ROBUST'98, 233 – 241 © JČMF 1998

O ROBUSTNOSTI KRITÉRIA PRO VYŘAZENÍ REGRESORU

Karel VLČEK

KMA ZČU, Plzeň

Abstract. The paper deals with a problem of omitting one regressor in a linear model with a nonstochastic matrix of a full rank. The value of the regression coefficient that makes the mean square errors of predictions for the full model and the corresponding submodel equal will be called the boundary coefficient. We consider a bayesian approach for which the square of the given regression coefficient has a dichotomic symmetric distribution the mean of which is the boundary coefficient. We also investigate an asymptotic case and robustness with regard to nonnormal distributions of errors.

Резюме: Рассматривается вопрос удаления одного регрессора в линейной регрессионной модели с постоянной матрицей полного ранга. Регрессионный коэффициент, для которого полная модель и модель без удаленого регрессора имеют одинакогую потерь MSE, будем называть предельным коэффициентом. Изучается байесовский подход, при котором квадрат регрессионного коэффициента имеет априорное распределение с двумя значениями симметрически расположенными около квадрата предельного коэффициента. Рассматривается также асимптотический случай и робастность по распределению ошибок.

1. Hlavní výsledky

Uvažujme lineární regresní model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\gamma + \mathbf{e}, \tag{1}$$

kde pro začátek předpokládejme, že $\mathbf{X}$ je konstantní matice typu $n \times p$, $\mathbf{Z}$ je konstantní matice typu $n \times 1$, matice $(\mathbf{X}, \mathbf{Z})$ má plnou hodnost $p_g = p+1 < n$ a $\mathbf{e}$ je vektor nekorelovaných náhodných chyb, $\mathrm{E}\,\mathbf{e} = \mathbf{0}$, $\mathrm{var}\,\mathbf{e} = \sigma^2\mathbf{I}$. Zároveň uvažujme model vzniklý vynecháním regresoru $\mathbf{Z}$, tj.

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}^*, \tag{2}$$

kde $\mathbf{e}^* = \mathbf{Z}\gamma + \mathbf{e}$. Odhady parametrů $\boldsymbol{\beta}$ a γ v modelu (1) získané metodou nejmenších čtverců (dále jen MNČ) označme $\mathbf{b}_g$ a c_g, tj.

$$\begin{pmatrix} \mathbf{b}_g \\ c_g \end{pmatrix} = \left[(\mathbf{X}, \mathbf{Z})'(\mathbf{X}, \mathbf{Z})\right]^{-1}(\mathbf{X}, \mathbf{Z})'\mathbf{y}. \tag{3}$$

Odhad parametru $\boldsymbol{\beta}$ v modelu (2) získaný MNČ označme $\mathbf{b}$, tj.

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \tag{4}$$

Zajímejme se o lineární parametrickou funkci $\Theta = \mathbf{r}'_{\boldsymbol{\beta}}\boldsymbol{\beta} + r_\gamma\gamma$, kde $\mathbf{r}_{\boldsymbol{\beta}}$ je typu $p \times 1$ a r_γ je reálné číslo. Odhady této parametrické funkce získané z modelů (1) a (2) označme $\hat{\Theta}^{(g)} = \mathbf{r}'_{\boldsymbol{\beta}}\mathbf{b}_g + r_\gamma c_g$, resp. $\hat{\Theta} = \mathbf{r}'_{\boldsymbol{\beta}}\mathbf{b}$.

Všímejme si středních kvadratických chyb těchto odhadů,

$$\begin{aligned} \mathrm{MSE}\,(\hat{\Theta}^{(g)}) &= \mathrm{E}\left[\hat{\Theta}^{(g)} - \Theta\right]^2, \\ \mathrm{MSE}\,(\hat{\Theta}) &= \mathrm{E}\left[\hat{\Theta} - \Theta\right]^2. \end{aligned}$$

V Reif a Vlček (1998) je dokázána věta:

Věta 1. *Pro výše definované odhady parametrické funkce* Θ *platí rovnost*

$$\mathrm{MSE}\,(\hat{\Theta}^{(g)}) - \mathrm{MSE}\,(\hat{\Theta}) = t^2\mathbf{Z}'\mathbf{M}\mathbf{Z}(\sigma^2 - \gamma^2\mathbf{Z}'\mathbf{M}\mathbf{Z}), \tag{5}$$

kde $\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ *a* t *je reálné číslo, které nezávisí na* σ^2 *a* γ.

