KOLOKVIJ 2 IZ RAČUNARSKE STATISTIKE

PMF-Matematika, Školska godina 2009/2010

Predavač: Doc.dr.sc. Vesna Lužar-Stiffler

Datum: _____

Ime, Prezime:_____

Potpis: _____

1. Sedam muškaraca i sedam žena je polazilo tečaj golfa. Na kraju tečaja je za svakog učesnika određen broj bodova (varijabla SCORE). Podaci su zapisani u dataset SCORES:

```
data scores;
   input Gender $ Score @@;
   datalines;
f 75  f 76  f 80  f 77  f 80  f 77  f 73
m 76  m 80  m 85  m 85  m 78  m 87  m 82
;
run;
```

   a) Testitajte hipotezu H0: $\mu_m = \mu_f$ (da je broj bodova jednak bez obzira na spol) nasuprot H1: $\mu_m > \mu_f$ (da je broj bodova muškaraca veći od broja bodova žena) na razini statističke značajnosti $\alpha = 0.01$ primjenom t testa za 2 nezavisna uzorka.

      UPUTA: Koristite proceduru TTEST. Prvo testirajte hipotezu o jednakosti varijanci, te u zavisnosti od ishoda, primijenite ili „pooled" ili „Satterthwaite" test za hipotezu H0: $\mu_m = \mu_f$. Može li se hipoteza H0 odbaciti?

   b) Testitajte hipotezu H0: $\mu_m = \mu_f$ nasuprot H1: $\mu_m > \mu_f$ na razini statističke značajnosti $\alpha = 0.01$ primjenom bootstrap testa.

      UPUTA: Koristite metodu B bootstrap uzorkovanja za testiranje hipoteza. Primijenite slijedeće vrijednosti makro varijabli SEED:
      %let seed=633921; * za grupu 1 (Gender=m);
      %let seed= 47558; * za grupu 2 (Gender=f);

      Za svaku grupu izvedite 1000 bootstrap ponavljanja (replikacija).

      Može li se hipoteza H0 odbaciti?

2. U datasetu FITNESS zapisani su rezultati mjera kondicije za 31 člana fitness kluba. Modelom višestruke linearne regresije treba predvidjeti vrijednosti varijable OXYGEN (kapacitet pluća, zavisna varijabla) na osnovu preostalih (prediktorskih, nezavisnih) varijabli, te odrediti vrijednost prilagođenog koeficijenta determinacije (engl. adjusted R2).
   a) Neparametarskom bootstrap metodom odredite 90% interval pouzdanosti za prilagođeni koeficijent determinacije. Koristite percentilnu metodu za određivanje bootstrap intervala pouzdanosti(engl. Bootstrap pctl method for confidence interval). Koristite 1000 bootstrap ponavljanja.
   b) Neparametarskom bootstrap metodom odredite 90% interval pouzdanosti za drugi korijen iz varijance pogreške (engl. root mean square error). Koristite percentilnu metodu za određivanje bootstrap intervala pouzdanosti(engl. Bootstrap pctl method for confidence interval). Koristite 1000 bootstrap ponavljanja.

   U a) i b) koristite slijedeću vrijednost seed makro varijable:
   %let seed=744956;

```
data fitness;
      input Age Weight Oxygen RunTime RestPulse RunPulse MaxPulse @@;
   datalines;
   44 89.47 44.609 11.37 62 178 182   40 75.07 45.313 10.07 62 185 185
   44 85.84 54.297  8.65 45 156 168   42 68.15 59.571  8.17 40 166 172
   38 89.02 49.874  9.22 55 178 180   47 77.45 44.811 11.63 58 176 176
   40 75.98 45.681 11.95 70 176 180   43 81.19 49.091 10.85 64 162 170
   44 81.42 39.442 13.08 63 174 176   38 81.87 60.055  8.63 48 170 186
   44 73.03 50.541 10.13 45 168 168   45 87.66 37.388 14.03 56 186 192
   45 66.45 44.754 11.12 51 176 176   47 79.15 47.273 10.60 47 162 164
   54 83.12 51.855 10.33 50 166 170   49 81.42 49.156  8.95 44 180 185
   51 69.63 40.836 10.95 57 168 172   51 77.91 46.672 10.00 48 162 168
   48 91.63 46.774 10.25 48 162 164   49 73.37 50.388 10.08 67 168 168
   57 73.37 39.407 12.63 58 174 176   54 79.38 46.080 11.17 62 156 165
   52 76.32 45.441  9.63 48 164 166   50 70.87 54.625  8.92 48 146 155
   51 67.25 45.118 11.08 48 172 172   54 91.63 39.203 12.88 44 168 172
   51 73.71 45.790 10.47 59 186 188   57 59.08 50.545  9.93 49 148 155
   49 76.32 48.673  9.40 56 186 188   48 61.24 47.920 11.50 52 170 176
   52 82.78 47.467 10.50 53 170 172
   ;
run;
```

UPUTA: U PROC REG naredbi je potrebno navesti opciju ADJRSQ za spremanje vrijednosti prilagođenog koeficijenta determinacije (varijabla _ADJRSQ_) u izlazni dataset EST (ako je navedena opcija OUTEST=EST u PROC REG naredbi).
Drugi korijen iz varijance pogreške (varijabla _RMSE_) se automatski sprema u izlazni dataset EST (ako je navedena opcija OUTEST=EST u PROC REG naredbi).

