Lazar Tenjović

Autlajeri – iznimci ili vangranične vrednosti (eng. outliers)

Možda je najbolje da tekst posvećen autlajerima započnemo jednim (za pisca ovih redova) zanimljivim zapažanjem: u indeksima pojmova u barem tri "klasična", i veoma poštovana udžbenika matematičke statistike iz šezdesetih godina XX veka (J.E. Freund: Mathematical statistics, Prentice-Hall, Inc., 1963; R. V. Hogg & A. T. Craig: Introduction to mathematical statistics, New York: The Macmillan Company, 1965; S.S. Wilks: Mathematical statistics, New York: John Wiley & Sons, Inc, 1962) nema pojma outlier! Neobičnosti radi navedimo i ovo: pisac jednog od udžbenika (S.S. Wilks) je 1962. godine objavio statistički test za detektovanje autlajera u multivarijacionim podacima! Nasuprot tome, čak i u onim savremenim udžbenicima primenjene statistike čiji naslovi ne zaslužuju da se navode barem na nekoliko mesta spominju se vangranične vrednosti, tj. autlajeri. Ova ilustracija možda na najbolji način odslikava postepeno naraslu svest o tome koliku važnost u statističkim analizama realnih podataka ima prisustvo autlajera.

Autlajere ili iznimke mogli bismo odrediti kao one vrednosti u skupu podataka koje su neuobičajeno daleko ili koje su veoma različite od glavnine podataka. Prisustvo autlajera u uzorku podataka najčešće se tumači kao posledica grešaka merenja, grešaka u unosu podataka ili (što je sa statističkog aspekta najvažnije) kao odraz intrinzične varijabilnosti izvora podataka (cf. Barnett, 1978). Određenje vangraničnih vrednosti kao odraza intrinzične varijabilnosti podrazumeva da postoji statistički model o izvoru koji generiše podatke, tj. o prirodi distribucije varijable u populaciji iz koje je generisan uzorak. Autlajer bi se u tom slučaju mogao posmatrati kao opservacija ili vektor podataka koji su

uzorkovani iz druge populacije u odnosu na onu iz koje je glavnina podataka. Moguće je, isto tako, postojanje iznimaka u podacima tumačiti i u smislu uzorkovanja iz populacije čija distribucija ima "teže repove" od pretpostavljene raspodele. Autlajeri u multivarijacionom slučaju mogu predstavljati opservacije koje odstupaju od glavnine podataka po veličini ("autlajeri pomerenosti" – eng. shift outliers) ili po strukturi. Dok prva vrsta autlajera prati osnovnu strukturu povezanosti među varijablama i predstavlja prosto daleke opservacije koje prate trend glavnine podataka, dotle druga vrsta vangraničnih vrednosti upravo narušava strukturu povezanosti između varijabli. Ukoliko su, na primer, dve osobine ličnosti u pozitivnoj linearnoj vezi, autlajer "po veličini" predstavljao bi ispitanika koji ima neuobičajeno visoke (ili niske) rezultate na obema crtama, dok bi autlajer "po strukturi" bila jedinica posmatranja sa neobično visokim rezultatom na jednoj, a neuobičajeno niskim rezultatom na drugoj osobini.

Otkrivanje autlajera u podacima neobično je važno. Pre svega, na taj način moguće je otkriti grube greške u merenju ili unosu podataka koje bi, ukoliko ostanu neuočene, mogle potpuno kompromitovati zaključke istraživanja. S druge strane, "prave" vangranične vrednosti, tj. rezultati izrazito netipičnih jedinica posmatranja mogu – budući da se u analizama podataka psiholoških istraživanja još uvek skoro isključivo koriste statistički postupci koji su nepostojani, tj. nerezistentni na autlajere – dovesti do nesrazmernog izobličenja rezultata i tako dovesti do neadekvatnih ocena ključnih parametara koji su od interesa za istraživača. Zavisno od toga koji matematički model distribucije, po pretpostavci, dobro opisuje raspodelu jedne ili više varijabli u populaciji postoje različiti statistički kriterijumi i raznovrsni grafički postupci za detekciju autlajera. Najčešće primenjivani klasični statistički kriterijumi za vangranične vrednosti u jednodimenzionalnom slučaju su postojanje podataka čije

odstupanje od aritmetičke sredine je veće od 2.5 ili 3 standardne devijacije (u zavisnosti od veličine uzorka) ili čija je udaljenost od gornje (ili donje) četvrti (eng. uper and lower forth, vrednosti koje grubo odgovaraju percentilima 75 i 25) veća od izraza koji se dobija množenjem međučetvrtnog raspršenja (razlike gornje i donje četvrti) vrednošću 1.5 (definicija i postupak računanja međučetvrtnog raspršenja mogu se videti, na primer, u Hoaglin, 1983. str. 38). U novije vreme se sve više koristi "robustni kriterijum" za deklarisanje jednodimenzionalnih autlajera koji se definiše na sledeći način: rezultat x_i je vangranična vrednost ako je

$$\frac{|x_i - \text{Mdn}|}{\text{MAD} / 0.6745} > 2.24$$

Pri tome je Mdn medijana, MAD je medijana apsolutnih odstupanja rezultata od medijane (eng. Median Absolute Deviation), a konstanta 0.6745 služi za reskaliranje mere MAD tako da u slučaju normalne raspodele MAD može poslužiti kao ocena standardne devijacije populacije (cf. Wilcox & Keselman, 2003).

