HEURISTIČKI ALGORITAM REGRESIONE STABILNOSTI

Branka Dimitrijević, Vladimir Simić
Saobraćajni fakultet Univerziteta u Beogradu

Sadržaj: *Pri formiranju raznovrsnih modela saobraćajnih tokova prisustvo multikolinearnosti može prouzrokovati višestruke negativne efekte. Iako su razvijene brojne metode kojima se pre svega "ublažava" efekat pomenutog problema, one nisu značajnije primenjivane u oblasti saobraćaja i transporta. Osnovni nedostatak postojećih pristupa za rešavanje problema multikolinearnosti predstavlja potpuno marginalizovanje procesa selekcije nezavisnih promenljivih. U ovom radu su prezentovani osnovni elementi i proces funkcionisanja razvijenog heurističkog algoritma za selekciju nezavisnih promenljivih regresionog modela, koji poseban akcenat stavlja na dostizanje stabilnosti i visoke preciznosti. Pokazano je da razvijeni heuristički algoritam ima efikasnost u procesu selekcije. Kao ilustrativni primer prikazan je izbor nezavisnih promenljivih u regresionom modelu koji određuje tražnju za uslugama ekspres prenosa pošiljaka.*

Ključne reči: *heuristički algoritam, multikolinearnost, višestruka regresija, post express model.*

1. Uvod

U oblasti planiranja i upravljanja saobraćajnim tokovima važan aspekt predstavlja kvalitetno modeliranje posmatrane pojave. Na kvalitet formiranog modela u značajnoj meri utiče uspešnost rešavanja problema multikolinearnosti. Štaviše, prisustvo multikolinearnosti u saobraćajnim modelima može prouzrokovati veoma ozbiljne posledice.

Pojam "multikolinearnost" označava pojavu kada je neka nezavisna promenljiva linearna kombinacija dve ili više drugih nezavisnih promenljivih i treba je pravovremeno detektovati i otkloniti. Pojava multikolinearnosti izaziva nestabilnost projektovanog modela, odnosno pojavu da male promene u vrednostima nezavisnih promenljivih uključenih u model mogu proizvesti velike promene vrednosti zavisne promenljive [1]. Jednostavniji pristupi za rešavanje problema multikolinearnosti se ili zasnivaju na prostom isključivanju potencijalno kolinearnih nezavisnih promenljivih iz dalje regresione analize ili nešto složenijoj modifikaciji nezavisnih promenljivih preko faktorske analize. Međutim, najkorišćenije metode za rešavanje problema multikolinearnosti su svakako one iz grupacije Biased regresije, čiji osnovni cilj je dobijanje "tačnijih" procena regresionih koeficijenata. Najpoznatiji metodi iz ove grupe su: Ridge regresija [2], koja "veštački" smanjuje nivo korelacije među zavisnim promenljivima da bi se dobile stabilnije procene; i Regresija osnovnih komponenata [3], koja predstavlja kombinaciju analize osnovnih komponenata (koja se prvo primenjuje u procesu stabilizacije modela) i metode poslednjeg kvadrata (koja se primenjuje naknadno u cilju određivanja vrednosti regresionih koeficijenata nezavisnih promenljivih za koje je u primarnoj analizi utvrđeno da su osnovne).

Generalno je rašireno shvatanje da se problem multikolinearnosti može najbolje prevazići iznalaženjem preciznijih vrednosti regresionih koeficijenata (u daljem tekstu estimatora). O tome svedoče aktuelni trendovi razvoja ove izuzetno važne i dinamične oblasti istraživanja, koji su pre svega samo nastavili razvoj već zastarelih estimatora poput Stein-ovog estimatora [4] i Ridge regresionog estimatora [5]. U skladu sa tim, u estimatore nove generacije se mogu ubrojati Liu-ov [6], kompleksni RLSE [7] i stohastički Liu-ov [8]. Posebno treba naglasiti da je osnovno ograničenje postojećih pristupa potpuno stavljanje u drugi plan procesa selekcije nezavisnih promenljivih, odnosno postojanje mogućnosti da važne nezavisne promenljive budu isključene iz modela. Prema tome, analizom pomenutih pravaca istraživanja, uočene zanemarenosti problema multikolinearnosti i identifikovanih ograničenja postojećih pristupa, ovaj rad ima za cilj razvoj postupka selekcije nezavisnih promenljivih regresionog modela sa posebnim fokusom na njegovu stabilnost.

