

Igor Jelaska¹
Marko Erceg¹
Danijela Kuna²

Izvorni znanstveni rad

¹Kineziološki fakultet sveučilišta u Splitu

²Srednja škola „Kupres”

KOMPARACIJA ALGORITAMA ODABIRA VARIJABLI REGRESIJSKOG MODELA U KINEZIOLOŠKIM ISTRAŽIVANJIMA

UVOD

Istraživanja u području selekcije varijabli u primijenjene statističke i matematičke modele počinju s Lewis (1962) i Sebestyen (1962). Na tim početnim istraživanjima su se nadogradili mnogi radovi u medicini i biologiji (Ganster i sur., 2001; Inza i sur., 2001a, 2001b). Važno je naglasiti da je konstrukcija najboljeg podskupa skupa varijabli koji bi trebali biti uključeni u model izuzetno kompleksno pitanje obzirom na činjenicu da broj mogućih odabira varijabli raste eksponencijalno u ovisnosti o broju varijabli. Preciznije, ako imamo n prediktorskih varijabli, tada imamo:

$$\sum_{k=1}^n \binom{n}{k} = 2^n - 1$$

mogućnosti za odabir prediktorskih varijabli u model. Pritom smo isključili mogućnost da model nema niti jednu varijablu. Navedeni problem je interesantan i sa samog aspekta teorijskog i praktičnog računarstva obzirom da spada u klasu NP-teških (*Nondeterministic Polynomial - Hard*) problema (Cotta i sur., 2004). Praktično to znači da je za velike skupove varijabli problem pronalaženja optimalnog rješenja nerješiv. Stoga su se vremenom razvila dva smjera istraživanja selekcije varijabli u regresijske modele: tehnike koje garantiraju optimalno rješenje, ali samo primijenjene na malim skupovima varijabli te heurističke tehnike koje pronalaze „dovoljno dobro” rješenje, ali u razumnom vremenu. Pritom, istraživanja u smjeru pronalaženja optimalnog rješenja problema selekcije varijabli u regresijski model su još uvijek znanstveni problemi na kojima se intenzivno radi. Tako primjerice, Yusta, Bonrosto i Letamendia (2007) predlažu novu metodu za selekciju varijabli, posebno u diskriminativnoj i logističkoj analizi. Pritom je metoda utemeljena na metaheurističkoj strategiji te u većini promatranih slučajeva generira bolje rezultate od klasičnih metoda *stepwise backward* i *stepwise forward*.

Naglasimo da svako povećanje broj prediktora u modelu višestruke regresije doprinosi povećanju koeficijenta determinacije, odnosno doprinosi objašnjenju kriterijske (regresand) varijable. To može biti objašnjeno formulom

$$\lim_{k \rightarrow \infty} R_k^2 = 1$$

pri čemu je R_k^2 koeficijent multiple determinacije uz k varijabli uključenih u model (Draper i Smith, 1991). Formalno matematički analizirano, svaki model višestruke regresije bit će 100% protumačen ako beskonačno mnogo varijabli uključimo u model. Suprotno tome, ukoliko raspoložemo s većim brojem prediktorskih varijabli, ponekad, nije potrebno sve varijable unijeti u model, pogotovo ako svaka značajno ne doprinosi objašnjavanju kriterijske varijable. Stoga se u praktičnim upotrebama regresijske analize, korištenjem principa parsimonije, raznim metodama ograničava broj regresorskih varijabli te se zadržavaju samo one koje djeluju značajno. Potrebno je naglasiti da je jedan od zahtjeva za ispravnu upotrebu regresijske analize nepostojanje multikolinearnosti, odnosno prediktorske varijable ne smiju biti međusobno zavisne. U skladu s tim, prisutnost većeg broja varijabli može uzrokovati probleme multikolinearnosti što može implicirati greške u interpretaciji i značajnosti parametara. Naglasimo da je selekcija varijabli pri generiranju modela važan korak kod više metodoloških postupaka: klasične diskriminativne analize, logističke regresije, neuronskih mreža, stabala odlučivanja, strojnog učenja te drugih njima srodnih metoda.

U znanstvenoj praksi u društvenim i humanističkim znanostima standardni algoritmi odabira varijabli u model su *backward stepwise*, *forward stepwise* i *all effects*.

