ODHAD BODU VZNIKU KVADRATICKÉHO TRENDU

DANIELA JARUŠKOVÁ

Abstrakt. The problem of least squares method estimation of the parameter τ^* in the regression model

$$Y_i = \beta^*(i/n - \tau^*)_+ + \gamma^*(i/n - \tau^*)_+^2 + e_i$$

is considered. Supposing $\gamma^* \neq 0$ and $\tau^* \in (0,1)$ it is shown that the distribution of $(\hat{\tau} - \tau^*)$ is largely affected by the value of β^*. In the case $\beta^* \neq 0$ the variable $\sqrt{n}\,(\hat{\tau} - \tau^*)$ is asymptotically normally distributed whereas in the case $\beta^* = 0$ the variable $\sqrt{n}\,(\hat{\tau} - \tau^*)^2$ has the same distribution as $\max(0, Z)$ where Z has a zero mean normal distribution.

The problem of least squares method estimation of the parameter τ^* in the regression model

$$Y_i = \beta^*(i/n - \tau^*)_+ + \gamma^*(i/n - \tau^*)_+^2 + e_i$$

is considered. Supposing $\gamma^* \neq 0$ and $\tau^* \in (0,1)$ it is shown that the distribution of $(\hat{\tau} - \tau^*)$ is largely affected by the value of β^*. In the case $\beta^* \neq 0$ the variable $\sqrt{n}\,(\hat{\tau} - \tau^*)$ is asymptotically normally distributed whereas in the case $\beta^* = 0$ the variable $\sqrt{n}\,(\hat{\tau} - \tau^*)^2$ has the same distribution as $\max(0, Z)$ where Z has a zero mean normal distribution.

Резюме: Уважается проблема оценивания параметра τ^* в модели линеинои регресии

$$Y_i = \beta^*(i/n - \tau^*)_+ + \gamma^*(i/n - \tau^*)_+^2 + e_i$$

при использовании метода наимењших квадратов. Показывается, что предпологая $\gamma^* \neq 0$ и $\tau^* \in (0,1)$, значение параметра β^* имеет бољшое влияние на распределение $(\hat{\tau} - \tau^*)$. В слушае когда $\beta^* \neq 0$, потом $(\hat{\tau} - \tau^*)$ имеет асимптотически нормаљное распределение. Наоборот, когда $\beta^* = 0$, потом $(\hat{\tau} - \tau^*)$ обладает темже паспределением как $(0, Z)$, кде Z следует нормаљное распределение с нуловым средним.

1. Úvod

V praxi občas narazíme na problém, kde se lineární závislost v neznámém časovém okamžiku změní v závislost kvadratickou. Sledujeme-li například u jistých typů slitin závislost napětí Y na zatížení X, ukazuje se, že při malém zatížení je zvyšování napětí lineární (elastická oblast), zatímco po překročení určitého kritického zatížení se mění kvadraticky (quasielastická oblast), tj. $Y = f(X)$, kde

$$f(x) = p + q\,x + \beta^*\,(x - \tau^*)_+ + \gamma^*\,(x - \tau^*)_+^2,$$

při označení $a_+ = \max(0, a)$.

V našem článku si poněkud zjednodušíme situaci tím, že budeme předpokládat, že parametry p a q jsou známé, a tedy je bez újmy na obecnosti můžeme položit rovny nule. Dále budeme předpokládat, že veličina X je měřena v equidistantních vzdálenostech Δ, $2\,\Delta, \ldots, n\,\Delta$, a tudíž ji lze transformovat na veličinu nabývající

2000 *Mathematics Subject Classification.* Primary 62F12.

Klíčová slova. Detekce bodu změny (change-point problem), nelineární regrese, odhady.

Práce byla částečně podporována granty GAČR 201/00/0769 a MSM 210000001.

hodnot $1/n$, $2/n, \ldots, 1$, to jest veličinu, která nabývá hodnot pouze v intervalu $[0, 1]$. Jestliže předpokládáme aditivní vliv náhodných chyb $\{e_i\}$ na naměřené hodnoty nezávisle proměnné Y, dospíváme k regresnímu modelu

$$Y_i = \beta^* \left(i/n - \tau^*\right)_+ + \gamma^* \left(i/n - \tau^*\right)_+^2 + e_i, \tag{1}$$

kde β^*, γ^* a τ^* jsou neznámé parametry.

