VLADIMIR VASIĆ[1]
E-mail: vladimir@ekof.bg.ac.rs

REŠAVANJE PROBLEMA MULTIVARIJACIONIH NEDOSTAJUĆIH ANKETNIH PODATAKA PRIMENOM EM ALGORITMA[2]

SOLVING PROBLEMS OF MULTIVARIATE INCOMPLETE SURVEYS DATA WITH EM ALGORITHM IMPLEMENTATION

JEL KLASIFIKACIJA: C10, C13, C18, C19

APSTRAKT:

Problem nedostajućih podataka dosta je prisutan kod anketnog istraživanja. Ukoliko se ne utvrdi tip mehanizma nedostajućih podataka, ocene nepoznatih parametara analiziranog statističkog modela, mogu biti pristrasne. Data neželjena osobina, može se preva-

1 Ekonomski fakultet, Univerzitet u Beogradu

2 Rad je nastao kao rezultat istraživanja u okviru projekta Ministarstva prosvete, nauke i tehnološkog razvoja Republike Srbije broj 179005.

zići pravilnim tretiranjem nedostajućih podataka, među kojima je svakako upotreba EM algoritma.

KLJUČNE REČI:
MULTIVARIJACIONI NEDOSTAJUĆI PODACI, MCAR TEST, EM ALGORITAM,
OCENE (OGRANIČENE) MAKSIMALNE VERODOSTOJNOSTI

ABSTRACT:

The problem of missing data is quite present in the survey research. If the type of missing data mechanism is not determined, the unknown parameter estimates of the analysed statistical model may be biased. An unwanted feature may be overcome by properly handling missing data, among which is the certainly using the EM algorithm.

KEY WORDS:
MULTIVARIATE INCOMPLETE DATA, MCAR TEST, EM ALGORITHM,
(RESTRICTED) MAXIMUM LIKELIHOOD ESTIMATES

1. UVOD

Često kod analiziranja prikupljenih anketnih podataka postoji problem neraspoloživosti, tj. ispitanici preskoče odgovore na pojedina pitanja[3]. Ova na izgled benigna situacija, može da prouzrokuje dosta ozbiljnije probleme, ukoliko se ne tretira na pravilan način. Naime, u zavisnosti od tipa mehanizma nedostajućih podatka, dati problem može biti rešiv ili ne. Takođe, u zavisnosti od tipa mehanizma nedostajućih podataka, upotreba tradicionalnih postupaka rešavanja problema nedostajućih podataka, može biti korektna ili ne (u smislu pristrasnog ocenjivanja nepoznatih parametara). Dok upotreba savremenih pristupa, kao što je EM algoritam (skraćenica od početnih slova izraza na engleskom jeziku *expectation-maximization*) u većini slučajeva daje superiornije rezultate statističke analize u smislu nepristrasnog ocenjivanja nepoznatih parametara. Iz navedenih razloga, veoma je bitno pažljivo pristupiti datom problemu, kako bi rezultati statističke analize bili korektni. Pojedini autori[4] protive se upotrebi tradicionalnih postupaka rešavanja nedostajućih podataka, čak i kada su oni, u odnosu na mehanizam nedostajućih podataka dozvoljeni, iz razloga postojanja i široke dostupnosti savremenih pristupa rešavanja pomenutog problema.

2. MEHANIZAM NEDOSTAJUĆIH PODATAKA

Ukoliko kod podataka, koji se statistički analiziraju, postoji problem nedostajućih podataka, veoma je važno da se utvrdi tip mehanizma koji dovodi do njih. Postoje tri tipa[5] mehanizma nedostajućih podataka. Prvi tip mehanizma je kada podaci nedostaju potpuno slučajno, i koji se označava sa MCAR (skr. od početnih slova izraza na engleskom jeziku *missing completely at random*). Drugi tip mehanizma je kada podaci nedostaju slučajno u oznaci MAR (skr. od početnih slova izraza na engleskom jeziku *missing at random*). Poslednji tip mehanizma je kada podaci nedostaju namerno. Za označavanje ovog tipa mehanizma upotrebljavaće se oznaka NMAR (skr. od početnih slova izraza na engleskom jeziku *not missing at random*).

Ukoliko promenljivu koja ima nedostajuće podatke obeležimo sa Y, zatim njen deo koji nedostaje sa Y^m, a deo koji je raspoloživ sa Y^o, onda se može uvesti indikator neraspoloživosti promenljive Y (u oznaci M) koji uzima vrednost 1 kod onih opservacija promenljive Y koje su neraspoložive (odnosi se na deo Y^m), i vrednost 0 kod onih opservacija koje su raspoložive (odnosi se na deo Y^o). Ukoliko su u analizi prisutne i druge promenljive, one će biti obeležene sa X.

Za mehanizam nedostajućih podataka, kaže se da je tipa MCAR, tj. da podaci koji nedostaju, nedostaju potpuno slučajno, ukoliko njihova neraspoloživost ne zavisi niti od dela Y^m, niti od dela Y^o, kao niti od ostalih prisutnih promenljivih u analizi X. Dati mehanizam

3 Little, R. & Rubin, D. (2002), str. 3.

4 Graham, J. (2012), 47-53.

5 Allison, D. (2001), str. 3-5.

nedostajućih podataka se može izraziti preko verovatnoće indikatora neraspoloživosti, kao[6]

$$\mathrm{P}\left(M\middle|Y,X\right)=P\left(M\right). \tag{1}$$

gde $\mathrm{P}\left(M\middle|Y,X\right)$ označava uslovnu verovatnoću promenljive M (koja predstavlja indikator neraspoloživosti opservacija promenljive Y) u odnosu na promenljivu Y (koja ima nedostajuće opservacije), kao i na ostale prisutne kompletno raspoložive promenljive X. $P\left(M\right)$ označava verovatnoću promenljive M, koja je dihotomnog tipa, i koja uzima vrednost 1 kod onih opservacija koje imaju neraspoložive vrednosti kod promenljive Y, i koja uzima vrednost 0 kod onih opservacija koje imaju raspoložive vrednosti kod promenljive Y.

