USEKNUTÉ ODHADY V MODELU NELINEÁRNÍ REGRESE.

Bohumír Procházka, IHE Praha

1. Úvod

Nejprve věnujme pozornost klasickému modelu nelineární regrese. Mějme náhodné veličiny Y_i

$$Y_i = f(x_i, \beta) + \varepsilon_i \qquad i = 1, \dots, n$$

kde funkce f je funkce daného analytického tvaru, závislá na vektoru nezávisle proměnných $x_i = (X_{i,1}, \dots, X_{i,m})'$ a vektoru neznámých parametrů $\beta = (\beta_1, \dots, \beta_p)'$, ε_i jsou nezávislé stejně rozdělené náhodné veličiny s distribuční funkcí F.

V klasickém modelu nelineární regrese se obvykle předpokládá, že ε_i jsou rozloženy normálně s nulovou střední hodnotou. Maximálně věrohodný odhad je pak odhad získaný metodou nejmenších čtverců, což je $\hat{\beta}$ minimalizující výraz

$$\min_{\beta} \sum_{i=1}^{n} (Y_i - f(x_i, \beta))^2$$

Tuto minimalizaci je nutno provést některou z iteračních metod, které mohou ale nemusí využívat znalost derivací funkce f. Některé z těchto metod popsal např. Y.Bard (3) nebo D.M.Himmelblau (4).

V praxi se ale často setkáváme s tím, že některá pozorování vybočují nějakým způsobem ze skupiny ostatních pozorování. Často se stává, že takovéto odlehlé pozorování způsobí silné vychýlení nebo selhání metody nejmenších čtverců. Proto jsou velmi užitečné metody omezující vliv odlehlých pozorování.

2. Useknutý odhad metodou nejmenších čtverců v lineárním modelu

Pro model lineární regrese bylo navrženo poměrně hodně metod robustních na odlehlá pozorování (viz. např. J.Antoch,G,Collomb,S.Hassani (1) nebo J.Antoch(2)). Jednou z těchto metod je useknutý odhad metodou nejmenších čtverců, kterým se zde budeme podrobněji zabývat.

Konstrukce useknutého odhadu nejmenších čtverců je založena na pojmu regresních kvantilů, jak je poprvé definovali R.Koenker a G.Bassett(8). Dříve než budeme definovat regresní α-kvantil, zvolme nejprve nějaké pevné $0 < \alpha < 1$ a označme

$$\varphi_\alpha(x) = \alpha x \qquad \text{pro } x \geq 0$$
$$= (\alpha - 1)\, x \qquad \text{pro } x < 0$$

Regresní α-kvantil β_α pak definujeme jako řešení minimalizační úlohy

$$\min_{\beta} \sum_{i=1}^{n} \varphi_\alpha (Y_i - x_i' \beta)$$

Pro takovýto regresní kvantil β_α lze dokázat, že počet záporných reziduí $(Y_i - x_i \beta_\alpha)$ je menší nebo roven $n\alpha$ a počet nekladných reziduí je větší nebo roven $n\alpha$ za předpokladu, že některá složka vektoru x_i je pro všechna $i=1, \dots, n$ rovna jedné (věta 3.4 v článku R.Koenker, G.Bassett (8)).

Na základě takto získaného odhadu regresního kvantilu β_α lze sestrojit useknutý nebo winsorisovaný odhad parametru β. Dále se budeme zabývat pouze useknutým odhadem metodou nejmenších čtverců.

Zvolme nejprve dvě konstanty $0 < \alpha_1 < \alpha_2 < 1$. Useknutý odhad β^{TLS} metodou nejmenších čtverců můžeme sestrojit následujícím způsobem:

1. Nechť pro $j=1,2$ je β_{α_j} řešení minimalizační úlohy

$$\min_{\beta} \sum_{i=1}^{n} \varphi_{\alpha_j}(Y_i - x_i'\beta)$$

2. Odstraňme z původního výběru všechna pozorování pro která je

$$Y_i - x_i'\beta_{\alpha_1} < 0 \quad \text{nebo} \quad Y_i - x_i'\beta_{\alpha_2} > 0$$

3. α-useknutý odhad β^{TLS} metodou nejmenších čtverců je odhad získaný metodou nejmenších čtverců ze zbylých pozorování.

