NEKA SVOJSTVA STUPIDNE REGRESIJSKE ANALIZE U STANDARDIZIRANOM ANTIIMAGE PROSTORU

M.Matečić i K.Momirović
Institut za kineziologiju, Zagreb

SAŽETAK

Metoda za robustnu regresijsku analizu koju su predložili Štalec i Momirović (1983) i detaljno opisali Dobrić, Štalec i Momirović (1984) primijenjena je na skupu regresora transformiranih u standardiziranu antiimage formu (Guttman,1953). Vrlo osobito ponašanje ovog tipa regresijske analize uočljivo je iz relacija LSR (least squares regression analysis) i ASRA (antiimage stupid regression analysis) reziduala, te iz samih svojstava parametara procijenjenih ovom amalizom.

Program koji omogućava praktičnu primjenu standardne regresijske analize pod modelom najmanjih kvadrata i antiimage robustnu regresijsku analizu napisan je u višem programskom jeziku GENSTAT verzija 4.04.

KLJUČNE RIJEČI: antiimage prostor / stupidna regresijska analiza / najmanji kvadrati

ABSTRACT

SOME PROPERTIES OF STUPID REGRESSION ANALYSIS IN STANDARDIZED ANTIIMAGE SPACE

Method for robust regression analysis, proposed by Štalec and Momirović (1983) and precisely described in Dobrić, Štalec and Momirović (1984) is applied to the regressor set transformed to standardized antiimage form (Guttman, 1953). Very peculier behavior of this type of regression analysis was showen, esspecialy concerning the relations between least squares and ASRA (antiimage stupid regression analysis) residuals and the properties of ASRA parametar estimates.

A program, alowing convenient application of standard least squares and ASRA regression, is written in GENSTAT 4.04 and described at the end of this paper.

KEYWORDS: antiimage space / stupid regression analysis / least squares

1.0 UVOD

Poznato je da je sociologija područje antropologijske znanosti u kojem su varijable najčešće slabo definirane, naime imaju loše metrijske karakteristike. Jedan od načina da se, donekle, umanje loša svojstva mijernih instrumenata je njihova transformacija u neki drugi metrički prostor. Ovakve transformacije su interesantne i zbog mogućnosti usporedbe tako dobijenih struktura i struktura u realnom prostoru, pri čemu se dobijaju značajne informacije o promijenama strukture nekog skupa socioloških varijabli ovisno o različitim eksperimentalnim modelima.

Stoga se u ovom radu predlaže postupak za analizu skupa socioloških varijabli, u smislu njegove prediktivne moći, nakon što je transformiran u standardiziranu antiimage formu (Guttman,1953)[1].Na tako definiranom skupu regresora bit će primijenjena metoda za robustnu regresijsku analizu koju su predložili Štalec i Momirović (1983), jer standardna regresijska analiza pod modelom najmanjih kvadrata, zbog nekih slabosti (nestabilnost regresijskog koeficijenta kod malih uzoraka i kod skoro singularne matrice kovarijanci regresora, te osijetljivost na utjecaj outlier-a) nije uvijek optimalna za sociološka istraživanja.

[1] antiimage varijable sadrže onaj dio varijabiliteta koji se ne može pripisata zajedničkom varijabilitetu analiziranog skupa regresora.

2.0 METODA

Neka je $S = \{s_i;\ i = 1,\dots,n\}$ slučajni uzorak iz populacije entiteta P i neka je $V = \{v_j;\ j = 1,\dots,m\}$ uzorak iz univerzuma regresora U za neku kriterijsku varijablu k.

Neka je

$$Z = S \otimes V$$

matrica podataka u standardnoj normalnoj formi definirana vrijednostima regresora u S i neka je

$$\kappa = S \otimes k$$

vektor vrijednosti, u standardnoj normalnoj formi, za kriterijsku varijablu k u S.

Definirajmo

$$R = Z^T Z$$
$$U^2 = (diag\ R^{-1})^{-1}$$
$$P = ZR^{-1}U$$
$$G = P^T P = U\ R^{-1}U\ ;$$

P je matrica regresora u standardnoj antiimage formi, pa je G korelaciona matrica regresora.