Pro jednoduchost předpokládejme, že náhodné chyby $\{e_i\}$ jsou nezávislé stejně rozdělené s rozdělením $N(0, \sigma^2)$, kde σ^2 je známé, a tudíž opět bez újmy na obecnosti můžeme položit $\sigma^2 = 1$.

Úlohou, kterou se budeme zabývat, je odhad neznámých parametrů β^*, γ^* a τ^*. Především nás bude zajímat odhad parametru τ^*, který se nazývá bod změny. Budeme předpokládat, že $\tau^* \in (0, 1)$. Parametr β^* odpovídá první derivaci zprava a $2\,\gamma^*$ druhé derivaci zprava regresní funkce v bodě τ^*.

Vzhledem k normalitě chyb $\{e_i\}$ lze maximálně věrohodné odhady získat metodou nejmenších čtverců.

2. Bodový odhad parametrů

Model (1) je speciálním modelem nelineární regrese, kterému se někdy říká semilineární model, viz Knowles et all. (1991). Kdybychom totiž znali hodnotu τ parametru τ^*, tj. $\tau^* = \tau$, pak by model (1) byl modelem lineární regrese s dvěma vysvětlujícími proměnnými, tj. lineární model s maticí plánu experimentu

$$\boldsymbol{X}_n(\tau) = \begin{pmatrix} 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \\ \frac{[n\tau]+1}{n} - \tau & \left(\frac{[n\tau]+1}{n} - \tau\right)^2 \\ \frac{[n\tau]+2}{n} - \tau & \left(\frac{[n\tau]+1}{n} - \tau\right)^2 \\ \vdots & \vdots \\ 1 - \tau & (1-\tau)^2 \end{pmatrix}.$$

Označme

$$(\widetilde{\beta}_\tau, \widetilde{\gamma}_\tau)^T = \left(\boldsymbol{X}_n^T(\tau)\boldsymbol{X}_n(\tau)\right)^{-1} \boldsymbol{X}_n^T(\tau)\boldsymbol{Y},$$

kde $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$, a

$$S_n(\tau, \beta, \gamma) = \sum_{i=1}^{n} \left(Y_i - \left(\beta\big(i/n - \tau\big)_+ + \gamma\big(i/n - \tau\big)_+^2\right)\right)^2.$$

Pak

$$S_n(\widehat{\tau}, \widehat{\beta}, \widehat{\gamma}) = \min_{\beta,\gamma,\tau \in (0,1)} S_n(\tau, \beta, \gamma) = \min_{\tau \in (0,1)} S_n(\tau, \widetilde{\beta}_\tau, \widetilde{\gamma}_\tau).$$

Nejčastějším numerickým postupem pro nalezení přibližných hodnot odhadů $\widehat{\tau}$, $\widehat{\beta}$, $\widehat{\gamma}$ spočívá v tom, že pro hodnoty τ z dosti husté mříže bodů T v intervalu (0,1) počítáme residuální součet čtverců $S_n(\tau, \widetilde{\beta}_\tau, \widetilde{\gamma}_\tau)$ a pak hledáme $\min_{\tau \in T} S_n(\tau, \widetilde{\beta}_\tau, \widetilde{\gamma}_\tau)$.

Statistik se však obvykle nespokojí s bodovým odhadem neznámých parametrů, ale zajímá se též o intervaly spolehlivosti, respektive oblasti spolehlivosti.

3. Oblasti spolehlivosti v nelineární regresi

Uvažujme obecný model nelineární regrese s k - rozměrným vektorem parametrů $\boldsymbol{\theta^*} = (\theta_1^*, \dots, \theta_k^*)^T$

$$Y_i = f_i(\boldsymbol{\theta^*}) + e_i, \quad i = 1, \dots, n,$$

kde $\{f_i(\,\cdot\,)\}$ jsou známé funkce a $\{e_i\}$ jsou nezávislé stejně rozdělené náhodné veličiny s rozdělením $N(0,1)$.

