4. Domaća zadaća

STATISTIČKI PRAKTIKUM 2

Zadatak 8. Bootstrap

Promatramo slučajni uzorak $X_1, \ldots, X_n \sim U(0, \theta)$ i procjenitelj za θ

$$\hat{\theta}_n = \max_{1 \leq i \leq n} \{X_i\} = X_{(n)}.$$

(a) Generirajte uzorak duljine 50 iz ovog modela za $\theta = 0.5$. Odredite pravu distribuciju procjenitelja $\hat{\theta}_n$.

(b) Procijenite distribuciju od $\hat{\theta}_n$ parametarskim i neparametarskim bootstrapom (B=1000) te ih usporedite s pravom distribucijom. Koja procjena je bolja i zašto?

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

(d) Ponovite ovu analizu za niz duljine 200 iz ovog modela, ovog puta za $\theta = 50$. Ima li razlike u zaključcima?

(10 bodova)

Zadatak 9. Bootstrap pouzdani intervali

Promatramo slučajni uzorak $X_1, \ldots, X_n \sim Exp(2)$ te koeficijent asimetrije θ. Pronađite procjenitelj za θ metodom momenata.

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

(b) Parametarskim i neparametarskim bootstrapom (B=500) procijenite očekivanje i varijancu procjenitelja $\hat{\theta}_n$.

(c) Izračunajte normalni, osnovni, percentilni i BC 90% 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 percentilni 90% pouzdani interval za očekivanje procjenitelja $\hat{\theta}_n$. Što iz toga zaključujete?

$(f)^*$ Uzmite jedan uzorak iz ove razdiobe duljine $n = 15$. **Bootstrap metodom** testirajte hipotezu da je koeficijent $\theta = 0$, nasuprot alternativi da je $\theta > 0$.

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

Zadatak 10. GLM

U datoteci `GLM.csv` dani su podaci o *defaultu* (prvi stupac), tj. neispunjavanju obaveza od strane klijenata banke po jednoj vrsti kredita u protekloj godini. Osim ove varijable, dostupne su još varijable (kovarijate) o spolu, županiji (21 vrijednost), trajanju radnog odnosa (u godinama), iznosu kredita (u eurima), udjelu rate kredita u mjesečnim primanjima pojedinog klijenta (u %) i broju uzdržavanih članova obitelji.

(a) Odredite relativne frekvencije defaulta po spolu, odnosno županiji.

(b) Ispitajte ovisnost defaulta o županiji, a zatim o spolu, prikladnim statističkim testom. Koja od te dvije varijable će snažnije utjecati na vjerojatnost defaulta klijenta?

(c) Grafički (možete koristiti npr. "kutije s brkovima" i stupčaste dijagrame, ali i razne druge grafičke prikaze) i inferencijalno, jednim parametarskim i jednim neparametarskim testom, ispitajte ovisnost defaulta o trajanju radnog odnosa.

(d) Prilagodite GLM za ove podatke koristeći logističku funkciju veze.

(e) Grafički prikažite ovisnost vjerojatnosti defaulta o broju uzdržavanih članova obitelji u tom modelu, a na istom grafu prikažite i dio podataka.

(f) Grafički prikažite i vjerojatnost defaulta u ovisnosti o iznosu kredita u tom modelu. Ponovno na istom grafu prikažite i dio podataka. Zatim ispitajte ovisi li taj omjer o županiji, prvo grafički, a potom i inferencijalno.

(g) Koristeći analizu devijance odredite "optimalni" model za podatke (ignorirajte interakcije). Navedite model koji ste odabrali te interpretirajte njegove parametre.

(h) Ima li u modelu utjecajnih točaka?

(i) Možemo li uklanjanjem nekih prediktora iz modela značajno poboljšati prediktivnu snagu modela?