Mehanizam nedostajućih podataka je tipa MAR, ukoliko podaci koji nedostaju ne zavise od vrednosti samih nedostajućih podataka, što se preko verovatnoće indikatora M, može formulisati kao

$$\mathrm{P}\left(M\middle|Y,X\right)=P\left(M\middle|Y^{o},X\right). \tag{2}$$

gde $P\left(M\middle|Y^{o},X\right)$ označava uslovnu verovatnoću promenljive M, koja zavisi samo od raspoloživog dela promenljive Y (u oznaci Y^{o}) kao i od prisutnih kompletno raspoloživih promenljivih X. Ukoliko je mehanizam nedostajućih podataka tipa NMAR, podaci koji nedostaju zavise i od vrednosti samih nedostajućih podataka, što se preko verovatnoće indikatora M, može formulisati kao

$$\mathrm{P}\left(M\middle|Y,X\right)=P\left(M\middle|Y^{o},\mathrm{Y}^{m},X\right). \tag{3}$$

gde $P\left(M\middle|Y^{o},\mathrm{Y}^{m},X\right)$ označava uslovnu verovatnoću promenljive M, koja zavisi od raspoloživog dela promenljive Y (u oznaci Y^{o}), zatim od neraspoloživog dela promenljive Y (u oznaci Y^{m}) kao i od prisutnih kompletno raspoloživih promenljivih X.

Važno je napomenuti da se tradicionalni postupci rešavanja problema nedostajućih podataka mogu upotrebljavati, samo ako je mehanizam nedostajućih podataka tipa MCAR. EM algoritam se može upotrebljavati, i ako je mehanizam nedostajućih podataka tipa MAR, dok ukoliko je mehanizam nedostajućih podataka tipa NMAR, nijedan od predloženih postupaka rešavanja nedostajućih podataka nije adekvatan. Sumirajući prethodno iskazano, konstatuje se da ukoliko podaci nedostaju potpuno slučajno (skr. MCAR), onda se problem nedostajućih podataka može ispravno rešavati, bilo tradicionalnim postupcima, bilo upotrebom EM algoritma. Međutim, ukoliko podaci nedostaju slučajno (skr. MAR), onda se problem nedostajućih podataka može ispravno rešavati samo upotrebom EM algoritma.

6 Fitzmaurice, G. at al. (2015), str. 7-11.

Ukoliko podaci nedostaju namerno (skr. NMAR), onda se problem nedostajućih podataka ne može ispravno rešavati ni upotrebom tradicionalnih postupaka, ni upotrebom EM algoritma. Iz tog razloga je važno ispitati tip mehanizma koji dovodi do neraspoloživosti podataka.

3. TESTIRANJE TIPA NEDOSTAJUĆIH PODATAKA

U prethodnom odeljku objašnjeno je da postoji više tipova mehanizma nedostajućih podataka. Veoma je važno utvrditi koji je tip mehanizma nedostajućih podataka u pitanju, naročito ako su podaci multivarijacionog tipa. U ovom odeljku, nakon detaljne teorijske elaboracije analiziranja i utvrđivanja tipa mehanizma nedostajućih multivarijacionih podataka, biće prezentirana i primena nad realnim podacima[7].

U uvodu je ukazano da ukoliko podaci koji nedostaju nisu tipa MCAR, onda svi uobičajeni postupci tretiranja nedostajućih podataka[8] davaće pristrasne ocene kod ocenjivanja konkretnog statističkog modela nad datim podacima, kao i kod ocenjivanja osnovnih statistika uzorka, kao što su: sredine promenljivih, njihove varijanse, kao i kovarijanse između promenljivih.

Kod utvrđivanja da li podaci koji nedostaju, nedostaju na potpuno slučajan način (što predstavlja nultu hipotezu) u odnosu na alternativnu hipotezu da podaci koji nedostaju, ne nedostaju na potpuno slučajan način; upotrebljava se MCAR test *Roderick Little*-a, čija statistika testa se može predstaviti kao[9]

$$\chi^2_{MCAR} = \sum_{svaki\ \ jedinstveni\ \ obrazac} \left(broj\ opservacija\ u\ obrascu\right) \cdot D^2 \tag{4}$$

gde D^2 predstavlja *Mahalanobis*-ovo odstojanje vektora sredina promenljivih datog jedinstvenog obrasca nedostajućih podataka od vektora sredina promenljivih, dobijenog u postupku ocenjivanja metodom maksimalne verodostojnosti. Jedinstveni obrazac nedostajućih podataka (ili skraćeno, samo jedinstveni obrazac) predstavlja kolekciju onih opservacija koje imaju nedostajuće podatke kod istih promenljivih. Na taj način, multivarijacioni set podataka, može imati nekoliko jedinstvenih obrazaca. χ^2_{MCAR} statistika testa ima broj stepeni slobode koji se izračunava po obrascu

7 Realni podaci su iz studije* *Istraživanje odnosa i saradnje u kanalima marketinga u Republici Srbiji*, koje je za potrebe Ministarstva trgovine, turizma i telekomunikacija sproveo istraživački tim NICEF-a sa Ekonomskog fakulteta Univerziteta u Beogradu. Jedan od zadataka bio je i kreiranje specifičnog indeksa zadovoljstva velikih veleprodavaca prema velikim maloprodavcima. Da bi mogao da se formira dati specifični indeks, potrebno je bilo da menadžeri velikih veletrgovina (koji su bili izabrani u reprezentativnom uzorku) odgovore na određenih 8 anketnih pitanja (koja su u ovome radu označena sa X1,...,X8). Odgovori na ponuđena pitanja nalazila su se na skali od 1 do 5.

* Petković, G. at al. (2017), str. 1-201

8 U uobičajene (tradicionalne) postupke rešavanja problematike nedostajućih podataka, ubraja se tehnika brisanja svih opservacija koji imaju barem jedan nedostajući podatak kod bilo koje promenljive (poznata i kao analiza kompletnih-podataka), zatim tehnika brisanja svih opservacija koje imaju barem jedan nedostajući podatak kod promenljivih koje se koriste u konkretnoj analizi (poznata i kao analiza raspoloživih-podataka), kao i umetanje vrednosti aritmetičke sredine promenljive umesto nedostajućeg podatka.