Označme $S_n(\boldsymbol{\theta}) = \sum\big(Y_i - f_i(\boldsymbol{\theta})\big)^2$ a $\widehat{\boldsymbol{\theta}}$ odhad $\boldsymbol{\theta^*}$ metodou nejmenších čtverců, tj. $S_n(\widehat{\boldsymbol{\theta}}) = \min_{\theta_1,\dots,\theta_k} \sum\big(Y_i - f_i(\theta_1, \theta_2, \dots, \theta_k)\big)^2$. Dále označme $\big(\widetilde{\theta}_2(\theta_1), \dots, \widetilde{\theta}_k(\theta_1)\big)$ odhad metodou nejmenších čtverců při pevné hodnotě prvního parametru rovnajícího se θ_1, tj. $S_n\big(\theta_1, \widetilde{\theta}_2(\theta_1), \dots, \widetilde{\theta}_k(\theta_1)\big) = \min_{\theta_2,\dots,\theta_k} \sum\big(Y_i - f_i(\theta_1, \theta_2, \dots, \theta_k)\big)^2$.

Užívané konfidenční oblasti pro celý vektor parametrů $\boldsymbol{\theta}$ jsou obvykle dvou typů:

1 a) $\{\boldsymbol{\theta}, (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T\, \boldsymbol{A}\, (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) < C_1\}$, kde $\boldsymbol{A}$ je nějaká symetrická matice,
2 a) $\{\boldsymbol{\theta}, S_n(\boldsymbol{\theta}) < S_n(\widehat{\boldsymbol{\theta}}) + C_2\}$.

První oblast je eliptická, zatímco druhá může mít naprosto obecný tvar a může být dokonce nesouvislá. Druhé oblasti se někdy říká exaktní, neboť je odvozena přímo z poměru věrohodnosti.

Analogické konfidenční oblasti pro jeden parametr, řekněme θ_1, mají tvar:

1 b) $\{\theta_1, |\theta_1 - \widehat{\theta}_1| < K_1\}$,
2 b) $\{\theta_1, S_n\big(\theta_1, \widetilde{\theta}_2(\theta_1), \dots, \widetilde{\theta}_k(\theta_1)\big) < S_n(\widehat{\boldsymbol{\theta}}) + K_2\}$.

Konstanty C_1, C_2, resp. K_1, K_2, jsou obvykle odvozeny z asymptotického rozdělení $\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*$ a $S_n(\widehat{\boldsymbol{\theta}}) - S_n(\boldsymbol{\theta}^*)$, resp. $\widehat{\theta}_1 - \theta_1^*$ a $S_n(\widehat{\boldsymbol{\theta}}) - S_n\big(\theta_1^*, \widetilde{\theta}_2(\theta_1^*), \dots, \widetilde{\theta}_k(\theta_1^*)\big)$.

Je-li splněna celá řada podmínek, viz například podmínky A(1), …, A(9) nebo B(1), …, B(8) v knize Seber & Wild (1989), kapitola 12, pak $\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)$ má asymptoticky normální rozdělení. Většinou mezi tyto podmínky patří existence spojitých prvních i druhých parciálních derivací $\frac{\partial f_i(\boldsymbol{\theta})}{\partial \theta_r}$ a $\frac{\partial^2 f_i(\boldsymbol{\theta})}{\partial \theta_r \partial \theta_s}$, $i = 1, \dots, n$, $r, s = 1, \dots, k$, na nějakém okolí správné hodnoty $\boldsymbol{\theta}^*$. Navíc se předpokládá, že matice $\frac{1}{n}\big(\boldsymbol{F}_n(\boldsymbol{\theta})\big)^T\big(\boldsymbol{F}_n(\theta)\big)$, kde

$$\boldsymbol{F}_n(\boldsymbol{\theta}) = \begin{pmatrix} \frac{\partial f_1}{\partial \theta_1} & \dots & \frac{\partial f_1}{\partial \theta_k} \\ \vdots & & \vdots \\ \frac{\partial f_n}{\partial \theta_1} & \dots & \frac{\partial f_n}{\partial \theta_k} \end{pmatrix},$$

konverguje stejnoměrně na nějakém okolí $\boldsymbol{\theta}^*$ k nesingulární matici $\boldsymbol{G}(\boldsymbol{\theta})$. Matice $\boldsymbol{G}^{-1}(\boldsymbol{\theta}^*)$ je pak limitní varianční maticí vektoru $\sqrt{n}\,(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)$.