(15 bodova)

Zadatak 11. Račun za mobitel

Tvrtku koja pruža telekomunikacijske usluge zanima utječe li *default* (nemogućnost pokrivanja kredita u banci) njihovih klijenata na iznos računa za mobitel. U tu svrhu, iskoristili su podatke iz datoteke `GLM.csv` i za te iste ljude uzeli prosječan iznos njihovih računa u toj godini: ti podaci nalaze se u datoteci `telefon.csv`.

Prvo grafički ispitajte je li default utjecao na iznos računa. Zatim taj odnos ispitajte prikladnim statističkim testom, odnosno ispitajte efekt defaulta na iznos računa na temelju logističkog modela koji ocjenjuje vjerojatnost defaulta (kao varijable poticaja koje utječu na vjerojatnost defaulta uzmite one varijable koje ste odabrali u (g) dijelu iz Zadatka 10.)

(5 bodova)

Zadatak 12. Udio rate kredita

Proučavamo podatke iz datoteke `GLM.csv` koji se odnose na udio rate kredita u mjesečnim primanjima i broj udržavanih članova u obitelji.

1. Grafički ispitajte postoji li veza između udjela rate kredita u primanjma i broju uzdržavanih članova.
2. Postavljenu hipotezu na temelju grafičke analize ispitajte prikladnim parametarskim testom. Koje su pretpostavke testa koji koristite? Jesu li zadovoljene?
3. Konstruirajte randomizacijski test kojim ćete testirati postoji li veza između ove dvije varijable.

(10 bodova)

Zadatak 13. Pismenost

Profesori hrvatskog jezika žele ispitati novu metodu za učenje djece pismenosti. U tu svrhu, 20 djece pisalo je ispit prije i nakon što su gradivo odslušali novom metodom. Na svakom ispitu su ukupno mogli ostvariti 10 bodova, a rezultati su dani u datoteci `pismenost.txt`.

Možemo li reći da se pismenost djece popravila nakon što su gradivo učili novom meodom? Možete li predložiti neku hipotezu na temelju grafičke analize? Testirajte svoju hipotezu prikladnim neparametarskim testom i donesite zaključke.

(5 bodova)

Zadatak 14. Rezultati trčanja u školi

4 škole sudjelovale su na državnom natjecanju u trčanju. Iz svake škole sudjelovalo je 6 učenika osmih razreda. Rezultati natjecanja (vrijeme u sekundama na 100m) dani su u datoteci `skole.txt`.

(a) Na temelju podataka svih učenika, odredite 90% pouzdani interval za medijan vremena potrebnog da se istrči 100m dvjema neparametarskim metodama. Pomoću tih rezultata testirajte hipotezu da su naši učenici u prosjeku brži od europskih kolega, ako znamo da je europski prosjek u školama prošle godine iznosio 14.4s na 100m. Zatim isto ispitajte jednom parametarskom metodom. Koje su pretpostavke testova koje koristite? Jesu li zadovoljene?

(b) Grafički usporedite rezultate učenika po školama. Postoji li razlika u rezultatima između škole koju pohađa pobjednik natjecanja i ostalih škola? Testirajte tu hipotezu jednom parametarskom i jednom neparametarskom metodom. Koje su pretpostavke testova koje koristite? Jesu li zadovoljene?

(10 bodova)

Zadatak 15.

Provedeno je istraživanje o utjecaju vrste glazbe na stres čovjeka. Podaci su dani u datoteci `stres.txt`. U istraživanju je sudjelovalo 50 ljudi identificiranih brojevima od 1 do 50 (stupac `ID`), prvih 20 ljudi su bile žene, a ostalo muškarci. Svaka osoba je odslušala 20 različitih skladbi podijeljenih u 3 skupine (stupac `glazba`), a nakon svake skladbe joj je izmjerena razina stresa izražena brojevima od 1 do 100 (stupac `stres`).

Grafički i odgovarajućim statističkim testom procijenite utjecaj vrste glazbe na čovjekov stres.

Zatim u model dodajte i varijablu o spolu i provjerite utječe li spol na čovjekov stres. Rezultate potkrijepite nekim grafičkim prikazom.

(5 bodova)