U nastavku, rad je organizovan na sledeći način. U poglavlju 2 dat je širok osvrt na problem modeliranja u prisustvu multikolinearnosti. U poglavlju 3 je prezentovan razvijeni heuristički algoritam, koji je u poglavlju 4 primenjen na rešavanje numeričkog primera iz domena poštanskog saobraćaja. Poglavlje 5 sadrži zaključna razmatranja.

2. Problem modeliranja u prisustvu multikolinearnosti

Multikolinearnost prouzrokuje povećanje vrednosti procenjenih regresionih koeficijenata (parametara), situaciju da procenjene parametre karakterišu nekonzistentni predznaci, formiranje regresionih modela koji ne obuhvataju "značajne" nezavisne promenljive, stvaranje situacija u kojima male promene vrednosti ulaznih podataka mogu prouzrokovati velike promene u vrednostima procenjenih parametara i sl. [9]. Multikolinearnost se može identifikovati na neki od sledećih načina:

- *Posmatranjem matrice koeficijenata korelacije nezavisnih promenljivih R*. Vrednosti veće od 0,80 su veoma problematične. Međutim, u slučaju višestruke regresione analize nije preporučljivo oslanjanje samo na ovaj jednostavan metod;
- *Određivanjem vrednosti determinante korelacione matrice det(R)*. Ona može da uzima vrednosti iz intervala od 0 (savršena multikolinearnosti) do 1 (nema multikolinearnosti));
- *Analizom VIF[1] vrednosti pojedinih nezavisnih promenljivih uključenih u model.* Ove vrednosti se smatraju najpouzdanijim pokazateljma postojanja multikolinearnosti u modelu;
- *Određivanjem uslovnog broja (UB), uslovnog indeksa (UI) i sopstvenih vrednosti*. Uslovni broj predstavlja uslovni indeks najveće vrednosti [1]:

[1] Engl. Variance Inflation Factor

$$UB = \sqrt{\lambda_{max} / \lambda_{min}} \qquad (1)$$

gde λ_{max} označava najveću sopstvenu vrednost, a λ_{min} označava najmanju sopstvenu vrednost. Pri tome, kada je: $UB \approx 1$ - nema korelacije; $1<UB<30$ - multikolinearnost postoji; $UB>30$ - postoji ozbiljna multikolinearnost; i $UB \approx \infty$ - postoji savršena multikolinearnost. S druge strane, uslovni indeks *i*-te nezavisne promenljive izračunavamo kao [1]:

$$UI_i = \sqrt{\lambda_{max} / \lambda_i}, \quad i=1,2,...,n \qquad (2)$$

pri čemu λ_i označava *i*-tu sopstvenu vrednost. Sopstvene vrednosti pružaju informaciju o broju različitih dimenzija među nezavisnim promenljivim. Tako na primer, više sopstvenih vrednosti bliskih 0 nagoveštavaju visoku multikolinearnost, dok bi idealan slučaj predstavljao $\lambda_i=1, \forall i$; i

- *Analizom osetljivosti.* Analiza promene vrednosti procenjenih regresionih koeficijenata može biti dobar pokazatelj nestabilnosti modela, jer ukoliko se utvrdi da procenjene vrednosti značajnije variraju tada će multikolinearnost biti sasvim izvesna.

U ovom radu je analiza stabilnosti modela bazirana na ispitivanju *VIF* vrednosti, kao najpouzdanijih indikatora multikolinearnosti.

2.1. Detekcija multikolinearnosti upotrebom VI faktora

VIF kvantifikuje "precenjenost" varijanse standardne greške procenjenih nestandardizovanih regresionih koeficijenata. Prema tome, *VIF* predstavlja faktor "precenjenosti" varijanse procenjenog koeficijenta regresije a_i ($i=1,2,\ldots,I$) [1]:

$$VIF_i = \left(1-R_i^2\right)^{-1} \qquad (3)$$

gde R_i predstavlja vrednost koeficijenta višestruke korelacije pri posmatranju uticaja nezavisnih promenljivih uključenih u model u slučaju *i*-te nezavisne promenljive. Kvadratni koren *VIF* *i*-te nezavisne promenljive $\sqrt{VIF_i}$, govori o tome koliko je varijansa standardne greške *i*-tog procenjenog koeficijenta "precenjena", odnosno koliko je ona veća nego što bi bila u slučaju da je $VIF_i=1$. Na primer, ako je $VIF_i=9$, tada bi standardna greška *i*-tog procenjenog koeficijenta bila 3 puta manja nego da je $VIF_i=1$.

