ACTA MATHEMATICA SPALATENSIA
Series didactica
Vol.7 (2024) 49–64
https://doi.org/10.32817/amssd.7.4

Zaprimljen: 12. travnja 2024.
Prihvaćen: 29. kolovoza 2024.

Multikolinearnost u višestrukoj regresiji: detekcija i moguća rješenja

Perišić Ana, Nakić Jelena, Beljo Ivana

Sažetak

Regresijska analiza jedna je od najčešćih metoda modeliranja veze između varijable odziva i jedne ili više eksplanatornih varijabli. Kada je glavni cilj analize opisati odnos između varijable odziva i eksplanatornih varijabli, jedan od problema koji može nastupiti jest slučaj kada su barem dvije eksplanatorne varijable linearno zavisne ili približno linearno zavisne. Kažemo da je tada prisutan problem multikolinearnosti. Prisutnost multikolinearnosti može dovesti do prevelikih procijenjenih standardnih pogrešaka, nestabilnih procjena parametara, a time i neispravnih zaključaka o odnosima varijabli, čime onemogućuje ocjenu važnosti individualnih varijabli u modelu. U ovom radu bavimo se problemom multikolinearnosti u višestrukoj linearnoj regresiji. Navodimo nekoliko metoda detekcije i uklanjanja ovoga problema, sve uz primjenu na stvarnom skupu podataka.

Ključni pojmovi: multikolinearnost, kolinearnost prediktora, višestruka linearna regresija

Abstract

Regression analysis is one of the most common methods of modeling the relationship between a response variable and one or more explanatory variables. When the main goal of the study is to describe the relationship between the response variable and the explanatory variables, one of the problems that can occur is the case when at least two explanatory variables are linearly dependent or approximately linearly dependent. We

then say that the problem of multicollinearity is present. The presence of multicollinearity can lead to excessive estimated standard errors, unstable parameter estimates, and thus incorrect conclusions about variable relationships. Also, it poses difficulties in evaluating the importance of individual variables in the model. In this paper, we deal with the problem of multicollinearity in multiple linear regression. We list several methods of detecting and removing this problem, supported by the application to a real data set.

Keywords: multicollinearity, predictor collinearity, multiple linear regression

1. Uvod

Regresijska analiza bavi se proučavanjem odnosa između odabrane varijable (ovisne varijable ili varijable odziva) i jednog ili više prediktora (eksplanatornih varijabli). Dva su temeljna cilja regresijske analize: (1) opisati odnos između varijable odziva i eksplanatornih varijabli i (2) predvidjeti vrijednosti varijable odziva za dane vrijednosti eksplanatornih varijabli. Varijablu odziva označavat ćemo s Y, dok ćemo eksplanatorne varijable označavati s $X_1, X_2, ..., X_{p-1}$. U ovome radu ograničit ćemo se na proučavanje slučajeva kada je varijabla odziva u linearnoj vezi s eksplanatornim varijablama. U tom slučaju pretpostavljamo da je uvjetno očekivanje od Y za dane vrijednosti $x_1, ..., x_{p-1}$ varijabli $X_1, ..., X_{p-1}$ linearna funkcija od $x_1, ..., x_{p-1}$, to jest

$$E[Y|X_1 = x_1, ..., X_n = x_p] = \beta_0 + \beta_1 x_1 + ... + \beta_{p-1} x_{p-1}.$$

Višestruki linearni regresijski model tada je dan s

$$Y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_{p-1} x_{p-1} + \varepsilon_i, \qquad i = 1, 2, ..., n,$$

pri čemu su $\beta_0, \beta_1, ..., \beta_{p-1}$ parametri regresijskog modela, dok ε_i predstavlja slučajni šum ili grešku. Za greške pretpostavljamo da su centrirane, nekorelirane i jednakih varijanci σ^2.

Višestruki regresijski model možemo elegantnije prikazati matrično

$$Y = X\beta + \varepsilon,$$

pri čemu je X matrica tipa $n \times p$ koja sadrži informacije o vrijednostima eksplanatornih varijabli. Nazivamo je matricom dizajna i defniramo s

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & ... & x_{1,p-1} \\ 1 & x_{21} & x_{22} & ... & x_{2,p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & ... & x_{n,p-1} \end{bmatrix}.$$

Vektor Y sadrži vrijednosti varijable odziva $y_i, i = 1, 2, ..., n$. Vektor $\beta = (\beta_0, \beta_1, ..., \beta_{p-1})$ predstavlja vektor nepoznatih parametara, dok je ε vektor slučajnih varijabli koje nisu opservabilne i predstavljaju slučajne greške. Vektor predviđenih vrijednosti, $\hat{Y}$, možemo pisati u obliku

$$\underbrace{\hat{Y}}_{n\times 1} = \underbrace{X}_{n\times p}\underbrace{\beta}_{p\times 1}.$$

Procjene parametara regresije vršimo na temelju uzorka, odnosno opaženih vrijednosti varijable odziva i eksplanatornih varijabli $(y_i, x_{i1}, x_{i2}..., x_{i,p-1})$, $i = 1, 2, ..., n$, primjerice metodom najmanjih kvadrata gdje rješavanjem problema

$$\min_{\hat{\beta}} (Y - X\hat{\beta})^T (Y - X\hat{\beta}) \tag{1}$$

dobivamo procjenitelja

$$\hat{\beta} = \left(\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p\right) = (X^T X)^{-1} X^T Y. \tag{2}$$