9 IBM (2017a), str. 682.

$$\sum_{svaki\ \ jedinstveni\ obrazac} \left(broj\ \ promenljivih\ koje\ nemaju\ nedostajuće\ podatke \right) - \upsilon \quad (5)$$

gde υ predstavlja ukupan broj posmatranih promenljivih.

Mahalanobis-ovo odstojanje koje je prikazano u izrazu (4) može se predstaviti kao[10]

$$\left(\hat{\mu}_j - \hat{\mu}_j^{(ML)} \right)^T \hat{\Sigma}_j^{-1} \left(\hat{\mu}_j - \hat{\mu}_j^{(ML)} \right) \quad (6)$$

gde se vektor $\hat{\mu}_j$ sastoji samo od sredina promenljivih koje su kompletno raspoložive u *j*-tom jedinstvenom obrascu nedostajućih podataka; dok vektor sredina $\hat{\mu}_j^{(ML)}$ se sastoji od ocena metodom maksimalne verodostojnosti, i odnosi se tj. obuhvata sve promenljive u analizi na osnovu raspoloživih podataka iz uzorka. Matrica $\hat{\Sigma}_j$ predstavlja ocene metodom ograničene maksimalne verodostojnosti kovarijacione matrice svih promenljivih u analizi na osnovu raspoloživih podataka iz uzorka. Kod ocenjivanja elemenata kovarijacione matrice upotrebljena je metoda ograničene maksimalne verodostojnosti, koja ustvari predstavlja metodu maksimalne verodostojnosti pomnožene sa $\frac{n}{n-1}$, gde n predstavlja veličinu uzorka.

Kod realnih podataka koji se koriste kao primer, veličina uzorka je 19, dok je broj promenljivih 8. Promenljive u okviru matričnog proračuna biće označene kao X1, X2, ..., X8. Ukupan broj podataka je 152, međutim u okviru ankete pojedini ispitanici nisu odgovorili na određena pitanja, tako da je broj nedostajućih podataka 14, što predstavlja 9.2% od ukupnog broja podataka. U stvari od 19 ispitanika 10 ispitanika je odgovorilo na sva pitanja, dok 9 ispitanika je imalo barem po jedan preskočen odgovor.

Kada se analiziraju pozicije nedostajućih podataka, konstatuje se da postoji 5 jedinstvenih obrazaca nedostajućih podataka. U okviru prvog obrasca (koji je ujedno i najbrojniji) i sastoji se od $n_1 = 10$ opservacija, sve promenljive su kompletno raspoložive. Kod drugog jedinstvenog obrasca nedostajućih podataka, koji je veličine $n_2 = 2$ opservacije, dva ispitanika nisu odgovorila na pitanje X4. Treći jedinstveni obrazac nedostajućih podataka se sastoji od odsustva odgovora 3 ispitanika na pitanja X2 i X4. Pretposlednji jedinstveni obrazac nedostajućih podataka se sastoji od neodgovora 2 ispitanika na pitanje X7; i poslednji peti jedinstveni obrazac nedostajućih podataka se sastoji od odsustva odgovora dva ispitanika na pitanja X5 i X6.

Pre početka izračunavanja vrednosti statistike testa date izrazom (4) potrebno je prethodno izračunati ocene maksimalne verodostojnosti vektora sredina promenljivih u analizi; kao i ocene ograničene maksimalne verodostojnosti kovarijacione matrice analiziranih

10 Enders (2010), str.20.

promenljivih. Do datih ocena se može doći raznim iterativnim postupcima kao što su: *Newton-Raphson* postupak, *Fisher scoring* postupak, *Quasi-Newton* postupak, *EM* algoritam[11] (koji je upotrebljen) i drugi postupci. Izrazima (7) i (8) prikazane su formule

$$\hat{\mu} = \left[\hat{\mu}_{X1}, \hat{\mu}_{X2}, \hat{\mu}_{X3}, \hat{\mu}_{X4}, \hat{\mu}_{X5}, \hat{\mu}_{X6}, \hat{\mu}_{X7}, \hat{\mu}_{X8} \right]^T = \\ = \left[3.84, 3.43, 3.68, 3.96, 2.98, 3.77, 4.18, 3.37 \right]^T \tag{7}$$

gde izračunate vrednosti predstavljaju ocene metodom maksimalne verodostojnosti vektora sredina analiziranih promenljivih na osnovu raspoloživih podataka

$$\hat{\Sigma} = \begin{bmatrix} \hat{\sigma}^2_{X1} & \hat{\sigma}_{X1,X2} & \hat{\sigma}_{X1,X3} & \hat{\sigma}_{X1,X4} & \hat{\sigma}_{X1,X5} & \hat{\sigma}_{X1,X6} & \hat{\sigma}_{X1,X7} & \hat{\sigma}_{X1,X8} \\ \hat{\sigma}_{X2,X1} & \hat{\sigma}^2_{X2} & \hat{\sigma}_{X2,X3} & \hat{\sigma}_{X2,X4} & \hat{\sigma}_{X2,X5} & \hat{\sigma}_{X2,X6} & \hat{\sigma}_{X2,X7} & \hat{\sigma}_{X2,X8} \\ \hat{\sigma}_{X3,X1} & \hat{\sigma}_{X3,X2} & \hat{\sigma}^2_{X3} & \hat{\sigma}_{X3,X4} & \hat{\sigma}_{X3,X5} & \hat{\sigma}_{X3,X6} & \hat{\sigma}_{X3,X7} & \hat{\sigma}_{X3,X8} \\ \hat{\sigma}_{X4,X1} & \hat{\sigma}_{X4,X2} & \hat{\sigma}_{X4,X3} & \hat{\sigma}^2_{X4} & \hat{\sigma}_{X4,X5} & \hat{\sigma}_{X4,X6} & \hat{\sigma}_{X4,X7} & \hat{\sigma}_{X4,X8} \\ \hat{\sigma}_{X5,X1} & \hat{\sigma}_{X5,X2} & \hat{\sigma}_{X5,X3} & \hat{\sigma}_{X5,X4} & \hat{\sigma}^2_{X5} & \hat{\sigma}_{X5,X6} & \hat{\sigma}_{X5,X7} & \hat{\sigma}_{X5,X6} \\ \hat{\sigma}_{X6,X1} & \hat{\sigma}_{X6,X2} & \hat{\sigma}_{X6,X3} & \hat{\sigma}_{X6,X4} & \hat{\sigma}_{X6,X5} & \hat{\sigma}^2_{X6} & \hat{\sigma}_{X6,X7} & \hat{\sigma}_{X6,X8} \\ \hat{\sigma}_{X7,X1} & \hat{\sigma}_{X7,X2} & \hat{\sigma}_{X7,X3} & \hat{\sigma}_{X7,X4} & \hat{\sigma}_{X7,X5} & \hat{\sigma}_{X7,X6} & \hat{\sigma}^2_{X7} & \hat{\sigma}_{X7,X8} \\ \hat{\sigma}_{X8,X1} & \hat{\sigma}_{X8,X2} & \hat{\sigma}_{X8,X3} & \hat{\sigma}_{X8,X4} & \hat{\sigma}_{X8,X5} & \hat{\sigma}_{X8,X6} & \hat{\sigma}_{X8,X7} & \hat{\sigma}^2_{X8} \end{bmatrix} =$$