Neka

$$\alpha = Z^T \kappa$$

i

$$\beta = R^{-1}\alpha\ ;$$

naravno, β je vektor parcijalnih regresijskih koeficijenata koji je dobijen, u okviru uobičajene regresijske analize pod modelom najmanjih kvadrata, kao riješenje problema

$$Z\beta = \kappa + E \quad \Big| \quad \sigma_E^2 = E^T E = \min$$

sa koeficijentom determinacije

$$\delta^2 = (\ \kappa - E\ )^T(\ \kappa - E) = \alpha^T\ R\ \alpha = \beta^T\ R\ \beta$$

i varijancom greške

$$\sigma_E^2 = (\ \kappa - Z\beta)^T(\kappa - Z\beta) = 1 - \delta^2\ ;$$

varijabla

$$\kappa^* = Z\beta$$

je linearni kompozit regresora, tj. latentna dimenzija koja je u maksimalnoj korelaciji $\rho = (\delta^2)^{1/2}$ sa kriterijem κ.

Definirajmo

$$\psi = \kappa^* \delta^{-1} = Z\ R^{-1}\alpha\ \delta^{-1}$$

kao standardiziranu latentnu dimenziju regresora i uočimo da je

$$\kappa^T\psi = \alpha^T R^{-1}\alpha\ \delta^{-1} = \rho\ .$$

Poznato je da

$$\kappa^T E = o\ ,\qquad \kappa^{*T} E = o \qquad i \qquad \kappa^T(\kappa - \psi) = 1 - \rho.$$

Razmotrimo problem

$$Px = L \quad \left|\begin{array}{l} \kappa^T L = c = \max \\ x^T x = 1 \end{array}\right.$$

tj. problem maksimiziranja kovarijance između kriterijske varijable κ i linearne kombinacije regresora uz uvijet $x^T x = 1$, što je, naravno, specijalni slučaj standardne robustne regresijske analize.

Funkcija koju treba maksimizirati je

$$f(x,\lambda) = \kappa^T P\ x \ - \ \lambda(x^T x - 1).$$

Definirajmo

$$v = P^T \kappa = U\ R^{-1} \alpha = U\ \beta\ ,$$

tada

$$f(x,\lambda) = \kappa^T L - \lambda(x^T x - 1)$$

ima jednostavno rješenje

$$v = x\ \lambda$$

tako da

$$x = v(v^T v)^{-1/2} = U\ \beta\ (\beta^T U^2 \beta)^{-1/2}$$

i

$$c = \lambda = \mathbf{k}^T\mathbf{L} = \mathbf{v}^T\mathbf{x} = (\mathbf{v}^T\mathbf{v})^{1/2} = (\beta^T U^2 \beta)^{1/2}.$$

Moguće je primjetiti da

$$\begin{aligned} \mathbf{L} = P\ \mathbf{x} &= Z\ R^{-1}U^2\beta(\beta^T U^2\beta)^{-1/2} \\ &= Z\ R^{-1}U^2R^{-1}\mathbf{a}\ c^{-1}\ , \end{aligned}$$

a varijanca linearnog kompozita regresora

$$\sigma_L^{\ 2} = \mathbf{L}^T\mathbf{L} = \mathbf{x}^T G\ \mathbf{x} = c^{-1}(\beta^T U^2 R^{-1} U^2 \beta)\ c^{-1}$$

tako da je standardizirana latentna varijabla definirana iz $\mathbf{L}$

$$\phi = \mathbf{L}\ \sigma_L^{\ -1} = Z\ R^{-1}U^2\beta(\beta^T U^2 R^{-1} U^2 \beta)^{-1/2}\ ,$$

a kvazimultipla korelacija je definirana kao

$$\eta = \mathbf{k}^T\phi = c\ \sigma_L^{\ -1} = (\beta^T U^2 \beta)\ \ (\beta^T U^2 R^{-1} U^2 \beta)^{-1/2}.$$

