4. Domaća zadaća

STATISTIČKI PRAKTIKUM 2

Zadatak 7. MLE1

Promatramo slučajni uzorak $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$, gdje je $\mu \in [0, +\infty)$. Procjenitelj maksimalne vjerodostojnosti za μ je

$$\hat{\mu}_n = \max\{\overline{X_n}, 0\}.$$

(a) Generirajte uzorak duljine 50 iz ovog modela za $\mu = 0$ i $\sigma^2 = 4$. Odredite pravu distribuciju procjenitelja $\hat{\mu}_n$.

(b) Procijenite distribuciju od $\hat{\mu}_n$ parametarskim i neparametarskim bootstrapom (B=500) te ih usporedite s pravom distribucijom. Koja procjena je bolja i zašto?

(c) Usporedite procijenjeno očekivanje i varijancu od $\hat{\mu}_n$ s pravim vrijednostima u oba slučaja.

(d) Ponovite ovu analizu za niz duljine 100 iz ovog modela, ovog puta za $\mu = 5$. Ima li razlike u zaključcima?

(15 bodova)

Zadatak 8. MLE2

Promatramo slučajni uzorak $X_1, \ldots, X_n \sim N(3, 2^2)$ te koeficijent spljoštenosti θ. Pronađite MLE procjenitelj za θ.

(a) Generirajte uzorak duljine $n = 100$.

(b) Parametarskim i neparametarskim bootstrapom (B=1000) procijenite očekivanje procjenitelja $\hat{\theta}_n$.

(c) Izračunajte normalni, osnovni, percentilni i BC 98% pouzdani interval za $\hat{\theta}_n$ u oba slučaja.

(d) Ponovite korake (a)-(c) 1000 puta i prikažite tablično za sva 4 tipa intervala pouzdanosti postotak intervala koji sadrže pravu vrijednost parametra θ.

(e) Na temelju dobivenih procjena u (b) dijelu, sami konstruirajte normalni i osnovni 80% pouzdani interval za očekivanje procjenitelja $\hat{\theta}_n$.

$(f)^*$ Uzmite jedan uzorak iz ove razdiobe duljine $n = 20$. Bootstrap metodom testirajte hipotezu da je koeficijent $\theta = 3$, nasuprot alternativi da je $\theta > 3$.

Napomena: podzadatak (f) ne morate riješiti, a nosi dodatnih 5 bodova.
$(10 + 5^*$ bodova)

Zadatak 9. Pušenje i smrtnost

Promatramo podatke `Whickam` iz paketa `mosaicData`. Za 1314 žena u Velikoj Britaniji, u periodu od 1972-1974 uzeta je informacija o tome je li pušač ili ne (kovarijata `smoker`) te koliko ima godina (age). Dvadeset godina nakon provjereno je jesu li još uvijek žive (odziv `outcome`). Želimo ispitati utjecaj pušenja na smrtnost koristeći generalizirani linearni (logistički) model.

(a) Prilagodite logistički model za dane podatke koristeći samo kovarijatu smoker. Interpretirajte dobivene koeficijente. Jesu li rezultati očekivani?

(b) Podijelite kovarijatu age u 5 grupa (primjerice koristeći naredbu `cut`). Izračunajte relativne frekvencije odziva, te kovarijate smoker, u odnosu na tako grupirane godine. Kako godine utječu na smrtnost dvadeset godina nakon? Kakva je proporcija pušača ovisno o godinama? Možete li sada objasniti što se dogodilo u (a)?

(c) Prilagodite logistički model za dane podatke koristeći obje kovarijate. Napišite model te interpretirajte dobivene koeficijente. Jesu li sada rezultati intuitivniji? Testirajte statističku značajnost kovarijate smoker na razini značajnosti 5% tako što ćete usporediti puni model i model u kojem je ta kovarijata ispuštena, koristeći test omjera vjerodostojnosti.

(d) Koja je vjerojatnost da žena koja je 1972. godine imala 45 godina i koja je bila pušač nije umrtla do 1974. godine?

(15 bodova)

Zadatak 10. Politika i potrošnja

U datoteci `expenditure.csv` dani su podaci prikupljeni pri istraživanju utjecaja različitih faktora (`Drzava` Ujedinjenog Kraljevstva, `Godina` prikupljanja podataka, `Dob`, `Spol`, `Prihod` subjekta unutar te godine, `Obrazovanje` koje je subjekt završio (osnovna škola - `ES`, srednja škola - `HS`, prvostupnik - `BC`, magisterij - `MS`, doktorat - `PhD`), `Broj_clanova_obitelji`) na varijablu `Potrosnja` koja označava iznos koji subjekt potroši godišnje na gorivo za automobil, u različitim državama Ujedinjenog Kraljevstva tijekom nekoliko godina. Varijabla `Skupina` označava pripadnost subjekta jednoj političkoj stranci Ujedinjenog Kraljevstva. Zanima nas postoji li značajna razlika u potrošnji obzirom na pripadnost subjekta političkoj stranci, uzimajući u obzir ostale karakteristike subjekta koje utječu na vjerojatnost njegovog učlanjenja u tu političku stranku. Analizirajte to na temelju dobivenih podataka i statistički testirajte svoje tvrdnje.
(10 bodova)

Zadatak 11. Cornflakes

U datoteci `cereal_new.csv` nalazi se 77 vrsta žitnih pahuljica (imena u stupcu `name`), oznaka njihovog proizvođača (stupac `mfr`) te prosječna ocjena, na skali 1-100, koju je skupina ispitanika dodijelila svakoj vrsti pahuljica (stupac `rating`).

(a) Odredite 90% pouzdani interval za 1. kvartil ocjena žitarica proizvođača s oznakom `K`, a zatim na razni značajnosti od 10% testirajte hipotezu da je barem 25% ispitanika bilo jako nezadovoljno žitaricama proizvođača s oznakom `K` i dalo im ocjenu manju od 35.

(b) Usporedite prosječne i medijalne ocjene za žitarice pojedinog proizvođača. Testirajte postoji li razlika između ocjena za žitarice različitih proizvođača ili sve te ocjene slijede istu distribuciju.

(c) Sada usporedite samo ocjene za žitarice proizvođača s oznakom `K` i onoga s oznakom `G`. Testirajte dolaze li te ocjene iz iste distribucije ili jedan od proizvođača ima u prosjeku više ocjene.

(d) U datoteci se nalazi i stupac `after` koji sadrži prosječnu ocjenu koju je skupina ispitanika dala pojedinoj vrsti žitarice nakon što su odradili zahtjevnu tjelovježbu. Testirajte nekim neparametarskim testom postoji li značajna razlika u ocjenama prije i nakon tjelovježbe.

(15 bodova)