$$= \begin{bmatrix} 1.140 & 0.751 & 0.447 & -0.237 & 0.630 & 0.575 & 0.854 & 0.784 \\ 0.751 & 1.840 & 0.761 & -0.397 & 0.564 & 0.711 & 0.989 & 0.874 \\ 0.447 & 0.761 & 1.117 & 0.192 & 0.534 & 0.240 & 0.746 & 0.401 \\ -0.237 & -0.397 & 0.192 & 1.437 & 0.187 & -0.126 & 0.130 & -0.646 \\ 0.630 & 0.564 & 0.534 & 0.187 & 1.954 & 1.016 & 0.034 & 1.251 \\ 0.575 & 0.711 & 0.240 & -0.126 & 1.016 & 1.357 & -0.202 & 1.387 \\ 0.854 & 0.989 & 0.746 & 0.130 & 0.034 & -0.202 & 1.757 & 0.097 \\ 0.784 & 0.874 & 0.401 & -0.646 & 1.251 & 1.387 & 0.097 & 2.135 \end{bmatrix} \tag{8}$$

i gde izračunate vrednosti predstavljaju ocene metodom ograničene maksimalne verodostojnosti kovarijacione matrice analiziranih promenljivih na osnovu raspoloživih podataka. Izračunate vrednosti koje se nalaze na glavnoj dijagonali predstavljaju varijanse promenljivih, dok kovarijanse između parova promenljivih, su date na vandijagonalnim pozicijama matrice.

Izračunavanje statistike testa date izrazom (4) započinjemo analizom opservacija koje pripadaju prvom jedinstvenom obrascu nedostajućih podataka. Već je pomenuto da analiziramo 8 promenljivih i da ispitanici nisu odgovorili na sva pitanja, i da prema indikatorima neraspoloživosti ukupno imamo pet oblika obrazaca nedostajućih podataka. Naime postoji pet različitih grupa ispitanika koji imaju (svaka grupa za sebe) jedinstven obrazac neraspoloživosti odgovora.

11 Webb i Copsey (2011), str. 55-60.

Kod datog prvog obrasca kome pripada 10 opservacija i koji ima odgovore na sva pitanja, tj. nema nedostajućih podataka, možemo izračunati njegov udeo χ^2 statistici testa. Naime formula (4) za podatke koje analiziramo se može razviti kao

$$\left(broj\ opservacija\ u\ obrascu\ 1\right)\cdot D^2_{obrasca\ 1}+\ldots+\left(broj\ opservacija\ u\ obrascu\ 5\right)\cdot D^2_{obrasca\ 5}$$

tako da imamo pet sabiraka, gde svaki sabirak participira u formiranju realizovane vrednosti statistike testa. Sledeći korak se sastoji od izračunavanja (ocenjivanja) vektora sredine kompletno raspoloživih promenljivih iz ovog prvog obrasca. Ocenjene vrednosti vektora sredine iznose

$$\hat{\mu}^{obrazac\ 1}=\left[\hat{\mu}^{obrazac\ 1}_{X1},\hat{\mu}^{obrazac\ 1}_{X2},\hat{\mu}^{obrazac\ 1}_{X3},\hat{\mu}^{obrazac\ 1}_{X4},\hat{\mu}^{obrazac\ 1}_{X5},\hat{\mu}^{obrazac\ 1}_{X6},\hat{\mu}^{obrazac\ 1}_{X7},\hat{\mu}^{obrazac\ 1}_{X8}\right]^T=$$

$$=\left[3.70,\,3.20,\,3.10,\,3.60,\,3.20,\,3.80,\,3.70,\,3.50\right]^T.$$

Konačno, udeo statistici testa od prvog jedinstvenog obrasca se dobija, izračunavanjem[12] sledećeg izraza: $\left(broj\ opservacija\ u\ obrascu\ 1\right)\cdot D^2_{obrasca\ 1}$ koji se može predstaviti i kao $n_1\left(\hat{\mu}^{obrazac\ 1}-\hat{\mu}_1^{(ML)}\right)^T\hat{\Sigma}_1^{-1}\left(\hat{\mu}^{obrazac\ 1}-\hat{\mu}_1^{(ML)}\right)$ i čija vrednost iznosi 5.4998 koja je dobijena matričnim računom koji sledi

$$10\left(\begin{bmatrix}3.70\\3.20\\3.10\\3.60\\3.20\\3.80\\3.70\\3.50\end{bmatrix}-\begin{bmatrix}3.84\\3.43\\3.68\\3.96\\2.98\\3.77\\4.18\\3.37\end{bmatrix}\right)^T\begin{bmatrix}1.140&0.751&0.447&-0.237&0.630&0.575&0.854&0.784\\0.751&1.840&0.761&-0.397&0.564&0.711&0.989&0.874\\0.447&0.761&1.117&0.192&0.534&0.240&0.746&0.401\\-0.237&-0.397&0.192&1.437&0.187&-0.126&0.130&-0.646\\0.630&0.564&0.534&0.187&1.954&1.016&0.034&1.251\\0.575&0.711&0.240&-0.126&1.016&1.357&-0.202&1.387\\0.854&0.989&0.746&0.130&0.034&-0.202&1.757&0.097\\0.784&0.874&0.401&-0.646&1.251&1.387&0.097&2.135\end{bmatrix}^{-1}\left(\begin{bmatrix}3.70\\3.20\\3.10\\3.60\\3.20\\3.80\\3.70\\3.50\end{bmatrix}-\begin{bmatrix}3.84\\3.43\\3.68\\3.96\\2.98\\3.77\\4.18\\3.37\end{bmatrix}\right)$$