Studiem statistických vlastností se zabývali například J.Jurečková (6),(7) nebo D.Ruppert a R.J.Carroll (10). V článku J.Jurečkové (5) jsou odvozeny testy v modelu polynomické regrese.

3. Useknuté odhady v modelu nelineární regrese

Vraťme se nyní k modelu nelineární regrese. Na prvý pohled se nabízí možnost zobecnění useknutého odhadu i na tento model. Nejprve zaveďme pro dannou funkci $f(x,\beta)$ novou funkci $\mathfrak{f}(x,\mathbb{b})$ následujícím způsobem :

1. Je-li možno funkci f zapsat ve tvaru $f(x,\beta) = g(x, \beta_1,\dots,\beta_{i-1},\beta_{i+1},\dots,\beta_p) + c\beta_i$ pro nějaké $i=1,\dots,p$ (kde c je libovolná konstanta), označíme $r = p$, $\mathbb{b} = (\beta_1,\dots,\beta_{i-1},\beta_{i+1},\dots,\beta_p,\beta_i/c)'$ a $\mathfrak{f}(x,\mathbb{b}) = f(x,\beta)$.
2. Jinak položme $r = p+1$, $\mathbb{b} = (\beta_1,\dots,\beta_p,\beta_{p+1})'$ a $\mathfrak{f}(x,\mathbb{b}) = f(x,\beta) + \beta_r$.

Definujme nynípro regresní funkci f regresní α-kvantil $\mathbb{b}_\alpha$ jako vektor minimalizující

$$\min_{\mathbb{b}} \sum_{i=1}^{n} \varphi_\alpha(Y_i - \mathfrak{f}(x_i,\mathbb{b}))$$

Pro takto definovaný α-kvantil $\mathbb{b}_\alpha$ lze dokázat větu podobnou větě 3.4 z článku R.Koenkera a G.Bassetta (8).

Věta :
Označme R^+ počet kladných reziduí $(Y_i - \mathfrak{f}(x_i,\mathbb{b}_\alpha))$, R^0 počet nulových reziduí a R^- počet záporných reziduí. Pak pro libovolné $0<\alpha<1$, pro každé řešení $\mathbb{b}_\alpha$ minimalizující

$$\min_{\mathbb{b}} \sum_{i=1}^{n} \varphi_\alpha(Y_i - \mathfrak{f}(x_i,\mathbb{b}))$$

platí

$$R^- \leqslant \alpha n \leqslant n - R^+ = R^- + R^0$$

Důkaz : Zaveďme nejprve následující označení

$$\psi_\alpha(\mathbb{b}) = \sum_{i=1}^{n} \varphi_\alpha(Y_i - \mathfrak{f}(x_i,\mathbb{b}))$$

$$\operatorname{sgn}^*(u, v) = \begin{cases} \operatorname{sgn}(u) & \text{je-li } u \neq 0 \\ \operatorname{sgn}(v) & \text{je-li } u = 0 \end{cases}$$

Zřejmě platí

$$\psi_\alpha(\mathbb{b}) = \sum_{i=1}^{n} \left(\frac{1}{2}\operatorname{sgn}(Y_i - f(\mathcal{X}_i,\mathbb{b})) + \alpha - \frac{1}{2}\right)(Y_i - f(\mathcal{X}_i,\mathbb{b}))$$

Uvažujme nyní derivaci této funkce ve směru w.