Generalno je pravilo da vrednosti *VIF* iz intervala od 4-10 upozoravaju na potrebu za daljim ispitivanjima moguće nestabilnosti modela, dok vrednosti $VIF_i>10$ ukazuje na ozbiljnu multikolinearnost za čije prevazilaženje su neophodne prepravke formiranog modela [10]. Pored toga, treba pomenuti da $VIF_i \approx 1$ označava da nema korelacije između promenjivih uključenih u model i posmatrane *i*-te nezavisne promenljive.

3. Heuristički algoritam regresione stabilnosti

Prilikom definisanja algoritamskog prikaza predložene heuristike korišćena je sledeća notacija.

Neka Ω_t ($\Omega_t \subseteq N$) predstavlja skup svih nezavisnih promenljivih, koje su zaključno sa t-tom iteracijom uključene u regresioni model, Ω^* ($\Omega_t \subseteq \Omega^*$) označava skup nezavisnih promenljivih uključenih u rešenje heuristike, Ω^{**} označava skup nezavisnih promenljivih uključenih u rešenje algoritma, Ψ_t ($\Psi_t \subseteq N$) skup svih nezavisnih promenljivih koje su zaključno sa t-tom iteracijom isključene iz rešenja, a A_t ($A_t \subseteq N$) skup aktivnih nezavisnih promenljivih u t-toj iteraciji (odnosno, promenljivih koje su preostale da u narednim iteracijama budu vrednovane). Takođe, važi da je: $N = \Omega_t \cup \Psi_t \cup A_t$ i $\Omega_t \cap \Psi_t \cap A_t = \varnothing$. Pošto algoritam pretpostavlja selekciju tipa "unapred", za početne uslove je usvojeno $\Omega_1 = \varnothing$, $\Psi_1 = \varnothing$ i $A_1 = N$. Pored toga, u radu je usvojeno da značaj i-te nezavisne promenljive v_i bude predstavljen parcijalnim koeficijentom korelacije, jer upravo ovaj statistički pokazatelj predstavlja doprinos koji će model imati ukoliko se pomenuta nezavisna promenljiva priključi regresionom modelu. Ova "interna" karakteristika nezavisne promenljive se ustvari ilustrativno može smatrati i njenim "težinskim koeficijentom". Prema tome, u t-toj iteraciji "težinski koeficijent" i-te nezavisne promenljive se izračunava kao:

$$V_i^t = r_{yi \cdot \Omega_{t-1}}, \; i \in A_t, \; t = 2,3,... \tag{4}$$

gde je pri određivanju korelacije između zavisne promenljive y i i-te nezavisne promenljive uzeto da su sve regresione promenljive konstantne. Inače, pošto na početku procesa selekcije nezavisnih promenljivih važi da je $\Omega_1 = \varnothing$, "težinske koeficijente" je potrebno izračunati pomoću jednostavnih koeficijenata korelacije:

$$V_i^1 = \left| r_{yi} \right|, \; i \in N \tag{5}$$

U jednačini (5) upotreba apsolutnih vrednosti koeficijenata korelacije je determinisana intervalom vrednosti koje on uzima, odnosno vrednostima od –1 do 1. Ukoliko se u t-toj iteraciji (t=2,3,...) utvrdi da i-tu ili i-te nezavisne promenljive karakterišu negativni koeficijenti korelacije, njih treba uvrstiti u skup Γ_t ($\Gamma_t = \left\{ i \mid V_i^t < 0 \right\}, \forall i \in A_t, t$=2,3,...) i isključiti iz dalje analize.

Kriterijumi za vrednovanje nezavisnih promenljivih modela su bazirani na konceptu stabilnosti, odnosno na sledeće tri karakteristike: $P^{t'}(i)$, $P^{t''}(i)$ i $P^{t'''}(i)$.