Poznámka 2. *Z předpokladu úplné hodnosti modelu* (1) *plyne* $\mathbf{Z}'\mathbf{M}\mathbf{Z} > 0$. *Můžeme tedy označit* $C = (\mathbf{Z}'\mathbf{M}\mathbf{Z})^{-1}$ *a počítat např.* $\sqrt{C}$.

Důsledek 3. *Za výše uvedených předpokladů platí implikace*

1) $\gamma^2 \le C\sigma^2 \;\Rightarrow\; \mathrm{MSE}\,(\hat{\Theta}^{(g)}) \ge \mathrm{MSE}\,(\hat{\Theta})$ *pro všechny par. funkce* Θ,

2) $\gamma^2 \ge C\sigma^2 \;\Rightarrow\; \mathrm{MSE}\,(\hat{\Theta}^{(g)}) \le \mathrm{MSE}\,(\hat{\Theta})$ *pro všechny par. funkce* Θ.

Abychom se rozhodli, zda z hlediska minimalizace MSE odhadů je vhodnější použít model (1) nebo (2), potřebujeme zjistit, zda $\gamma^2 \ge C\sigma^2$ nebo $\gamma^2 \le C\sigma^2$, kde C je známé číslo, parametry γ^2 a σ^2 však neznáme. Nabízí se možnost použít místo γ jeho odhad c_g zjištěný MNČ z většího modelu (1) a za σ^2 příslušný reziduální rozptyl s_g^2 (z většího modelu). Statistika s_g^2 je sice nevychýleným odhadem σ^2, avšak c_g^2 je vychýleným odhadem γ^2. Z klasické teorie lze snadno odvodit, že nevychýleným odhadem γ^2 je statistika $\hat{\gamma^2} = c_g^2 - Cs_g^2$. Potom platí ekvivalence

$$\hat{\gamma^2} \ge Cs_g^2 \quad\Leftrightarrow\quad c_g^2 - Cs_g^2 \ge Cs_g^2 \quad\Leftrightarrow\quad c_g^2 \ge 2Cs_g^2 \quad\Leftrightarrow\quad \frac{|c_g|}{s_g\sqrt{C}} \ge \sqrt{2}$$

a analogicky

$$\hat{\gamma^2} \le Cs_g^2 \quad\Leftrightarrow\quad \frac{|c_g|}{s_g\sqrt{C}} \le \sqrt{2} \tag{6}$$

Zatím jsme neměli žádný specifický předpoklad o rozdělení náhodných odchylek e_i, $i = 1, \ldots, n$. Předpokládejme dále, že $\mathbf{e} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$. Výraz na pravé straně ekvivalence (6) připomíná oboustranný t-test hypotézy $\gamma = 0$, pouze namísto kvantilu $t_{1-\alpha/2}(n - p_g)$ srovnáváme absolutní hodnotu známé t-statistiky s číslem $\sqrt{2}$. Pro zajímavost poznamenejme, že „test“ (6) pro $n \to \infty$ odpovídá hladině významnosti $\alpha \approx 0.16$.

Otázkou zůstává, zda dosazování nevychýlených odhadů parametrů γ^2 a σ^2 do nerovností v důsledku 3 při rozhodování o vynechání regresoru $\mathbf{Z}$ je vhodnou volbou, zda se raději neřídit jinými kritérii.

Přistupme nyní k našemu problému jako k bayesovskému rozhodovacímu problému s určitým apriorním rozdělením koeficientu γ.

Označíme-li $\tilde{\gamma} = \sigma\sqrt{C}$, lze důsledek 3 přepsat do tvaru

$$\begin{aligned} \mathrm{MSE}\,(\hat{\Theta}^{(g)}) &\geq \mathrm{MSE}\,(\hat{\Theta}) \quad \text{pro } \gamma^2 \leq \tilde{\gamma}^2, \\ \mathrm{MSE}\,(\hat{\Theta}^{(g)}) &\leq \mathrm{MSE}\,(\hat{\Theta}) \quad \text{pro } \gamma^2 \geq \tilde{\gamma}^2. \end{aligned}$$