Ako je očekivana vrijednost grešaka jednaka nuli, procjenitelj dobiven metodom najmanjih kvadrata (2) jest nepristran procjenitelj za β, a uz pretpostavku konstantne varijance i nekoreliranosti grešaka, procjenitelj (2) ima najmanju varijancu od svih linearnih nepristranih procjenitelja. Matrica kovarijanci procjenitelja $\hat{\beta}$ dana je s $\sum_{\hat{\beta}\hat{\beta}} = \sigma^2 (X^T X)^{-1}$, a varijancu procjenitelja $\hat{\beta}_l, l = 1, 2, \ldots, p$, možemo zapisati u obliku

$$Var(\hat{\beta}_l) = \sigma^2 (X^T X)^{-1}_{ll} = \frac{\sigma^2}{\sum_{i=1}^{n} (x_{il} - \overline{x}_l)^2 (1 - R_l^2)} = \frac{\sigma^2}{S_{ll}(1 - R_l^2)}, \tag{3}$$

gdje je $S_{ll} = \sum_{i=1}^{n} x_{il}^2 - n\bar{x}_l^2$, a R_l^2 predstavlja kvadrat koeficijenta višestruke korelacije između x_l i ostalih eksplanatornih varijabli, odnosno koeficijent determinacije modela gdje je x_l varijabla odziva, a ostali su prediktori eksplanatorne varijable. R_l^2 poprima vrijednosti između 0 i 1, gdje veće vrijednosti upućuju na čvršću povezanost varijabli. Detaljnije o izvodu te općenito o višestrukoj regresiji može se pronaći, na primjer, u [**rice2006**], [**natasa**].

Primijetimo da je za postojanje procjenitelja dobivenog metodom najmanjih kvadrata (2) nužan uvjet invertibilnost matrice $X^T X$, što će biti zadovoljeno u slučaju kada je zadovoljen uvjet linearne nezavisnosti eksplanatornih varijabli (matrica dizajna mora biti punog ranga). Naime, u slučaju kada prediktori nisu nezavisni nije moguće odrediti vrijednosti regresijskih parametara, čak ni u slučaju kada nije prisutan šum. Prikažimo ovo na jednostavnom primjeru gdje prikazujemo slučaj kada između dviju eksplanatornih varijabli postoji funkcionalna (linearna) veza. Ovaj slučaj poznat je kao egzaktna ili savršena kolinearnost.

Primjer 1. *Pretpostavimo da je dana jednadžba regresije*

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + u$$

pri čemu su x_1 *i* x_2 *povezani relacijom*

$$x_1 - 3x_2 = -2$$

Tada je $x_1 = 3x_2 - 2$ *pa jednadžbu regresije možemo zapisati na sljedeći način:*

$$y = (\alpha - 2\beta_1) + (3\beta_1 + 2\beta_2)x_2 + u.$$

Stoga možemo procijeniti $(\alpha - 2\beta_1)$ *i* $(3\beta_1 + 2\beta_2)$*, ali ne možemo procijeniti* α, β_1, β_2 *zasebno.*

S matematičkog gledišta, multikolinearnost će biti problem samo ako se radi o savršenoj multikolinearnosti. U praksi ćemo rijetko naići na slučaj savršene (multi)kolinearnosti prediktora, ali ćemo nerijetko biti suočeni sa skupovima podataka gdje bilježimo visoke razine koreliranosti (dijela) prediktora, a u većini slučajeva će prediktori biti u nekoj mjeri korelirani. Na primjer, u primijenjenim istraživanjima ekonomske prirode, gotovo je nemoguće pronaći dvije ili više ekonomskih varijabli koje nisu korelirane, makar u nekoj manjoj mjeri [**gujarati**]. Multikolinearnost može biti rezultat *stanja prirode* (neke su varijable po svojoj prirodi jako korelirane, neke varijable prate isti trend). Izvori multikolinearnosti mogu biti i primijenjena metoda prikupljanja podataka, specifikacija modela ili prevelik broj varijabli u modelu u odnosu na veličinu uzorka [**montgomery2021introduction**]. Multikolinearnost se može javiti u slučaju pogrešne specifikacije modela koji uključuje kategorijalnu varijablu. Naime, kada je u regresijski model uključen kategorijalni prediktor s m kategorija, isti je potrebno zamijeniti s $m-1$ *dummy* varijabli. Inače je moguća pojava savršene multikolinearnosti. Ovaj slučaj još nazivamo *dummy variable trap*, a zanimljiv primjer može se naći u [**gujarati**].

Kako ocijeniti je li koreliranost prediktora prevelika u smislu da narušavanje pretpostavke nepostojanja multikolinearnosti otežava procjenu parametara i donošenje zaključaka o modelu? U idućem poglavlju predstavljamo nekoliko metoda detekcije multikolinearnosti.