$$\psi'_\alpha(\mathbb{b},\mathbf{w}) = \sum_{i=1}^{n} \left(\frac{1}{2} - \alpha - \frac{1}{2}\operatorname{sgn}^*(Y_i - f(\mathcal{X}_i,\mathbb{b}), -\mathbf{w}'\mathbf{d})\right)(\mathbf{w}'\mathbf{d})$$

kde $\mathbf{d} = \left(\dfrac{\partial f(\mathcal{X}_i,\mathbb{b})}{\partial b_j}\right)_{j=1,\dots,r}$. Protože každý regresní α-kvantil minimalizuje funkci $\psi_\alpha(\mathbb{b})$, platí pro libovolné w

$$\psi'_\alpha(\mathbb{b}_\alpha, \mathbf{w}) \geqslant 0.$$

Dosadíme-li do této nerovnosti vektory $\mathbf{w}^+ = (0,\dots,0,1)'$ a $\mathbf{w}^- = (0,\dots,0,-1)'$, je

$$\sum_{i=1}^{n} \pm\left(\frac{1}{2} - \alpha - \frac{1}{2}\operatorname{sgn}^*(Y_i - f(\mathcal{X}_i,\mathbb{b}_\alpha), \mp 1)\right) \geqslant 0$$

což je

$$-\alpha R^+ + (1-\alpha)R^0 + (1-\alpha)R^- \geqslant 0 \quad \text{a} \quad -\alpha R^+ - \alpha R^0 - (1-\alpha)R^- \leqslant 0$$

je tedy $R^- \leqslant \alpha n \leqslant R^- + R^0$. ∎

Nyní definujeme α-useknutý odhad metodou nejmenších čtverců následujícím způsobem :

1. Nechť pro předem zvolená $0 < \alpha_1 < \alpha_2 < 1$ je $\mathbb{b}_{\alpha_j}$ $j=1,2$ řešení minimalizace
$$\min_{\mathbb{b}} \sum_{i=1}^{n} \varphi_{\alpha_j}(Y_i - f(\mathcal{X}_i,\mathbb{b}))$$
2. Odstraňme z původního výběru všechna pozorování pro která je
$$Y_i - f(\mathcal{X}_i, \mathbb{b}_{\alpha_1}) < 0 \quad \text{nebo} \quad Y_i - f(\mathcal{X}_i,\mathbb{b}_{\alpha_2}) > 0$$
3. α-useknutý odhad β^{TLS} metodou nejmenších čtverců položíme rovný odhadu získanému metodou nejmenších čtverců ze zbylých pozorování s regresní funkcí $f(\mathcal{X},\beta)$.

V další části se budeme zabývat pouze symetrickým useknutím tj. nechť $0<\alpha<.5$ pak $\alpha_1 = \alpha$, $\alpha_2 = 1-\alpha$. Je nutno si ale uvědomit, že v mnohdy může být vhodné jednostranné nebo nesymetrické useknutí.

Poznámka : lineárním modelu je možno použít na konstrukci regresních kvantilů metod lineárního programování a odhad metodou nejmenších čtverců je numericky též velmi snadný. V modelu nelineární regrese se všechny tyto výhody ztrácejí. Je nutno použít nějakou iterační metodu. Jedna z metod vhodných jak pro výpočet regresních kvantilů, tak i pro odhad metodou nejmenších čtverců je simplexová metoda nelineárního programování popsaná v knize D.Himmelblaua (4).

