Použití technik Jackknife a Bootstrap při odhadu modelových parametrů metodami nelineární regrese.

Jiří Militký, Výzkumný ústav zušlechťovací
544 28 Dvůr Králové n.L.

1. Úvod

Odhad parametrů v nelineárních regresních modelech patří mezi základní úlohy řešené v mnoha oblastech vědy a techniky.
V naprosté většině případů se používá metody nejmenších čtverců (MNČ) v nejjednodušším tvaru (bez statistických vah měření).
Tato metoda však často vede, zejména pro malé výběry (malý počet experimentálních bodů), k vychýleným odhadům. Důvody jsou v tom, že:

a) regresní model nevystihuje přesně chování sledovaného systému (je více či méně zdařilou aproximací jisté neznámé funkční závislosti)

b) jednotlivá data ve výběru se neřídí předpoklady, za nichž byla MNČ odvozena (jde zejména heteroskedasticitu a porušení podmínek normality)

c) kromě náhodných chyb měření se v datech vyskytují také hrubé odchylky (jejichž zdrojem může být i vnitřní variabilita sledovaného systému).

Pro odhad parametrů lze v těchto případech použít robustní metody (numerické postupy pro nelineární modely uvádí např. Dennis a Welch /1/).
V poslední době se v literatuře objevují práce pojednávající o použití technik snižujících vychýlenost bodových odhadů (metody typu Jackknife a Bootstrap) pro lineární i nelineární regresní modely /14-16/.
V tomto příspěvku je uveden přehled základních variant technik Jackknife a Bootstrap. Je popsán program umožňující stanovení regresních odhadů metodou nejmenších čtverců a pomocí těchto technik. Na jednom modelu (se simulovanými daty) je provedeno porovnání odhadů získaných různými postupy s ohledem na jejich vychýlení.

2. Snižování vychýlenosti bodových odhadů.

Mějme náhodnou veličinu X, která má frekvenční funkci $f(x)=f(x,\Theta)$. Účelem je nalézt odhad $\hat{\Theta}$ neznámého parametru Θ na základě výběru sestaveného z N realizací $(x_1,\dots x_N)$ této náhodné veličiny. Odhad $\hat{\Theta}$ se počítá podle jisté funkce $\hat{\Theta}=S(x_1,\dots x_N)$. Pokud je $\hat{\Theta}$ vychýlený, platí, že

$$E(\hat{\Theta}-\Theta) = b(\hat{\Theta}) \neq 0 \qquad (1)$$

kde $E(\cdot)$ je operátor matematického očekávání a $b(\hat{\Theta})$ je vychýlení, které je funkcí také velikosti výběru N. Obecně lze jednotlivým pozorováním ve výběru přiřadit kladné váhy $(w_1,\dots w_N)$. Základní problém je v tom, že často neznáme přesný tvar $f(x)$. Na základě výběru však můžeme definovat empirickou frekvenční funkci

$$g(x)=\left(1/\sum_{(i)} w_i\right)\cdot\sum_{(i)} w_i\cdot\delta(x-x_i) \qquad (2)$$

$$i=1,\dots N$$

kde $\delta(x-x_i)$ je Diracova funkce.
Parametr $\Theta=S(f(x))$ pak odhadujeme pomocí funkce $S(g(x))=\hat{\Theta}=S(x_1,\dots x_N,w_1,\dots w_N)$
Je zřejmé, že čím bude $g(x)$ blíže k $f(x)$ (např. ve smyslu Prochorovovy metriky /6/), tím se bude více $\hat{\Theta}\to\Theta$.
Pro stanovení vychýlení $b(\hat{\Theta})$ uvažujme frekvenční funkci $h(x)$, která je dána vztahem

$$h(x)=(1-t)f(x)+t\,g(x) \qquad (3)$$

Odhad $S(h(x))$ založený na této frekvenční funkci lze vyjádřit Taylorovým rozvojem /6/

$$S(h(x))=S(f(x))+t A_1+t^2/2\,A_2+\dots \qquad (4)$$

kde A_1, A_2 jsou derivace $S(h(x))$. Např. parametr $A_1=\int\psi(x)g(x)dx$ obsahuje tzv. von Miessovu derivaci prvního řádu $\psi(x)$.
Pokud je $h(x)$ blízké $f(x)$, lze pro $t=1$ vyjádřit rozvoj (1) ve tvaru

$$S(g(x))=\Theta+\int\psi(x)g(x)dx+\frac{1}{2}\iint\psi(x,y)g(x)g(y)dx\cdot dy+\cdots \qquad (5)$$