2. Odabrane metode detekcije multikolinearnosti

2.1. Simptomi multikolinearnosti

Jedan od najvažnijih problema koje uzrokuje multikolinearnost jest velika varijanca procjenitelja, odnosno standardna pogreška procjene parametara. Lako je vidjeti (izraz (3)) da velike vrijednosti R_l^2 uzrokuju

velike vrijednosti $Var(\hat{\beta}_l)$, to jest, što je prediktor jače koreliran s ostalim prediktorima, varijanca procjenitelja biti će veća. Ako je varijanca procjenitelja velika, procjenitelj će biti neprecizan, u smislu da procijenjene vrijednosti parametara leže daleko od stvarnih vrijednosti. Pripadni pouzdani intervali za β_j bit će jako široki, a testovi hipoteza o parametru bit će male snage, odnosno hipoteze tipa $H_0 : \beta_j = b_j$ neće biti odbačene za različite vrijednosti b_j. Napomenimo kako je procjenitelj (2) i dalje nepristran, najmanje varijance među svim linearnim nepristranim procjeniteljima, ali to ne znači da će varijanca procjenitelja biti mala (u odnosu na procijenjenu vrijednost parametra). Upravo će zbog velike standardne pogreške moći nastupiti slučaj kada je u regresijskom modelu vrijednost nekog parametra procijenjena pogrešnim predznakom. Istraživač može prepoznati da je predznak u višestrukom modelu procijenjen pogrešnim predznakom ako je u univarijatnom modelu predznak procijenjene vrijednosti parametra istog prediktora bio suprotnog predznaka ili ako ima teorijsko znanje o odnosu prediktora i varijable odziva. Nadalje, posljedica multikolinearnosti može biti slučaj kada je prediktor u univarijatnom modelu značajan, ali nije značajan u višestrukom regresijskom modelu. Stoga usporedba procijenjenih vrijednosti parametara u univarijatnim i multivarijatnim modelima može biti korisna pri detekciji prisutnosti multikolinearnosti.

Problem multikolinearnosti možemo usporediti s problemom malih uzoraka pri procjeni očekivanja univarijatne populacije [**goldberger**]. Naime, poznato je kako je nepristrani procjenitelj najmanje varijance za univarijatno očekivanje uzoračka sredina $\overline{y}$, a standardna pogreška procjene jednaka je $\frac{\sigma^2}{n}$. Male vrijednosti veličine uzorka n tako će dovesti do velikih vrijednosti standardne pogreške i širokih pouzdanih intervala za procjenu očekivanja. Problem savršene multikolinearnosti tada možemo usporediti sa slučajem kada je veličina uzorka jednaka nuli. Tada je, kao i u slučaju savršene multikolinearnosti, procjena parametara nemoguća.

Multikolinearnost uzrokuje nestabilnost procjena parametara u smislu velikih razlika u procijenjenim vrijednostima parametara na temelju različitih (pod)uzoraka. Naime, prilikom uzimanja različitih uzoraka (na primjer izostavljanjem opservacija ili uzimanjem različitih poduzoraka iz uzorka) prirodno dobivamo različite procjene varijanci i kovarijanci. Zbog prisutnosti multikolinearnosti, iako bilježimo male promjene u procijenjenim vrijednostima varijanci i kovarijanci, dolazi do velikih promjena u procijenjenim vrijednostima parametara. Dakle, procjene parametara bit će jako osjetljive na male promjene uzorka. Zanimljiv primjer i diskusija dani su u [**maddala**].

2.2. Indikatori multikolinearnosti

Budući da je multikolinearnost vezana uz postojanje linearne zavisnosti prediktora, kao polazišna točka ispitivanja postojanja multikolinearnosti prirodno se nameće proučavanje korelacijeske matrice prediktora. Visoki koeficijenti korelacije upućivat će na koreliranost prediktora, odnosno na postojanje multikolinearnosti. Pri tome, ne postoji zlatno pravilo o graničnoj vrijednosti koja ukazuje na prisutnost multikolinearnosti. Napomenimo da nije dovoljno osloniti se na koeficijente korelacije prediktora jer visoki koeficijenti korelacije dovoljan su, ali ne i nužan uvjet postojanja multikolinearnosti [**gujarati**].
Nadalje, visoke vrijednosti koeficijenata korelacije ne moraju nužno biti problem, osim ako su po apsolutnoj vrijednosti veće od koeficijenta višestruke korelacije (korijen koeficijenta determinacije). Ovaj pristup osnova je prvog Kleinovog kriterija koji ocjenjuje kako je u modelu prisutan problem multikolinearnosti ako je barem jedan koeficijent korelacije između dva prediktora po apsolutnoj vrijednosti veći od koeficijenta višestruke korelacije.
Drugi Kleinov kriterij oslanja se također na koeficijent determinacije i ocjenjuje kako je problem multikolinearnosti prisutan ako je koeficijent determinacije velik, a istovremeno su vrijednosti testnih statistika $t_j = \frac{\hat{\beta}_j}{\sqrt{Var[\hat{\beta}_j]}}$ male. U tom slučaju male vrijednosti testnih statistika ukazuju na neznačajnost prediktora u modelu, dok velika vrijednost koeficijenta determinacije ukazuje na izvrsnu prilagodbu modela.

Jedan od najčešće korištenih indikatora multikolinearnosti jest faktor inflacije varijance (VIF, engl. *Variance Inflation Factor*). Za procjenitelja $\hat{\beta}_l$, faktor inflacije varijance definiramo kao