Analognim proračunavanjem, izračunava se i udeo drugog jedinstvenog obrasca statistici testa, koji iznosi 6.9879. Nakon izračunavanja doprinosa preostalih jedinstvenih obrazaca nedostajućih podataka statistici testa, kao i njihovog sabiranja, dobija se realizovana vrednost statistike testa $\chi^2_{MCAR}=42.22$. Za izračunavanje *p* vrednosti potrebno je odrediti broj stepeni slobode statistike testa, na osnovu primene izraza (5):

$$\left(8+7+6+7+6\right)-8=26.$$

12 Sva izračunavanja u ovome radu sprovedena su upotrebom statističkog softvera IBM SPSS Statistics 25. Za sprovođenje matričnog računa upotrebljeno je u sintaksi programa komanda MATRIX-END MATRIX. Za više informacija o ovoj komandi pogledati kod IBM (2017b), str. 1109-1141.

p vrednost testa iznosi 0.022 tako da se odbacuje nulta hipoteza da podaci nedostaju potpuno slučajno, tj. da je mehanizam nedostajućih podataka tipa MCAR.

U zavisnosti koji se tip statističkog modela koristi u daljoj analizi, mehanizam nedostajućih podataka onda može biti MAR ili NMAR, ili samo MAR. Naime, ukoliko statistički model koji se planira primeniti, pripada modelima zavisnosti[13], onda je moguće da mehanizam nedostajućih podataka bude tipa MAR ili NMAR (nakon odbacivanja nulte hipoteze o mehanizmu nedostajućih podataka tipa MCAR). No, ukoliko statistički model pripada modelima međusobne zavisnosti, onda nakon odbacivanja nulte hipoteze o mehanizmu nedostajućih podataka tipa MCAR, sledi da je mehanizam nedostajućih podataka tipa MAR.

Nakon utvrđivanja da podaci koji nedostaju, ne nedostaju na potpuno slučajan način; već da nedostaju slučajno, upotreba uobičajenih postupaka za rešavanje problematike nedostajućih podataka, bi bila pogrešna, iz razloga pristrasnosti prilikom statističkog ocenjivanja bilo osnovnih statistika uzorka, bilo nepoznatih parametara određenog statističkog modela. Jedan od postupaka koji omogućava nepristrasno ocenjivanje nepoznatih parametara u uslovima kada je mehanizam nedostajućih podataka tipa MAR je EM algoritam, koji u svom postupku vrši ocenjivanje metodom maksimalne verodostojnosti, kao i ocenjivanje metodom ograničene maksimalne verodostojnosti.

4. EM ALGORITAM

EM algoritam predstavlja jedno od najčešćih postupaka savremenog rešavanja problema nedostajućih podataka. U odnosu na tradicionalne postupke rešavanja problema nedostajućih podataka, on ima prednosti, čak i kada je mehanizam nedostajućih podataka tipa MCAR. Ukoliko je mehanizam nedostajućih podataka tipa MAR, EM algoritam (za razliku od tradicionalnih metoda) i dalje omogućava nepristrasno ocenjivanje nepoznatih parametara.

Najbolje rezultate daje kada u okviru statističke analize koja se planira upotrebiti, nema testiranja značajnosti ocena nepoznatih parametara, a takvih analiza ima, kao što su: faktorska analiza, analiza glavnih komponenata, klaster analiza, analiza pouzdanosti, i njima slične analize.

Rešavanje problema multivarijacionih nedostajućih podataka, gotovo uvek zahteva iterativne algoritme optimizacije. Dati iterativni algoritmi su dosta zahtevni, u smislu da u svom postupku izračunavaju prve i druge izvode. Iterativni postupak, koji umesto izračunavanja prvih i drugih izvoda[14], koristi izračunavanje regresionih koeficijenata, predstavlja algoritam, koji zahteva dosta manje kapaciteta i resursa, te se stoga dosta češće upotrebljava. EM algoritam ima dato prethodno opisano svojstvo, kao i osobinu da kroz svaku svoju iteraciju funkcija verodostojnosti ne opada[15] te stoga predstavlja jednu stabilnu proceduru.

13 Tabachnick i Fidell (2014), str. 97.

14 Raghunathan, T. (2016), str. 146.

15 Kim i Shao (2014), str. 36.

EM algoritam predstavlja dvostepeni iterativni postupak, koji se sastoji od koraka E i koraka M. Korak E je dobio ime od prvog slova engleske reči *expectation*, što u prevodu znači očekivanje, dok je korak M, dobio ime od prvog slova engleske reči *maximization*, što u prevodu znači maksimizacija.

U prvom koraku E upotrebljavaju se elementi (inicijalno ocenjenih) vektora sredina i kovarijacione matrice promenljivih da bi se kreirali modeli višestruke linearne regresije, na osnovu kojih bi se ocenili nedostajući podaci. Na osnovu ocenjenih nedostajućih podataka, u prvom koraku M se izračunavaju (ocenjuju) elementi vektora sredina, kao i elementi kovarijacione matrice.

U drugom koraku E nad ažuriranim vrednostima elemenata vektora sredina i kovarijacione matrice, ponovo se kreiraju modeli višestruke linearne regresije, na osnovu kojih se ponovo ocenjuju nedostajući podaci.