Protože si přejeme volit apriorní hustotu tak, aby byla co nejjednodušší a nezvýhodňovala předem ani bohatší, ani chudší model, budeme volit diskrétní rozdělení pro γ tak, aby

$$\gamma^2 = \begin{cases} \gamma_A^2 & \text{s pravděpodobností } 1/2 \\ \gamma_B^2 & \text{s pravděpodobností } 1/2, \end{cases}$$

kde γ_A^2, γ_B^2 jsou symetricky rozloženy kolem $\tilde{\gamma}^2$. Volme

$$\begin{aligned} \gamma_A &= \tilde{\gamma}\sqrt{1+\Delta}, \\ \gamma_B &= \tilde{\gamma}\sqrt{1-\Delta}, \end{aligned}$$

kde $0 < \Delta < 1$. Tato volba zaručuje, že pro $\gamma^2 = \gamma_A^2$ má výraz na pravé straně rovnosti (5) přesně opačnou hodnotu než pro $\gamma^2 = \gamma_B^2$. Motivací této volby je, že později budeme místo ztrátové funkce MSE uvažovat jednodušší ztrátovou funkci s hodnotami $\{\,0; 1\,\}$.

Při dalším apriorním předpokladu $\gamma > 0$ je tedy naše apriorní rozložení koeficientu γ definováno následovně:

$$\begin{aligned} \gamma &= \gamma_A \quad \text{s pravděpodobností } 1/2, \\ \gamma &= \gamma_B \quad \text{s pravděpodobností } 1/2. \end{aligned}$$

Tuto variantu budeme nazývat jednostrannou.

Později budeme uvažovat též variantu, kterou budeme nazývat oboustrannou, kdy γ může nabývat hodnot γ_A, γ_B, $-\gamma_A$, $-\gamma_B$, každé s pravděpodobností $1/4$.

Volme pro začátek velmi jednoduchou ztrátovou funkci zapsanou zatím obecně takto:

$$L = \begin{cases} 0 & \text{je-li } \gamma^2 = \gamma_A^2 \text{ a volíme model (1)} \\ & \text{nebo} \\ & \text{je-li } \gamma^2 = \gamma_B^2 \text{ a volíme model (2)}, \\ & \\ 1 & \text{je-li } \gamma^2 = \gamma_A^2 \text{ a volíme model (2)} \\ & \text{nebo} \\ & \text{je-li } \gamma^2 = \gamma_B^2 \text{ a volíme model (1)}. \end{cases}$$

K úplnému popisu funkce L je samozřejmě třeba definovat, pro jaký vektor $\mathbf{y}$ se rozhodneme pro model (1), ev. (2).

Jednou z možností volby rozhodnutí je využít opět statistiku

$$Q := \frac{c_g}{s_g\sqrt{C}}, \tag{7}$$

respektive statistiku

$$Q^2 := c_g^2 \Big/ s_g^2 C.$$

V jednostranném případě přijmeme chudší model (2) pro $Q \leq q^{(1)}$, bohatší model (1) pro $Q > q^{(1)}$, kde konstantu $q^{(1)}$ budeme specifikovat později. V oboustranném případě přijmeme chudší model (2) pro $Q^2 \leq q^{(2)}$, bohatší model (1) pro $Q^2 > q^{(2)}$.

Omezili jsme se tedy zatím jen na rozhodovací funkce typu „t-testu“. Konstanty $q^{(1)}$, $q^{(2)}$ budeme volit tak, abychom minimalizovali *bayesovskou rizikovou funkci* ϱ.

V jednostranném případě chceme minimalizovat

$$\begin{aligned} \varrho^{(1)} &= \big[1 \cdot \mathrm{P}(Q \leq q^{(1)} \mid \gamma = \gamma_A) + 0 \cdot \mathrm{P}(Q > q^{(1)} \mid \gamma = \gamma_A)\big]\, \mathrm{P}(\gamma = \gamma_A) + \\ &+ \big[0 \cdot \mathrm{P}(Q \leq q^{(1)} \mid \gamma = \gamma_B) + 1 \cdot \mathrm{P}(Q > q^{(1)} \mid \gamma = \gamma_B)\big]\, \mathrm{P}(\gamma = \gamma_B), \end{aligned}$$

tzn. minimalizovat

$$\varrho^{(1)} = \frac{1}{2}\big[\mathrm{P}(Q \leq q^{(1)} \mid \gamma = \gamma_A) + \mathrm{P}(Q > q^{(1)} \mid \gamma = \gamma_B)\big].$$

Za předpokladu normality, $\mathbf{e} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$, má Q necentrální t-rozdělení pravděpodobnosti s parametrem necentrality

$$\delta_A = \sqrt{1 + \Delta} \quad \text{pro } \gamma = \gamma_A, \tag{8}$$

respektive

$$\delta_B = \sqrt{1 - \Delta} \quad \text{pro } \gamma = \gamma_B, \tag{9}$$

a s počtem stupňů volnosti $f := n - p_g$.