Kod *forward* metode na početku nema niti jedne varijable u regresiji. Prva varijabla koja ulazi u model je ona koja ima najveći koeficijent korelacije sa kriterijskom varijablom. Zatim se ispituje njena značajnost F-testom. Pritom možemo fiksirati F-omjer kojeg treba premašiti i nivo signifikantnosti koji F-omjer treba zadovoljiti. Potrebno je naglasiti da fiksiranje F-omjera ne daje uvijek isti nivo signifikantnosti jer se zbog mijenjanja broja varijabli mijenja broj stupnjeva slobode. Ukoliko je prva varijabla uvrštena u model zadovoljila F-test, iz preostalog skupa prediktora bira se ona koja ima najveći koeficijent parcijalne korelacije sa kriterijskom varijablom. Ponovo se testira F-testom jeli ta varijabla značajno doprinijela povećanju protumačenosti modela. Postupak se ponavlja dok sve varijable koje zadovoljavaju navedene kriterije ne budu uvrštene u regresijski model.

Backward metoda ide u suprotnom smjeru u odnosu na metodu *forward*. Na početku selektiranja varijabli u model se uključe sve prediktorske varijable, a pritom se umjesto kriterija za ulazak varijabli u model postavljaju uvjeti za izlazak varijable

iz modela. Prvi kriterij je minimalna vrijednost F-omjera koji neka varijabla mora zadovoljiti da bi ostala u modelu a drugi kriterij je da je maksimalna signifikantnost za isključenje koja obično iznosi 0,1. U daljnjem tijeku algoritma promatra se varijabla koja ima najmanji koeficijent parcijalne korelacije sa kriterijskom varijablom. Ako ta varijabla pri tom i zadovoljava kriterijima za isključenje tada se isključuje iz modela. Postupak se nastavlja sve dok ima varijabli koje zadovoljavaju sve uvjete za isključenje. U praksi je ipak najzastupljenija *all effects* metoda - uzima sve prediktorske varijable u regresijski model. Unutar programskog paketa Statistica 8.0 unutar modula *Advanced Linear/Nonlinear Models – General Regression Models – Multiple Regression*, integrirani su *all effects, backward stepwise, forward stepwise, forward entry, backward removal* i *best subsets* algoritmi selekcije varijabli u regresijski model.

CILJ ISTRAŽIVANJA

Cilj je ovog istraživanja napraviti usporednu analizu efikasnosti različitih algoritama odabira varijabli u model višestruke regresijske analize u izabranom kineziološkom istraživanju. Pritom će se koristiti različiti algoritmi: *all effects, backward stepwise, forward stepwise, forward entry, backward removal* i *best subsets* odabira varijabli u regresijski model. U skladu s tim, također je cilj dati opće metodološke smjernice znanstvenicima i praktičarima u kineziologiji za korištenje algoritama selekcije prediktorskih varijabli u regresijski model.

METODE ISTRAŽIVANJA

Uzorak entiteta sastojao se od 97 učenika koji su mjereni u varijablama: visina tijela (AVIS), dužina noge (ADN), dužina ruke (ADR), širina ramena (AŠR), širina zdjelice (AŠZ), dijametar ručnog zgloba (ADRZ), dijametar koljena (ADK), težina tijela (ATT), opseg podlaktice (AOPL), opseg potkoljenice (AOPK), opseg grudnog koša (AOGK), kožni nabor nadlaktice (AKNN), kožni nabor leđa (AKNL), kožni nabor trbuha (AKNT) i podizanje trupa iz ležanja s pogrčenim nogama (MDTR), a svaka varijabla mjerena je 3 puta, a u razmatranje je uzeta srednja vrijednost. Sukladno s ciljem istraživanja, korištena je višestruka regresijska analiza te su se pritom koristili različiti algoritmi odabira varijabli u model: *all effects, backward stepwise, forward stepwise, forward entry, backward removal* i *best subsets*. Pritom je varijabla podizanje trupa iz ležanja s pogrčenim nogama uzeta kao kriterij dok je prediktorske varijable odabran skup navedenih morfoloških varijabli. Za svaku varijablu i za svaki odabir algoritma pronađeni su koeficijenti beta i njihova signifikantnost, koeficijent multiple korelacije, koeficijent multiple determinacije, korigirani koeficijent multiple determinacije te signifikantnost čitavog modela.

