Robustní odhady v lineárním regresním modelu

$$Y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_{p-1} x_{ip-1} + \varepsilon_i \,, \quad i = 1, \dots, n$$

$$Y_i = x_i^T \beta + \varepsilon_i$$

$$\mathbb{X} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1p-1} \\ 1 & x_{21} & \dots & x_{2p-1} \\ \vdots & & & \\ 1 & x_{n1} & \dots & x_{np-1} \end{pmatrix} \quad x_i^T$$

$x_i = (1, x_{i1}, \dots, x_{ip-1})^T$ ... p-rozměrný vektor regresorů (pevná čísla)

$\beta = (\beta_0, \beta_1, \dots, \beta_{p-1})^T$ .. p-rozměrný vektor neznámých parametrů

$$\mathbb{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$$

ε_i jsou chyby modelu, i.i.d. s distribuční funkcí F (obecně neznámá)

$E\varepsilon_i = 0, \; D\varepsilon_i = \sigma^2$.

Odhad parametru β metodou nejmenších čtverců: $\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i - x_i^T \beta)^2 = (\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T Y$ ... maximálně výhodný odhad β, jsou-li chyby modelu ε_i normální

M-odhady

M-odhad parametru β je řešením minimalizace: $\hat{\beta}^M = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} \rho(Y_i - x_i^T \beta)$, kde ρ je nějaká vhodná funkce (symetrická)

Je-li ρ spojitá derivace $\psi(x) = \rho'(x)$, pak $\hat{\beta}^M$ řeší soustavu rovnic $\sum_{i=1}^{n} x_i \cdot \psi(Y_i - x_i^T \hat{\beta}^M) = 0$.

Poznámky

- volba ρ, resp. ψ stejně jako v modelu polohy
- špatně definovaný M-odhad není ekvivariantní vůči měřítku → studentizace

- odhadneme σ pomocí vhodné měřítkové statistiky S_n (klasický odhad $\sqrt{\frac{1}{n-p}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2}$).

- studentizovaný M-odhad: $\hat{\beta}^{SM} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} \rho\left(\frac{Y_i - x_i^T\beta}{S_n}\right)$, příp. řeší soustavu $\sum_{i=1}^{n} x_i \cdot \psi\left(\frac{Y_i - x_i^T\hat{\beta}^{SM}}{S_n}\right) = 0$.

Influenční funkce M-odhadu je $IF(\underbrace{x_1, y_1}_{(p+1)\text{ proměnných}}; T, P) = \underbrace{B^{-1}}_{\text{nesnězená funkce}} x \underbrace{\psi(y - x^T T(P))}_{\text{musíme } \psi \text{ volit omezenou}}$, kde $B = \int_{\mathbb{R}^{p+1}} x x^T \psi'(y - x^T T(P))\, dP(x,y)$

Bod selhání je 0.

GM-odhady

$\hat{\beta}^{GM} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{n} \rho(Y_i - x_i^T\beta) \cdot w(x_i)$, kde $w: \mathbb{R}^p \to \mathbb{R}$ je vhodná váhová funkce snižující vliv odlehlých hodnot v x_i.

$\hat{\beta}^{GM}$ řeší soustavu: $\sum_{i=1}^{n} x_i \cdot w(x_i)\, \psi(Y_i - x_i^T\beta) = 0$.

$IF(x_1, y_1; T, P) = B^{-1} \underbrace{x \cdot w(x)}_{\text{volíme } \psi \text{ omezenou}} \psi(y - x^T T(P))$

w volíme tak, aby funkce $x \cdot w(x)$ byla omezená

např. $w(x_i) = \frac{1}{h_{ii}}$, kde h_{ii} je diagonální prvek matice $H = X(X^TX)^{-1}X^T$

L-odhady

1978 Koenker, Bassett definovali α-regresní kvantil $\hat{\beta}(\alpha)$ jako řešení minimalizace:

$$\hat{\beta}(\alpha) = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_\alpha (Y_i - x_i^T \beta) \quad , \text{kde } 0 < \alpha < 1 \text{ je pevné a}$$

$$\rho_\alpha(x) = \left\{ \begin{array}{ll} \alpha \cdot x & , x > 0 \\ (\alpha - 1)\, x & , x \leq 0 \end{array} \right\} = |x| \cdot \left(\alpha \cdot \mathbb{1}_{\{x > 0\}} + (1-\alpha) \cdot \mathbb{1}_{\{x \leq 0\}} \right).$$

Poznámka

$\alpha = 0{,}5$, $\rho_\alpha(x) = 0{,}5 \cdot |x|$, $\hat{\beta}(0{,}5) = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{n} 0{,}5 \cdot |Y_i - x_i^T \beta| = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{n} |Y_i - x_i^T \beta|$ je regresní medián.

regresní kvantily se hledají pomocí lineárního programování - simplexová metoda

řešení duální úlohy $\hat{a}(\alpha) = (a_1(\alpha), \dots, a_n(\alpha))^T$ ... regresní pořadové skóry