Požadavek $\dfrac{d\varrho^{(1)}}{dq^{(1)}}$ vede na rovnici (pro neznámou $x \equiv q^{(1)}$)

$$p_f^1(x, \delta_A) - p_f^1(x, \delta_B) = 0, \tag{10}$$

kde $p_f^1(x, \delta)$ je hustota necentrálního t-rozdělení s parametrem necentrality δ a počtem stupňů volnosti f, tedy

$$p_f^1(x, \delta) = \Gamma_f \cdot \Xi_f(x) \cdot \Pi_f(x, \delta), \tag{11}$$

kde

$$\Gamma_f = \left(2^{\frac{f-1}{2}}\Gamma\left(\frac{f}{2}\right)\sqrt{\pi f}\right)^{-1}, \tag{12}$$

$$\Xi_f(x) = \left(1+\frac{x^2}{f}\right)^{-\frac{f+1}{2}}, \tag{13}$$

$$\Pi_f(x,\delta) = \Theta_f(x,\delta)\cdot\Upsilon_f(x,\delta), \tag{14}$$

$$\Theta_f(x,\delta) = \exp\left(-\frac{1}{2}\,\frac{f\delta^2}{f+x^2}\right) \tag{15}$$

$$\Upsilon_f(x,\delta) = \int_0^\infty v^f e^{-\frac{1}{2}(v-\kappa_f)^2}dv, \tag{16}$$

$$\kappa_f(x,\delta) = x\delta/\sqrt{f+x^2}. \tag{17}$$

V oboustranném případě se minimalizuje ztrátová funkce

$$\begin{aligned}
\varrho^{(2)} &= \left[\mathrm{P}(Q^2 \le q^{(2)} \mid \gamma=\gamma_A) + 0\cdot\mathrm{P}(Q^2 > q^{(2)} \mid \gamma=\gamma_A)\right]\mathrm{P}(\gamma=\gamma_A) + \\
&+ \left[0\cdot\mathrm{P}(Q^2 \le q^{(2)} \mid \gamma=\gamma_B) + \mathrm{P}(Q^2 > q^{(2)} \mid \gamma=\gamma_B)\right]\mathrm{P}(\gamma=\gamma_B) + \\
&+ \left[\mathrm{P}(Q^2 \le q^{(2)} \mid \gamma=-\gamma_A) + 0\cdot\mathrm{P}(Q^2 > q^{(2)} \mid \gamma=-\gamma_A)\right]\mathrm{P}(\gamma=-\gamma_A) \\
&+ \left[0\cdot\mathrm{P}(Q^2 \le q^{(2)} \mid \gamma=-\gamma_B) + \mathrm{P}(Q^2 > q^{(2)} \mid \gamma=-\gamma_B)\right]\mathrm{P}(\gamma=-\gamma_B) \\
&= \frac{1}{4}\big[\mathrm{P}(Q^2 \le q^{(2)} \mid \gamma=\gamma_A) + \mathrm{P}(Q^2 > q^{(2)} \mid \gamma=\gamma_B) + \\
&+ \mathrm{P}(Q^2 \le q^{(2)} \mid \gamma=-\gamma_A) + \mathrm{P}(Q^2 > q^{(2)} \mid \gamma=-\gamma_B)\big].
\end{aligned}$$

Odtud lze $\varrho^{(2)}$ vyjádřit pomocí hodnot distribučních funkcí necentrálního t-rozdělení pravděpodobnosti s parametry necentrality $\pm\delta_A$, $\pm\delta_B$ v bodech $\pm\sqrt{q^{(2)}}$ a derivováním podle $q^{(2)}$ dostaneme, že neznámá $x \equiv q^{(2)}$ splňuje rovnici

$$p_f^2(x,\delta_A) - p_f^2(x,\delta_B) + p_f^2(x,-\delta_A) - p_f^2(x,-\delta_B) = 0, \tag{18}$$

kde

$$p_f^2(x,\delta) = \left(p_f^1(\sqrt{x},\delta) + p_f^1(-\sqrt{x},\delta)\right)/(2\sqrt{x}). \tag{19}$$

