

Student	1. vježba	2. vježba
D.B.	16	17
S.B.	19	16
T.B.	10	10
T.K.	9	11
A.M.	12	13

   (a) (2 boda) Izračunajte Spearmanov koeficijent korelacije. Što zaključujete iz njega?

   (b) (2 bod) Pretpostavite da imate veći uzorak ($n > 30$). Postavite hipoteze i prikladnu testnu statistiku za dvostrani test koreliranosti.

   (c) (2 bod) Što je glavna prednost, a što mana neparametarskih testova naspram parametarskih? Nabrojite dva neparametarska testa i njihove parametarske inačice.

5. (6 bodova) S kolegama u pauzi od predavanja kupujete kavu na aparatu koji se nalazi pored glavnog ulaza FER-a. Apriorna distribucija proporcije prolivenih kava p (zbog nedostatka čaša) na automatu je sljedeća:

p	0.01	0.05	0.1
$\pi(p)$	0.2	0.3	0.5

   Ako se dvije od pet narednih kava proliju u nepovrat:

   (a) (4 boda) Izračunajte aposteriornu distribuciju proporcije prolivenih kava p.

   (b) (2 boda) Izračunajte procjenu parametra p koristeći Bayesovski pristup.

Rješenja zadataka

1. (a) Matrica dizajna X:

$$X = \begin{bmatrix} 1 & 2 & 4 \\ 1 & 7 & 49 \\ 1 & 12 & 144 \\ 1 & 20 & 400 \end{bmatrix}$$

Regresijski koeficijenti u matričnom obliku:

$$\begin{aligned} \beta &= \left(X^\intercal X\right)^{-1} X^\intercal y \\ &= \begin{bmatrix} 2.05 & -0.4 & 0.02 \\ -0.4 & 0.1 & -0.0 \\ 0.02 & 0.0 & 0.0 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 2 & 7 & 12 & 20 \\ 4 & 49 & 144 & 400 \end{bmatrix} \begin{bmatrix} 7 \\ 12 \\ 13 \\ 14 \end{bmatrix} \\ &= \begin{bmatrix} 42.46 \\ 35.00 \\ 0.92 \end{bmatrix} \end{aligned}$$

Predikcija modela u slučaju $x = 2$ je 116.14.

(b) Pretpostavke: reziduali $\epsilon_i \sim \mathcal{N}\left(0, \sigma^2\right)$ i nezavisni $\forall i$.

- Nezavisnost: rezidualni graf (na x-osi se nalaze predviđene vrijednosti modela; na y-osi se nalaze reziduali modela; promatramo uočavamo li kakvu pravilnost / uzorak)
- Normalnost: Q-Q plot (na x-osi se nalaze teoretski kvantili normalne slučajne varijable sa parametrima 0 i 1; na y-osi se nalaze uzorački kvantili iz danog skupa podataka; želimo da i-ti uzorački kvantil i i-ti teoretski kvantil prate pravac

$$y = \hat{\mu} + \hat{\sigma} x.$$

- Homogenost: rezidualni graf (na x-osi se nalaze predviđene vrijednosti modela; na y-osi se nalaze reziduali modela; promatramo raste li varijanca reziduala sa rastom predviđenih vrijednosti)

2. (a) Uz $\Lambda(x^\intercal \beta) = \dfrac{1}{1 + e^{-x^\intercal \beta}}$ derivacija je jednaka

$$\frac{\partial P(Y = 1|x)}{\partial x_1} = \Lambda(x^\intercal \beta) \cdot (1 - \Lambda(x^\intercal \beta)) \cdot \beta_1 = -7 \cdot \Lambda(x^\intercal \beta)(1 - \Lambda(x^\intercal \beta))$$

Budući da vrijedi $\Lambda(x^\intercal \beta) \in (0, 1)$, član $\Lambda(x^\intercal \beta) \cdot (1 - \Lambda(x^\intercal \beta))$ je pozitivan, ali zbog množenja s $\beta_1 = -7$ derivacija poprima negativnu vrijednost. Stoga možemo ustvrditi da će povećanje x_1 smanjiti izlaznu varijablu.

(b) Utjecaj promjene x_1 na zavisnu varijablu dan je derivacijom $\Lambda(x^\intercal \beta) \cdot (1 - \Lambda(x^\intercal \beta)) \cdot \beta_1$ koja je izračunata u prethodnom podzadatku. Ipak, ne možemo odrediti koliko će se promijeniti zavisna varijabla jer ta promjena ovisi i o samom vektoru x, odnosno razinama na kojima se nalazi x_1 i x_2.

(c) Odnos utjecaja parametara

$$\frac{\frac{\partial P(Y=1|x)}{\partial x_1}}{\frac{\partial P(Y=1|x)}{\partial x_2}} = \frac{\Lambda(x^\intercal\beta)\cdot(1-\Lambda(x^\intercal\beta))\cdot\beta_1}{\Lambda(x^\intercal\beta)\cdot(1-\Lambda(x^\intercal\beta))\cdot\beta_2} = \frac{\beta_1}{\beta_2} = \frac{-7}{2}.$$

Možemo zaključiti da varijabla x_1 ima veći utjecaj na zavisnu varijablu. Odnos utjecaja pojedinih parametara možemo kvantificirati samim odnosom parametara, i na njega ne utječu razine na kojima se nalaze x_1 i x_2.

