TESTY A ODHADY POLOHY A REGRESE PŘI PORUŠENÍ PŘEDPOKLADU NEZÁVISLOSTI

Jana Jurečková, MFF UK Praha

1. Úvod. Je známa řada postupů, robustních vzhledem k odchylkám od předpokládaného tvaru rozdělení pravděpodobností, z něhož pozorování pocházejí. Z těchto postupů jmenujme např. pořadové testy a robustní odhady (M-, L- a R-odhady) parametru polohy a regrese.

Z praxe víme, že často bývá porušen předpoklad o nezávislosti pozorování. Jak ukazují teoretické i numerické studie řady autorů, jsou na porušení tohoto předpokladu citlivé jak klasické i pořadové testy, tak klasické i neklasické odhady. Např. se ukazuje, že velikost pořadových testů přestává být nezávislé na rozdělení pravděpodobností, že se zcela mění rozptyl odhadu apod.

Proto je žádoucí modifikovat standardní postupy tak, aby se citlivost vzhledem k porušení předpokladu nezávislosti maximálně snížila. Jde o obtížný problém a neexistuje jednotný návod na jeho řešení, už proto, že závislost mezi náhodnými veličinami může být různých typů. Proto neuvedeme nějaký univerzální postup, jak modifikovat běžné statistické postupy, jako spíše ukážeme, jaká modifikace byla navržena v některých důležitých speciálních případech. Přesněji řečeno, budeme uvažovat situaci, kdy předpoklad nezávislosti posloupnosti pozorování $X_1, X_2, \ldots$ je nahrazen předpokladem m-závislosti. Pro tuto situaci popíšeme modifikace dvou statistických postupů:
(1) modifikaci klasického t-testu, kterou navrhl Albers (1978a) a
(2) modifikaci Huberova M-odhadu parametru polohy, kterou navrhl Portnoy (1977). Další postupy a úvahy lze nalézt v seznamu literatury.

2. m-závislost. Nechť $X_1, X_2, \ldots$ je posloupnost náhodných veličin se stejným rozdělením pravděpodobností. Řekneme, že veličiny $X_1, X_2, \ldots$ jsou m-závislé, jestliže korelační koeficienty $\rho(X_i, X_j)$, $i,j = 1,2,\ldots$ vyhovují podmínce

$$(2.1)\qquad \rho(X_i, X_j) = \begin{cases} \rho_{|i-j|} & \text{pro } 1 \le |i-j| \le m \\ 0 & \text{pro } |i-j| > m \end{cases}$$

kde ρ_k, $k = 1,2,\ldots$ jsou konstanty takové, že korelační matice

libovolných n po sobě jdoucích veličin jsou positivně definitní a takové, že

$$(2.2) \qquad \sigma^{*2} = 1 + 2 \sum_{k=1}^{m} \rho_k > 0 ;$$

m je dané nezáporné celé číslo.

3. Modifikace t-testu o průměru normálního rozdělení při m-závislých pozorováních (Albers (1978a)).

Nechť $X_1,\dots,X_n$ jsou náhodné veličiny, všecky s normálním rozdělením $N(\mu, \sigma^2)$. Jestliže jsou X_i nezávislé, pak obvyklým testem hypotézy $H_0 : \mu = 0$ proti alternativě $H_1 : \mu > 0$ je Studentův t-test s kritickým oborem

$$(3.1) \qquad T_n > t_\alpha(n-1) \quad (\text{asymptoticky } T_n > u_\alpha), \quad \text{kde}$$

$$(3.2) \qquad T_n = \frac{\bar{X}_n}{S_n}\sqrt{n}, \quad S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X}_n)^2, \quad \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

$t_\alpha(n-1)$ je 100α-procentní kritická hodnota rozdělení t o $(n-1)$ stupních volnosti a u_α je 100α-procentní kritická hodnota normálního rozdělení $N(0,1)$.

Předpokládejme, že $X_1,\dots,X_n$ jsou m-závislé, kde m je dané přirozené číslo. Pak

$$\sqrt{n}\,\frac{\bar{X}_n - \mu}{\sigma}$$

má normální rozdělení $N(0, \sigma_n^{*2})$, kde

$$(3.3) \qquad \sigma_n^{*2} = 1 + 2\sum_{k=1}^{m}\left(1 - \frac{k}{n}\right)\rho_k .$$

Protože platí $S_n^2 \xrightarrow{p} \sigma^2$ při $n \to \infty$, dostáváme ze Sluckého věty, že $\sqrt{n}(\bar{X}_n - \mu)/S_n$ má při $n \to \infty$ asymptoticky normální rozdělení $N(0, \sigma^{*2})$, kde σ^{*2} je definováno ve (2.2).

Kdybychom použili t-test nebo asymptotický t-test (viz (3.1)), byla by asymptotická velikost testu nikoli α, ale byla by rovna

$$(3.4) \qquad P_0(T_n > u_\alpha) = P_0\left(\frac{T_n}{\sigma^*} > \frac{u_\alpha}{\sigma^*}\right) = 1 - \Phi\left(\frac{u_\alpha}{\sigma^*}\right) + o(1),$$

což se může značně lišit od α, jsou-li $\varrho_1, \ldots, \varrho_m$ nenulové.