Dosadíme-li poslední vztah do (18) a vzniklou rovnici vynásobíme $2\sqrt{x}$, máme podmínku

$$\begin{aligned}
0 = {} & p_f^1(\sqrt{x},\delta_A) + p_f^1(-\sqrt{x},\delta_A) - \left[p_f^1(\sqrt{x},\delta_B) + p_f^1(-\sqrt{x},\delta_B)\right] + \\
& + p_f^1(\sqrt{x},-\delta_A) + p_f^1(-\sqrt{x},-\delta_A) - \left[p_f^1(\sqrt{x},-\delta_B) + p_f^1(-\sqrt{x},-\delta_B)\right].
\end{aligned} \tag{20}$$

Rovnici (10) lze upravit do tvaru

$$0 = \Pi_f(x,\delta_A) - \Pi_f(x,\delta_B), \tag{21}$$

kde $\Pi_f(x,\delta)$ je definována vztahem (14), rovnici (20) do tvaru

$$\begin{aligned} 0 &= \Theta_f(\sqrt{x},\delta_A)\cdot\left(\Upsilon_f(\sqrt{x},\delta_A)+\Upsilon_f(-\sqrt{x},\delta_A)\right) \\ &- \Theta_f(\sqrt{x},\delta_B)\cdot\left(\Upsilon_f(\sqrt{x},\delta_B)+\Upsilon_f(-\sqrt{x},\delta_B)\right), \end{aligned} \tag{22}$$

kde funkce $\Theta_f(x,\delta)$ a $\Upsilon_f(x,\delta)$ jsou definovány vztahy (15) – (17). Při úpravě rovnice (10) jsme pouze dělili nenulovými výrazy (12) a (13), při úpravě rovnice (20) jsme dělili výrazy Γ_f a $\Xi(\sqrt{x}) = \Xi(-\sqrt{x})$, vytkli jsme členy Θ_f (funkce (15) je sudá v obou proměnných) a uvážili jsme, že $\Upsilon(\sqrt{x},\delta_A) = \Upsilon(-\sqrt{x},-\delta_A)$, $\Upsilon(-\sqrt{x},\delta_A) = \Upsilon(\sqrt{x},-\delta_A)$, $\Upsilon(\sqrt{x},\delta_B) = \Upsilon(-\sqrt{x},-\delta_B)$ a $\Upsilon(-\sqrt{x},\delta_B) = \Upsilon(\sqrt{x},-\delta_B)$, viz definiční vztahy (16) a (17).

Rovnice (21) a (22) lze numericky řešit a najít tak optimální konstanty $q^{(1)}$, $q^{(2)}$, které pro dané $f = n - p_g$ a hodnotu parametru Δ označíme $q_f^{(1)}(\Delta)$ a $q_f^{(2)}(\Delta)$. Numerickou integrací lze dokonce přibližně určit i odpovídající hodnoty ztrátové funkce — označme je $L_f^{(1)}(\Delta)$ a $L_f^{(2)}(\Delta)$. Tak například

$$q_1^{(1)}(0.5) \approx 0.7996, \quad L_1^{(1)}(0.5) \approx 0.4127, \tag{23}$$

$$q_1^{(2)}(0.5) \approx 1.3936, \quad L_1^{(2)}(0.5) \approx 0.4417, \tag{24}$$

$$q_8^{(2)}(0.5) \approx 1.4151, \quad L_8^{(2)}(0.5) \approx 0.4163. \tag{25}$$

Článek Vlček (1996) uvádí jiné hodnoty, neboť se tam (na straně 88 sborníku) uvažují jiná δ_A, δ_B než v (8) a (9).