$$VIF(\hat{\beta}_l) = \frac{1}{1 - R_l^2}, \tag{4}$$

gdje je R_l^2 kvadrat koeficijenta višestruke korelacije između x_l i ostalih eksplanatornih varijabli. Usporedbom izraza za VIF (4) i izraza za varijancu procjenitelja $Var[\hat{\beta}_l]$ danu u (3), VIF možemo interpretirati kao omjer stvarne varijance od $\hat{\beta}_l$ i varijance od $\hat{\beta}_l$ koju bismo dobili kada x_l ne bi bio koreliran s $x_1, x_2, ..., x_{l-1}, x_{l+1}, ..., x_n$. Idealna situacija bila bi kada bi svi $x_i, i = 1, ..., n$ bili nekorelirani. Dakle, VIF uspoređuje stvarno stanje s idealnim stanjem [**maddala**]. Napomenimo da velike vrijednosti pokazatelja VIF nisu ni nužan ni dovoljan uvjet za dobivanje velikih vrijednosti varijance procjenitelja jer ista ovisi i o varijanci grešaka i varijanci prediktora. Već smo naveli kako posljedica multikolinearnosti može biti prevelika varijanca procjenitelja i široki po-

uzdani intervali za parametre modela. Jedna korisna interpretacija povezana s ovom posljedicom dana je u [**montgomery2021introduction**]: $\sqrt{VIF(\hat{\beta}_l)}$ indicira koliko puta je širi pouzdani interval za β_l zbog prisutnosti multikolinearnosti.
Ponekad se uz pokazatelj VIF koristi pokazatelj $TOL(\hat{\beta}_l) = \frac{1}{VIF(\hat{\beta}_l)}$ (engl. *Tolerance*). Ne postoji konsenzus oko granične vrijednosti ovih pokazatelja koja upućuje na postojanje multikolinearnosti. Na primjer, može se smatrati da je ozbiljan problem multikolinearnosti prisutan ako je $R_l^2 > 0.8$, odnosno $VIF > 5$ [**natasa**], a ponekad se uzima i stroža granica gdje vrijednost $VIF > 10$, što odgovara vrijednosti $R_l^2 > 0.9$ [**gujarati**].

Faktor inflacije varijance generaliziran je na slučaj kada u linearnom modelu promatramo podskupove parametara. U tom slučaju, predložen je generalizirani faktor inflacije varijance (GVIF) kao mjera multikolinearnosti [**foximonette**]. Tipičan primjer korištenja generaliziranog faktora inflacije varijance vezan je za slučaj kada su prediktori kategorijalne varijable. Naime, za jednu kategorijalnu varijablu s k kategorija u regresijski model uključeno je $(k-1)$ dummy varijabli. Stoga je uz jednu kategorijalnu varijablu povezano $(k-1)$ parametara pa postojanje multikolinearnosti nije moguće ispitati primjenom faktora inflacije varijance. Nadalje, s ciljem usporedivosti vrijednosti pokazatelja GVIF s obzirom na različite podskupove varijabli predlaže se korištenje modificirane, odnosno skalirane verzije $GVIF^{(1/(2Df))}$, gdje je Df broj procijenjenih koeficijenata u podskupu varijabli od interesa. Napomenimo kako u slučaju $Df = 1$ vrijedi GVIF=VIF. Više o ovom pokazatelju moguće je pronaći u [**foximonette**].

U detekciji multikolinearnosti često se koriste i svojstvene vrijednosti matrice X^TX. Neka su $\lambda_1, ..., \lambda_n$ svojstvene vrijednosti matrice X^TX. Promatramo odnos najveće svojstvene vrijednosti i ostalih svojstvenih vrijednosti kroz definiranje kondicionih indeksa

$$CI_j = \sqrt{\frac{\lambda_{max}}{\lambda_j}}.$$

Kondicioni indeksi (CI engl. *Condition index*) mjere osjetljivost procjenitelja regresije na male promjene u podatcima. Da bismo ocijenili je li u modelu prisutna multikolinearnost, dovoljno je proučiti odnos najveće i najmanje svojstvene vrijednosti, pa se u praksi najčešće koristi kondicioni indeks

$$CI = \sqrt{\frac{\lambda_{max}}{\lambda_{min}}}.$$

Što je CI bliže 1, u većoj mjeri možemo zaključiti kako multikolinearnost nije prisutna. Ni u slučaju primjene kondicionog indeksa kao pokazatelja postojanja multikolinearnosti ne postoji konsenzus oko granične vrijednosti pokazatelja koja upućuje na postojanje multikolinearnosti [**natasa**]. Obično se kao granična vrijednost koja ukazuje na postojanje multikolinearnosti uzima $CI > 10$, gdje vrijednosti između 10 i 30 sugeriraju da postoji umjerena do jaka multikolinearnost, dok vrijednosti iznad 30 upućuju na postojanje ozbiljnog problema multikolinearnosti. Napomenimo da istraživači ponekad koriste kondicioni broj $CN = \frac{\lambda_{max}}{\lambda_{min}}$, gdje vrijednosti $CN > 100$ ukazuju na prisutnost multikolinernosti, odnosno vrijednosti između 100 i 1000 na prisutnost umjerene do jake multikolinearnosti. Napomenimo da je primjenom kondicionog broja ili indeksa moguće utvrditi postojanje multikolinearnosti u modelu, dok je primjenom faktora inflacije varijance moguće utvditi postojanje multikolinearnosti vezano uz individualni prediktor.

