UDK 519.1
Pregledni članak

**Dr. sc. Mirta Benšić**
**Dr. sc. Miljenko Crnjac**

# OPTIMALAN $L_p$ PROCJENITELJ NEPOZNATIH PARAMETARA REGRESIJSKOG MODELA

*Osjetljivost procjenitelja regresijskih parametara dobivenog metodom najmanjih kvadrata na jako odstupajuće vrijednosti u podacima upućuje na potrebu za proučavanjem robusnih statistika za procjenu regresijskih parametara u linearnim i nelinearnim modelima. Jednu klasu robusnih statistika koja je primjenjiva u tu svrhu čine $L_p$ procjenitelji za $p \in [1,\infty)$. Ovdje su izložene metode za izbor eksponenta p, $L_p$ procjenitelja kojim će biti izvršena procjena u danom modelu da bi se postigao minimum varijance $L_p$ procjenitelja (tzv. optimalan $L_p$ procjenitelj).*

## 1. UVOD

Za procjenjivanje parametara u regresijskim modelima najčešće se koristi procjenitelj dobiven metodom najmanjih kvadrata (MNK procjenitelj). Međutim, poznato je da je taj procjenitelj osjetljiv na odstupanje od normalnosti u distribuciji reziduala. Ova činjenica ilustrirana je u mnogim Monte Carlo studijama (npr. [2], [9], [7]), gdje je također pokazano da MNK procjenitelj ima neka nepoželjna statistička svojstva posebno u slučaju kada gustoće rezidualnih distribucija ne opadaju dovoljno brzo ("distributions with long tails", tj. pojava tzv. stršećih vrijednosti među podacima). Kod takvih distribucija korisno je proučiti svojstva drugih tipova procjenitelja za nepoznate parametre. Klasu procjenitelja koji su se pokazali vrlo dobri u slučajevima simetričnih rezidualnih distribucija čine tzv. $L_p$ procjenitelji. (Ovdje je $p$ neki pozitivan realan broj ili beskonačno). Ova klasa procjenitelja sadrži procjenitelja dobivenog metodom najmanjih kvadrata (MNK procjenitelj) za $p = 2$, procjenitelja dobivenog metodom minimalne apsolutne greške za $p = 1$ i Čebiševljevog procjenitelja za $p = \infty$.

U ovom članku dane su neke metode koje se mogu koristiti prilikom određivanja optimalnog $L_p$ procjenitelja.[1]

## 2. DEFINICIJA $L_p$ PROCJENITELJA

Pretpostavimo da zavisne varijable $y_i$ ($i = 1, ..., n$), mjerimo s nepoznatim greškama $e_i$ ($i = 1, ..., n$) i pokušavamo ih uskladiti s $m$ fiksnih nezavisnih varijabli $x_{i1}, ..., x_{im}$ ($x_i = [x_{i1}, ..., x_{im}]^r$), $i = 1, ..., n$ koristeći funkciju - model $f(x_i;\theta)$. Ovdje je $\theta = [\theta_1, ..., \theta_k]^r$ vektor $k$ nepoznatih parametara. Funkcija $f$ može biti linearna i nelinearna u nepoznatim parametrima.

Kada su greške $e_i$ aditivne slučajne varijable, zavisna varijabla može biti prikazana kao:

$$y_1 = f(x_1;\theta) + e_1$$
$$\vdots$$
$$y_n = f(x_n;\theta) + e_n.$$

[1] Pojam optimalnog $L_p$ procjenitelja definiran je u 3. poglavlju.

U cijelom tekstu pretpostavljamo da su **greške** $e_1, ..., e_n$ **nezavisne i jednako distribuirane slučajne varijable.**

**Primjer 2.1** *Neka je funkcija* $f(x; a, b) = ax + b$. *Procjena parametara a i b na osnovi n izmjerenih vrijednosti zavisne varijable* $y_1, ..., y_n$ *za pripadne vrijednosti nezavisne varijable* $x_1, ..., x_n$:

$$y_1 = ax_1 + b + e_1$$
$$y_2 = ax_2 + b + e_2$$
$$\vdots$$
$$y_n = ax_n + b + e_n,$$