Dosazením za $g(x)$ z rov. (2) a úpravami vyjde

$$\hat{\Theta}=\Theta+\frac{\sum_{(i)} w_i\psi(x_i)}{\sum_{(i)} w_i}+\frac{1}{2}\,\frac{\sum_{(i)}\sum_{(j)} w_i w_j\psi(x_i,x_j)}{\left(\sum_{(i)} w_i\right)^2}+\dots \qquad (6)$$

Poznámka: Funkce $\psi(x_i)$ vyjadřuje vliv hodnoty x_i na odhad $\hat{\Theta}$. Proto ji Hampel /9/ nazval vlivová funkce. Jde vlastně o první von Miessovu derivaci ve směru Diracovy funkce. Lze dokázat, že $\psi(x_i) = \lim_{t \to 0} \{[S(g(x)) - S(f(x))]/t\}$.

Z rov. (6) je patrné, že vychýlení $b(\hat{\Theta})$ lze pro případ, kdy $u_i = 1, i = 1..N$ vyjádřit vztahem

$$b(\hat{\Theta}) = \frac{a_1}{N} + \frac{a_2}{N^2} + \dots \tag{7}$$

V roce 1956 navrhl Quenouille /2/ jednoduchý způsob snížení možné vychýlenosti bodových odhadů, který vychází z rozvoje (7). Vyšel přitom z odhadu $\hat{\Theta}_i = S(x_1, \dots x_{i-1}, x_{i+1}, \dots x_N)$, který získal po vynechání i-tého pozorování x_i. Jeho vychýlení lze vyjádřit tvarem

$$b(\hat{\Theta}_i) = \frac{a_1}{(N-1)} + \frac{a_2}{(N-1)^2} + \dots \tag{8}$$

Snadno pak dokázal, že odhad

$$\tilde{\Theta}_i = N\hat{\Theta} - (N-1)\hat{\Theta}_i \tag{9}$$

je méně vychýlený, protože již neobsahuje první člen rozvoje. Vychýlení $b(\tilde{\Theta}_i)$ je dáno vztahem

$$b(\tilde{\Theta}_i) = -\frac{a_2}{N(N-1)} + \dots \approx -\frac{a_2}{N^2} + O(N^{-3}) \tag{10}$$

Postupným opakováním této techniky lze eliminovat i další členy rozvoje (7). Aby nedošlo ke ztrátě eficience odhadu, je nutné nahradit $\tilde{\Theta}_i$ průměrnou hodnotou $\tilde{\Theta}_J$.

$$\tilde{\Theta}_J = \frac{1}{N} \sum_{i=1}^{N} \tilde{\Theta}_i \tag{11}$$

Tukey označil $\tilde{\Theta}_i$ jako pseudohodnoty a $\tilde{\Theta}_J$ jako Jackknife odhad. V abstraktu /4/ předpokládal, že $\tilde{\Theta}_i$ jsou přibližně nezávislé náhodné veličiny, pro které lze stanovit výběrový rozptyl

$$var(\tilde{\Theta}_J) = \frac{1}{N(N-1)} \sum_{i=1}^{N} (\tilde{\Theta}_i - \tilde{\Theta}_J)^2 \tag{12}$$

Pro odhad konfidenčních intervalů lze použít statistiky

$$\sqrt{N}(\tilde{\Theta}_J - \Theta)\left\{\frac{1}{N-1} \sum_{(i)} (\tilde{\Theta}_i - \tilde{\Theta}_J)^2\right\}^{-1/2} \tag{13}$$

která má přibližně t-rozdělení s N-1 stupni volnosti.

Poznámka: Miller /8/ uvádí některé příklady, které ukazují, že odhad konfindenčních mezí podle rov. (13) nemusí být vždy lepší než obvyklé odhady (založené např. na předpokladu normality).