4. Oblasti spolehlivosti v modelu (1)

Je patrné, že v modelu (1) nemá regresní funkce vzhledem k parametru τ^* již ani první derivaci. Přesto se však dá ukázat, že v případě, že $\beta^* \neq 0$, $\gamma^* \neq 0$ a $\tau^* \in (0,1)$, má $\sqrt{n}\,(\widehat{\tau} - \tau^*, \widehat{\beta} - \beta^*, \widehat{\gamma} - \gamma^*)$ asymptoticky normální rozdělení se symetrickou varianční maticí $\boldsymbol{G}^{-1}$, kde

$$\boldsymbol{G} = \begin{pmatrix} \beta^{*2}(1-\tau^*) + 4\beta^*\gamma^*\dfrac{(1-\tau^*)^2}{2} + 4\gamma^{*2}\dfrac{(1-\tau^*)^3}{3} & \cdots & \cdots \\ -\beta^*\dfrac{(1-\tau^*)^2}{2} - 2\gamma^*\dfrac{(1-\tau^*)^3}{3} & \dfrac{(1-\tau^*)^3}{3} & \cdots \\ -\beta^*\dfrac{(1-\tau^*)^3}{3} - 2\gamma^*\dfrac{(1-\tau^*)^4}{4} & \dfrac{(1-\tau^*)^4}{4} & \dfrac{(1-\tau^*)^5}{5} \end{pmatrix}$$

Zřejmě $\boldsymbol{G} = \lim_{n\to\infty} \frac{1}{n} \boldsymbol{F}_n^{*\,T} \boldsymbol{F}_n^*$, kde

$$\boldsymbol{F}_n^* = \begin{pmatrix} 0 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 0 \\ -\beta^* - 2\gamma^*\left(\frac{[n\tau^*]+1}{n} - \tau^*\right) & \frac{[n\tau^*]+1}{n} - \tau^* & \left(\frac{[n\tau^*]+1}{n} - \tau^*\right)^2 \\ \vdots & \vdots & \vdots \\ -\beta^* - 2\gamma^*(1-\tau^*) & 1-\tau^* & (1-\tau^*)^2 \end{pmatrix}$$

K důkazu lze použít stejný postup, který použila Hušková (1998) pro jednodušší model typu

$$Y_i = \beta^* \left(i/n - \tau^*\right)_+ + e_i, \tag{2}$$

a který spočívá v tom, že na okolí správné hodnoty aproximujeme funkci nejmenších čtverců kvadratickou funkcí. Speciálně ukázala, že na okolí bodu τ^* lze aproximovat

$$S_n(\tau, \widetilde{\beta}_\tau) - S_n(\tau^*, \widetilde{\beta}_{\tau^*})$$

funkcí

$$-C\,n\,(\tau - \tau^*)^2 + 2\,X\,\sqrt{n}\,(\tau - \tau^*),$$

kde X/C má normální rozdělení $N(0, 1/C)$.

