O JEDNÉ VARIANTĚ LINEÁRNÍ EIV REGRESE S OMEZENÝMI CHYBAMI

M. ČERNÝ, M. HLADÍK A J. ANTOCH

ROBUST 2016
16. ZÁŘÍ 2016

Problém

Uvažujeme lineární regresní model

$$y_i = x_i'\beta + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

- $\beta \in \mathbb{R}^p$ je vektor neznámých regresních parametrů
- $x_i = (x_{i1}, \ldots, x_{ip})'$ jsou nepozorovatelné stochastické regresory
- ε_i, $i = 1, \ldots, n$, jsou aditivní náhodné chyby (v pozorování závisle proměnné)

v němž

- místo matice regresorů pozorujeme matici regresorů zatíženou aditivním stochastickým šumem (tzv. strukturální EIV model)

EIV-model

Místo

$$y_i = x_i'\beta + \varepsilon_i, \quad i = 1, \ldots, n,$$

uvažujeme (EIV) model, v němž pozorovaná data jsou $\left(y_i, z_{ij}\right)$, kde

$$z_{ij} = x_{ij} + \nu_{ij}, \quad i = 1, \ldots, n, \; j = 1, \ldots p,$$

ν_{ij} jsou „náhodné chyby regresorů“

Značení

$$Z^n = \left(z_{ij}\right)_{i=1,\ldots,n}^{j=1,\ldots p} \qquad \text{and} \qquad y^n = (y_1, \ldots, y_n)'. \tag{2}$$

EIV-model

Místo

$$y_i = x_i'\beta + \varepsilon_i, \quad i = 1, \ldots, n,$$

uvažujeme (EIV) model, v němž pozorovaná data jsou $\left(y_i, z_{ij}\right)$, kde

$$z_{ij} = x_{ij} + \nu_{ij}, \quad i = 1, \ldots, n, \; j = 1, \ldots p,$$

ν_{ij} jsou „náhodné chyby regresorů“

Značení

$$Z^n = \left(z_{ij}\right)_{i=1,\ldots,n}^{j=1,\ldots p} \qquad \text{and} \qquad y^n = (y_1, \ldots, y_n)'. \tag{2}$$

Poznámka

Za jistých „tradičních“ předpokladů lze (je zvykem) regresní parametry „dobře“ odhadovat pomocí metody úplných nejmenších čtverců (Total Least Squares, TLS).

Total Least Squares (TLS)

Připomeňme, že TLS problém může být algebraicky formulován následovně

Pro dané $A \in \mathbb{R}^{n\times p}$ a $w \in \mathbb{R}^n$ najděme $\Delta A \in \mathbb{R}^{n\times p}$ a $\Delta w \in \mathbb{R}^n$ tak, že:

- lineární systém $(A + \Delta A)\xi = (w + \Delta w)$ je řešitelný
- $\|(\Delta A, \Delta w)\|_F$ is minimal, where $\|\cdot\|_F$ označuje Frobeniovu normu ($\|A\|_F = \sum_i \sum_j A_{ij}^2$)

[1] Maticová norma $\|\cdot\|$ je orthogonálně invariantní jestliže $\|UAV\| = \|A\|$ $\forall\ A \in \mathbb{R}^{n\times p}$ and všechny unitární matice $U \in \mathbb{R}^{n\times n}$ a $V \in \mathbb{R}^{p\times p}$

Total Least Squares (TLS)

Připomeňme, že TLS problém může být algebraicky formulován následovně
Pro dané $A \in \mathbb{R}^{n \times p}$ a $w \in \mathbb{R}^n$ najděme $\Delta A \in \mathbb{R}^{n \times p}$ a $\Delta w \in \mathbb{R}^n$ tak, že:

- lineární systém $(A + \Delta A)\xi = (w + \Delta w)$ je řešitelný
- $\|(\Delta A, \Delta w)\|_F$ is minimal, where $\|\cdot\|_F$ označuje Frobeniovu normu ($\|A\|_F = \sum_i \sum_j A_{ij}^2$)

Poznámka
Někteří autoři studovali TLS při použití jiných maticových norem, především tzv. ortogonálně invariantních norem (OIN)[1].
Pro ně, podobně jako pro TLS, $\widehat{\beta}$ existuje s pstí 1, asymptorické vlastnosti jsou tytéž (podobné) jako o TLS řešeni, ...
Problém: Čebyševova norma mezi OIN bohužel nepatří, a celá situace se začíná překvapivě značně komplikovat.