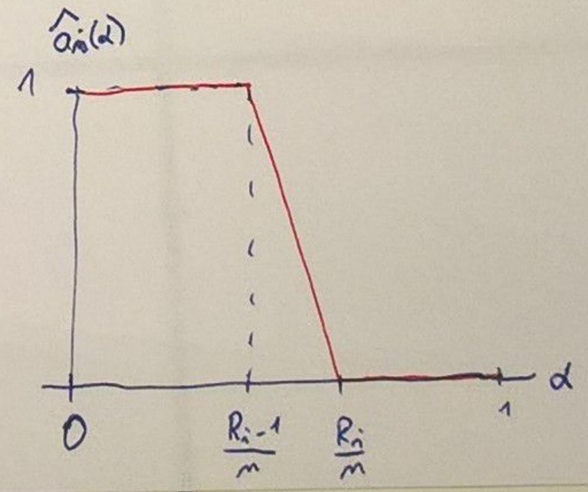
Poznámky

Regresní pořadové skóry jsou invariantní vůči regresi X, tj. $\hat{a}(\alpha, Y + X^T b) = \hat{a}(\alpha, Y)$, $\forall b \in \mathbb{R}^p$

speciálně v modelu polohy: $X = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$ $Y_i = \beta_0 + e_i$

$$\hat{a}_i(\alpha) = \left\{ \begin{array}{ll} 1 & , \alpha \leq \frac{R_i - 1}{n} \\ R_i - \alpha n & , \frac{R_i - 1}{n} < \alpha < \frac{R_i}{n} \\ 0 & , \alpha \geq \frac{R_i}{n} \end{array} \right.$$

R_i je pořadí Y_i mezi $Y_1, \dots, Y_n$.



o dvecném modelu (díky dualitě) platí:

$$\hat{a}_i(\alpha) = \begin{cases} 1 & , Y_i > x_i^T \hat{\beta}(\alpha) \\ 0 & , Y_i < x_i^T \hat{\beta}(\alpha) \end{cases} \quad \text{a} \quad \hat{a}_i(\alpha) \in (0,1) \text{, pokud } Y_i = x_i^T \hat{\beta}(\alpha)$$

$\hat{a}_i(\alpha)$ je spojitá, po částech lineární $\hat{a}_i(0) = 1$, $\hat{a}_i(1) = 0$.

Vrátíme se zpět k regresním kvantilům:

- L_1 - odhad (regresní medián, 0,5-regresní kvantil, LAD)
  - bod selhání 0 (robustní jen v odlehlých Y_i)

- trimmed least squares (TLS) - useknuté nejmenší čtverce

$0 < \underline{\alpha}_1 < \underline{\alpha}_2 < 1$ a spočteme jen ta pozorování, která leží mezi α_1 a α_2 - regresními kvantily a s jejich pomocí spočteme odhad metodou nejmenších čtverců

$$a_i = \mathbb{1}\{ x_i^T \hat{\beta}(\alpha_1) \le Y_i \le x_i^T \hat{\beta}(\alpha_2) \} \quad \dots i = 1, \dots, n$$

$$A = \operatorname{diag}\{a_i\}$$

$$\hat{\beta}^{TLS} = (X^T A X)^{-1} X^T A Y .$$

- preliminary estimate (PE)

• preliminary estimate (PE)

- zvolíme počáteční odhad $\hat{\beta}^{(0)}$ parametru β. Jako $\hat{\beta}^{(0)}$ se volí $\begin{cases} \text{LSE} \\ \text{LAD} \\ \frac{1}{2}(\hat{\beta}(\alpha) + \hat{\beta}(1-\alpha)) \end{cases}$

- odstraníme $\lfloor \alpha n \rfloor$ pozorování s nejmenšími a $\lfloor n\alpha \rfloor$ pozorování s největšími rezidui.

- ze zbylých pozorování určíme odhad metodou nejmenších čtverců

Odhady s vysokým bodem selhání

• least median of squares (LMS) : $\hat{\beta}^{LMS} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \underset{i=1,\dots,n}{\operatorname{med}} \{(Y_i - x_i^T\beta)^2\}$ $\quad \varepsilon^* = 0{,}5$, není však příliš efektivní

• least trimmed squares (LTS): $\hat{\beta}^{LTS} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{h} r^2_{(i)}(\beta)$, kde $r_i^2(\beta) = (Y_i - x_i^T\beta)^2$, $r^2_{(1)} \le r^2_{(2)} \le \dots \le r^2_{(n)}$ $\quad h = \lfloor \frac{n}{2} \rfloor + \lfloor \frac{p+1}{2} \rfloor \dots \varepsilon^* = 0{,}5$

• least trimmed absolute deviation (LTA): $\hat{\beta}^{LTA} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{n-2\lfloor n\alpha \rfloor} |r(\beta)|_{(i)}$, kde $|r_i(\beta)| = |Y_i - x_i^T\beta|$, $|r(\beta)|_{(1)} \le |r(\beta)|_{(2)} \le \dots \le |r(\beta)|_{(n)}$

model regresní přímky: $Y_i = \beta_0 + \beta_1 x_i + e_i$

• repeated median : $\hat{\beta}_1 = \underset{i=1,\dots,n}{\operatorname{med}} \underset{\substack{j=1,\dots,n \\ j \neq i}}{\operatorname{med}} \frac{Y_i - Y_j}{x_i - x_j}$ $\quad \varepsilon^* = 0{,}5$

• Theilův - Senův odhad: $\hat{\beta}_1 = \underset{i<j}{\operatorname{med}} \frac{Y_i - Y_j}{x_i - x_j}$ $\quad \varepsilon^* = 1 - \frac{1}{\sqrt{2}} \doteq 0{,}293$

Robustní odhady parametru σ

- median absolute deviation (MAD)

$\beta^{(0)}$ počáteční odhad parametru β

$r_i = y_i - x_i^T \beta^{(0)}$

$\tilde{r} = \text{med}\{r_1, \dots, r_n\}$

$S_n = \text{med}\{|r_1 - \tilde{r}|, |r_2 - \tilde{r}|, \dots, |r_n - \tilde{r}|\}$.

- L-statistiky založené na regresních kvantilech

$S_n = \| \hat{\beta}(\alpha_2) - \hat{\beta}(\alpha_1) \|$, $0 < \alpha_1 < \alpha_2 < 1$ a $\| . \|$ je eukleidovská vzdálenost v $\mathbb{R}^p$.