Jiný postup použil Feder (1975), který modely (1) i (2) považuje za speciální případy regresní funkce

$$\begin{aligned} f(x, \boldsymbol{\theta}) &= f_1(x, \boldsymbol{\theta}_1) \quad \text{pro} \quad 0 \le x \le \tau^* \\ &= f_2(x, \boldsymbol{\theta}_2) \quad \text{pro} \quad \tau^* \le x \le 1, \end{aligned}$$

kde $f_i(x, \boldsymbol{\theta}) = \sum_{j=1}^{K(i)} \theta_{ij} f_{ij}(x)$. Za přípustnou množinu parametrů Θ pak uvažuje jen takové parametry, při kterých se funkce $f_1(x, \boldsymbol{\theta}_1)$ a $f_2(x, \boldsymbol{\theta}_2)$ protínají uvnitř intervalu (0,1). Ukazuje, že limitní rozdělení $\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*$ závisí na tom, zda $\boldsymbol{\theta}^*$ je vnitřním nebo krajním bodem množiny Θ. Parametr τ^* pak odhaduje jako průsečík funkcí $f_1(x, \widehat{\boldsymbol{\theta}}_1)$ a $f_2(x, \widehat{\boldsymbol{\theta}}_2)$. Řád konvergence $\widehat{\tau}$ k τ^* závisí na tom, v kolika derivacích se shodují funkce $f_1(x, \boldsymbol{\theta}^*)$ a $f_2(x, \boldsymbol{\theta}^*)$ v bodě τ^*.

V případě modelu (1) s $\beta^* \neq 0$, $\gamma^* \neq 0$ a $\tau^* \in (0, 1)$ lze z přístupů Huškové (1998) i Federa (1975) odvodit asymptotickou normalitu $(\widehat{\tau} - \tau^*, \widehat{\beta} - \beta^*, \widehat{\gamma} - \gamma^*)$. Speciálně platí

$$\sqrt{n}\,(\widehat{\tau} - \tau^*) \sim N\left(0,\, \frac{9}{\beta^{*2}(1-\tau^*)}\right) \tag{3}$$

a $S_n(\tau^*, \widetilde{\beta}_{\tau^*}, \widetilde{\gamma}_{\tau^*}) - S_n(\widehat{\tau}, \widehat{\beta}, \widehat{\gamma})$ má asymptoticky χ^2 rozdělení o 1 stupni volnosti.

Všimněme si, že asymptotický rozptyl odhadu $\widehat{\tau}$, viz (3), záleží na poloze τ^*, což je jakási globální vlastnost regresní funkce. Je přirozené, že parametr τ^* lépe

odhadneme, máme-li možnost pozorovat kvadratický trend delší dobu. Dále však je asymptotický rozptyl $\widehat{\tau}$, a tedy i délka intervalu spolehlivosti typu 2 a), ovlivněna první derivací kvadratické funkce v bodě τ^*, což je lokální vlastnost regresní funkce.

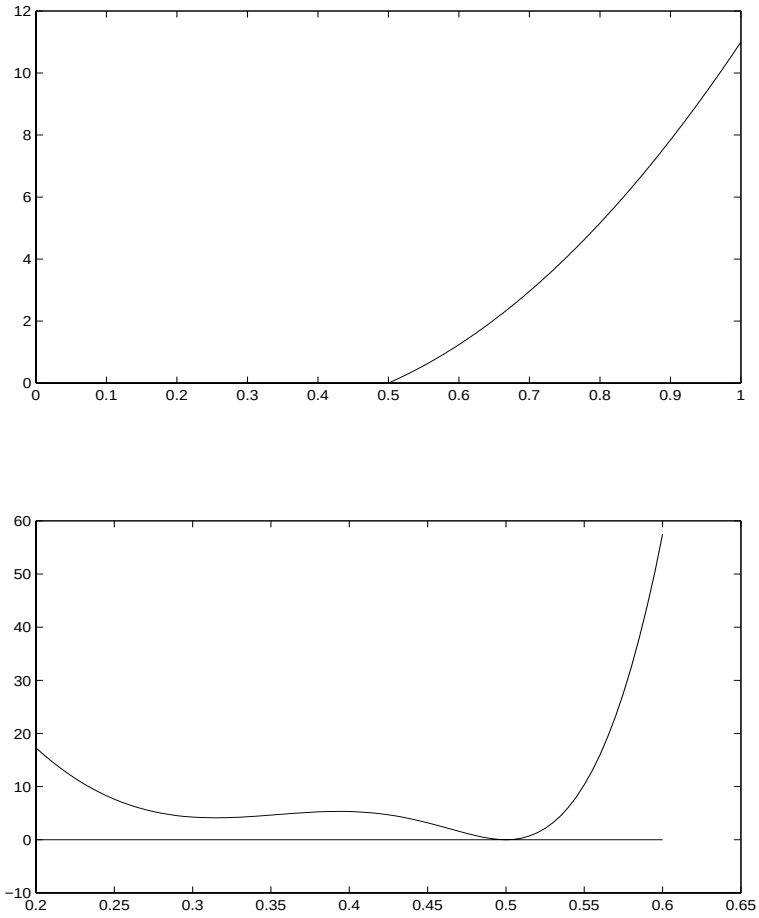
Je-li β^* malé, můžeme (použijeme-li postup 1 b)) dostat pro konečná n dokonce interval spolehlivosti, jehož krajní body leží mimo interval $(0, 1)$. V teorii nelineární regrese se tento jev nazývá špatná podmíněnost. Špatná podmíněnost je způsobena tvarem regresní funkce.