2.3. Testovi multikolinearnosti

Dosad su u tekstu predstavljeni indikatori multikolinearnosti kod kojih je potrebno postaviti neka praktična pravila, odnosno granične vrijednosti indikatora koje upućuju na postojanje multikolinearnosti. U ovom potpoglavlju ukratko ćemo predstaviti nekoliko statističkih testova koji se koriste u detekciji multikolinearnosti. Testiramo hipoteze

$$H_0 : \text{multikolinearnost je prisutna}$$
$$H_1 : \text{multikolinearnost nije prisutna.}$$

Jedan od poznatijih, ali i često kritiziranih, testova jest Farrar-Glauberov test. Test se provodi u tri koraka. U prvom koraku utvrđuje se postojanje multikolinearnosti analizom korelacijske matrice i primjenom Bartlettovog testa. U drugom koraku promatraju se koeficijenti višestruke korelacije i testira se njihova značajnost, dok se u trećem koraku promatraju koeficijenti parcijalne korelacije među parovima varijabli i testira se hipoteza o nepostojanju korelacije između dva prediktora temeljem kojih je moguće detektirati prediktore odgovorne za postojanje multikolinearnosti.

Razvijeni su i neparametarski testovi koji daju statističku potporu dvjema najpoznatijima metodama za otkrivanje multikolinearnosti u primjeni: Kleinovo pravilo i faktor inflacije varijance (VIF) [**mtest**]. Temelje se na računanju procjena koeficijenta determinacije R_g^2 modela s uključenim svim eksplanatornim varijablama i koeficijenata determinacije R_j^2 modela u kojima je j-ti prediktor uzet kao varijabla odziva, dok

su ostali prediktori eksplanatorne varijable, pri čemu se vrijednosti procjenjuju iz n bootstrap uzoraka dobivenih iz skupa podataka, $R^2_{g_{boot}}$ i $R^2_{j_{boot}}$ redom. Ovaj pristup omogućuje formuliranje VIF-a i Kleinova pravila u smislu testiranja statističkih hipoteza:

(VIF pristup)	(Kleinov kriterij)
$H_0 : \mu_{R^2_{j_{boot}}} \geq 0.9$	$H_0 : \mu_{R^2_{j_{boot}}} \geq \mu_{R^2_{g_{boot}}}$
$H_1 : \mu_{R^2_{j_{boot}}} < 0.9$	$H_1 : \mu_{R^2_{j_{boot}}} < \mu_{R^2_{g_{boot}}}$.

Pri tome, moguće je odabrati drugačiju graničnu vrijednost umjesto 0.9.

Napomenimo da je u ovom poglavlju naveden tek dio indikatora i procedura razvijenih za analizu multikolinearnosti. U svrhu detekcije multikoliearnosti moguće je koristiti i proporcije dekompozicije varijance [**belsley**], Stewartov indeks [**Stewart**] i druge procedure, a uz dostupnost izračuna najčešće korištenih pokazatelja multikolinearnosti kroz osnovne regresijske procedure, razvijeni su i specijalizirani programski paketi. Primjerice, u R-u su dostupni paketi multiColl [**multicoll**], Mtest [**mtest**], te paket mcvis [**mcvis**] kojim je moguće pripremiti specijalizirane grafičke prikaze pri analizi multikolinearnosti.

3. Odabrane metode uklanjanja multikolinearnosti

Kada je u višestrukom regresijskom modelu prisutan (ozbiljan) problem multikolinearnosti, najjednostavnije rješenje je izostaviti varijablu ili varijable koje su kolinearne. Ipak, s ovom strategijom treba biti jako oprezan jer može dovesti do pristranosti, odnosno greške u specifikaciji (engl. *specification bias*). Stoga za izostavljanje varijabli treba imati snažno teorijsko uporište. Jedan od mogućih pristupa jest i procjena parametara sa i bez uključenog prediktora te definiranje procjenitelja uvjetno na vrijednost t-statistike kojom ispitujemo značajnost prediktora. Nadalje, može se definirati vagani procjenitelj koji je linearna kombinacija procjenitelja dobivenih u modelu sa i bez uključenog prediktora. Više o ova dva pristupa može se naći u [**maddala**].

Jedna od često korištenih metoda uklanjanja problema multikolinearnosti jest i analiza glavnih komponenata (engl. *Principal Component Analysis* (PCA)) gdje je osnovna ideja zamijeniti skup linearno zavisnih varijabli njihovim linearnim kombinacijama. PCA se koristi kako bi se smanjila dimenzionalnost podataka uz najmanji mogući gubitak

informacija i često se koristi s ciljem uklanjanja problema multikolinearnosti. Glavni je cilj PCA odrediti nekoliko komponenti koje objašnjavaju najveći mogući dio varijance mjerenih varijabli. Pri tome su komponente linearne kombinacije eksplanatornih varijabli. Ono što istraživač treba odrediti jest broj komponenti koje je potrebno zadržati, o čemu su razvijene različite smjernice. Na primjer, kriterij latentnog korijena (zadržavaju se samo komponente koje imaju svojstvenu vrijednost veću od 1), apriorni kriterij, scree test i kriterij udjela varijance [**Hair**]. Prilikom odabira komponenti istraživač treba biti oprezan jer ne mora nužno postojati povezanost između redoslijeda glavnih komponenti (redoslijed važnosti u smislu udjela objašnjene varijance) i stupnja korelacije s varijablom odziva. Glavna zamjerka ovom pristupu jest gubitak interpretabilnosti. Naime, originalne varijable predstavljaju pojave koje imaju jasno značenje i interpretaciju, dok njihove linearne kombinacije ne moraju imati jasnu interpretaciju. Dakle, primjenom PCA-e možemo izgraditi robustnije modele, ali uz gubitak interpretabilnosti modela.