[1] Maticová norma $\|\cdot\|$ je orthogonálně invariantní jestliže $\|UAV\| = \|A\|$ $\forall\ A \in \mathbb{R}^{n \times p}$ and všechny unitární matice $U \in \mathbb{R}^{n \times n}$ a $V \in \mathbb{R}^{p \times p}$

Modifikovaný EIV-problém

V modelu

$$z_{ij} = x_{ij} + \nu_{ij}, \quad i = 1, \ldots, n, \; j = 1, \ldots p,$$

předpokládejme, že

- chybové distribuce mají omezený nosič v intervalu $(-\gamma, \gamma)$, kde $\gamma > 0$ (tzv. *poloměr chyb*) je neznámá konstanta
- víme, které regresory měříme přesně a které s chybou
- Frobeniovu normu nahradíme Čebyševovou normou ($\|A\|_{max} = \max_{i,j} |A_{ij}|$)
- předpokládáme řadu technických, mnohdy pro statistika poněkud neobvyklých, podmínek regularity

z nichž vyjímáme

- chyby nemusí být nutně stejně rozdělené
- nepotřebujeme ani nulové střední hodnoty chyb ani jejich „nezávislost“ (s vyjímkou eliminace „patologických“ případů)

Assumptions (in detail)

Assumption 1. There exists an unknown constant $\gamma > 0$, called *radius*, such that

(i) $|\varepsilon_i| \leq \gamma$ a.s., $i = 1, \ldots, n$;

and there is a known index set $\Gamma \subseteq \{1, \ldots, p\}$ such that for all $j = 1, \ldots, p$:

(ii) if $j \notin \Gamma$, then $\nu_{ij} = 0$ a.s. for all $i = 1, \ldots, n$;
(iii) if $j \in \Gamma$, then $|\nu_{ij}| \leq \gamma$ a.s. for all $i = 1, \ldots, n$.

Assumption 2 (asymptotic properties of regressors and errors). Let $\|\cdot\|$ be any vector norm. We assume that

$$(\forall \alpha > 0)\ (\exists c > 0)\ (\forall u \in \mathbb{R}^p \text{ s.t. } \|u\| = 1)$$
$$\Pr[\mathcal{A}_n(\alpha, c, u)] \overset{n\to\infty}{\longrightarrow} 1,$$

where $\mathcal{A}_n(\alpha, c, u)$ is the following event: there exists $i_0 \in \{1, \ldots, n\}$ such that

(i) $|x_{i_0}{}' u| \geq c$; and
(ii) $-\operatorname{sgn}(x'_{i_0} u) \cdot \varepsilon_{i_0} \geq \gamma - \alpha$; and
(iii) $(\forall j \in \Gamma)\ \operatorname{sgn}(x'_{i_0} u) \cdot \operatorname{sgn}(\beta_j + u_j) \cdot \nu_{i_0 j} \geq \gamma - \alpha$,

where $\operatorname{sgn}(\xi) = 1$ if $\xi \geq 0$ and $\operatorname{sgn}(\xi) = -1$ if $\xi < 0$.

Reformulace pro případ Čebyševovy normy

CNP problém může být algebraicky formulován následovně
Pro dané $A \in \mathbb{R}^{n \times p}$, $w \in \mathbb{R}^n$, *a* $\Gamma \subseteq \{1, \ldots, p\}$, *najděmež* $\Delta A \in \mathbb{R}^{n \times p}$ *a* $\Delta w \in \mathbb{R}^n$ *takové, že:*

- lineární systém $(A + \Delta A)\xi = (w + \Delta w)$ je řešitelný
- jestliže $j \notin \Gamma$, potom j-tý sloupec ΔA je nulový
- $\|(\Delta A, \Delta w)\|_{\max}$ je minimální

Co (ne)umíme

Co umíme?

- sestrojit konsistentní estimátor jak pro vektor regresních parametrů, tak pro poloměr chyb γ
- spočítat tento odhad pomocí systému zobecněných lineárně-frakcionálních programů (GLFP)
- víme něco o složitosti celé procedury, neboť stačí vyřešit $p \leq 2^p$ GLFP (2^{20} je hranice pro „rozumná“ n)
- výpočet umíme poměrně snadno paralelizovat

Co (ne)umíme

Co umíme?

- sestrojit konsistentní estimátor jak pro vektor regresních parametrů, tak pro poloměr chyb γ
- spočítat tento odhad pomocí systému zobecněných lineárně-frakcionálních programů (GLFP)
- víme něco o složitosti celé procedury, neboť stačí vyřešit $p \leq 2^p$ GLFP (2^{20} je hranice pro „rozumná“ n)
- výpočet umíme poměrně snadno paralelizovat

Co neumíme?

- najít asymptotické rozdělení tohoto odhadu
- přesvědčit „zákazníky“, že náš přístup má smysl