Osnovna ideja I karakteristike je da se za svaku regresionu promenljivu $i \in A_t$ (nezavisnu promenljivu već uključenu u regresioni model) ustanovi koje nezavisne promenljive $j \in \{A_t \backslash i\}$ ne bi destabilizovale postojeći regresioni model, odnosno kada je VIF_{ij}^t<10. Prema tome, može se reći da $P^{t'}(i)$ predstavlja skup svih onih nezavisnih promenljivih $j \in \{A_t \backslash i\}$ koje uključivanjem u model u t-toj iteraciji ne bi izazvale pojavu multikolinearnosti (odnosno, ne bi ga destabilizovale), ako je nezavisna promenljiva $i \in A_t$ regresiona promenljiva:

$$P^{t'}(i) = \left\{ j \mid VIF_{ij}^t < 10, \; j \in \{A_t \backslash i\} \right\}, \; t=2,3,... \tag{6}$$

gde je VIF_{ij}^{t} skraćeno od $VIF_{j\cdot[i\cup\Omega_{t-1}]}^{t}$ i predstavlja *VIF* vrednost u *t*-toj iteraciji koju promenljiva $i \in A_t$ ima prema nezavisnoj promenljivoj $j \in \{A_t \backslash i\}$ u slučaju kada se regresionim mogu smatrati promenljive $i \cup \Omega_{t-1}$.

Ideja II karakteristike je da se za svaku regresionu promenljivu utvrdi koje bi to nezavisne promenljive $i \in A_t$ održale stabilnost regresionog modela. Prema tome, može se reći da $P^{t''}(i)$ predstavlja skup svih onih regresionih promenljivih $j \in \{A_t \backslash i\}$ za koje model ostaje stabilan, ako u *t*-toj iteraciji u model uključimo nezavisnu promenljivu $i \in A_t$:

$$P^{t''}(i) = \left\{ j \mid VIF_{ji}^{t} < 10,\ j \in \{A_t \backslash i\} \right\},\ t=2,3,... \tag{7}$$

Konačno, III karakteristika $P^{t'''}(i)$ predstavlja skup svih onih regresionih promenljivih $j \in \{A_t \backslash i\}$, koje vode "indirektnoj" nestabilnosti modela. Ona se javlja u situacijama kada za svaku regresionu promenljivu $i \in A_t$ uključivanje u model nezavisnih promenljivih $j \in \{A_t \backslash i\}$ neće direktno narušiti uslov stabilnosti (odnosno, biće i dalje ispunjeno $VIF_{j\cdot[i\cup\Omega_{t-1}]}^{t} < 10$), ali će model ipak učiniti nestabilnim, pošto će postojati makar jedna regresiona promenljiva iz skupa rešenja ($k \in \Omega_{t-1}$), koja neće više ispunjavati uslov stabilnosti. Sa druge strane, "direktnu" nestabilnost modela prouzrokuju oni parovi nezavisnih promenljivih koji ne ispunjavaju uslove stabilnosti modela iz jednačina (6) i (7). $P^{t'''}(i)$ se određuje kao:

$$P^{t'''}(i) = \left\{ j \mid \exists\ vif_{kj\cdot[i\cup(\Omega_{t-1}\backslash k)]}^{t} < 10,\ j \in \{A_t \backslash i\},\ k \in \Omega_{t-1} \right\},\ t=2,3,... \tag{8}$$

Nakon određivanja elemenata skupova $P^{t'}(i)$, $P^{t''}(i)$ i $P^{t'''}(i)$ treba izračunati združenu karakteristiku $P^{t}(i)$ kao:

$$P^{t}(i) = \left[P^{t'}(i) \cap P^{t''}(i) \right] \backslash P^{t'''}(i),\ \ \forall i \in A_t \tag{9}$$

Vrednovanje nezavisnih promenljivih $i \in A_t$ se vrši na bazi projektovanih kriterijuma selekcije. Na osnovu združene karakteristike (jednačina (9)) su ustanovljeni kriterijumi $\boldsymbol{\alpha}$ i $\boldsymbol{\beta}$, a njihovom daljom modifikacijom i kombinacijom kriterijumi $\boldsymbol{\alpha\beta}$, **w**$\boldsymbol{\alpha}$, ***w***$\boldsymbol{\beta}$ i **w**$\boldsymbol{\alpha\beta}$.