Na osnovu ponovo ocenjenih nedostajućih podataka, u drugom koraku M se ponovo izračunavaju (ocenjuju) elementi vektora sredina, kao i elementi kovarijacione matrice.

Dati iterativni postupak se ponavlja, sve dok se ne izvrši maksimalno zadat broj iteracija, ili dok se ne ispuni kriterijum konvergencije[16] koji je dat izrazom

$$\frac{\left|\hat{\sigma}_{j,j}^{m-ta\ iteracija} - \hat{\sigma}_{j,j}^{(m-1)-va\ iteracija}\right|}{\hat{\sigma}_{j,j}^{m-ta\ iteracija}} \leq 0.0001 \text{ za svako } j, \tag{9}$$

tj. sve dok relativna promena (u okviru iteracija EM algoritma) svih varijansi promenljivih u analizi ne bude zanemarljivo mala. Napomenimo, da $\hat{\sigma}_{j,j}^{m-ta\ iteracija}$ označava ocenjenu vrednost varijanse promenljive *j* u *m*-toj iteraciji EM algoritma, dok $\hat{\sigma}_{j,j}^{(m-1)-va\ iteracija}$ označava ocenjenu vrednost varijanse promenljive *j* u (*m-1*)-voj iteraciji EM algoritma.

Primena EM algoritma u ovome radu, sprovešće se nad prikupljenim podacima u okviru ankete, koji predstavljaju multivarijacione nedostajuće podatke, jer pojedini ispitanici nisu odgovorili na neka pitanja. Naglasimo još jednom da se kod primene EM algoritma, u koraku E, za svaki jedinstveni obrazac nedostajućih podataka kreira model višestruke linearne regresije za svaku promenljivu koja ima nedostajuće podatke. U datom modelu, objašnjena promenljiva predstavlja promenljivu sa nedostajućim podacima, a objašnjavajuće promenljive su one koje su u datom jedinstvenom obrascu kompletno raspoložive. Na taj način se kod promenljive koja ima nedostajuće podatke, oni ocenjuju pomoću linearne povezanosti sa kompletno raspoloživim promenljivama. Kod datih modela višestruke linearne regresije nepoznati parametri se ocenjuju pomoću elemenata vektora sredina promenljivih, kao i kovarijacione matrice promenljivih, upotrebom sledećih obrazaca[17]:

$$\hat{\beta}_1 = \hat{\Sigma}_{xx}^{-1}\hat{\sigma}_{yx} \text{ i } \hat{\beta}_0 = \hat{\mu}_y - \hat{\sigma}'_{yx}\hat{\Sigma}_{xx}^{-1}\hat{\mu}_x \tag{10}$$

16 IBM (2017a), str. 681.

17 Rencher i Christensen (2012), str. 345-346.

gde je pretpostavljeno da je u pitanju model višestruke linearne regresije sa q objašnjavajućih promenljivih[18], zatim da je slobodan član modela označen sa β_0 dok parametri (koeficijenti) uz objašnjavajuće promenljive su označeni vektorom $\hat{\beta}_1 = \left[\beta_1, \beta_2, \ldots, \beta_q\right]^T$. Zatim, sredina objašnjene promenljive označena je sa μ_y dok je vektor sredina nezavisnih promenljivih označen sa μ_x dok ocene $\hat{\sigma}'_{yx}$ i $\hat{\Sigma}_{xx}$ predstavljaju podmatrice proširene kovarijacione matrice

$$\hat{\Sigma} = \begin{bmatrix} \hat{\sigma}_y^2 & \hat{\sigma}_{y1} & \cdots & \hat{\sigma}_{yq} \\ \hat{\sigma}_{1y} & \hat{\sigma}_1^2 & \cdots & \hat{\sigma}_{1q} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\sigma}_{qy} & \hat{\sigma}_{q1} & \cdots & \hat{\sigma}_q^2 \end{bmatrix} = \begin{bmatrix} \hat{\sigma}_y^2 & \hat{\sigma}'_{yx} \\ \hat{\sigma}_{xy} & \hat{\Sigma}_{xx} \end{bmatrix}.$$

Za sprovođenje EM algoritma, potrebne su inicijalne ocene vektora sredina i kovarijacione matrice. Kod inicijalnog ocenjivanja (što inače predstavlja nulti korak maksimizacije, u oznaci M_0) upotrebljavaju se tradicionalni postupci ocenjivanja vektora sredina i kovarijacione matrice promenljivih u prisustvu nedostajućih podataka, i jedan od često upotrebljavanih je i postupak ocenjivanja na osnovu raspoloživih-podataka. Date inicijalne ocene ($\hat{\mu}_0, \hat{\Sigma}_0$) su date izrazima koji slede:

$$\hat{\mu}_0 = \left[\hat{\mu}_{X1}, \hat{\mu}_{X2}, \hat{\mu}_{X3}, \hat{\mu}_{X4}, \hat{\mu}_{X5}, \hat{\mu}_{X6}, \hat{\mu}_{X7}, \hat{\mu}_{X8}\right]^T =$$

$$= \left[3.84, 3.25, 3.68, 3.64, 2.88, 3.71, 3.94, 3.37\right]^T \qquad (11)$$

gde izračunate vrednosti predstavljaju ocene vektora sredina analiziranih promenljivih na osnovu raspoloživih podataka. Izrazom (12) prikazana je formula