*primjer je linearnog modela.*

**Primjer 2.2** *Neka je funkcija* $f(x; A, b, \gamma) = \frac{A}{be^{-\gamma x}+1}$ *(logistička funkcija). Procjena parametara A, b i* $\gamma$ *na osnovi n izmjerenih vrijednosti zavisne varijable* $y_1, ..., y_n$ *za pripadne vrijednosti nezavisne varijable* $x_1, ..., x_n$:

$$y_1 = f(x_1, a, b, \gamma) + e_1$$
$$y_2 = f(x_2, a, b, \gamma) + e_2$$
$$\vdots$$
$$y_n = f(x_n; a, b, \gamma) + e_n,$$

*primjer je nelinearnog modela.*

$L_p$ procjenitelj parametra $\theta$ je vrijednost $\hat{\theta}^{(p)} = [\hat{\theta}_1^{(p)}, ..., \hat{\theta}_k^{(p)}]$ koja minimizira sumu $p$-te potencije apsolutnih vrijednosti reziduala, $p \in [1,\infty)$. Dakle, ako označimo reziduale:

$$r_i(\theta) = y_i - f(x_i;\theta), \qquad i = 1, ..., n,$$

a sumu $p$-tih potencijala apsolutnih vrijednosti reziduala:

$$S_p(\theta) = \sum_{i=1}^{n} |r_i(\theta)|^p,$$

$\hat{\theta}^{(p)}$ je vektor koji zadovoljava

$$S_p(\hat{\theta}^{(p)}) = \min_{\theta \in \Theta} S_p(\theta)$$

ako takav minimum postoji. Ovdje je $\Theta$ skup svih mogućih vrijednosti vektora parametara $\theta$.

Ako je $p = \infty$ definira se $\theta(\infty)$ kao ona vrijednost vektora parametara koja minimizira $S_\infty(\theta)$,

$$S_\infty(\theta) = \max_i |r_i(\theta)|,$$

u skupu svih dozvoljenih vrijednosti nepoznatog vektora parametara.

Kao što se može vidjeti, za $p = 2$ dobijemo upravo MNK procjenitelja kao specijalan slučaj iz klase $L_p$ procjenitelja.

## 3. IZBOR $L_p$ PROCJENITELJA

Izbor optimalnog $L_p$ procjenitelja zasniva se na rezultatima dobivenim teoretskim proučavanjem statističkih svojstava $L_p$ procjenitelja u linearnim modelima (npr. [6], [8]) i simulacijskim studijama (npr. [7], [10]).

Linearan regresijski model podrazumijeva da funkcija $f(x, \theta)$ ima oblik:

$$f(x;\theta) = f(x_1, ..., x_k;\theta_1, ..., \theta_k) = \theta_1 x_1 + \theta_2 x_2 + ... + \theta_k x_k.$$

Dakle, u ovakvim modelima problem se svodi na procjenu $k$ nepoznatih parametara $\theta_1, ..., \theta_k$ na osnovi eksperimentalnih vrijednosti stanja zavisne varijable $y_1, ..., y_n$:

$$y_1 = \theta_1 x_{11} + ... + \theta_k x_{1k} + e_1$$
$$y_2 = \theta_1 x_{21} + ... + \theta_k x_{2k} + e_2$$
$$\vdots$$
$$y_n = \theta_1 x_{n1} + ... + \theta_k x_{nk} + e_n.$$

Označimo li matricu

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & ... & x_{1k} \\ x_{21} & x_{22} & ... & x_{2k} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & ... & x_{nk} \end{bmatrix},$$

onda se linearan regresijski model može matrično prikazati u obliku

$$\mathbf{y} = \mathbf{X}\theta + \mathbf{e},$$
$$\mathbf{y} = [y_1, ..., y_n]^T, \qquad \mathbf{e} = [e_1, ..., e_n]^T.$$

Za linearne modele pokazano je da je u slučaju simetrične rezidualne distribucije oko nule, distribucija $L_p$ procjenitelja $\hat{\theta}^{(p)} = [\hat{\theta}_1^{(p)}, ..., \hat{\theta}_k^{(p)}]^T$ simetrična oko stvarne vrijednosti parametra $\theta$ i ako postoji očekivanje od $\hat{\theta}^{(p)}$, onda je $\hat{\theta}^{(p)}$ nepristran procjenitelj za $\theta$ ([8]).