Rey /6/ navrhl zobecnění techniky Jackknife na případ, kdy váhy měření w_i jsou nekonstantní. Místo vynechávání bodů provádí úpravu w_i pomocí faktoru (1+t). Odhad $\hat{\Theta}_i$ se pak určí ze vztahu

$$\hat{\Theta}_i(t) = f(x_1, \ldots x_N, w_1, \ldots w_{i-1}, (1+t)w_i, w_{i+1} \ldots w_N) \tag{14}$$

Odpovídající pseudohodnoty $\tilde{\Theta}_{w_i}$ jsou definovány vztahem

$$\tilde{\Theta}_{w_i} = \left[\left(tw_i + \sum_{j=1}^{N} w_j\right)\hat{\Theta}_i(t) - \left(\sum_{j=1}^{N} w_j\right)\hat{\Theta}\right] / t \tag{15}$$

Jackknife odhad $\tilde{\Theta}_{wJ}$ je pak vážený průměr

$$\tilde{\Theta}_{wJ} = \left(1 / \sum_{(i)} w_i\right) \sum_{(i)} \tilde{\Theta}_{w_i} \tag{15a}$$

Odpovídající vychýlení $b(\tilde{\Theta}_{wJ})$ má tvar

$$b(\tilde{\Theta}_{wJ}) = \frac{\sum w_i \varphi_i}{\sum w_i} + \frac{1}{2}(1+t)\frac{\sum\sum w_i w_j \varphi(x_i, x_j)}{(\sum w_i)^2} + \ldots \tag{16}$$

Volbou t=-1 vypadne v rov. (16) druhý člen. Odhad $\tilde{\Theta}_{wJ}$ je pak váženou variantou původní Jackknife techniky. Pro odhad rozptylu platí

$$var(\tilde{\Theta}_{wJ}) = \sum_{(i)}\left(\tilde{\Theta}_{w_i} - w_i\tilde{\Theta}_{wJ}\right)^2 / \left[\left(\sum_{(i)} w_i\right)^2 - \sum_{(i)} w_i^2\right] \tag{17}$$

Pro malá záporná t lze vyjádřit odhad vychýlení $\hat{\Theta}$ vůči $\tilde{\Theta}_{wJ}$ vztahem /7/

$$\hat{b}(\hat{\Theta}, t) = (\hat{\Theta} - \tilde{\Theta}_{wJ}) = \frac{1-t}{Nt^2}\left\{\frac{1}{N}\sum \hat{\Theta}_i(t) - \hat{\Theta}\right\} \tag{18}$$

Jaeckel (viz /7/) navrhl tzv. infinitezimální Jackknife odhad $\tilde{\Theta}_{IJ}$ ve tvaru

$$\tilde{\Theta}_{IJ} = \hat{\Theta} - \lim_{t \to 0} \hat{b}(\hat{\Theta}, t) \tag{18a}$$

Rozšíření techniky váženého Jackknife odhadu na případ, kdy odhadované parametry tvoří vektor, je publikováno v práci /6/.

Schucany, Gray a Owen /3/ stanovili, že ze dvou vychýlených odhadů $\hat{\Theta}_1$ a $\hat{\Theta}_2$ lze vypočítat nevychýlený odhad $\tilde{\Theta}_S$ podle vztahu

$$\tilde{\Theta}_S = (\hat{\Theta}_1 - R\hat{\Theta}_2)(1-R)^{-1} \tag{19}$$

kde $R = b(\hat{\Theta}_1)/b(\hat{\Theta}_2)$ je poměr vychýlení odhadů. Lze snadno dokázat, že pro $\hat{\Theta}_1 = \hat{\Theta}$, $\hat{\Theta}_2 = \sum \hat{\Theta}_i / N$ a $R = (N-1)/N$ je $\tilde{\Theta}_S = \tilde{\Theta}_J$.
V téže práci je navržen postup pro tvorbu odhadů, jejichž vychýlení $b(\tilde{\Theta}_S)$ neobsahuje prvních k členů v mocninném rozvoji /7/. Dalšího snížení vychýlenosti odhadů lze docílit rozdělením výběru do g skupin o velikosti m $(N = m g)$ a vylučováním celých skupin při výpočtu pseudohodnot. Pak

$$\hat{\Theta}_j = S(x_1 \ldots x_{(j-1)m}, x_{jm+1}, \ldots x_N) \qquad j = 1 \ldots g \tag{20}$$

$$\tilde{\Theta}_{GJ} = g\hat{\Theta} - \frac{g-1}{g}\sum_{j=1}^{g}\hat{\Theta}_j$$

Aziz /5/ navrhl vylučovat všechny skupiny kromě ℓ-té. Pak

$$\hat{\Theta}_\ell = S(x_{(\ell-1)n+1} \ldots x_{\ell n}) \qquad \ell = 1 \ldots g \tag{21}$$

$$\tilde{\Theta}_{AJ} = \frac{g}{g-1}\hat{\Theta} - \frac{1}{g(g-1)}\sum_{\ell=1}^{g}\hat{\Theta}_\ell$$

