K. Momirović, D. Dugić, M. Gredelj

OPTIMALNO SKALIRANJE U STUPIDNOJ REGRESIJSKOJ ANALIZI

OPTIMAL SCALING IN STUPID REGRESSION ANALYSIS:

Let $E = \{e_i;\ i = 1,\ldots,n\} \subset P$ is a set of objects selected from some population P, $V = \{v_j;\ j = 1,\ldots,m\} \subset U$ a set of regressors selected from some universum U, and K a continuous criterion variate. Let $Y = E \otimes K$, $Y^T 1 = 0$, $Y^T Y \frac{1}{n} = 1$; $Z = E \otimes V$, $Z^T 1 = 0$, $R = Z^T Z \frac{1}{n}$, diag $R = I$, and $c = Y^T Z x \frac{1}{n}$; let $Q = Z^T Y \frac{1}{n}$, and let X, $X^T X < \infty$, is some vector chosen so that $c = Q^T X = \max$. Let $K = ZX$ and $\sigma^2 = K^T K \frac{1}{n} = X^T R X$. Define the quasimultiple correlation as $\rho = E(Y, ZX^*)$, $\sigma^2 = X^{*T} R X^* = 1$, so that $\rho = c\sigma^{-1}$ Then the following theoreme is stated and proved:

Optimal scaling for error of variable $K^* = ZX\beta^*$ is $L = ZX\beta$ with $\beta = \sigma^{-1}\rho$, so that error variance $\varepsilon^2 = (Y-L)^T(Y-L)\frac{1}{n} = 1 - \rho^2$ is minimum.

This theoreme is valid, for any standardized K, for any type of regression analysis. Consequently, for stupid regression analysis (Štalec and Momirović, 1983; Dobrić, Štalec and Momirović, 1984; Dobrić, 1986), with $X = Q(Q^T Q)^{-1/2}$ and $\sigma^2 = (Q^T Q)^{-1}(Q^T R Q)$,

$$L = ZQ(Q^T Q)^{-1}(Q^T R Q)^{-1}.$$

KEYWORDS: ROBUST REGRESSION, CANONICAL COVARIANCE ANALYSIS, OPTIMAL SCALING

0. UVOD

Standardni model regresijske analize pod kriterijem najmanjih kvadrata, koji je, u stvari, poseban slučaj kanoničke korelacijske analize, nije pogodan u mnogim sociološkim, kao, uostalom, i u istraživanjima u mnogim drugim znanostima, zbog više razloga, od kojih su najvažniji nestabilnost regresijskih koeficijenata ako je matrica krosprodukata regresora slabo uvjetovana, sistematska pristranost u funkciji stupnjeva slobode, i jaka osjetljivost na objekte veoma udaljene od očekivanih vrijednosti u regresorskim i kriterijskoj varijabli. Stoga je predložen niz robustnih metoda, dijelom usmjerenih na smanjenje osjetljivosti na skoro singularne matrice krosprodukata regresora, a dijelom na redukciju osjetljivosti na ekstremne entitete. U klasu robustnih metoda spada i stupidna regresijska analiza, konstruirana tako da reducira sva tri glavna izvora osjetljivosti regresijske analize pod modelom najmanjih kvadrata, ili ekvivalentnim modelom maksimiziranja korelacija linearne kombinacije regresora i kriterijske varijable. Naime, stupidna regresijska analiza maksimizira kovarijancu linearne kombinacije regresora i kriterijske varijable, pa kako stoga rezultati ne zavise od inverza matrice krosprodukata, slabo je osjetljiva na pseudosingularitet, estremne vrijednosti i stupnjeve slobode analiziranog sistema varijabli.

Nažalost, za razliku od kanoničkog regresijskog modela, u kome je optimalno skaliranje implicitno (jer, u stvari, operira na lijevim svojstvenim vektorima matrice podataka), nijedna robustna metoda, pa stoga ni stupidna regresijska analiza, ne skalira automatski rezultate obzirom na varijancu pogreške. Stoga je problem optimalnog skaliranja obzirom na varijancu pogreške problem koji treba riješiti za sve nestandardne regresijske procedure. U ovom je radu dokazana jedna opća teorema o optimalnom skaliranju pod bilo kojim modelom regresijske analize, i iz te je teoreme izveden

korolar o optimalnom skaliranju u stupidnoj regresijskoj analizi.