Također je pokazano da ako su zadovoljeni uvjeti:

A1: $L_1$ i $L_\infty$, procjenitelji su jedinstveni;

A2:

$$Q = \lim_{n\to\infty} \frac{1}{n} \mathbf{X}^T\mathbf{X}$$

je pozitivno definitna matrica;

A3: Ako je $p = 1$, distribucija reziduala je neprekidna s neprekidnom gustoćom $\phi(x)$ različitom od nule u nuli;

A4: Ako je $1< p < \infty$, postoje sljedeća očekivanja:

$$E(|e_1|^{p-2})$$
$$E(|e_1|^{2p-2}),$$

dok je $E(|e_1|^{p-1}) = 0$;

onda je $\sqrt{n}(\hat{\theta}^{(p)}-\theta)$ ima asimptotski $k$-varijantnu normalnu distribuciju s očekivanjem nula i kovarijacijskom matricom $\omega_p^2 Q^{-1}$, gdje je:

$$\omega_p^2 = \begin{cases} [2\phi(0)]^{-2} & , \quad p = 1 \\ E[|e_1|^{2p-2}]\{(p-1)E[|e_1|^{p-2}]\}^{-2}, & 1 < p < \infty. \end{cases}$$

Iz navedenih rezultata vidljivo je da varijanca $L_p$ procjenitelja kao funkcija eksponenta $p$, $L_p$ norme, ovisi prvenstveno o rezidualnoj distribuciji.

Optimalan $L_p$ procjenitelj definiramo kao onaj za koji je varijanca od $\sqrt{n}(\hat{\theta}^{(p)}-\theta)$ minimalna po $p \in [1, \infty)$. Možemo, dakle, odrediti optimalan $p$ (u oznaci $p_o$) kao minimum funkcije $\omega(p)$.

Na ovaj način je 1983. godine pokazano ([8]) da za greške koje imaju normalnu distribuciju $p_o$ iznosi 2, za greške koje imaju simetričnu uniformnu distribuciju oko nule $p_o = \infty$, za greške koje imaju Cauchyjevu distribuciju $p_o = 1$, za greške koje imaju Laplaceovu distribuciju $p_o = 1$, itd. Navedeni teoretski rezultati podudaraju se s rezultatima simulacijskih studija na istu temu ([7], [5]).

Valja naglasiti da ovi rezultati potvrđuju da je u slučaju normalnih grešaka uputno koristiti MNK procjenitelja nepoznatih parametara s obzirom da je optimalan $L_p$ procjenitelj upravo za vrijednost $p_o = 2$ dok za ostale distribucije grešaka to očigledno ne mora biti slučaj.

Želimo li iskoristiti navedene treoretske rezultate za odabir $L_p$ norme kojom ćemo procijeniti nepoznate regresijske parametre, moramo unaprijed poznavati vrstu rezidualne distribucije. Ova činjenica predstavlja velik problem ukoliko je u postupku procjene parametara regresijski model nepoznat do te mjere da se ne može apriori tvrditi koja je distribucija reziduala, već se ona istražuje tek nakon eliminiranja trenda. Međutim, da bismo eliminirali trend, potrebno je prvo procijeniti parametre, što ne možemo bez odabira eksponenta $p$ norme kojom ćemo procjenu izvršiti.

Budući da su stvarni regresijski modeli vrlo često upravo takvi da se rezidualna distribucija ne može unaprijed utvrditi, R. Gonin i A.H. Money sugeriraju 1985. godine ([5]) algoritam za utvrđivanje optimalne $L_p$ norme. Ovaj algoritam temelji se na simulacijskim studijama. Koristeći primjere kod kojih je optimalan $L_p$ procjenitelj poznat, primjenom ovog algoritma postignuta je optimalna vrijednost već u četvrtom koraku. Moramo, međutim, naglasiti da konvergencija algoritma prema stvarnoj optimalnoj vrijednosti eksponenta $p$ još uvijek nije teoretski dokazana.