Jak je zřejmé, je odhad $\tilde{\Theta}_{AJ}$ výpočetně nesložitý (zejména pro velké N).
V práci /5/ je dokázáno, že pro všechna reálná $p > 0$ taková, že

$$0 < \lim_{N \to \infty} N^p \, b(\hat{\Theta}) < \infty$$

je $b(\tilde{\Theta}_{AJ}) < b(\hat{\Theta})$ jen v případě, že $p < \ln(2g-1)/\ln g$.
Za těchže podmínek je však $b(\tilde{\Theta}_{AJ}) > b(\tilde{\Theta}_{GJ})$ (pro $p \neq 1$).
Aziz /5/ také navrhl, jak určit optimální g s ohledem na minimalizaci střední kvadratické odchylky odhadů.
Další možností je odstranění n prvků ve výběru všemi $\binom{N}{N-n}$ možnými způsoby.
Výhody a nevýhody jednotlivých způsobů dělení výběru do skupin diskutuje např. Rey /6/.
Jak je zřejmé z výše uvedeného, umožňují techniky Jackknife snížení vychýlenosti odhadů a stanovení jejich rozptylů bez detailních znalostí o výběrové frekvenční funkci. Podmínkou je však splnění předpokladů, za nichž byla odvozena rov. (7).

Velmi blízká technikám Jackknife je metoda Bootstrap /10/. Postup při jejím použití lze vyjádřit posloupností těchto kroků /11/

I. Vytvoření empirické výběrové frekvenční funkce $g(x)$, např. podle rov. (2) položením $n_i = 1 \; (i = 1 \ldots N)$

II. Sestavení náhodného výběru o velikosti N z $g(x)$, tedy tzv. Bootstrap výběru $(x_{1j}^*, x_{ij}^*, x_{Nj}^*)$

III. Stanovení odhadu $\tilde{\Theta}_{Bj} = \int (x_{1j}^*, \ldots x_{Nj}^*)$

IV. Opakování kroků II a III pro $j = 1 \ldots M$ a pak výpočet Bootstrap odhadu

$$\tilde{\Theta}_B = \frac{1}{M} \sum_{j=1}^{M} \tilde{\Theta}_{Bj} \tag{22}$$

Na rozdíl od Jackknife se náhodné výběry tvoří z množiny $(x_1 \ldots x_N)$ s vracením (tj. některá x_i se v Bootstrap výběru vyskytují víckrát a některá vůbec ne). Vztah mezi $\tilde{\Theta}_B$ a $\tilde{\Theta}_J$ je diskutován v práci /10/.

3. Odhad parametrů v regresních modelech.

Mějme modelovou funkci $\varphi(\bar{x}, \bar{\Theta})$ a výběrová data (experimentální body) $(x_i, y_i) \; i = 1 \ldots N$. Účelem je nalézt odhady $\hat{\bar{\Theta}}$ modelových parametrů $\bar{\Theta}$. V nejjednodušším případě lze uvažovat model měření ve tvaru

$$y_i = \varphi(\bar{x}_i, \bar{\Theta}) + \varepsilon_i \tag{23}$$

kde ε_i jsou náhodné chyby, o nichž se obvykle předpokládá, že mají stejné rozdělení s frekvenční funkcí $f(\varepsilon)$, a že

$$E(\varepsilon_i) = 0, \quad E(\varepsilon_i, \varepsilon_j) = 0 \qquad i \neq j = 1 \ldots N$$

Odhad parametrů $\bar{\Theta}$ lze pro známé $f(\varepsilon)$ získat maximalizací věrohodnostní funkce

$$\hat{\bar{\Theta}} = \max_{\bar{\Theta} \in E^p} \prod_{i=1}^{N} f(\varepsilon_i) \tag{24}$$

Pokud $\varepsilon_i \in N(0, \sigma^2)$, vede řešení problému (24) na metodu nejmenších čtverců odchylek, kdy

$$\hat{\bar{\Theta}} = \min_{\bar{\Theta} \in E^p} \sum_{i=1}^{N} [y_i - \varphi(x_i, \bar{\Theta})]^2 \tag{25}$$

Lze dokázat, že normální rozdělení je stabilní (definice stabilnosti viz /12/) ve třídě všech hustot s ohraničeným rozptylem.
Na druhé straně však bylo stanoveno, že MNČ poskytuje v případě, že $f(\mathcal{E})$ je Laplaceovo oboustranné rozdělení, odhady s dvojnásobným rozptylem než metoda nejmenších absolutních odchylek (L_1 aproximace) /13/.
Je také známo, že $\hat{\bar{\Theta}}$ určené MNČ jsou citlivé na výskyt hrubých chyb a porušení předpokladu homoskedasticity.
Obvykle používané robustní metody regrese nahrazují součet čtverců v rov. (25) jinou, méně rychle rostoucí funkcí. To vede zejména u nelineárních modelů ke značným numerickým obtížím.