Ukazuje se, že optimální konstanty $q^{(1)}$, $q^{(2)}$ se příliš nemění, měníme-li hodnotu pomocného parametru Δ. Lze očekávat (a numerické experimenty to potvrzují), že rozhodování mezi oběma regresními modely bude problematičtější pro malé hodnoty parametru Δ. Podívejme se tedy, co se stane v limitě pro $\Delta \to 0_+$ v nejjednodušším jednostranném případě pro jeden stupeň volnosti. Vydělme rovnici (21) rozdílem $\delta_A - \delta_B \neq 0$ a vezměme $\Delta \to 0_+$. Dostaneme rovnici (pro $x \equiv q^{(1)}$)

$$\left(\frac{\partial \Pi_f}{\partial \delta}\right)_{\delta=1}(x) = 0. \tag{26}$$

Derivováním a běžnými úpravami lze pro $f = 1$ po chvíli dospět k rovnici

$$(1+x^2)\left(\frac{\partial \Pi_1}{\partial \delta}\right)_{\delta=1}(x)$$

$$= -e^{-\frac{1}{2}} + \frac{x^3 e^{-\frac{1}{2}/(1+x^2)}}{\sqrt{1+x^2}}\sqrt{2\pi}\cdot\Phi\left(\frac{x}{\sqrt{1+x^2}}\right) = 0,$$

jejímž řešením je číslo $q_1^{(1)}(0) \approx 0.8304$ (argument nula označuje limitní případ).

Uvažujme limitní případy pro $n \to \infty$, tzn. $f \to \infty$. Necentrální t-rozdělení s hustotou (11) pak přechází v normální rozdělení se střední hodnotou δ_A

nebo δ_B, viz (8) a (9), a rozptylem 1. Všiměme si, že do stejné situace se dostaneme pro konečný počet stupňů volnosti, budeme-li znát rozptyl σ^2, tj. nahradíme-li ve vztahu (7) statistiku s_g směrodatnou odchylkou σ.

V dalším bude φ značit hustotu normálního normalizovaného rozdělení pravděpodobnosti.

V obou případech ($n \to \infty$ nebo známý rozptyl σ^2) přejde rovnice (10) v rovnici

$$\varphi(x - \delta_A) - \varphi(x - \delta_B) = 0,$$

jejímž řešením je $x = (\delta_A + \delta_B)/2 = (\sqrt{1+\Delta} + \sqrt{1-\Delta})/2$. Rovnice (20) podobně přejde v rovnici

$$\begin{aligned} 0 \;&=\; \varphi(\sqrt{x} - \delta_A) + \varphi(-\sqrt{x} - \delta_A) - \big[\varphi(\sqrt{x} - \delta_B) + \varphi(-\sqrt{x} - \delta_B)\big] + \\ &+\; \varphi(\sqrt{x} + \delta_A) + \varphi(-\sqrt{x} + \delta_A) - \big[\varphi(\sqrt{x} + \delta_B) + \varphi(-\sqrt{x} + \delta_B)\big]; \end{aligned}$$

protože hustota φ je sudá funkce, shodují se první čtyři sčítance se zbývajícími, takže můžeme rovnici po dělení dvěma přepsat do tvaru

$$f(x, \delta_A) - f(x, \delta_B) = 0, \tag{27}$$

kde $f(x, \delta) = \varphi(\sqrt{x} - \delta) + \varphi(\sqrt{x} + \delta)$. Rovnici (27) můžeme nyní upravit stejným způsobem, jako jsme upravili rovnici (21). Vydělíme ji rozdílem $\delta_A - \delta_B$, vezmeme limitu pro $\Delta \to 0_+$ a po úpravě dospějeme k rovnici

$$2\sqrt{x} = \ln \frac{\sqrt{x} + 1}{\sqrt{x} - 1},$$

jejímž řešením je číslo $q_\infty^{(2)}(0) \approx 1.4392$ (index ∞ spolu s argumentem nula označují dvojnásobně limitní případ $n \to \infty$, $\Delta \to 0_+$, horní index (2) připomíná, že jde o případ, který byl v předchozím nazván oboustranným).

Na závěr si uvedeme výsledky simulačních studií, které byly provedeny za účelem testování robustnosti našeho postupu ohledně volby regresního modelu. Úspěšnost našeho rozhodovacího postupu typu „t-testu“ zároveň porovnáme se standardními postupy:

(1) s t-testem,

(2) s minimalizací statistiky

$$\mathrm{PRESS} = \sum_{i=1}^{n} [y_i - \hat{y}_{(i)}]^2, \tag{28}$$

kde $\hat{y}_{(i)}$ je předpověď i-té hodnoty závisle proměnné podle modelu (1) nebo (2), získaná po vynechání i-tého pozorování x_i, y_i,

(3) s minimalizací reziduálního rozptylu.