3. a) $N = 10$, $k = 3$
$SSA = 10\cdot[(4-4)^2 + (4.1-4)^2 + (3.9-4)^2] = 0.2$, uz $k-1 = 2$ stupnja slobode
$SSE = 9\cdot(0.12^2+0.15^2+0.17^2) = 0.5922$ (priznavalo se i rješenje sa pristranim estimatorom koji množi varijance s 10), uz $k(n-1) = 27$ stupnjeva slobode
$s_1^2 = 0.1$, $s^2 = 0.0219$, $f = 4.559$, kritična vrijednost: $f_{\alpha=0.05}(2,27) = 3.35$
$f > f_{\alpha=0.05}(2,27)$ - odbacujemo H_0.

b) Primjer dva ortogonalna kontrasta:
- $w_1 = \mu_1 + \mu_2 - 2\mu_3$, ispituje $H_0 : \mu_1 + \mu_2 - 2\mu_3 = 0$
- $w_2 = \mu_1 - \mu_2$, ispituje $H_0 : \mu_1 - \mu_2 = 0$

4. a)

Student	1. vježba – rang	2. vježba – rang	d
D.B.	4	5	-1
S.B.	5	4	1
T.B.	2	1	1
T.K.	1	2	-1
A.M.	3	3	0

$$r_s = 1 - \frac{6\cdot 4}{5\cdot(25-1)} = 0.8$$

Zaključujem da postoji velika korelacija između uspjeha na prvoj i uspjeha na drugoj laboratorijskoj vježbi.

b) Hipoteze su

$$H_0 : \rho = 0$$
$$H_1 : \rho \neq 0$$

Budući da za $n > 30$ vrijedi $r_s \sim \mathcal{N}(0, \frac{1}{n-1})$, testna statistika je:

$$z = \frac{r_s - 0}{\frac{1}{\sqrt{n-1}}} = r_s\sqrt{n-1}$$

c) Prednost je da su primjenjivi kada je distribucija populacije nepoznata, mana da su slabije statističke snage od odgovarajućih parametarskih testova. Kruskal-Wallisov test (ANOVA), Wilcoxonov test (t-test), Mann-Whitney-Wilcoxonov test (t-test za nezavisne uzorke), ...

5. (a) Aposteriorna distribucija parametra p, uz dani x definirana je kao:

$$\pi(p|x) = \frac{f(x|p)\pi(p)}{g(x)}.$$

Funkciju izglednosti modeliramo binomnom distribucijom (gdje je k=5, a x=2):

$$f(x|p) = \binom{k}{x} p^x (1-p)^{k-x}.$$

Za izračun aposteriorne distribucije potrebna nam je i vjerojatnost od x:

$$g(x) = \sum_p f(x|p)\pi(p).$$

Funkciju izglednosti računamo za diskretno definiranu apriornu distribuciju parametra p kao:

$$\begin{aligned}
\text{Za } p = 0.01 &\rightarrow f(x|p_1) = \binom{5}{2} \cdot 0.01^2 \cdot 0.99^3 = 0.00097 \\
\text{Za } p = 0.05 &\rightarrow f(x|p_2) = \binom{5}{2} \cdot 0.05^2 \cdot 0.95^3 = 0.027 \\
\text{Za } p = 0.1 &\rightarrow f(x|p_3) = \binom{5}{2} \cdot 0.1^2 \cdot 0.9^3 \quad = 0.073
\end{aligned}$$

$$\begin{aligned}
g(x) &= f(x|p=0.01) \cdot \pi(p=0.01) + f(x|p=0.05) \cdot \pi(p=0.05) + f(x|p=0.1) \cdot \pi(p=0.1) \\
&= 0.043
\end{aligned}$$

Uvrštavajući izračunate izraze u formulu za aposteriornu distribuciju dobivamo:

$$\begin{aligned}
\pi(p_1|x=2) &= \frac{0.2 \cdot 0.00097}{0.043} \approx 0 \\
\pi(p_2|x=2) &= \frac{0.3 \cdot 0.0.021}{0.043} \approx 0.15 \\
\pi(p_3|x=2) &= \frac{0.5 \cdot 0.073}{0.043} \approx 0.85
\end{aligned}$$

p	0.01	0.05	0.1
$\pi(p)$	0	0.15	0.85

(b) Računamo očekivanje aposteriorne distribucije (procjena parametra p):

$$E[\pi(p|x)] = \sum p \cdot \pi(p|x) = 0.01 \cdot 0 + 0.05 \cdot 0.15 + 0.1 \cdot 0.85 = 0.0925.$$

Priznaju se i drugi točno definirani točkasti procjenitelji za 1 bod, a za još 1 bod točno izračunati točkasti procjenitelj za aposteriornu distribuciju parametra p.