Při použití technik Jackknife a Bootstrap se vychází z opakovaného řešení rov. (25) pro různě modifikovaná data.
První práci týkající se aplikace Jackknife metody při odhadu parametrů v lineárních regresních modelech publikoval Miller /14/.
Ten vyšel z modelu

$$\bar{Y} = X\bar{\Theta} + \bar{\mathcal{E}} \tag{25a}$$

kde $\bar{Y} = (y_1 \ldots y_N)^T$, $\bar{\Theta} = (\Theta_1 \ldots \Theta_p)^T$, $\bar{\mathcal{E}}$ je chybový vektor a X je matice ($N \times p$) obsahující jako řádky vektory $\bar{x}_i = (x_{1,i}, \ldots x_{p,i})^T$. Metodou nejmenších čtverců lze určit $\hat{\bar{\Theta}}$ podle výrazu

$$\hat{\bar{\Theta}} = (X^T X)^{-1} X^T \bar{Y} \tag{26}$$

Odhad $\hat{\bar{\Theta}}_i$ vzniklý vyloučením bodu $(y_i, \bar{x}_i^T)$ je možno vypočítat podle relace (viz /14/)

$$\hat{\bar{\Theta}}_i = \hat{\bar{\Theta}} - \frac{(X^T X)^{-1} \bar{x}_i^T (y_i - \bar{x}_i \hat{\bar{\Theta}})}{1 - \bar{x}_i (X^T X)^{-1} \bar{x}_i^T} \tag{27}$$

Rov. (27) tedy umožňuje snadný výpočet $\hat{\bar{\Theta}}_i$ při znalosti $\hat{\bar{\Theta}}$ z MNČ. Z odhadu $\hat{\bar{\Theta}}_i$ se dle rov. (9) snadno určí pseudohodnoty $\tilde{\bar{\Theta}}_{LJ}$ a z nich pomocí rov. (11) linearizovaný Jackknife odhad $\tilde{\Theta}_{LJ}$ (pro lineární regresní modely je $\tilde{\Theta}_{LJ} = \tilde{\Theta}_J$). Hinkley /15/ nazývá tento postup jako rovnovážný Jackknife.

Pro případ, že jsou "váhy" $\bar{x}_i(X^T X)^{-1}\bar{x}_i^T$ v jednotlivých bodech "regresního plánu" X nestejné, navrhuje tzv. nerovnovážný Jackknife, pro který lze pseudohodnoty $\tilde{\Theta}_{Ni}$ vyjádřit tvarem

$$\tilde{\bar{\Theta}}_{Ni} = \hat{\bar{\Theta}} + N(X^T X)^{-1}\bar{x}_i(y_i - \bar{x}_i\hat{\bar{\Theta}}) \tag{28}$$

Odpovídající Jackknife odhad $\tilde{\bar{\Theta}}_{NJ}$ však nesnižuje pro lineární modely žádné vychýlení, protože platí

$$\tilde{\bar{\Theta}}_{NJ} = \frac{1}{N}\sum_{i=1}^{N}\tilde{\bar{\Theta}}_{Ni} = \hat{\bar{\Theta}} \tag{29}$$

Použití nerovnovážného Jackknife odhadu je tedy vhodné pouze při výpočtu rozptylu odhadů podle rov. (12).

Miller /14/ dokázal, že Jackknife odhad $\tilde{\bar{\Theta}}_J$ funkce $\Theta = \varphi(\bar{\beta})$ regresních parametrů $\bar{\beta}$ v lineárním modelu $\bar{Y} = X\bar{\beta} + \bar{\mathcal{E}}$ je asymptoticky normální i pro případ, že $\mathcal{E}_i$ jsou negaussovské náhodné veličiny.