$$\hat{\Sigma}_0 = \begin{bmatrix} \hat{\sigma}_{X1}^2 & \hat{\sigma}_{X1,X2} & \hat{\sigma}_{X1,X3} & \hat{\sigma}_{X1,X4} & \hat{\sigma}_{X1,X5} & \hat{\sigma}_{X1,X6} & \hat{\sigma}_{X1,X7} & \hat{\sigma}_{X1,X8} \\ \hat{\sigma}_{X2,X1} & \hat{\sigma}_{X2}^2 & \hat{\sigma}_{X2,X3} & \hat{\sigma}_{X2,X4} & \hat{\sigma}_{X2,X5} & \hat{\sigma}_{X2,X6} & \hat{\sigma}_{X2,X7} & \hat{\sigma}_{X2,X8} \\ \hat{\sigma}_{X3,X1} & \hat{\sigma}_{X3,X2} & \hat{\sigma}_{X3}^2 & \hat{\sigma}_{X3,X4} & \hat{\sigma}_{X3,X5} & \hat{\sigma}_{X3,X6} & \hat{\sigma}_{X3,X7} & \hat{\sigma}_{X3,X8} \\ \hat{\sigma}_{X4,X1} & \hat{\sigma}_{X4,X2} & \hat{\sigma}_{X4,X3} & \hat{\sigma}_{X4}^2 & \hat{\sigma}_{X4,X5} & \hat{\sigma}_{X4,X6} & \hat{\sigma}_{X4,X7} & \hat{\sigma}_{X4,X8} \\ \hat{\sigma}_{X5,X1} & \hat{\sigma}_{X5,X2} & \hat{\sigma}_{X5,X3} & \hat{\sigma}_{X5,X4} & \hat{\sigma}_{X5}^2 & \hat{\sigma}_{X5,X6} & \hat{\sigma}_{X5,X7} & \hat{\sigma}_{X5,X6} \\ \hat{\sigma}_{X6,X1} & \hat{\sigma}_{X6,X2} & \hat{\sigma}_{X6,X3} & \hat{\sigma}_{X6,X4} & \hat{\sigma}_{X6,X5} & \hat{\sigma}_{X6}^2 & \hat{\sigma}_{X6,X7} & \hat{\sigma}_{X6,X8} \\ \hat{\sigma}_{X7,X1} & \hat{\sigma}_{X7,X2} & \hat{\sigma}_{X7,X3} & \hat{\sigma}_{X7,X4} & \hat{\sigma}_{X7,X5} & \hat{\sigma}_{X7,X6} & \hat{\sigma}_{X7}^2 & \hat{\sigma}_{X7,X8} \\ \hat{\sigma}_{X8,X1} & \hat{\sigma}_{X8,X2} & \hat{\sigma}_{X8,X3} & \hat{\sigma}_{X8,X4} & \hat{\sigma}_{X8,X5} & \hat{\sigma}_{X8,X6} & \hat{\sigma}_{X8,X7} & \hat{\sigma}_{X8}^2 \end{bmatrix} =$$

18 Objašnjavajuće promenljive označene su sa $x_1,\ldots,x_q$, dok je objašnjena promenljiva označena sa y.

$$= \begin{bmatrix} 1.140 & 0.717 & 0.447 & -0.132 & 0.728 & 0.632 & 0.637 & 0.784 \\ 0.717 & 1.800 & 0.617 & 0.115 & 0.418 & 0.670 & 0.791 & 0.783 \\ 0.447 & 0.617 & 1.117 & 0.011 & 0.441 & 0.165 & 0.794 & 0.401 \\ -0.132 & 0.115 & 0.011 & 1.016 & 0.576 & 0.242 & 0.106 & -0.011 \\ 0.728 & 0.418 & 0.441 & 0.576 & 1.985 & 1.088 & 0.662 & 1.210 \\ 0.632 & 0.670 & 0.165 & 0.242 & 1.088 & 1.471 & 0.400 & 1.430 \\ 0.637 & 0.791 & 0.794 & 0.106 & 0.662 & 0.400 & 1.059 & 0.654 \\ 0.784 & 0.783 & 0.401 & -0.011 & 1.210 & 1.430 & 0.654 & 2.135 \end{bmatrix} \quad (12)$$

gde izračunate vrednosti predstavljaju ocene kovarijacione matrice analiziranih promenljivih na osnovu raspoloživih podataka. Nakon inicijalnog koraka M_0, u prvom koraku E ocenjujemo nedostajuće podatke sledećim postupkom: za svaki jedinstveni obrazac nedostajućih podataka, formiramo model višestruke linearne regresije, gde promenljiva koja ima nedostajuće podatke predstavlja objašnjenu promenljivu, a promenljive koje su u tom jedinstvenom obrascu kompletno raspoložive predstavljaće objašnjavajuće promenljive. Napomenimo, da ako u jedinstvenom obrascu ima nekoliko promenljivih sa nedostajućim podacima, za svaku od njih se kreira poseban model višestruke linearne regresije.

Na taj način ukoliko analizirani multivarijacioni nedostajući podaci, imaju nekoliko jedinstvenih obrazaca nedostajućih podataka i ukoliko nekoliko promenljivih ima nedostajuće podatke, onda će u koraku E biti potrebno formirati više modela višestruke linearne regresije za ocenjivanje nedostajućih podataka.

Kod analiziranih podataka imamo pet jedinstvenih obrazaca nedostajućih podataka, kod kojih je potrebno formirati ukupno šest modela višestruke linearne regresije. Naime kod prvog jedinstvenog obrasca nedostajućih podataka, nemamo nedostajuće podatke, tako da se kod tog obrasca neće formirati modeli višestruke linearne regresije za ocenjivanje nedostajućih podataka.

Kod drugog jedinstvenog obrasca nedostajućih podataka, promenljiva X4 ima nedostajuće podatke, tako da se samo za nju kreira model višestruke linearne regresije (gde ona predstavlja objašnjenu promenljivu) a sve ostale promenljive biće u ulozi objašnjavajućih promenljivih.

Kod trećeg jedinstvenog obrasca nedostajućih podataka, promenljive X2 i X4 imaju nedostajuće podatke, tako da se u tom obrascu kreiraju dva modela višestruke linearne regresije gde sve preostale promenljive su objašnjavajuće, i gde u jednom modelu X2 predstavlja objašnjenu promenljivu, a u drugom modelu X4 je objašnjena promenljiva. Kod preostalih jedinstvenih obrazaca biće formirana još tri modela višestruke linearne regresije.

Kod prethodno formiranih šest modela višestruke linearne regresije, nepoznati parametri biće ocenjeni upotrebom formule (10). Data formula koristi inicijalne ocene vektora sredina i kovarijacione matrice date izrazima (11) i (12). Nakon formiranja i upotrebe

pomenutih šest modela višestruke linearne regresije, ocenjuju se nedostajući podaci, čime se završava prvi E korak.