Kriterijum $\boldsymbol{\alpha}$ se bazira na tzv. "α koristi" koja se ostvaruje izborom nezavisne promenljive, odnosno njenim uključivanjem u rešenje. Pri tome, odgovarajuća vrednost "α koristi" se izražava kao razlika težinskog koeficijenta *i*-te nezavisne promenljive i sume težinskih koeficijenata svih onih nezavisnih promenljivih koje bi ovaj izbor isključio iz rešenja:

$$\alpha_t(i) = v_i^t - \sum_{i \in \left\{ A_t \backslash [P^t(i) \cup i] \right\}} v_i^t,\ i \in A_t \tag{10}$$

Kriterijum $\boldsymbol{\beta}$ se bazira na tzv. "β koristi" koja se ostvaruje uključivanjem nezavisne promenljive u regresioni model. Pri tome, odgovarajuća vrednost "β koristi" se izražava kao zbir težinskog koeficijenta *i*-te nezavisne promenljive i sume težinskih koeficijenata svih onih nezavisnih promenljivih koje ovaj izbor ne bi isključio iz rešenja:

$$\beta_t(i)=v_i^t + \sum_{i\in P^t(i)} v_i^t,\ i\in A_t \tag{11}$$

Kriterijum $\alpha\beta$ je svojevrsna sublimacija kriterijuma α i β. On praktično predstavlja “ukupnu korist” ostvarenu izborom određene nezavisne promenljive:

$$\alpha\beta_t(i)=v_i^t + \sum_{i\in P^t(i)} v_i^t - \sum_{i\in\left\{A_t\setminus[P^t(i)\cup i]\right\}} v_i^t,\ i\in A_t \tag{12}$$

Kriterijumi $w\alpha$ (jednačine (13)), $w\beta$ (jednačine (14)) i $w\alpha\beta$ (jednačine (15)), predstavljaju svojevrsne modifikacije kriterijuma α, β i $\alpha\beta$, respektivno. Oni veći značaj pridaju samim nezavisnim promenljivima (odnosno, njihovim težinskim koeficijentima) nego združenim karakteristikama:

$$w\alpha_t(i)=v_i^t - \left(v_i^t \Big/ \sum_{i\in A_t} v_i^t \right) \sum_{i\in\left\{A_t\setminus[P^t(i)\cup i]\right\}} v_i^t,\ i\in A_t \tag{13}$$

$$w\beta_t(i)=v_i^t + \left(v_i^t \Big/ \sum_{i\in A_t} v_i^t \right) \sum_{i\in P^t(i)} v_i^t,\ i\in A_t \tag{14}$$

$$w\alpha\beta_t(i)=v_i^t + \left(v_i^t \Big/ \sum_{i\in A_t} v_i^t \right)\cdot\left(\sum_{i\in P^t(i)} v_i^t - \sum_{i\in\left\{A_t\setminus[P^t(i)\cup i]\right\}} v_i^t \right),\ i\in A_t \tag{15}$$

Algoritam za heuristike Hα, Hβ, H$\alpha\beta$, H$w\alpha$, H$w\beta$ i H$w\alpha\beta$, zasnovane na koristima α, β, $\alpha\beta$, $w\alpha$, $w\beta$ **i** $w\alpha\beta$, respektivno, dat je u integralnoj formi, pri čemu je u pojedinim koracima algoritma eksplicitno naznačeno kada se nešto odnosi na konkretnu heuristiku:

KORAK 1 Inicijalizacija t=1

KORAK 2 Inicijalizacija $\Omega_t=\varnothing, \Psi_t=\varnothing, A_t=N$

KORAK 3 Za svaku nezavisnu promenljivu $i\in A_t$ ako je $t\neq1$ odrediti V_i^t u skladu sa jednačinom (4), u suprotnom V_i^t odrediti u skladu sa jednačinom (7)

KORAK 4 Ako je $t\neq1$ preći na KORAK 5, u suprotnom preći na KORAK 6

KORAK 5 Ako postoji V_i^t<0 ($i\in A_t$) preći na KORAK 5.1, u suprotnom preći na KORAK 6

KORAK 5.1 Odrediti skup Γ_t: $\Gamma_t=\left\{i \mid V_i^t<0\right\}, \forall i\in A_t$

KORAK 5.2 Postaviti $t=t+1$

KORAK 5.3 Ažurirati skup Ω_t: $\Omega_t=\Omega_{t\text{-}1}$

KORAK 5.4 Ažurirati skup Ψ_t: $\Psi_\mathrm{t}=\Psi_{t-1}\cup\Gamma_{t\text{-}1}$

KORAK 5.5 Ažurirati skup A_t: $A_t=A_{t\text{-}1}\setminus\Psi_t$

KORAK 5.6 Preći na KORAK 17

KORAK 6 Ako je $\left|A_t\right|=1$ preći na KORAK 6.1, u suprotnom preći na KORAK 7