Uočavanje potencijalnih iznimaka u univarijacionom i bivarijacionom slučaju relativno je jednostavno i za to je ponekad dovoljno pažljivo i znalačko posmatranje valjanog grafičkog prikaza podataka (npr. kutijasti dijagram – boxplot, dijagram raspršenja). Detektovanje autlajera u multivarijacionom "roju tačaka" (dakle, u prostoru sa više od dve dimenzije) predstavlja veoma komplikovan zadatak jer je oslanjanje na vizuelnu percepciju u tom slučaju praktično nemoguće (premda u tom smislu postoje predlozi veoma sofisticiranih algoritama koji koriste dinamičko grafičko programiranje, a koji bi mogli biti od koristi u slučaju eliptičnih ili skoro eliptičnih distribucija, cf. Bartkowiak & Szustalewicz, 1997). Za otkrivanje postojanja *jednog* autlajera u multivarijacionom slučaju moguće bi bilo primeniti varijantu "bez jednog" (eng. leaving-one-out) postupka "univerzalnog noža" (eng. jackknife). Postupak "univerzalnog noža" izvorno je razvijen u radovima

Quenouillea i Tukeya za ocenu stabilnosti (tj. standardne greške) i pristrasnosti različitih ocenitelja parametara (cf. Efron, 1981; Rodgers, 1999). Preuzorkovanjem (eng. resampling) bez vraćanja jedinica posmatranja iz postojećeg uzorka tako da se u svakom koraku izostavi po jedna jedinica iz uzorka i analiza u svakom koraku izvede na preostalih n-1 opservacija može biti od pomoći u "lociranju" autlajera. Ipak, osim toga što bi za velike i složene uzorke ovaj postupak zahtevao ogroman broj koraka (što s obzirom na brzinu današnjih računara nije ozbiljan problem) veliki nedostatak ovog postupka je u tome što se njime ne mogu otkriti višestruke vangranične vrednosti ili klasteri autlajera.

Klasični postupak za otkrivanje autlajera u multivarijacionom slučaju zasniva se na računanju Mahalanobisove distance za i-tu jedinicu posmatranja (vektor podataka $\mathbf{x}_i$), u oznaci MD_i:

$$MD_i = \sqrt{(\mathbf{x}_i - \mathbf{m})^t \mathbf{C}^{-1} (\mathbf{x}_i - \mathbf{m})}$$

pri čemu je **m** vektor aritmetičkih sredina, a **C** matrica kovarijansi varijabli. Mahalanobisova distanca očigledno pruža informaciju o udaljenosti tačke koja predstavlja datu jedinicu posmatranja od središta multidimenzionalnog roja tačaka (vektor **m**) uzimajući pri tome u obzir oblik roja (matrica **C**), tj. strukturu povezanosti među varijablama. Jedinica posmatranja sa visokom vrednošću MD_i, tj. vrednošću Hi-kvadrat statistika većom od kvantila 0.975 iz Hi-kvadrat distribucije za p stepeni slobode (pri čemu je p broj varijabli) može biti kandidat za autlajera[1]. Međutim, tzv. čiste "autlajere pomerenosti" (vangranične vrednosti koje potiču iz multivarijacione distribucije sa istim oblikom matrice raspršenja kao glavnina podataka, ali sa pomerenim parametrom lokacije) praktično je nemoguće otkriti ovim postupkom zbog efekata *maskiranja* i *pretegnuća* (teorema o očekivanoj kvadriranoj

[1] Kvadrirana Mahalanobisova distanca D_i^2, ima, ako uzorak potiče iz multivarijacione normalne raspodele, Hi-kvadrat distribuciju sa p stepeni slobode, pri čemu je p broj varijabli (Gnanadesikan & Kettenring, 1972).