Na osnovu ocenjenih nedostajućih podataka, analizirani anketni podaci su sada kompletno raspoloživi, tako da se u prvom koraku M, ocene vektora sredina i kovarijacione matrice, metodom maksimalne verodostojnosti i ograničene maksimalne verodostojnosti, mogu ponovo oceniti, upotrebom izraza

$$\hat{\mu} = \begin{bmatrix} \hat{\mu}_1 \\ \vdots \\ \hat{\mu}_q \end{bmatrix} = \frac{1}{n} \begin{bmatrix} \sum_{i=1}^{n} X_{i1} \\ \vdots \\ \sum_{i=1}^{n} X_{iq} \end{bmatrix} \text{ i } \hat{\Sigma} = \frac{1}{n-1} \begin{bmatrix} \sum_{i=1}^{n} \left(X_{i1} - \hat{\mu}_1\right)^2 & \cdots & \sum_{i=1}^{n} \left(X_{i1} - \hat{\mu}_1\right)\left(X_{iq} - \hat{\mu}_q\right) \\ \vdots & \ddots & \vdots \\ \sum_{i=1}^{n} \left(X_{iq} - \hat{\mu}_q\right)\left(X_{i1} - \hat{\mu}_1\right) & \cdots & \sum_{i=1}^{n} \left(X_{iq} - \hat{\mu}_q\right)^2 \end{bmatrix} \quad (13)$$

Prethodnim postupcima, završena je prva iteracija EM algoritma.

Druga iteracija, započinje drugim korakom E, u okviru koga se nad novim ocenama sredine vektora i kovarijacione matrice promenljivih, izračunavaju nove ocene nedostajućih podataka upotrebom novih šest modela višestruke linearne regresije, kod kojih su ocene nepoznatih parametara, ocenjene pomoću izraza (10).

Drugi korak M, na osnovu novih (ažuriranih) ocena nedostajućih podataka, ponovo izračunava ocene sredine vektora i kovarijacione matrice promenljivih, upotrebom izraza (13). Ovim je završena i druga iteracija EM algoritma.

Ukažimo da se broj iteracija EM algoritma izvršava, sve dok se ne dostigne konvergencija definisana izrazom (9) koja se u ovom primeru ispunjava u 257. iteraciji. Ocene vektora sredina i kovarijacione matrice promenljivih u datoj poslednjoj iteraciji EM algoritma, predstavljaju ocene metodom maksimalne verodostojnosti i prikazane su izrazima (7) i (8). Takođe ocene nedostajućih podataka iz poslednje iteracije, biće upotrebljene kako bi anketni podaci bili kompletno raspoloživi, što je veoma korisno kod narednih statističkih analiza.

5. ZAKLJUČAK

U radu je prikazan pravilan način rešavanja problematike nedostajućih podataka, koji se sastoji prvenstveno od provere tipa mehanizma nedostajućih podataka. Ukoliko podaci nedostaju na potpuno slučajan način, nedostajući podaci se mogu rešavati i na tradicionalan način, iako bi primena EM algoritma dala bolje osobine određenih ocena osnovnih statistika uzorka, kao i veću snagu testova u odnosu na tradicionalne postupke (koji se baziraju na isključivanju opservacija koje imaju nedostajuće podatke).

Ukoliko je mehanizam nedostajućih podataka tipa MAR, onda upotreba tradicionalnih postupaka za rešavanje nedostajućih podataka predstavlja pogrešan pristup, iz razloga pristrasnosti u ocenjivanju osnovnih statistika uzorka. Kao optimalno rešenje uzima se

EM algoritam, naročito kada se u analizi primenjuju statistički modeli međuzavisnosti (kao što je npr. analiza glavnih komponenata, faktorska analiza, klaster analiza, analiza pouzdanosti i dr.).

Takođe u radu je prikazan detaljan teorijsko-praktičan pristup primene EM algoritma, kod rešavanja problema multivarijacionih nedostajućih podataka (koji po običaju imaju po nekoliko oblika jedinstvenih obrazaca nedostajućih podataka). Postupak je iterativan i dosta kompleksan, ali je detaljnim razrađivanjem svakog koraka objašnjen, i na realnom primeru je sproveden ceo postupak.

LITERATURA

Allison, D. (2001), *Missing Data*, Sage Publications, Thousand Oaks.

Enders, C. (2010), *Applied Missing Data Analysis*, The Guilford Press, New York.

Fitzmaurice, G. at al. (2015), "Missing Data: Introduction and Statistical Preliminaries", in Molenbergs, G. et al. (ed.) (2015). *Handbook of Missing Data Methodology*, CRC Press, Boca Raton, pp. 3-22

Graham, J. (2012), *Missing Data - Analysis and Design*, Springer, New York.

Hardle, W. K. & Simar, L. (2015), *Applied Multivariate Statistical Analysis* (4 ed.), Springer, Berlin.

IBM (2017a), *IBM SPSS Statistics Algorithms*, IBM Corporation, Armonk.

IBM (2017b), *IBM SPSS Statistics 25 Command Syntax Reference*, IBM Corporation, Armonk.

Johnson, R. & Wichern, D. (2014), *Applied Multivariate Statistical Analysis* (6 ed.), Pearson, Harlow.

Kim, J. K. & Shao, J. (2014), *Statistical Methods for Handling Incomplete Data*, CRC Press, Boca Raton.

Little, R. & Rubin, D. (2002), *Statistical Analysis with Missing Data* (2 ed.), Wiley, Chichester.

Petković, G. at al. (2017), *Istraživanje odnosa i saradnje u kanalima marketinga u Republici Srbiji* (studija), NICEF, Beograd

Raghunathan, T. (2016), *Missing Data Analysis in Practice*, CRC Press, Boca Raton.

Rencher, A. & Christensen, W. (2012), *Methods of Multivariate Analysis* (3 ed.), Wiley, Hoboken.

Tabachnick, B. & Fidell, L. (2014), *Using Multivariate Statistics* (6 ed.), Pearson, Harlow.

Webb, A. & Copsey, K. (2011), *Statistical Pattern Recognition* (3 ed.), Wiley, Chichester.