REZULTATI I RASPRAVA

Unutar tablice 1 nalaze se rezultati regresijske analize s pripadnom signifikantnošću za različite algoritme odabira varijabli u regresijski model: *all effects*, *backward stepwise*, *forward stepwise*, *forward entry*, *backward removal* i *best subsets*. Pritom je korištena opcija uključivanja *intercept* koeficijenta u sve modele. Možemo uočiti da su na ovom uzorku varijabli *forward stepwise* i *forward entry* algoritmi generirali isti odabir varijabli, kao i *backward stepwise* i *backward removal*. To je vjerojatno posljedica sličnosti navedenih algoritama kao i relativne homogenosti uzorka. Važno je uočiti da različiti algoritmi generiraju modele s različitim brojem varijabli. Standardno korištena metoda uključivanja svih varijabli u regresijski model – *all effects* – ima najveći koeficijent determinacije ali sve varijable osim ADRZ, ADK i AOGK ne pridonose značajno modelu (Tablica 1). Uočimo da *forward* i *backward* algoritmi generiraju modele u kojima su algoritmom uključene varijable u model i sam model značajni (osim *intercepta*), ali je koeficijent determinacije manji nego u odnosu na *best subsets* i *all effects*.

Tablica 1. Beta koeficijenti regresijske jednadžbe sa pripadnom značajnošću za različite algoritme odabira varijabli u regresijski model: *all effects*, *backward stepwise*, *forward stepwise*, *forward entry*, *backward removal* i *best subsets*

	All effects		Forward stepwise		Backward stepwise		Forward entry		Backward removal		Best subsets	
	Beta	p	Beta	p	Beta	p	Beta	p	Beta	p	Beta	p
Intercept	-0.43	0.99	-0.03	1.00	4.62	0.75	-0.03	1.00	4.62	0.75	-1.74	0.92
AVIS	-0.12	0.72										
AND	0.46	0.21									0.32	0.23
ADR	-0.37	0.42									-0.36	0.31
AŠR	0.27	0.75										
AŠZ	-0.19	0.82										
ADRZ	8.26	0.02	7.61	0.00	9.03	0.00	7.61	0.00	9.03	0.00	8.70	0.01
ADK	-6.68	0.03			-6.22	0.01			-6.22	0.01	-6.88	0.01
ATT	0.02	0.98										
AOPL	0.58	0.61										
AOPK	-0.28	0.73										
AOGK	0.78	0.05			0.63	0.01			0.63	0.01	0.81	0.00
AKNN	-0.60	0.17	-0.83	0.00	-0.90	0.00	-0.83	0.00	-0.90	0.00	-0.60	0.13
AKNL	-0.71	0.26									-0.59	0.23
AKNT	0.07	0.88										

Iz tablice 2 vidljivo je, algoritam *all effects* kao posljedicu uključenja svih varijabli u model ima veći nivo signifikantnosti nego ostali algoritmi kojima je nivo signifikantnosti 0,00 odnosno koji su statistički 100% pouzdani. Potrebno je naglasiti da je algoritam *best subset* realiziran uz opciju „stop=7”, što je opravdano u kontekstu rezultata ostalih algoritama.

Tablica 2. Koeficijent multiple korelacije (*R*), koeficijent multiple determinacije (*R*²), korigirani koeficijent multiple determinacije (Korigirani *R*²), broj stupnjeva slobode (*df*) i nivo signifikantnosti regresijskog modela (*p*) za različite algoritme odabira varijabli u regresijski model: *all effects*, *backward stepwise*, *forward stepwise*, *forward entry*, *backward removal* i *best subsets*

	R	R²	Korigirani R²	df	p
All effects	0,51	0,26	0,13	14	0,02
Forward stepwise	0,39	0,15	0,13	2	0,00
Backward stepwise	0,48	0,23	0,20	4	0,00
Forward entry	0,39	0,15	0,13	2	0,00
Backward removal	0,48	0,23	0,20	4	0,00
Best subsets	0,50	0,25	0,20	7	0,00

ZAKLJUČAK

U kineziološkim istraživanjima standardno koristi *all effects* postupak selekcije varijabli u regresijski model – sve se varijable uključuju u model. Taj postupak generira najveći koeficijent determinacije, ali nivo signifikantnosti je veći u odnosu na druge algoritme. Iako često nedostatno tretirana, algoritamska selekcija varijabli unutar pojedinog regresijskog modela bi trebala biti prvi korak pri primjeni višestruke regresijske analize, u znanstvenim istraživanjima. To naročito stoga jer:

1. Preciznost predikcije i značajnost samog modela može biti poboljšana ako iz istraživanja izuzmemo nebitne i redundantne varijable.
2. Prediktor je postaje jednostavniji za interpretaciju te kasniju primjenu.
3. Znanje o tome koje varijable su relevantne pri tumačenju varijance kriterija daje „čisti” uvid u prirodu veze kriterija sa prediktorima te pruža mogućnost boljeg interpretiranja samog problema.
4. Jeftinije je vršiti mjerenja na manjem skupu varijabli.