Vektor pondera za procjenu standardiziranog kvazioptimalnog linearnog kompozita varijabli iz P je

$$\gamma = \mathbf{x}\ \sigma_L^{\ -1} = \mathbf{v}\ \ (\beta^T U^2 R^{-1} U^2 \beta)^{-1/2}$$

iz čega slijedi

$$\eta = \mathbf{v}^T\gamma\ .$$

Struktura kvazioptimalnog regresijskog faktora je

$$\begin{aligned} \mathbf{F} = P^T\phi &= U\ R^{-1}U^2\beta\ \ (\beta^T U^2 R^{-1} U^2 \beta)^{-1/2} \\ &= U\ R^{-1}U\ \mathbf{v}\ \ (\beta^T U^2 R^{-1} U^2 \beta)^{-1/2} \\ &= G\ \gamma \end{aligned}$$

3.0 NEKA SVOJSTVA PREDLOŽENOG MODELA

Usporedujući parametre procijenjene regresijskom analizom pod modelom najmanjih kvadrata (LSR) s parametrima koji su procijenjeni antiimage stupidnom regresijskom analizom (ASRA) moguće je uočiti slijedeće razlike:

(1) $$\beta = R^{-1}\alpha$$
$$\gamma = v\ (\beta^T U^2 R^{-1} U^2 \beta)^{-1/2}$$

(2) $$S = R\ \beta\ \delta^{-1} = \alpha\ \delta^{-1}$$
$$F = G\ \gamma$$

(3) $$\beta^T\beta = \alpha^T R^{-2}\alpha$$
$$\gamma^T\gamma = \sigma_L^{-2}$$

(4) $$K^{*T}K^{*}\ 1/n = \delta^2 = \alpha^T R^{-1}\alpha$$
$$L^T L\ \ 1/n = \sigma_L^{\ 2} = c^{-2}(\beta^T U^2 R^{-1} U^2 \beta)\ .$$

Razmotrimo slijedeće relacije između LSR i ASRA procijenjenih parametara:

(1) Kovarijanca linearnog kompozita regresora izvedenog antiimage robustnom regresijskom analizom i latentne dimenzije procijenjene regresijskom analizom pod modelom najmanjih kvadrata je ujedno i kovarijanca koja je podvrgnuta maksimizaciji u antiimage stupidnoj regresijskoj analizi, naime

$$L^T K^* 1/n = x^T P^T Z \ \beta = (v^T v)^{-1/2} v^T U \ \beta$$

$$= (\beta^T U^2 \beta)^{-1/2} (\beta^T U^2 \beta)$$

$$= c \ .$$

Korelacija istih parametara

$$\sigma_L^{-1} L^T K^* \delta^{-1} = \phi^T K^* \delta^{-1} = \eta \ \delta^{-1},$$

je isto što i omjer kvazimultiple korelacije procijenjene u proceduri ASRA i multiple korelacije procijenjene LSR analizom.

(2) Definirajmo

$$E = K - Z\beta \ ,$$
$$\tilde{E} = K - Z \ R^{-1} U \ \gamma$$

rezidual izveden u okviru LSR analize, odnosno rezidual izveden ASRA procedurom. Tada je kovarijanc tih reziduala

$$\begin{aligned} E^T \tilde{E} \ 1/n &= (K - Z \ \beta)^T (K - Z \ R^{-1} U \ \gamma) \\ &= 1 - \beta^T U \ \gamma - \beta^T \alpha + \beta^T U \ \gamma \\ &= 1 - \beta^T \alpha \\ &= 1 - \beta^T R \ \beta \\ &= 1 - \delta^2 \end{aligned}$$

isto što i varijanca greške procijenjene kriterijske varijable koja je dobijena regresijskom analizom pod modelom najmanjih kvadrata.

Korelacija spomenutih reziduala je omjer standardne devijacije greške dobijene LSR analizom i standardne devijacije greške dobijene u okviru ASRA procedure

$$\sigma_{\tilde{E}}^{-1}\tilde{E}^T E\ \delta^{-1} 1/n = (1 - \delta^2)^{1/2}(1 - 2\eta)^{-1/2}$$

pri čemu je

$$\begin{aligned}\sigma_{\tilde{E}}^2 = \tilde{E}^T\tilde{E}\ 1/n &= (K - Z\ R^{-1}U\ \gamma)^T(K - Z\ R^{-1}U\ \gamma)\\ &= 1 - \beta^T U\ \gamma - \gamma^T U\ \beta + \gamma^T U\ R^{-1}U\ \gamma\\ &= 1 - 2\eta\end{aligned}$$

varijanca greške u okviru antiimage stupidne regresijske analize.