Horní část obrázku 1 představuje regresní funkci

$$f(x, \tau^*, \beta^*, \gamma^*) = \beta^*(x - \tau^*)_+ + \gamma^*(x - \tau^*)_+^2 \tag{4}$$

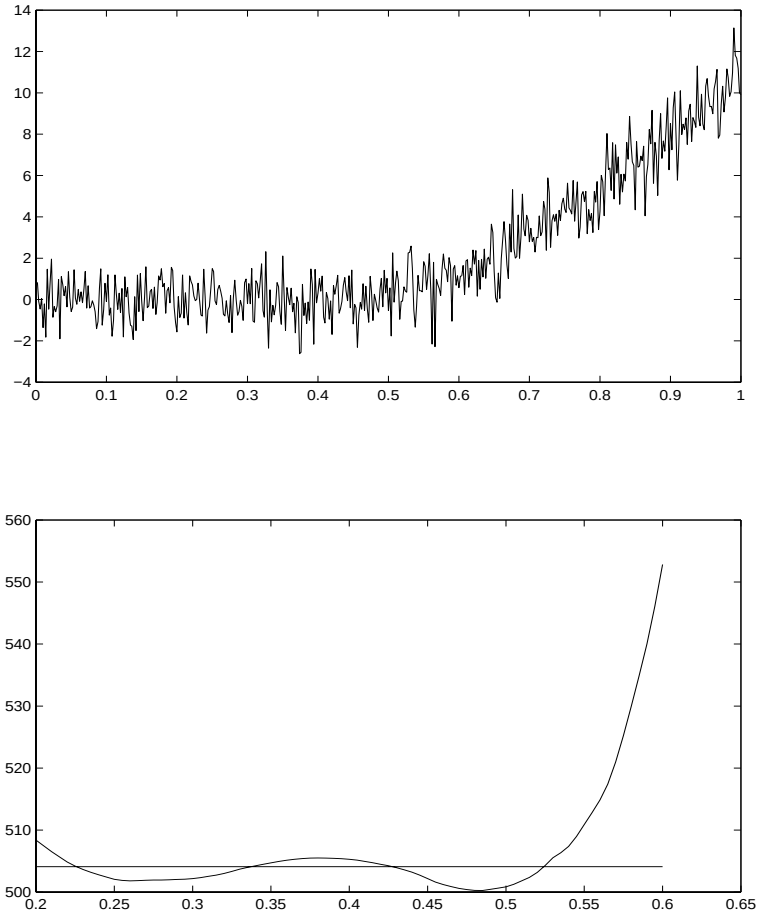
pro $\tau^* = 0.5$, $\beta^* = 10$ a $\gamma^* = 24$ a dolní část obrázku zobrazuje odpovídající funkci nejmenších čtverců $B_n(\tau, \tau^*)$ pro $n = 500$ a $\tau \in (0.2, 0.6)$ za předpokladu, že model neobsahuje žádné chyby, tj.

$$B_n(\tau, \tau^*) = \frac{1}{n} \sum_{i=1}^{n} \left(f\Big(\frac{i}{n}, \tau, \widetilde{\beta}_\tau, \widetilde{\gamma}_\tau\Big) - f\Big(\frac{i}{n}, \tau^*, \widetilde{\beta}_{\tau^*}, \widetilde{\gamma}_{\tau^*}\Big) \right)^2.$$



Obrázek 1

Jestliže uvažujeme model (1) s regresní funkcí (4) včetně náhodných chyb, pak může funkce nejmenších čtverců a 95% oblast spolehlivosti typu 2 b) vypadat například jako na obrázku 2.