Mahalanobisovoj distanci iz koje sledi ovaj zaključak i dokaz te teoreme može se naći u Rocke & Woodruff, 1996). Efekat maskiranja događa se kada u multivarijacionim podacima postoji određeni skup vangraničnih vrednosti tipa "pomerenosti" koje pomeraju centar glavnine podataka i povećavaju elemente matrice kovarijansi. Efekat pretegnuća (eng. swamping) dešava se kada skup autlajera toliko izobliči matricu kovarijansi da podaci iz glavnine podataka koji u stvari nisu iznimci počinju po svojoj Mahalanobisovoj distanci da liče na autlajere, dok pravi autlajeri bivaju neotkriveni. U kojoj meri je otkrivanje iznimaka u multivarijacionim podacima složeno najbolje se može naslutiti iz ogromnog broja članaka (i, prema tome, različitih algoritama) koji su posvećeni tom problemu a objavljeni su u novije vreme u vodećim svetskim statističkim časopisima (cf. Gnanadesikan & Kettenring, 1972; Campbell, 1978; Rousseeuw & van Zomeren, 1990; Hadi, 1992; Atkinson, 1994; Rocke & Woodruff, 1996; Kosinski, 1999; Hardin & Rocke, 2004). Veoma složeni problemi koji postoje u otkrivanju autlajera u multivarijacionim podacima mogu se svakako umanjiti korišćenjem multivarijacionih statističkih postupaka koji poseduju određeni stepen rezistentnosti na autlajere. S obzirom na opasnost izobličenja rezultata dobijenih "klasičnim" multivarijacionim postupcima zbog prisustva nedetektovanih vangraničnih vrednosti u podacima čini se veoma važnim ispitati robustnost na autlajere svakog novog postupka koji se primenjuje u analizama podataka. Ukoliko se postojeći postupci pokažu nepostojanima u prisustvu iznimaka veoma je korisno pristupiti njihovom upostojanjavanju (eng. robustification). Korist od takvog pristupa u primeni statističkih postupaka može biti dvostruka: s jedne strane, velike razlike u rezultatima koji su dobijeni primenom nepostojanih i postojanih metoda mogu ukazivati na postojanje skrivenih autlajera u podacima; s druge strane, ishodi primene postupaka rezistentnih na vangranične vrednosti mogli bi biti više u

skladu sa osnovnim zakonitostima ili složajevima odnosa u ispitivanoj realnosti koje bi, inače, mogle biti maskirane prisustvom autlajera u podacima. Pri upostojanjavanju postupka potrebno je postići istovremeno dva često suprostavljena cilja:

1. postojan postupak bi trebalo da bude – u smislu efikasnosti – veoma blizu optimalnom postupku kada su ispunjene pretpostavke modela na kojem se zasniva optimalni ali nepostojani postupak;
2. postojan postupak bi trebalo da pokazuje robustnost, tj. neosetljivost na umerene perturbacije modela na kojem se zasniva klasični optimalni i, za dati model, najefikasniji postupak. Robustnost postupka na umerene perturbacije modela često se u praksi analize podataka proveravaju ispitivanjem rezistentnosti na autlajere.

Reference:

Atkinson, A. (1994). Fast very robust methods for the detection of multiple outliers, *Journal of the American Statistical Association*, *89*, 1329–1339.

Barnett, V. (1978). The study of outliers: purpose and model, *Applied statistics*, *27*, 242-250.

Bartkowiak, A., & Szustalewicz, A. (1997). The Grand Tour as a method for detecting multivariate outliers, *Mashine Graphics & Vision*, *6*, 487–505.

Campbell, N.A. (1978). The influence function as an aid in outlier detection in discriminant analysis, *Applied Statistics*, *27*, 251–258.

Efron, B. (1981). Nonparametric estimates of standard error: the jackknife, the bootstrap and other methods, *Biometrika*, *68*, 589–599.

Gnanadesikan, R. & Kettenring, J. (1972). Robust estimates, residuals, and outlier detection with multiresponse data, *Biometrics*, *28*, 81–124.

Hadi, A.S. (1992). Identifying multiple outliers in multivariate data, *Journal of the Royal statistical society. Series B (Methodological)*, *54*, 761–771.

Hardin, J., & Rocke, D. M. (2004). The distribution of robust distances, Preprint, Skinuto 8. juna 2006. sa URL adrese: http://www.cipic.ucdavis.edu/~dmrocke/.

Hoaglin, D. C. (1983). Letter values: a set of selected order statistics, In: D.C. Hoaglin, F. Mosteller, & J.W. Tukey, Understanding Robust and Exploratory Data Analysis (pp.33–57), New York: John Wiley & Sons., Inc.

Kosinski, A. S. (1999). A procedure for the detection of multivariate outliers, *Computational statistics & Data* analysis, *29*, 145–161.

Rocke & Woodruff (1996). Identification of outliers in multivariate data, *Journal of the American Statistical Association*, *91*, 1047–1061.

Rodgers, J. L. (1999). The bootstrap, the jackknife, and the randomization test: a sampling taxonomy, *Multivariate Behavioural research*, *34*, 441–456.

Rousseeuw, P.J., & van Zomeren, B.C.(1990). Unmasking multivariate outliers and leverage points, *Journal of the American Statistical Association*, *85*, 633–639.

Wilcox, R.R, & Keselman, H.J. (2003). Modern robust data analysis methods: measures of central tendency, *Psychological Methods*, *8*, 254–274.