Najčešće korištena metoda procjene parametara u višestrukoj regresiji jest metoda najmanjih kvadrata. Bez obzira na prisutnost multikolinearnosti, dobiveni su procjenitelji nepristrani i imaju najmanju varijancu među svim nepristranim linearnim procjeniteljima. No, to ne garantira da će njihova varijanca ujedno biti i mala. Stoga je ponekad korisno dopustiti korištenje pristranih procjenitelja, koji će imati manju varijancu. Hrbatna regresija (engl. *ridge regression*) [**ridge**] metoda je procjene parametara koja se često koristi upravo kod problema multikolinearnosti. Princip procjene parametara isti je kao kod metode najmanjih kvadrata, ali uz uvođenje ograničenja na vrijednosti procijenjenih parametara. Procjenitelje dobivamo rješavanjem problema koji kažnjava velike vrijednosti parametara

$$\hat{\beta}^{ridge} = \arg\min_{\beta} (y - \hat{\beta}X)^T (y - \hat{\beta}X) + \lambda \beta^T \beta.$$

Dakle, uz sumu kvadrata odstupanja stvarnih vrijednosti od procijenjenih vrijednosti, u funkciju cilja koju minimiziramo dodajemo i kvadriranu normu vektora β pomnoženu s regularizacijskim parametrom λ. Na istraživaču je odrediti vrijednost regularizacijskog parametra koju će primijeniti, gdje odabirom većih vrijednosti λ u većoj mjeri preferiramo manje vrijednosti parametara β_j. Primijetimo da se u slučaju $\lambda = 0$ radi o metodi najmanjih kvadrata. Procjenitelj metodom hrbatne regresije dan je s

$$\hat{\beta}^{ridge} = (X^T X + \lambda I)^{-1} X^T Y.$$

Procjenitelj $\hat{\beta}^{ridge}$ nije nepristran te je

$$E[\hat{\beta}^{ridge}] = (X^T X + \lambda I)^{-1}(X^T X)\hat{\beta},$$
$$Var[\hat{\beta}^{ridge}] = \sigma^2(X^T X + \lambda I)^{-1} X^T X (X^T X + \lambda I)^{-1}.$$

Odabirom većih vrijednosti parametra λ povećava se pristranost procjenitelja, ali se smanjuje varijanca.

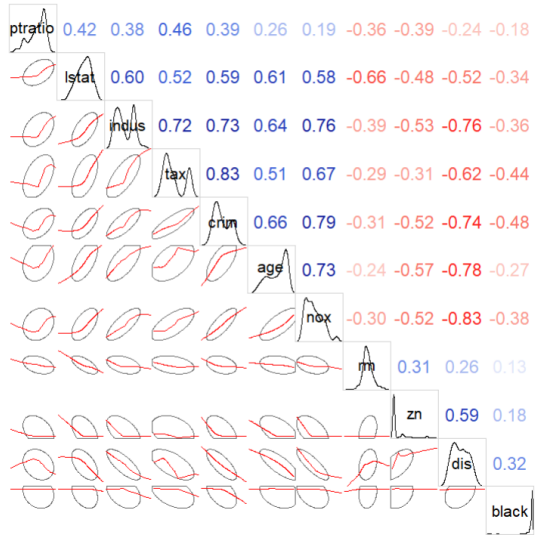
Budući da izvori multikolinearnosti mogu biti različiti, ne postoji univerzalan recept kojim je moguće ukloniti ili ublažiti problem multikolinearnosti. U ovisnosti o izvoru multikolinearnosti, ponekad je problem multikolinearnosti moguće ublažiti i uzimanjem većih uzoraka. Nadalje, u slučaju analize vremenskih nizova i prisutnosti zajedničkog trenda, multikolinearnost je moguće ublažiti transformiranjem varijabli u oblik diferencija ili omjera.

Ako je cilj izgradnje regresijskog modela izgradnja modela s najvećom prediktivnom moći, bez ulaženja u analizu i interpretaciju odnosa varijable odziva i eksplanatornih varijabli, prisutnost multikolinearnosti u modelu ne predstavlja problem. Naime, prisutnost multikolinearnosti ne narušava prediktivnu moć regresijskih modela, stoga u slučajevima kada nas interesira isključivo predikcija varijable odziva, multikolinearnost nije potrebno uklanjati.

4. Praktičan primjer

U ovom odjeljku prikazujemo slučaj izgradnje regresijskog modela uz prisutnost multikolinearnosti, uz nekoliko metoda detekcije i primjenu PCA-e u otklanjanju multikolinearnosti. Primjer je preuzet iz diplomskog rada [**jelena**]. Koristimo skup podataka *Boston* koji se može pronaći u R paketu $MASS$ [**mass**]. Ovaj skup podataka daje informacije o vrijednosti stanova u predgrađu Bostona. Varijabla koju želimo procijeniti jest *medv*, odnosno srednja vrijednost naseljenih stanova (u 1000 USD), a koristimo 11 kontinuiranih prediktora pri čemu su tri varijable log-transformirane. 506 observacija smo podijelili na podatke za treniranje i podatke za testiranje u omjeru $75 : 25$.

Na slici 1 prikazan je korelogram prediktora pri čemu su prediktori grupirani prema koreliranosti. Među prediktorima postoji jako pozitivno povezana grupa, a u njoj se nalaze varijable *age*, *nox*, *indus* i *tax*. S druge strane, možemo uočiti da je prediktor *dis* u jakoj negativnoj korelaciji s prediktorima *age*, *nox* i *indus*.



Slika 1. Prediktori grupirani prema korelaciji.

Za početak smo procijenili univarijatne modele linearne regresije. Rezultati su dani u tablici 1.