KORAK 6.1 Postaviti $t=t+1$

KORAK 6.2 Ažurirati skup Ω_t: $\Omega_t = \Omega_{t\text{-}1} \cup A_{t\text{-}1}$

KORAK 6.3 Ažurirati skup Ψ_t: $\Psi_t = \Psi_{t\text{-}1}$

KORAK 6.4 Ažurirati skup A_t: $A_t = \varnothing$ i preći na KORAK 18

KORAK 7 Odrediti matricu VIF: $VIF^t = \left\| VIF_{ij}^t \right\|_{|A_t| \times |A_t|}$ u skladu sa jednačinom (3)

KORAK 8 Ako je t=1, preći na KORAK 9, u suprotnom preći na KORAK 10

KORAK 9 $\forall i \in A_t$ odrediti skup $P^t(i)$: $P^t(i)=\left\{ j \mid VIF_{ij}^t < 10,\ j \in \{A_t \backslash i\} \right\}$

KORAK 9.1 Preći na KORAK 11

KORAK 10 $\forall i \in A_t$ odrediti skup $P^{t'}(i)$ u skladu sa jednačinom (6)

KORAK 10.1 $\forall i \in A_t$ odrediti skup $P^{t''}(i)$ u skladu sa jednačinom (7)

KORAK 10.2 $\forall i \in A_t$ odrediti skup $P^{t'''}(i)$ u skladu sa jednačinom (8)

KORAK 10.3 $\forall i \in A_t$ odrediti skup $P^t(i)$ u skladu sa Jednačinom (9)

KORAK 11 $\forall i \in A_t$ izračunati vrednosti kriterijuma: $\alpha_t(i)$ prema jednačini (10) za $\boldsymbol{H\alpha}$; $\beta_t(i)$ prema jednačini (11) za $\boldsymbol{H\beta}$; $w\alpha_t(i)$ prema jednačini (12) za $\boldsymbol{Hw\alpha}$; $w\beta_t(i)$ prema jednačini (13) za $\boldsymbol{wH\beta}$; $\alpha\beta_t(i)$ prema jednačini (14) za $\boldsymbol{H\alpha\beta}$; $w\alpha\beta_t(i)$ prema jednačini (15) za $\boldsymbol{Hw\alpha\beta}$

KORAK 12 Odrediti nezavisnu promenljivu i^* za koju važi: $\alpha_t^*(i) = \max_{i \in A_t}[\alpha_t(i)]$ za $\boldsymbol{H\alpha}$; $\beta_t^*(i) = \max_{i \in A_t}[\beta_t(i)]$ za $\boldsymbol{H\beta}$; $\alpha\beta_t^*(i) = \max_{i \in A_t}[\alpha\beta_t(i)]$ za $\boldsymbol{H\alpha\beta}$; $w\alpha_t^*(i) = \max_{i \in A_t}[w\alpha_t(i)]$ za $\boldsymbol{wH\alpha}$; $w\beta_t^*(i) = \max_{i \in A_t}[w\beta_t(i)]$ za $\boldsymbol{wH\beta}$; $w\alpha\beta_t^*(i) = \max_{i \in A_t}[w\alpha\beta_t(i)]$ za $\boldsymbol{wH\alpha\beta}$. Ako dve ili više nezavisnih promenljivih imaju istu maksimalnu vrednost: $\alpha_t^*(i)$; $\beta_t^*(i)$; $\alpha\beta_t^*(i)$; $w\alpha_t^*(i)$; $w\beta_t^*(i)$; $w\alpha\beta_t^*(i)$; uključiti ih u skup A_t', pa za i^* usvojiti onu nezavisnu promenljivu za koju važi da je $V_i^{t*} = \max_{i \in A_t'} V_i^t$.

KORAK 13 Postaviti t=t+1

KORAK 14 Ažurirati skup Ω_t: $\Omega_t = \Omega_{t\text{-}1} \cup i_{t\text{-}1}^*$

KORAK 15 Ažurirati skup Ψ_t: $\Psi_t = \Psi_{t\text{-}1} \cup \{ j \mid j \in [A_{t\text{-}1} \backslash (P^{t\text{-}1}(i_{t\text{-}1}^*) \cup i_{t\text{-}1}^*)]\}$

KORAK 16 Ažurirati skup A_t: $A_t = A_{t\text{-}1} \backslash [\Psi_t \cup i_{t\text{-}1}^*]$