1. STUPIDNA REGRESIJSKA ANALIZA

Neka je $E = \{e_i;\ i = 1,\ldots,n\}$ skup objekata slučajno izabranih iz neke populacije P. Neka je $V = \{v_j;\ j = 1,\ldots,m\}$ skup varijabli sa nekom eliptičnom funkcijom raspodjele koje imaju logički status regresora, izabranih i nekog univerzuma varijabli U u skladu sa nekim, ne nužno eksplicitnim, logičkim ili matematičkim modelom ponašanja neke kontinuirane, eliptično distribuirane kriterijske varijable

Neka je

$$Z = E \boxtimes V \qquad Z^T 1 = 0$$

$$R = E(Z,Z) \qquad \operatorname{diag} R = I$$

i

$$Y = E \boxtimes K \qquad Y^T 1 = 0$$

$$\phi^2 = E(Y,Y) \qquad \phi^2 = 1$$

i neka je

$$Q = E(Z,Y) \qquad Q \neq 0$$

Sada se stupidna regresijska analiza (Štalec i Momirović, 1983; Dobrić, Štalec i Momirović, 1984; Dobrić, 1986), koja je, u stvari, poseban slučaj kanoničke analize kovarijanci (Momirović, Dobrić i Karaman, 1983), može definirati kao rješenje problema

$$ZX = K \qquad c = E(K,Y) = \max$$
$$X^T X = 1.$$

Očito,

$$X = Q(Q^T Q)^{-1/2}$$

i, kako je

$$\sigma^2 = E(K,K) = X^T R X = (Q^T Q)^{-1}(Q^T R Q),$$

kvazimultipla korelacija izmedju Y i funkcije $\Theta(K)$, $\Theta^2 = 1$, je

$$\rho = c\sigma^{-1}.$$

Ako je $K = ZX$ pogreška prognoze je $(Y-K)$, sa varijancom pogreške

$$\varepsilon^2 = E((Y-K),(Y-K)) = 1+\sigma^2-2c;$$

za $\Theta(K)$, $\Theta^2=1$, ta je varijanca

$$\varepsilon^2 = E((Y-\Theta(K)),(Y-\Theta(K))) = 2(1-\rho).$$

Kako u oba slučaja $\varepsilon^2 \neq$ min, cilj je ovog rada da postavi i dokaže jednu opću teoremu o optimalnom skaliranju pod bilo kojim modelom regresijske analize*, i da izvede optimalno skaliranje u stupidnoj regresijskoj analizi kao korolar ove teoreme.

2. OPTIMALNO SKALIRANJE U REGRESIJSKOJ ANALIZI

Iako je, naravno, moguće neposredno izvesti optimalno skaliranje u stupidnoj regresijskoj analizi, mnogo je pogodnije najprije dokazati jednu opću teoremu pod bilo kojim regresijskim modelom koji se može svesti na generalnu kanoničku formu, a zatim izvesti optimalno skaliranje u stupidnoj regresijskoj analizi kao korolar ove teoreme.

Teorema 1.

Neka je X, $X^T X<\infty$, neki vektor koji maksimizira funkciju $c = E(Y,ZX)$; neka je $\sigma^2 = E(ZX,ZX)$ i neka je $\rho = c\sigma^{-1}$. Tada

* *Jednu sličnu teoremu su postavili i dokazali Breiman i Friedman (1985) za poseban slučaj regresijskog algoritma definiranog iterativnom maksimizacijom alternativnog uvjetnog očekivanja (Teorema 5.1, str. 590).*

je optimalno skaliranje obzirom na pogrešku $\varepsilon^2 = E((Y-L),(Y-L)$. $L = \psi(ZX)$,

$$L = ZX\beta$$

gdje je

$$\beta = \sigma^{-1}\rho$$

Dokaz:

Neka je $K^* = ZX$. Tada je $\varepsilon^{*2} = 1+\sigma^2-2c$. Za $L = ZX\sigma^{-1}\rho$,

$$\begin{aligned}\varepsilon^2 &= (Y-L)^T(Y-L)\frac{1}{n}\\ &= 1-2(Q^TX\sigma^{-1}\rho) + \rho\sigma^{-1}(X^TRX)\sigma^{-1}\rho\\ &= 1-2c\sigma^{-1}+\rho^2\\ &= 1-\rho^2\end{aligned}$$

što je očito minimum, ako je $E(Y,Y) = 1$, za bilo koji koeficijent kvazimultiple korelacije ρ.

Korolar 1.

Optimalno skaliranje u stupidnoj regresijskoj analizi, gdje je $X = Q(Q^TQ)^{-1/2}$ i $\sigma^2 = (Q^TQ)^{-1}(Q^TRQ)$ je

$$ZX\sigma^{-1}\rho = ZQ(Q^TQ)^{-1}(Q^TRQ)^{-1}$$

jer je $c = Q^TX = (Q^TQ)^{1/2}$.

LITERATURA

1. Breiman, L. and J.H. Friedman (1985): Estimating optimal transformation for multiple regression and correlation. Journal of American Statistical Association, 80, 391:580-598.

2. Dobrić, V., J. Štalec i K. Momirović (1984): Note on some relationships between least squares and stupid regression analysis. Proceedings of 6th International Symposium "Computer at the University", Dubrovnik, 507:1-7.