Nužno je naglasiti da se dobiveni zaključci u ovom radu mogu primjenjivati ne samo u kineziološkim već i u svim drugim znanstvenim istraživanjima. Zaključno, u daljnjim metodološkim istraživanjima ovog tipa, a posebice u primijenjenoj kineziologiji, trebalo bi koristiti homogenije uzorke ispitanika, ispitivati algoritme za više kriterijskih varijabli te usporediti kakve bi efekte na pojedini algoritam imali drugačiji odabiri uključenja/isključenja *intercept* koeficijenta kao i modifikacije p odnosno F levela u *stepwise* algoritmima za uključenje i isključenje varijabli u model.

LITERATURA

1. Cotta, C., Sloper, C. & Moscato, P. (2004). Evolutionary Search of Thresholds for Robust Feature Set Selection: Application to the Analysis of Microarray Data. In *EvoWorkshops*, 21-30.
2. Draper, N. & Smith, H. (1981). *Applied Regression Analysis, 2nd Edition*, New York: John Wiley & Sons, Inc.
3. Ganster, H., Pinz, A., Rohrer, R., Wilding, E., Binder, M. & Kittler, H. (2001). Automated Melanoma Recognition. *IEEE Transactions On Medical Imaging*, 20 (3), 233-239.
4. Inza, I., Merino, M., Larranaga, P., Quiroga J., Sierra B. & Giralda M. (2001a). Feature Subset Selection by Genetic Algorithms and Estimation of Distribution Algorithms - A Case Study in the Survival of Cirrhotic Patients Treated with TIPS. *Artificial Intelligence In Medicine*, 23 (2): 187-205.
5. Inza, I., Larranaga, P. & Sierra, B. (2001b). Feature Subset Selection by Bayesian Networks: A Comparison with Genetic and Sequential Algorithms. *International Journal of Approximate Reasoning*, 27 (2): 143-164.
6. Lewis, P.M. (1962). The Characteristic Selection Problem in Recognition Systems, *Information Theory*, 8(2), 171-178.
7. Yusta, C.S., Bonrostro, P.J. & Letamendia, N.L. (2007). ASEPUMA. Asociación Española de Profesores Universitarios de Matematicas aplicadas a la Economía y la Empresa, *Recta*, 15(1).
8. Sebestyen, G. (1962). Decision-Making Processes. Pattern Recognition. NewYork: MacMillan.

THE COMPARISON OF ALGORITHMS FOR VARIABLE SELECTION IN THE REGRESSION MODEL IN KINESIOLOGICAL RESEARCH

ABSTRACT

In the regression analysis, while dealing with multiple choices of variables selection in the scientific research, appropriate algorithms for the selection of variables into the model exist with the aim to generate optimal regression model. In this paper, comparative efficiency study of different algorithms for the selection of variables into the multiple regression model in the chosen kinesiological research has been made. On the sample consisted of 97 pupils aged 6-7, a dependence of repetitive strength through morphological status has been analyzed by using different algorithms for the selection of variables into the regression models: all effects, backward stepwise, forward stepwise, forward entry, backward removal and best subsets. Variables of longitudinal and transversal skeleton dimensionality, voluminosity and subcutaneous fat tissue measures have been used as a set of predictor variables. The results imply that in the applied kinesiological research, from purely methodological point of view, before the model is finally generated, the problem of which variables should be included into the model has to be analyzed. The obtained conclusions can be applied, not only in kinesiological research, but also in research in other applied sciences.

Key words: *methodology, multiple regression analysis, variable selection, algorithm*

Rad je napisan u okviru projekta 315-1773397-3332 (Laboratorijski mjerni instrumenti u kineziologiji) kojeg podupire Ministarstvo znanosti, obrazovanja i športa Republike Hrvatske.