Varijabla odziva	Procijenjene vrijednosti parametara	R^2	Standardna greška	t-statistika	p-vrijednost
$log(dis)$	4.44	0.07	0.86	5.16	0.00
nox	-31.34	0.16	3.69	-8.49	0.00
$indus$	-0.60	0.21	0.06	-10.08	0.00
tax	-0.02	0.20	0.00	-9.87	0.00
age	-0.12	0.12	0.02	-7.26	0.00
$log(lstat)$	-12.36	0.65	0.47	-26.25	0.00
zn	0.12	0.1	0.02	6.50	0.00
rm	8.91	0.45	0.51	17.53	0.00
$ptratio$	-2.05	0.23	0.19	-10.68	0.00
$log(crim)$	-1.73	0.17	0.20	-8.77	0.00
$black$	0.03	0.11	0.01	7.06	0.00

Tablica 1. Tablica procijenjenih vrijednosti u univarijatnim modelima.

Procijenjen je model linearne regresije sa svih navedenih 11 prediktora:

$$\begin{aligned} medv =& 60.99 + 0.53log(crim) + 0.003zn - 0.05indus - 20.23nox \\ &+ 2.56rm + 0.02rm + 0.02age - 5.93log(dis) - \\ &- 0.004tax - 0.83ptratio + 0.01black - 9.85log(lstat) \end{aligned}$$

Rezultati su sistematizirani u tablici 2.

Prediktor	$\hat{\beta}$	VIF	Standardna greška	t-statistika	p-vrijednost
$log(crim)$	0.53	5.52	0.25	2.13	0.03
zn	0.00	2.03	0.01	0.20	0.84
$indus$	-0.05	3.54	0.06	-0.73	0.46
nox	-20.23	4.60	4.21	-4.81	0.00
rm	2.56	1.94	0.47	5.51	0.00
age	0.02	3.51	0.02	1.12	0.26
$log(dis)$	-5.93	4.78	0.95	-6.25	0.00
tax	-0.00	4.46	0.00	-1.39	0.17
$ptratio$	-0.83	1.72	0.14	-5.90	0.00
$black$	0.01	1.36	0.00	3.63	0.00
$log(lstat)$	-9.85	3.12	0.68	-14.49	0.00

Tablica 2. Tablica procijenjenih vrijednosti u multivarijatnom modelu.

Modelom je objašnjeno 76.22 % varijabilnosti varijable odziva, te je model značajan ($F = 111.7, p = 0.00$). Model je primijenjen na podatke iz test seta te RMSE iznosi 3.82.
U tablicama 1 i 2 možemo primijetiti da prediktori *crim*, *age* i *dis* imaju koeficijente različitih predznaka u univarijatnom i multivarijatnom modelu. To bi moglo ukazivati na postojanje multikolinearnosti. Primijetimo kako se značajnost prediktora *zn*, *indus* i *tax* promijenila u multivarijatnom modelu. Nadalje, iz tablice 2 možemo iščitati da prediktor $log(crim)$ ima najveću VIF vrijednost (5.52). Stoga ispitujemo postojanje multikolinearnosti. U ovome radu služili smo se paketima *Mtest*, *car* i *mctest*. Kondicioni indeks izračunat za slučaj modela s uključenih 11 prediktora iznosi 89.117, što ukazuje na postojanje problema multikolinearnosti. Proveden je Farrar–Glauberov test te rezultati ($\chi^2 = 3071.93$) također sugeriraju kako je multikolinearnost prisutna. Provedeni su testovi multikolinearnosti na bootstrap poduzrocima te hipoteze $H_0 : \mu_{R^2_{j_{boot}}} \geq 0.8, j = 1, ..., 11$, i $H_0 : \mu_{R^2_{j_{boot}}} \geq \mu_{R^2_{g_{boot}}}, j = 1, ..., 11$,

nisu odbačene u slučaju prediktora *crim*, *indus*, *nox*, *age*, *dis*, *tax*, *lstat*, što ukazuje na prediktore koji uzrokuju multikolinearnost.

Predstavljamo dvije mogućnosti korekcije: isključivanje varijabli iz modela i analiza glavnih komponenti. Napomenimo kako je svrha ovog primjera nije izgradnja najboljeg modela, već prikaz slučaja prisutnosti multikolinearnosti te prikaz nekoliko mogućnosti detekcije i uklanjanja.

1. Isključivanje varijabli iz modela.
   U ovome slučaju iz modela s uključenim kolinearnim varijablama kolinearnost otklanjamo na način da iz modela isključujemo prediktore koji uzrokuju kolinearnost. Odluku o tome koji prediktor isključiti iz modela istraživač može donijeti na temelju vlastite prosudbe (na primjer, teorijske smjernice, pouzdanost prediktora) ili na temelju statističkih performansi (na primjer, isključujemo prediktor koji nije bio značajan u univarijatnim modelima ili je imao najmanji postotak objašnjene varijabilnosti u univarijatnom modelu). Iz multivarijatnog modela s 11 prediktora isključujemo varijable $log(crim)$, nox, $indus$, tax i rm temeljem VIF kriterija. Pri tome isključujemo varijable koje su manje značajne. Također, iz modela isključujemo varijable age i zn zbog velike p-vrijednosti koja nam govori da navedene varijable nisu značajne u modelu.
   Procijenjeni model sa samo 4 prediktora dan je s

   $$medv = 68.66 - 4.13log(dis) - 0.82ptratio + 0.01black - 12.54log(lstat).$$

   Isključivanjem navedenih varijabli nije došlo do velikog smanjenja korigiranog koeficijenta determinacije - ovim modelom objašnjeno je 71.9 % varijabilnosti varijable odziva. Model je primijenjen na podatke iz test seta te RMSE iznosi 4.34.
   Najveća VIF vrijednost u ovom modelu jest 1.61, što je zadovoljavajuće. Primjena različitih testova dostupnih u korištenim paketima sugerira kako više nije prisutan problem multikolinearnosti.