Pro nelineární regresní modely $\varphi(\bar{x}, \bar{\Theta})$ se obyčejně využívá linearizace pomocí Taylorova rozvoje /14, 16/

$$\bar{\Theta} - \hat{\bar{\Theta}} = (Z^T Z)^{-1} Z^T \bar{r} \tag{30}$$

kde $\bar{r} = (y_1 - \varphi(\bar{x}_1, \hat{\bar{\Theta}}), \ldots y_i - \varphi(\bar{x}_i, \hat{\bar{\Theta}}), \ldots y_N - \varphi(\bar{x}_N, \hat{\bar{\Theta}}))$

a Z je Jakobián, jehož řádky tvoří vektory

$$\bar{z}_i^T = \left(\frac{\partial}{\partial\Theta_1}\varphi(\bar{x}_i, \hat{\bar{\Theta}}), \ldots \frac{\partial}{\partial\Theta_j}\varphi(\bar{x}_i, \hat{\bar{\Theta}}), \ldots \frac{\partial}{\partial\Theta_p}\varphi(\bar{x}_i, \hat{\bar{\Theta}})\right)$$

Nahradíme-li v rov. (27) a (28) matici X maticí Z, a vektory $\bar{x}_i$ pomocí vektorů $\bar{z}_i$, lze snadno určit odhady $\tilde{\bar{\Theta}}_{LJ}$ i $\tilde{\bar{\Theta}}_{NJ}$, aniž je nutno řešit (n+1)krát iterativně rov. (25) (pro výpočet $\tilde{\bar{\Theta}}_J$). I pro odhady parametrů nelineárních modelů lze určit odhady rozptylů podle rov. (12) /17/. Rozdíl mezi $\tilde{\bar{\Theta}}_J$ a $\tilde{\bar{\Theta}}_{LJ}$ je měřítkem toho, nakolik je linearizace (30) dobrou aproximací $\varphi(\bar{x}, \bar{\Theta})$.

Chambers /19/ navrhl pro aproximaci pseudohodnot $\tilde{\Theta}_i$ v nelineárních modelech vztah

$$\tilde{\bar{\Theta}}_{Ci} = \hat{\bar{\Theta}} + (N-1)(Z_{(i)}^T Z_{(i)})^{-1}\bar{r}_{(i)} \tag{31}$$

Zde $Z_{(i)}$ resp. $\bar{r}_{(i)}$ jsou matice Z resp. vektor $\bar{r}$ s vynechaným i-tým bodem $(\bar{x}_i^T, y_i)$. Dosazením $\tilde{\bar{\Theta}}_{Ci}$ do rov. (11)

rezultuje aproximativní Jackknife odhad $\tilde{\hat{\Theta}}_{CJ}$.

Pro stanovení zrobustněných Jackknife odhadů se doporučuje nahradit aritmetický průměr v rov. (11) např. 10%ním useknutým průměrem nebo jiným robustním odhadem polohy /15/.

Při odhadu regresních parametrů technikou Bootstrap se postupuje takto /10/.

I. Na základě původního výběru (x_i^T, y_i) $i = 1..N$ se metodou MNČ (podle rov. (25)) stanoví odhad $\hat{\Theta}$.

II. Určí se rezidua $\hat{\varepsilon}_i = y_i - f(x_i, \hat{\Theta})$ a jejich empirická frekvenční funkce $g(\hat{\varepsilon})$ podle rov. (2) (při $w_i = 1$, $i = 1..N$).

III. Z empirické frekvenční funkce $g(\hat{\varepsilon})$ se pomocí generátoru náhodných čísel sestaví Bootstrap výběr o velikosti N $(\hat{\varepsilon}_1^{(j)} \ldots \hat{\varepsilon}_N^{(j)})$. Jde o výběr s vracením.

IV. Určí se nové "náhodné" proměnné $y_i^{(j)} = f(x_i, \hat{\Theta}) + \hat{\varepsilon}_i^{(j)}$ $(i = 1..N)$

V. Provede se odhad parametrů Θ z rov. (25) pro výběr $(x_i^T, y_i^{(j)})$ $i = 1, \ldots N$. Získaný odhad označme $\tilde{\Theta}^{(j)}$.

VI. Provede se M opakování ($j = 1, \ldots M$) kroků III až V.

Bootstrap odhad $\tilde{\hat{\Theta}}_B$ je pak aritmetickým průměrem

$$\tilde{\hat{\Theta}}_B = \frac{1}{M} \sum_{j=1}^{M} \tilde{\Theta}^{(j)} \tag{32}$$

a odpovídající rozptyl lze vyčíslit z výrazu

$$var(\tilde{\hat{\Theta}}_B) = \frac{RSC}{N-p} (\mathcal{X}^T \mathcal{X})^{-1} \tag{33}$$

kde RSC je reziduální součet čtverců.