KORAK 17 Ako je $A_t \neq \varnothing$, vratiti se na KORAK 3, u suprotnom preći na KORAK 18

KORAK 18 Usvojiti $\Omega^* = \Omega_t$, $\Psi^* = \Psi_t$

KORAK 19 Izračunati S_y^2 vrednosti rezidijumskih disperzija za svaku od heuristika:

$$S_y^2 = \sum_{j=1}^{J} \left(y_j - y_{rj}\right)^2$$

KORAK 20 Izračunati $S_{y_{\min}}^2$: $S_{y_{\min}}^2 = \min[S_{y\_\mathrm{H}\alpha}^2, S_{y\_\mathrm{H}\beta}^2, S_{y\_\mathrm{H}\alpha\beta}^2, S_{y\_\mathrm{Hw}\alpha}^2,$

$S_{y\_\mathrm{Hw}\alpha\beta}^2, S_{y\_\mathrm{Hw}\alpha\beta}^2]$ i odgovarajući skup Ω^{**}

KORAK 21 Kraj algoritma.

4. Numerički primer - modeliranje tražnje usluge Post Express

Ako se ima u vidu da je ekspres prenos pošiljaka tokom svog razvoja postao moćna industrija, koja ima značajan uticaj na globalna ekonomska kretanja i poslovanje preduzeća u svim privrednim granama, razumljivo je interesovanje za ovu oblast poštanskog saobraćaja. Ekspres industrija beleži stalni rast obima ostvarenih usluga, a samim tim i prihoda, što implicira njen dalji razvitak i povećanje udela koji ona ima u globalnoj ekonomiji [11].

Predvuđanje obima saobraćaja je oduvek bila jedna od glavnih preokupacija poštanskih uprava. U okviru faze strateškog planiranja određene usluge, veoma je važno proceniti faktore, koji imaju ili bi mogli imati pozitivan ili negativan uticaj na njen razvoj. U ovom numeričkom primeru, zavisnu promenljivu je predstavljao broj Post Express pošiljaka. Sa druge strane, od mnogobrojnih uticajnih faktora tražnje Post Express pošiljaka, za nezavisne promenljive Post Express modela odabrane su 1. vreme, 2. BDP, 3. obim saobraćajne grupe aktivnosti u okvirima BDP-a, 4. doprinos “saobraćajne aktivnosti” rastu BDP-a, 5. indeks rasta cena na malo, 6. indeks troškova života, 7. indeks bazne inflacije, 8. izvoz, 9. uvoz, 10. zarade po zaposlenom na republičkom nivo, 11. zarade po zaposlenom u Beogradu, 12. ocena postojećeg privredog stanja, 13. očekivanja srpskih privrednika, 14. broj pravnih lica prema Zakonu o klasifikaciji delatnosti, 15. broj “svih” pravnih lica, 16. indeks broja zaposlenih u RS, 17. indeks produktivnosti rada i 18. marketing. Inače, detaljan pregled korišćenih vrednosti se može naći u [12].Rezultati algoritma po pojedinim heuristikama su priloženi u tabeli 1.

Tabela 1. *Rezultati algoritma*

Heuristika	Ω^*	Vrednosti rezidijumske disperzije
α	{18,8,15,4,17}	9,952,867,617.13
β	{18,8,15,4,17}	9,952,867,617.13
αβ	{18,8,15,4,17}	9,952,867,617.13
wα	{18,17,13,15}	8,694,483,108.73
wβ	{18,17,13,15}	8,694,483,108.73
wαβ	{18,8,17,15}	10,046,446,287.23

Iz pomenute tabele može se primetiti da se minimalna vrednost rezidijumske disperzije postiže u slučaju heuristika ***wα*** i ***wβ***, kao i da ona iznosi 8,694,483,108.73. U skladu sa tim, skup nezavisnih promenljivih dobijenih kao rešenje algoritma regresione stabilnosti čine: marketing, indeks produktivnosti rada, očekivanja srpskih privrednika i broj “svih” pravnih lica.

5. Zaključak

Prilikom projektovanja raznovrsnih saobraćajnih modela prisustvo multikolinearnosti može prouzrokovati ozbiljne probleme i imati veoma negativne efekte. Iako su razvijene brojne metode kojima se pre svega “ublažava” efekat pomenutog problema, one nisu značajnije primenjivane u saobraćaju i transportu.