**Algoritam:**

1. Postaviti $i$: = 0 s $p_i = 2$ (tj. MNK); $E = 10^{-4}$
2. Izračunati vrijednost regresijskih parametara koristeći normu $Lp_i$.
3. Izračunati zakrivljenost $K$ (kurtosis) tako dobivene rezidualne distribucije. Iskoristiti formule:

$$p_{i+1} = \frac{9}{K^2} + 1, \quad 1 \le p_i < \infty \qquad (1)$$

$$p_{i+1} = \frac{6}{\overset{(2)}{K}} \qquad 1 \le p_i < 2$$

za računanje $p_{i+1}$.

4. Ponoviti korak 1, 2. i 3. sve dok ne bude

$$|p_{i+1} - p_i| < E.$$

Formule (1) i (2) kojima se određuje vrijednost eksponenta $p$ također su rezultat simulacijskih studija. Usporedbom s teoretski dobivenim vrijednostima za optimalan eksponent $p_o$ u simuliranim modelima pokazano je da ove formule daju dobre rezultate u velikom broju rezidualnih distribucija. Međutim, s obzirom da njihova statistička svojstva i teoretska veza s $p_o$ još uvijek nisu dovoljno izučeni, uputno je nakon provedene procedure procjene optimalnog eksponenta $p_o$ i izvršene procjene regresijskih parametara eliminirati

trend i analizirati distribuciju reziduala te usporediti vrijednost $p_o$ s teoretskom vrijednošću za dobivenu distribuciju da bismo bili sigurni da je optimalna vrijednost zaista postignuta u danom slučaju.

Navedeni algoritam može biti primijenjen i u nelinearnom regresijskom modelu s aditivnim greškama ([4]).

*LITERATURA*

(1) Benšić, M. (1997). Confidence regions and intervals in nonlinear regresion. *Mathematical Communications 2*, 71-76.

(2) Blattberg, R. and Sargent, T. (1971) Regression With Non-Gaussian Stable Disturbances, *Econometrica 39*, 501-510.

(3) Donaldson, J.R. and Schnabel, R.B. (1987). Computational expirience with confidence regions and confidence intervals for nonlinear least squares. *Technometrics* **29**, pp 67-82.

(4) Gonin, R., and Money, A.H. (1989). Nonlinear $L_p$-norm estimation, *Marcel Dekker, Inc., New York and Basel*

(5) Gonin, R. and Money, A.H. (1985). Nonlinear $L_p$-norm estimation: Part I - On the choice of the exponent, $p$, where the errors are additive. *Commun. Statist. - Theor. Meth.* **14**, 827-840.

(6) Huber, P.J. (1973). Robust Regression: Asymptotics, Conjectures and Monte Carlo, *The Annals of Statistics*, **1**, 5 799-821

(7) Money, A.H., Affleck-Graves, J.F., Hart, M.L. and Barr, G.D.I. (1982) The Linear Regression Model: $L_p$-norm Estimation and the choice of $p$, *Comm. Statist. - Simula. Computa. 11*, 89-109.

(8) Nyquist, H. (1983). The optimal $L_p$ norm estimator in linear regression models. *Commun. Statist. - Theor. Meth.* **12**, pp 2511-2524.

(9) Smith, V.K. and Hall, T.W. (1972) A Comparation of Maximum Likelihood versus BLUE Estimators, *Rev. Econ. Statist. 54*, 186-190.

(10) Sposito, V.A., Hand, M.L. and Skarpness, B. (1983). On the efficiency of using the sample kurtosis in selecting optimal $L_p$ norm estimators. *Commun. Statist. - Simula. Computa.* **12**, pp 265-272.

**Mirta Benšić**
**Miljenko Crnjac**

**OPTIMAL $L_p$ VALUER OF UNKNOWN PARAMETERS OF THE REGRESSION MODEL**

*Summary*

Sensitivity of the regression parameters valuer achieved by the method of the smallest squares to very sharp deviation value in the data refers to the need of the robust statistics study to value the regression parameters in the linear and nonlinear models. One class of the robust statistics which is applicable for this purpose is made up of $L_p$ valuers for $p \in [1, \infty)$. Here we put forward the methods for the choice of exponents p, $L_p$ valuers by which the evaluation will be carried out in the given model to achieve the minimum of the variance of $L_p$ valuer (so called optimal $L_p$ valuer).