4. Program pro odhad modelových parametrů.

Na základě výše uvedených vztahů byl sestaven program v jazyce HPL pro stolní kalkulátor HP 9825 A. Ten umožňuje výpočet odhadů $\hat{\Theta}$ (MNČ z rov. (25)), $\tilde{\hat{\Theta}}_J$ (z rov. (9) a (11)), $\tilde{\hat{\Theta}}_{LJ}$ (z rov. (27) a (11)), $\tilde{\hat{\Theta}}_{NJ}$ (z rov. (29) a (11)), $\tilde{\hat{\Theta}}_{CJ}$ (z rov. (31) a (11)) a $\tilde{\hat{\Theta}}_B$ (z rov. (32)). Při výpočtu odhadu $\tilde{\hat{\Theta}}_B$ je použito $M = N$ Bootstrap výběrů. Pro průměrování pseudohodnot je použito jak aritmetického tak i 10%ního useknutého průměru. Pro odhady $\hat{\Theta}$ a $\tilde{\hat{\Theta}}_B$ jsou kovarianční matice určeny z rov. (33) pro ostatní z rov. (12).

Rov. (25) (MNČ) je řešena Marquardtovou metodou s modifikacemi navrženými Nagelem a Wolfem /18/ a adaptivní korekcí maximální délky přírůstků v každé iteraci.

5. Příklad

V literatuře existuje zatím velmi málo prací, ve kterých byly techniky Jackknife a Bootstrap prakticky použity při odhadu parametrů v nelineárních modelech $\varphi(x, \Theta)$ v případě malého počtu měření N. Zde je pro porovnání výše uvedených odhadů použit jednoduchý Arrheniův model vyjadřující závislost rychlostní konstanty K na teplotě T vztahem

$$K = exp(K^* - E/RT) \tag{34}$$

kde K^* a E jsou modelové parametry (E má význam aktivační energie daného děje) a R je univerzální plynová konstanta. Protože má E známý fyzikální význam, je přirozené požadovat, aby její odhad byl co nejméně vychýlený. Z praktického hlediska však často nelze experimentálně stanovit dostatečně velký výběr $(K_i, T_i)_{i=1..N}$ ani provést opakování měření v bodech T_i.

Aby bylo možno porovnat chování jednotlivých odhadů i za těchto podmínek, byly připraveny simulované výběry o velikosti 10. Bylo zvoleno $K^* = ln(2{,}51.10^9)$ a $E = 59$ a rovnoměrné dělení teploty $T_i = 323{,}15 + 10(i-1)$ $i = 1..10$

Jednotlivé rychlostní konstanty byly určeny ze vztahu

$$K_i = exp(2{,}51.10^9 - 59/RT_i) + \varepsilon_i$$

Náhodná proměnná ε_i byla volena tak, aby bylo možno stanovit vliv heteroskedasticity i pro případy, kdy jde o součet dvou gaussovsky rozdělených veličin. Jednotlivá ε se počítala ze vztahu

$$\varepsilon_i = (1-G)\cdot N(0, \sigma_i^2) + G\cdot N(0, \sigma^2) \tag{35}$$

Pomocí parametru G (bylo voleno G = 0, 0.25, 0.5, 0.75 a 1) lze získat ε_i, která se více či méně liší od předpokladu $\varepsilon_i \in N(0, \sigma^2)$.

První člen v rov. (35) odpovídá heteroskedastickým datům. Rozptyl σ_i^2 se určoval z výrazu $\sigma_i^2 = 0{,}1 K_i$ a bylo voleno $\sigma^2 = 15$. Celý simulační experiment není ještě ukončen. Zatím byly pro každé G generovány čtyři náhodné výběry. Vzhledem k tomu, že tento počet nepostačuje pro statistické zpracování výsledků, bylo postupováno tak, že pro každý výběr bylo stanovano pořadí podle vzdálenosti odhadů aktivační energie E od zadané hodnoty 59. V tab. I jsou pro jednotlivá G uvedena vždy průměrná pořadí (čím nižší pořadí, tím je odhad méně vychýlený).

Tab. I Pořadí odhadů aktivační energie (průměry ze čtyř náhodných výběrů).