3. Provedite slijedeću Monte Carlo (MC) studiju za ispitivanje pogrešaka krosvalidacije linearne diskriminativne analize:

a) Generirajte ponavljano (nrep=500) bivarijatne uzorke za grupu 1 (dataset NORMAL1) veličine n=100 sa slijedećim karakteristikama:
varijabla x slijedi normalnu distribuciju sa sredinom 0 i varijancom 1,
varijabla y slijedi normalnu distribuciju sa sredinom 0 i varijancom 1,
varijable x i y potiču iz populacije sa korelacijskim koeficijentom $\rho_{xy} = 0.40$
Koristite slijedeću vrijednost seed makro varijable:
%let seed=1235;

Generirajte ponavljano (nrep=500) bivarijatne uzorke za grupu 2 (dataset NORMAL2) veličine n=100 sa slijedećim karakteristikama:
varijabla x slijedi normalnu distribuciju sa sredinom 1 i varijancom 1,
varijabla y slijedi normalnu distribuciju sa sredinom 1 i varijancom 1,
varijable x i y potiču iz populacije sa korelacijskim koeficijentom $\rho_{xy} = 0.40$
Koristite slijedeću vrijednost seed makro varijable:
%let seed=5786;

Spojite datasetove NORMAL1 i NORMAL2, sortirajte po replikacijskoj varijabli, te za svako ponavljanje provedite linearnu diskriminativnu analizu te izračunajte vrijednosti pogrešaka klasifikacije podataka metodom krosvalidacije u grupu 1 i u grupu 2.

Izračunajte MC prosječne vrijednosti i standardne pogreške pogrešaka klasifikacije podataka metodom krosvalidacije u grupu 1 i u grupu 2.

b) Generirajte ponavljano (nrep=500) bivarijatne uzorke za grupu 1 (dataset NORMAL1) veličine n=100 sa slijedećim karakteristikama:
varijabla x slijedi normalnu distribuciju sa sredinom 0 i varijancom 1,
varijabla y slijedi normalnu distribuciju sa sredinom 0 i varijancom 1,
varijable x i y potiču iz populacije sa korelacijskim koeficijentom $\rho_{xy} = 0.40$
Koristite slijedeću vrijednost seed makro varijable:
%let seed=1235;

Generirajte ponavljano (nrep=500) bivarijatne uzorke za grupu 2 (dataset NORMAL2) veličine n=100 sa slijedećim karakteristikama:
varijabla x slijedi normalnu distribuciju sa sredinom 2 i varijancom 1,
varijabla y slijedi normalnu distribuciju sa sredinom 2 i varijancom 1,
varijable x i y potiču iz populacije sa korelacijskim koeficijentom $\rho_{xy} = 0.40$
Koristite slijedeću vrijednost seed makro varijable:
%let seed=5786;

Spojite datasetove NORMAL1 i NORMAL2, sortirajte po replikacijskoj varijabli, te za svako ponavljanje provedite linearnu diskriminativnu analizu i izračunajte vrijednosti pogrešaka klasifikacije podataka metodom krosvalidacije u grupu 1 i u grupu 2.

Izračunajte MC prosječne vrijednosti i standardne pogreške pogrešaka klasifikacije podataka metodom krosvalidacije u grupu 1 i u grupu 2.

Što uočavate?

UPUTA: Ispred PROC DISCRIM navedite naredbu:
ods output ErrorCrossVal=ErrorCrossVal ;

(za spremanje pogrešaka klasifikacije metodom krosvalidacije u izlazni dataset ErrorCrossVal).

4. Koristite podatke iz zadatka 1.

a) Parametarskom bootstrap metodom odredite bootstrap standardnu pogrešku i 90% bootstrap interval pouzdanosti za apsolutnu vrijednost razlike između prosječnih vrijednosti broja bodova za muškarce i prosječnih vrijednosti broja bodova za žene.

a) Parametarskom bootstrap metodom odredite bootstrap standardnu pogrešku i 90% bootstrap interval pouzdanosti za apsolutnu vrijednost razlike između medijana broja bodova za muškarce i medijana broja bodova za žene.

U a) i b):

Primijenite slijedeće vrijednosti makro varijabli SEED:
%let seed=633921; * za grupu 1 (Gender=m);
%let seed= 47558; * za grupu 2 (Gender=f);

Pretpostavite da varijabla SCORE slijedi normalnu distribuciju.

Za svaku grupu izvedite 1000 bootstrap ponavljanja (replikacija).