Poslední postup je ekvivalentní volbě modelu s větší hodnotou upraveného koeficientu determinace (odtud označení `adj.R2` v násl. tabulce), který je klesající funkcí rezid. rozptylu, a v našem případě, kdy vynecháváme nebo

ponecháváme jediný regresor (sloupec) **Z**, je navíc ekvivalentní námi navrženému postupu s konstantou $q^{(2)} = 1$, je-li ovšem regresní matice konstantní, jak jsme až dosud předpokládali.

Násl. tabulka uvádí výsledky simulačních experimentů s regresní přímkou a alternativou konstantní „závislosti“ pro $n = 10$ ($f = 8$) a $n = 3$ ($f = 1$) s nezávisle proměnnou $x_i = -1 + 2(i-1)/(n-1)$, $i = 1, \ldots, n$. Ve sloupcích zleva doprava rostou odchylky od normality (rozdělení dvojitě exp., rovnoměrné, ve tvaru U a exponenciální; v posledním sloupci je zcela zvláštní případ, kdy kromě normálních chyb v závisle proměnné máme též normální šum v bodech x_i s rozptylem 0.01), v řádcích pak máme postupně přibližný 95 %-ní int. spolehlivosti pro optimální konstantu $q_f^{(2)}(0.5)$, dále odhad optimální konstanty (ovlivněný chybou diskretizace 0.1 a metodou Monte Carlo) a konečně odhady ztrátové funkce pro jednotlivé rozhodovací postupy — srov. též s analyt. výsl. (24) a (25). Poznamenejme, že směrodatné odchylky průměrných ztrát vycházely pro všechna uvažovaná kritéria a rozdělení chyb přibližně 1.8×10^{-4} pro $n = 10$ a 1.4×10^{-4} pro $n = 3$.

n = 10	normální	dvoj. exp.	rovnoměr.	U-tvar	exponenc.	šum v X
int. :	(1.1,1.7)	(1.4,2.1)	(1,1.5)	(1,1.3)	(1.7,2.4)	(1.2,2)
q	1.4	1.7	1.2	1.1	1.9	1.6
L=L(q)	0.4164	0.4094	0.4197	0.4214	0.4045	0.4254
t-test	0.4524	0.439	0.4592	0.4654	0.4293	0.4527
PRESS	0.4234	0.409	0.4333	0.4444	0.3997	0.4302
adj.R2	0.4184	0.4147	0.4205	0.4222	0.413	0.3549

n = 3	normální	dvoj. exp.	rovnoměr.	U-tvar	exponenc.	šum v X
int. :	(1.1,1.8)	(1.7,2.8)	(1.3,2.4)		(2.5,?)	(1.1,1.8)
q	1.4	2.2	1.8	161.4	2.8	1.4
L=L(q)	0.4417	0.4337	0.4532	0.4713	0.4321	0.4419
t-test	0.4894	0.4849	0.491	0.4713	0.4811	0.4894
PRESS	0.452	0.4386	0.4574	0.4894	0.4309	0.4526
adj.R2	0.4424	0.4381	0.4546	0.5187	0.4418	0.4418

Jak je vidět, pro $n = 10$ dostáváme ve všech sloupcích hodnotu optimální konstanty podstatně nižší než 5.318, což by odpovídalo t-testu. V normálním případě je optimální konstanta blízko 1.5, při větších odchylkách od normality se „stěhuje“ k 1 (viz rozdělení tvaru U) nebo naopak ke 2 (exponenciální rozdělení). Pro $n = 3$ se zdá, že optimální konstanty jsou větší než pro $n = 10$, v případě rozdělení tvaru U dokonce zvítězil t-test (konstanta 161.4). Ve sloupci exponenciálního rozdělení chyb je zřejmé vítězství statistiky (28) pro $n = 10$ i $n = 3$. V případě dvojitě exp. rozdělení dává pro $n = 10$ statistika PRESS prakticky stejnou ztrátovou funkci (rozdíl je vzhledem k výše uvedené směrodatné odchylce statisticky nevýznamný) jako optimální konstanta. Totéž lze říci o upraveném koeficientu determinace

v posledním sloupci pro $n = 3$; pro $n = 10$ je pak jeho úspěšnost skutečně překvapující.

Literatura

[1] Reif J. a Vlček K. (1998), *MSE-Improvement of Least Squares by Dropping Variables*. Článek v přípravě.
[2] Vlček K. (1996). *Linear Regression*. Sborník 4. konference studentů VŠTEZ, str. 83 – 94.