Obrázek 2

Ačkoliv k tomu nemáme žádný teoretický důvod doporučujeme ze zkušeností používat spíše oblast spolehlivosti typu 2 b), kde K_2 je příslušný kvantil χ^2 - rozdělení o 1 stupni volnosti. V každém případě doporučujeme vždy při statistické analýze vykreslit průběh funkce $S_n(\tau, \widetilde{\beta}_\tau, \widetilde{\gamma}_\tau)$.

V krajním případě, kde $\beta^* = 0$, $\gamma^* \neq 0$ a $\tau^* \in (0,1)$, lze ukázat, že v okolí bodu τ^* lze rozdíl funkcí čtverců

$$S_n(\tau, \widetilde{\beta}_\tau, \widetilde{\gamma}_\tau) \; - \; S_n(\tau^*, \widetilde{\beta}_{\tau^*}, \widetilde{\gamma}_{\tau^*})$$

aproximovat funkcí

$$-C\,n\,(\tau - \tau^*)^4 + 2\,X\,\sqrt{n}\,(\tau - \tau^*)^2,$$

kde $X/C \sim N(0, 1/C)$ a $C = \gamma^{*2}\,(1 - \tau^*)/9$. Je zřejmé, že funkce

$$-Cx^2(x^2 - 2\,X/C)$$

nabývá maxima v nule, pokud X je záporné, a hodnoty X/C, pokud X je kladné. Odtud vyplývá, že $\sqrt{n}\,(\widehat{\tau} - \tau^*)^2$ má asymptoticky stejné rozdělení jako $\max(0, Z)$, kde veličina Z má normální rozdělení

$$Z \sim N\big(0, \frac{9}{\gamma^{*2}(1 - \tau^*)}\big). \tag{5}$$

Výsledek (5) lze použít pro konstrukci symetrického intervalu spolehlivosti 2 a). Můžeme též zkonstruovat oblast typu 2 b) s využitím toho, že $S_n(\tau^*, \widetilde{\beta}_{\tau^*}, \widetilde{\gamma}_{\tau^*}) - S_n(\widehat{\tau}, \widehat{\beta}, \widehat{\gamma})$ má asymptoticky stejné rozdělení jako $(\max(0, Y))^2$, kde Y má standardní normální rozdělení.

Pro zajímavost uveďme, že pokud odhadujeme parametr τ^* v modelu

$$Y_i = \gamma^*(i/n - \tau^*)_+^2 + e_i, \quad i = 1, \ldots, n,$$

pak $\sqrt{n}\,(\widehat{\tau} - \tau^*)$ má asymptoticky normální rozdělení $N\Big(0, 12/\big(\gamma^{*2}(1-\tau^*)^3\big)\Big)$. Odtud je zřejmé, že informace, zda-li je první derivace v bodě změny nulová, je při odhadování tohoto bodu velmi velmi důležitá.

Literatura

Feder P. J. (1975). *On asymptotic distribution theory in segmented regression problems - identified case.* The Annals of Statistics **3**, 49 – 63.

Hušková M. (1998). *Estimators in the location model with gradual changes.* Comment. Math. Univ. Carolinae **39**, 147 – 157.

Knowles M., Siegmund D. a Zhang H. (1991). *Confidence regions in semilinear regression.* Biometrika **78**, 1991, 15 – 31.

Seber G. A. F. a Wild C. J. (1989). *Nonlinear regression.* J. Wiley, New York.

ČVUT v Praze, Stavební fakulta, Katedra matematiky, Thákurova 7, 166 29 Praha 6
E-mail: jarus@mat.fsv.cvut.cz