2. Analiza glavnih komponenata.
   Drugi način uklanjanja multikolinearnosti koji primjenjujemo jest metoda glavnih komponenata. Primjenom metode glavnih komponenata formirali smo tri linearne kombinacije prediktora koje objašnjavaju 76 % ukupne varijabilnosti prediktora. Tri formirane linearne kombinacije (u oznaci $PC1, PC2$ i $PC3$) dalje koristimo u izgradnji regresijskog modela. Glavne komponente procijenjene su primjenom paketa *psych*. Nakon izgradnje tri linearne kombinacije originalnih varijabli, iste koristimo za izgradnju regresijskog modela s varijablom odziva $medv$. Procijenjeni model dan je s:

   $$medv = 22.49 - 2.27PC1 - 3.71PC2 + 1.61PC3.$$

Koeficijent determinacije jednak je 0.67. Vrijednost F-statistike jednaka je 254 i p-vrijednost je 0.00. Dakle, model je značajan. Model je primijenjen na podatke iz test seta te RMSE iznosi 6.3. U slučaju odbira 4 komponente procijenjeni model dan je s:

$$medv = 22.49 - 2.27PC1 - 3.71PC2 + 1.61PC3 + 2.05PC4.$$

Ovaj model ima veći korigirani koeficijent determinacije (71 %), a RMSE (na test setu) iznosi 5.47. Primijetimo kako su procijenjeni koeficijenti prve tri komponente stabilni u slučaju dodavanja četvrte komponente. U svakom slučaju, problema multikolinearnosti nema, ali model više nema jasnu interpretaciju.

5. Zaključak

Multikolinearnost prediktora česta je pojava prilikom izgradnje regresijskih modela. U ovome radu nastojali smo dati odgovore na tri temeljna pitanja vezana za multikolinearnost u regresijskoj analizi: (1) koje posljedice očekivati, (2) kako ocijeniti razinu prisutne multikolinearnosti, i (3) što učiniti kada je u regresijskom modelu prisutna multikolinearnost.

Multikolinearnost može uzrokovati nepreciznost i nestabilnost procjena parametara, a time i dovesti do pogrešnih zaključaka o odnosima analiziranih pojava. Stoga je prilikom izgradnje eksplanatornih modela potrebno ispitati postojanje multikolinearnosti. U radu je predstavljeno nekoliko indikatora i testova detekcije multikolinearnosti. No, naglasimo kako često bez primjene sofisticiranih metoda, već samo iscrpnim proučavanjem i razumijevanjem procijenjenih modela, možemo naslutiti da multikolinearnost postoji. Ako su usporedbom univarijatnih i multivarijatnog modela, dodavanjem i uklanjanjem prediktora te procjenom modela na podskupovima uzorka, razlike u vrijednostima procijenjenih parametara velike, možemo naslutiti da je prisutan problem multikolinearnosti. Nadalje, od istraživača se očekuje da razumije i teorijske pretpostavke vezane uz model koji gradi, pa ako procijenjene vrijednosti parametara u multivarijatnom modelu teoretski nemaju smisla, možda je to upravo posljedica multikolinearnosti.

U današnje vrijeme razvijeni su napredni algoritmi i statistički paketi kojima je kroz nekoliko klikova moguće izgraditi regresijske modele. Korištenje moćnih alata zahtjeva i odgovornost u korištenju istih, razumijevanje teorijske pozadine i kritički pristup rezultatima. Na primjeru multikolinearnosti možemo vidjeti kako površna primjena može rezultirati neispravnim interpretacijama koje mogu dovesti do pogrešnih od-

luka.

U radu je predstavljeno nekoliko metoda otklanjanja ili ublažavanja problema multikolinearnosti. Ipak, predstavlja li multikolinearnost uvijek problem i treba li je uklanjati? Ako je cilj analize izgradnja modela velike prediktivne moći, bez ulaženja u odnos varijabli, tada multikolinearnost ne predstavlja problem te model u kojem je prisutna multkolinearnost može biti koristan. Ovo je najlakše objasniti kroz poznati primjer neprimjerenog korištenja korelacijske analize. Sposobnost čitanja i veličina stopala djece pozitivno je povezana s njihovom dobi. Moguće je izgraditi model velike prediktivne moći u kojem će veličina stopala biti značajan prediktor sposobnosti čitanja. Korištenje takvog modela u potpunosti je opravdano ako ga koristimo isključivo u svrhu predikcije, ali bilo bi u potpunosti suludo donositi zaključke o uzročnosti.

Perišić Ana
Prirodoslovno-matematički fakultet u Splitu, Sveučilište u Splitu, Split, Hrvatska
Veleučilište u Šibeniku, Šibenik, Hrvatska
E-mail: aperisic1@pmfst.hr

Nakić Jelena
E-mail: jelenanakic5@gmail.com

Beljo Ivana
Veleučilište u Šibeniku, Šibenik, Hrvatska
E-mail: ibeljo@vus.hr