U ovom radu je predložen algoritam regresione stabilnosti, tzv. heuritički algoritam selekcije “dozvoljenih” nezavisnih promenljivih. Iako on ne garantuje optimalnost modela, prilikom njegovog testiranja je potvrđena zadovoljavajuća preciznost procene. Kao ilustrativni primer prikazan je regresioni model za određivanje tražnje usluge ekspres prenosa pošiljaka. Predloženi algoritam obezbeđuje efikasnost procesa selekcije nezavisnih promenljivih, što je i pokazano na ilustrativnom primeru rešenom posle samo 7 iteracija. Inače, formirani Post Express model je obuhvatio sledeće četiri nezavisne promenljive: marketing, indeks produktivnosti rada, očekivanja srpskih privrednika i broj “svih” pravnih lica, a karakterisala ga je i zadovoljavajuća preciznost procene broja prenetih pošiljaka.

U daljem razvoju opisanog heurističkog algoritma stabilnosti neophodno je proširiti ga i ostalim fazama formiranja strukture regresionog modela, kako bi njegov razvojni proces bio upotpunjen. Prema tome, potrebno je implementirati faze preliminarne analize podataka, izbora modela veze između zavisne i nezavisnih promenljivih i validacije modela.

Ovaj rad je podržan od strane Ministarstva za nauku i tehnološki razvoj Republike Srbije, kroz projekat TR 36006, za period 2011-2014.

Literatura

[1] D. A. Belsley, “A guide to using the collinearity diagnostics”, *Computer Science in Economics and Management*, vol. 4, pp. 33-50, 1991.

[2] D. W. Marquardt, and R. D. Snee, “Ridge regression in practice”, *American Statistician*, vol. 29, pp. 3-20, 1975.

[3] W. F. Massy, “Principal Components Regression in Exploratory Statistical Research”, *Journal of the American Statistical Association*, vol. 60, pp. 234-246, 1965.

[4] C. Stein, “Inadmissibility of the usual estimator for mean of multivariate normal distribution”, *Proceedings of the third Berkley symposium on mathematical and statistics probability*, Berkley, USA, pp. 197-206, 1956.

[5] A. E. Hoerl, and R. Kennard, “Ridge regression: Applications to nonorthogonal problems”, *Technometrics*, vol. 12, pp. 55-82, 1970.

[6] K. Liu, “A new class of biased estimate in linear regression”, *Comm Statistical Theory and Methods*, vol. 22, pp. 393-402, 1993.

[7] N. Sarkar, “A new estimator combining the ridge regression and the restricted least squares methods of estimation”, *Comm Statistical Theory and Methods* , vol. 21, pp. 1987-2000, 1992

[8] M. Hubert, and P. Wijekoon, “Improvement of the Liu estimator in the linear regression model”, *Stat Pap*, vol. 47, pp. 471-479, 2006.
[9] B. Bowerman, and R. O’Connell, *Linear Statistical Models an Applied Approach*, Boston, MA: PWS-Kent Publishing Co., 1990.
[10] D. A. Belsley, E. Kuh, and R. E. Welsch, *Regression diagnostics: Identifying influential data and sources of collinearity*, New York, NY, 1980.
[11] B. Dimitrijević, i V. Simić, “Neuro-fazi pristup pri proceni broja Post Express pošiljaka”, *Zbornik radova: PosTel 2008*, Beograd, Srbija, Decembar 16-17, str. 37-46, 2008.
[12] B. Dimitrijević, i V. Simić, “Modeliranje tražnje usluge ekspres prenosa pošiljaka primenom inovativnog postupka strukturiranja fazi regresionog modela”, *Zbornik radova: PosTel 2009*, Beograd, Serbia, Decembar 15-16, str. 85-94, 2009.

Abstract: *The presence of multicollinearity may cause different negative effects during modeling process of traffic flows. Numerous methods which reduce the effects of the aforementioned problem were developed, but they did not find significant application in the field of traffic and transportation. Marginalization of independent variables selection process is major disadvantage of available approaches for solving multicollinearity problem. In this paper we present basic elements and steps of developed heuristics algorithm for selecting independent variables in regression model which is focused on stability and high precision. It is shown that developed heuristic algorithm has an efficiency in selection process. As an illustrative example we present independent variable selection process in case of regression model which determining demand for express mail service.*

Keywords: *heuristic algorithm, multicollinearity, multiple regression, EMS model.*

A HEURISTIC ALGORITHM FOR REGRESSION STABILITY

Branka Dimitrijević, Vladimir Simić