Odhad \ G	0	0.25	0.5	0.75	1
$\hat{\Theta}$ (MNČ)	3-4	5-6	6-7	3-4	2-3
$\tilde{\Theta}_{CJ}$	5-6	8	2-3	6	10
$\tilde{\Theta}_{CJ}$ *)	7	10	5	5	5-6
$\tilde{\Theta}_{LJ}$	9	9	9-10	10	5-6
$\tilde{\Theta}_{LJ}$ *)	1-2	1	4	8	7
$\tilde{\Theta}_{NJ}$	3-4	5-6	6-7	3-4	2-3
$\tilde{\Theta}_{NJ}$ *)	1-2	4	2-3	1-2	8
$\tilde{\Theta}_{B}$	8	2	8	1-2	9
$\tilde{\Theta}_{J}$	10	7	9-10	9	1
$\tilde{\Theta}_{J}$ *)	5-6	3	1	7	4

*) je použito 10%ního useknutého průměru

Vzhledem k malému počtu simulací zde nejsou ani porovnávány jednotlivé kovarianční matice odhadů.

I když pořadí není kvantitativní mírou vychýlenosti jednotlivých odhadů, lze i z těchto údajů zjistit zajímavé závěry:

a) Odhady $\hat{\Theta}$ a $\tilde{\Theta}_{NJ}$ jsou prakticky stejné. To svědčí o ustálenosti experimentálního "plánu" /15/.

b) Odhady $\tilde{\Theta}_J$ a $\tilde{\Theta}_{LJ}$ resp. $\tilde{\Theta}_{CJ}$ se liší. To svědčí o faktu, že linearizace vyjádřená rov. (30) není příliš dobrou aproximací rov. (34).

c) Náhradou aritmetického průměru useknutým průměrem dochází většinou ke snížení vychýlení. Jsou však případy (zejména u $\tilde{\Theta}_{CJ}$), kdy dojde k malému zhoršení.

d) Počet výběru $M = N = 10$ zřejmě nestačí pro získání méně vychýlených odhadů $\tilde{\Theta}_B$.

e) Chování odhadů je silně ovlivněno typem chyb ε_2 .

Ve většině případů se však ukázalo, že pokud byla $\hat{\Theta}$ podstatněji vychýlená, došlo u všech Jackknife i Bootstrap techniky buď k malému snížení vychýlenosti nebo dokonce i ke zvýšení $b(\hat{\Theta})$. Výhodnější by zřejmě bylo vypuštění "bodů", jimž odpovídající pseudohodnoty nebyly zařazeny do výpočtu useknutého průměru.
Definitivnější závěry však bude možno stanovit až po provedení celého zamýšleného simulačního experimentu.

6. Závěr:

V příspěvku byly uvedeny postupy pro snižování vychýlenosti bodových odhadů. Bylo ukázáno, jak lze využít při odhadu parametrů v nelineárních regresních modelech.

Literatura:

/1/ Dennis J., Welch R.E.: Commun. Statist. B7, 345 (1978)

/2/ Quenouille M.H.: Biometrika 43, 353 (1956)

/3/ Schucany W.R., Gray H.L., Owen D.B.: J. Amer. Stat. Assoc. 66, 524 (1971)

/4/ Tukey J.W.: Ann. Math. Statist. 29, 614 (1958)

/5/ Aziz E.S.: On Jackknife Estimators Based on Dependent Samples, Rept. Rouen University No 4, July 1979

/6/ Rey W.J.J.: Robust Statistical Methods, Lect. Notes in Matematics, Springer Verlag, Berlin 1978

/7/ Miller R.G.: Biometrika 61, 1 (1974)

/8/ Miller R.G.: Ann. Math. Statist. 35, 1594 (1964)

/9/ Hampel F.R.: J. Amer. Stat. Assoc. 69, 383 (1974)

/10/ Efron B.: Ann. Statist. 7, 1 (1979)

/11/ Efron B.: SIAM Rev. 21, 460 (1979)

/12/ Vapnik V.N.: Vosstanovlenije zavisimostěj po empiričeskim dannym, Nauka Moskva 1979

/13/ Mudrov V.I., Kuško V.L.: Metody obrabotki izmerenij, Sov. Radio, Moskva 1976

/14/ Miller R.G.: Ann. Statist. 2, 880 (1974)

/15/ Hinkley D.V.: Technometrics 19, 285 (1977)

/16/ Fox T., Hinkley D.V., Larntz K.: Technometrics 22, 29 (1980)

/17/ Duncan G.T.: Technometrics 20, 123 (1978)

/18/ Nagel G., Wolf O.: Biometrische Ztsch. 16, 431 (1974)

/19/ Chambers J.M.: Biometrika 60, 1 (1973)