4. Příklad

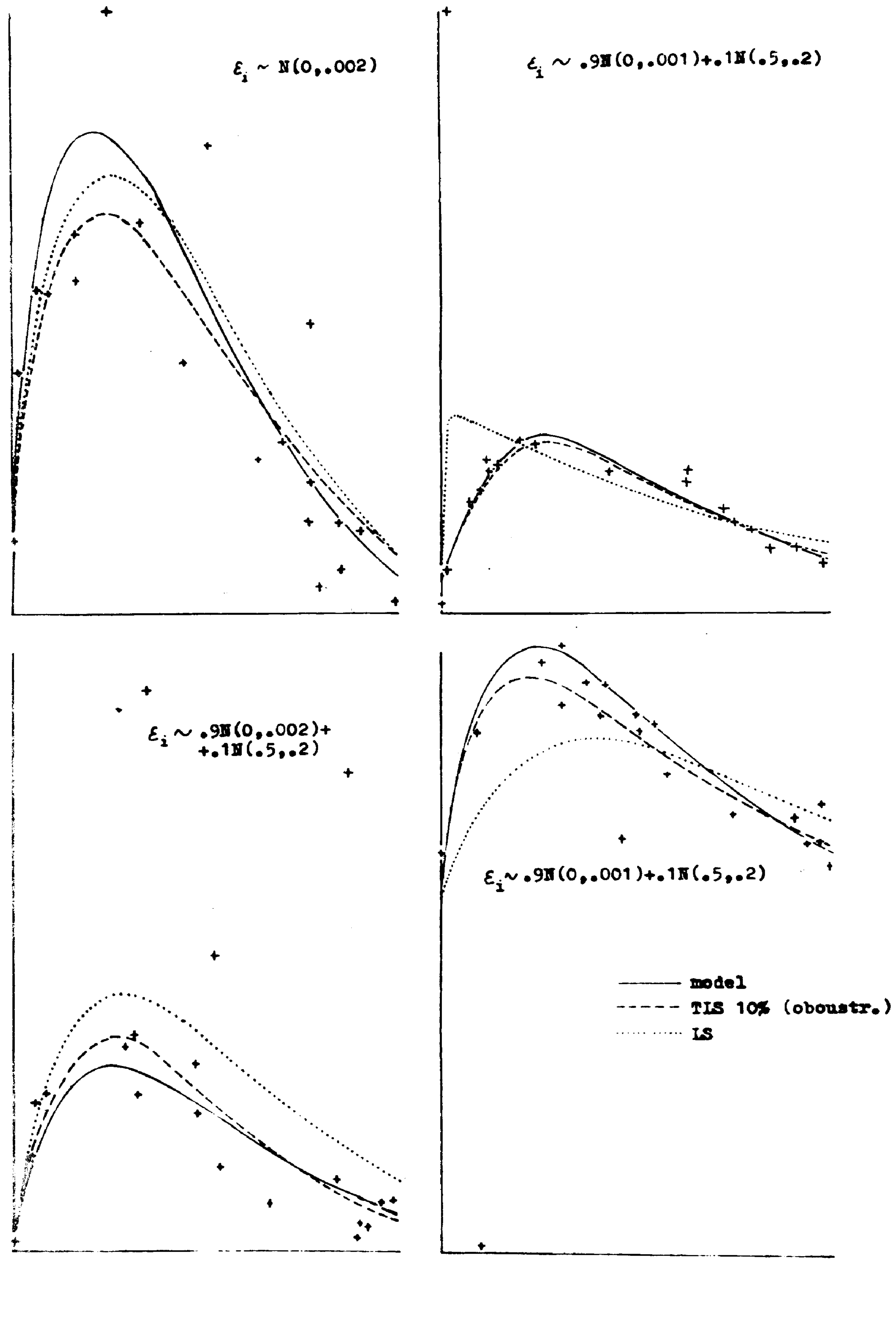
Pro simulační pokus byl zvolen model s regresní funkcí

$$f(x,\beta) = \beta_3(\exp(-x\beta_2) - \exp(-x\beta_1)).$$

Všechny simulace byly provedeny s vektorem $\beta = (0.6, 0.2, 1.)'$. Pro výpočet odhadů byl vytvořen program v jazyce Fortran na počítači HP 2100. Jako iterační procedura byla zvolena simplexová metoda (D.Himmelblau(4)).

Pro tento model bylo nagenerováno několik výběrů o rozsahu 40 , kde ε_i mělo rozložení normální, normální kontaminované cauchyovým a cauchyovo. Pozorování byla simulována pronezávisle proměnnou $x_i = 10i/40$ $i=1,\dots,40$. Případná kontaminace byla provedena 15% nebo 20%. Výsledky simulací jsou otištěny v tabulce. Jako měřítko přiblížení odhadu skutečné hodnotě parametru bylo zvoleno supremum

Příloha : Ukázky odhadů v modelu s regresní funkcí z příkladu. odhady byly sestrojeny z výběrů o rozsahu 20. x byly voleny náhodně (rovnoměrně na ⟨0,12⟩).