(3) $$\beta^T\gamma = \beta^T U\ \beta\ (\beta^T U^2 R^{-1} U^2 \beta)^{-1/2}$$

Pogledajmo neke od relacije ASRA parametara i parametara procijenjenih generalnim modelom stupidne regresijske analize (SRA).

(1) Kovarijanca linearnog kompozita regresora procijenenog ASRA metodom i regresijske varijable procijenjene SRA procedurom je

$$\begin{aligned}L^T\hat{\psi} &= (\beta^T U^2 \beta)^{-1/2}\beta^T U\ \alpha\ (\alpha^T\alpha)^{-1/2}\\ &= X^T Y\end{aligned}$$

pri čemu je vektor Y skaliran tako da kompozit varijabil

$$\hat{\psi} = Z\ Y$$

ima maksimalnu kovarijancu sa kriterijskom varijablom

$$\hat{\psi}^T K = Y^T\alpha = \max = c$$

te da je zadovoljen uvijet

$$\gamma^T \gamma = 1.$$

Ekstremizacijom zadane funkcije c vektor pondera

$$\gamma = \alpha \ (\alpha^T \alpha)^{-1/2}.$$

Korelacija između linearnog kompozita procjenjenog ASRA metodom i linearnog kompozita procijenjenog SRA prosedurom je

$$\sigma_L^{-1} L^T \hat{\psi} \ \sigma_{\hat{\psi}}^{-1} = \gamma^T U \ \Gamma = r$$

pri čemu je

$$\sigma_{\hat{\psi}}^{-1} = c \ (\alpha^T R \ \alpha)^{-1/2}$$

standardna devijacija linearnog kompozita procijenjenog u okviru SRA procedure, a

$$\Gamma = \hat{\psi} \ \sigma_{\hat{\psi}}^{-1}$$

je standardizirani linearni kompozit iz SRA procedure.

(2) Definirajmo reziduale u okviru SRA procedure kao

$$\hat{\varepsilon} = (\kappa - Z \ \Gamma),$$

pa je njihova kovarijanca sa rezidualima dobijenim u ASRA proceduri

$$
\begin{aligned}
\tilde{E}^T\hat{E}\ 1/n &= (\kappa - Z\ R^{-1}U\ \gamma)^T(\kappa - Z\ \Gamma) \\
&= 1 - \alpha^T\Gamma - \gamma^T U\ \beta + \gamma^T U\ \Gamma \\
&= 1 - \sigma_\psi^{-1} - \eta + r.
\end{aligned}
$$

(3) $$\gamma^T\Gamma = (\beta^T U^2 R^{-1} U^2 \beta)^{-1/2} v^T \alpha\ (\alpha^T R^{-1} \alpha)^{-1/2}.$$

Iz prikazanih relacija parametara procijenjenih antiimage stupidnom regresijskom analizom i parametara procijenjenih standardnom regresijskom analizom pod modelom najmanjih kvadrata očigledno je da je moguća procijena nekih LSR parametara na osnovu ASRA parametara (naravno i obrnuto, ali je to od manjeg značaja).

Pri primjeni ASRA procedure treba imati u vidu da procijenjeni paametri nisu invarijantni na metriku, jer se maksimizira kovarijanca, a ne standardizirana kovarijanca izmedu regresora i kriterijske varijable.

4.0 LITERATUTA

Dobrić, V., Štalec, J. i Momirović, K.:
Note on some relationships between least squares and stupid regression analysis. 5.Međunarodni simpozij "Kompjuter na Sveučilišti", Cavtat,1984,507.1-507.7

Štalec, J. i K. Momirović:
Some properties of a very simple model for robust regression analysis. 6.Međunarodni simpozij"Kompjuter na Sveučilištu", Cavtat,1983,453-461.