Tabulka

	n	$\beta^{TLS10\%}$	$\|f-\hat{f}\|$	$\beta^{TLS5\%}$	$\|f-\hat{f}\|$	β^{LS}	$\|f-\hat{f}\|$
N	40	.47 .25 1.75	.0118	.48 .24 1.57	.0112	.46 .25 1.78	.0113
		.64 .18 .87	.0091	.68 .17 .79	.0130	.62 .19 .91	.0093
		.50 .23 1.35	.0141	.58 .20 1.00	.0079	.48 .24 1.56	.0101
.85N+.15C	40	.73 .17 .77	.0102			.89 .13 .59	.0275
		.51 .23 1.38	.0066			.43 .29 2.86	.0369
		.74 .17 .75	.0133			.43 .24 1.74	.0366
.8N+.2C	40	.61 .19 .92	.0087	.54 .21 1.16	.0115	.34 .15 .85	.1633
		.56 .21 1.12	.0046	.58 .20 1.03	.0030	.47 .26 1.76	.0106
C	40	.70 .17 .77	.0122			.51 .23 1.44	.0279
		1.28 .16 .71	.1424			31. .20 1.00	.7964
		.75 .17 .82	.0293			.48 .24 1.80	.0589

$$\|f-\hat{f}\| = \sup_{x \in \langle 0,10\rangle} |f(x,\beta)-f(x,\hat{\beta})| , \quad \hat{\beta} = \beta^{TLS10\%}, \beta^{TLS5\%}, \beta^{LS}.$$

Kde β^{LS} je odhad parametru β metodou nejmenších čtverců a $\beta^{TLS10\%}$, $\beta^{TLS5\%}$ jsou jsou useknuté odhady metodou nejmenších čtverců s oboustranným 10% , 5% useknutím.

Poznámka : Při konstrukci useknutých odhadů v modelu nelineární regrese, stejně jako u ostatních odhadů bývají problémy spojené s konstrukcí počátečního odhadu, s možností existence více lokálních minim, s možností divergence nebo oscilace iteračního procesu. Jiné nesnáze mohou nastat při špatném plánu experimentu (při nevhodně volené nezávisle proměnné x), může se totiž snadno stát, že mohou být vyloučena pozorování z některé klíčové části křivky kde bylo málo pozorování náhodné veličiny Y. Přes všechny tyto problémy lze tento odhad v kontaminovaném modelu jen doporučit.

Literatura :

(1) J.Antoch,G.Collomb,S.Hassani : Robustness in parametric and non-parametric regression estimation. Sborník COMPSTAT 1984, Physica Verlag, Viena, 1984

(2) J.Antoch : Některé postupy pro výpočet robustních odhadů v lineárním regresním modelu. Sborník ROBUST 84, JČMF, 1984

(3) Y.Bard : Nonlinear parameter estimation. Academia Press, 1974

(4) D.M.Himmelblau : Process analysis by statistical methods. J.Wiley sons,1970

(5) J.Jurečková : Trimmed polynomial regression. CMUS, 24.4.1983

(6) J.Jurečková : Linear statistical inference based on L-estimators. Sborník ROBUST 84, JČMF, 1984

(7) J.Jurečková : Regressions Quantiles and trimmed least squares estimator under a general design. Kybernetika 20,5 (1984)

(8) R.Koenker,G.Bassett : Regression quantiles. Econometrika 46,1 (1978)

(9) B.Procházka : Nelineární regrese vefarmakokinetice. Dipl.práce MFF UK 1977

(10) D.Ruppert,R.J.Carroll : Trimmed least squares in estimation in the linear model. JASA 75,372 (1980)