Naším cílem je modifikovat t-test tak, aby jeho asymptotická velikost byla rovna α pro všecky hodnoty $\varrho_1, \ldots, \varrho_m$ a nikoli jen v případě $\varrho_1 = \ldots = \varrho_m = 0$. K tomu účelu nejprve odhadneme $\varrho_1, \ldots, \varrho_m$ pomocí odhadů

$$(3.5) \qquad \hat{\varrho}_k = \sum_{i=1}^{n} (X_i - \bar{X}_n)(X_{i+k} - \bar{X}_n) / ((n-1)S^2), \quad k = 1, \ldots, m$$

(kde dodefinujeme $X_{n+k} = X_k$, $k = 1, \ldots, n$) nebo pomocí odhadů

$$(3.6) \qquad \hat{\varrho}_k = \sum_{i=1}^{n-k} (X_i - \bar{X}_n)(X_{i+k} - \bar{X}_n) / ((n-1)S_n^2), \quad k = 1, \ldots, m.$$

Oba typy odhadů (3.5) a (3.6)) vedou ke stejným asymptotickým výsledkům. Definujme statistiku W_m

$$(3.7) \qquad W_m^2 = \begin{cases} S_n^2 \left(1 + 2 \sum_{k=1}^{m} \hat{\varrho}_k\right) & \text{pro } \underset{\sim}{x} = (x_1, \ldots, x_1) \in B \\ \text{libovolná ohraničená} & \text{pro } \underset{\sim}{x} \notin B, \end{cases}$$

kde

$$(3.8) \qquad B = \left\{ \underset{\sim}{x} = (x_1, \ldots, x_n) : 1 + 2 \sum_{k=1}^{m} \hat{\varrho}_k > 0 \right\}.$$

t-test daný ve (3.1) a (3.2) nahradíme modifikovaným testem s kritickým oborem

$$(3.9) \qquad V_m = \sqrt{n}\, \bar{X}_n / W_m > u_\alpha \quad .$$

Následující věta ukazuje, že test (3.9) má asymptotickou velikost α i při m-závislých pozorováních.

<u>VĚTA 3.1.</u> <u>Nechť</u> $X_1, X_2, \ldots, X_n$ <u>jsou m-závislé náhodné veličiny, všecky s normálním rozdělením</u> $N(\mu, \sigma^2)$. <u>Pak pro lib.</u> $x \in R^1$ <u>platí</u>

$$(3.10) \qquad P\left\{ \sqrt{n}\, \frac{\bar{X}_n - \mu}{W_m} \leq x \right\} \longrightarrow \bar{\Phi}(x) \quad \text{<u>při</u>} \quad n \to \infty ,$$

kde Φ je distribuční funkce $N(0,1)$. Konvergence (3.10) je stejnoměrná na množině

$$A = \left\{ (\sigma, \rho_1, \ldots, \rho_m) : \sigma^2 > C_1 > 0 \text{ a } 1 + 2\sum_{k=1}^{m} \rho_k > C_2 > 0 \right\},$$

kde C_1 a C_2 jsou libovolné kladné konstanty. Test hypotézy $H_0 : \mu = 0$ proti $H_1 : \mu > 0$ tvaru

$$(3.11)\quad \psi_V^{(m)}(\underset{\sim}{x}) = \begin{cases} 1 \ldots & V_m \geq u_\alpha \quad [= \Phi^{-1}(1-\alpha)] \\ 0 \ldots & V_m < u_\alpha \end{cases}$$

má asymptotickou velikost α pro vš. $(\sigma, \rho_1, \ldots, \rho_m) \in A$.

Důkaz : viz Albers (1978a).

Poznámka. Jestliže je $m=0$, je test $\psi_V^{(0)}$ shodný s asymptotickým t-testem.

Jestliže víme, že $X_1, \ldots, X_n$ jsou m-závislé a kromě toho známe $\rho_1, \ldots, \rho_m$, můžeme použít t-testu, optimálního pro tuto situaci (Scheffé (1959)). Test je založen na statistice

$$(3.12)\qquad L = \left(\tilde{s}_n^2 \sum_{i=1}^{n}\sum_{j=1}^{n} u_{ij}\right)^{-1/2} \sum_{i=1}^{n}\left(\sum_{j=1}^{n} u_{ij}\right) X_i$$

kde $U_n = \left[u_{ij}\right]_{i,j=1}^{n}$ je inverzní korelační matice veličin $X_1, \ldots, X_n$ (a tedy známá) a

$$(3.13)\quad \tilde{s}_n^2 = \frac{1}{n-1}\left\{ \sum_{i=1}^{n}\sum_{j=1}^{n} u_{ij} X_i X_j - \left[\left(\sum_{i=1}^{n}\sum_{j=1}^{n} u_{ij} X_i \right)^2 / \sum_{i=1}^{n}\sum_{j=1}^{n} u_{ij} \right] \right\}.$$

Albers (1978a) dokázal, že test $\psi_V^{(m)}$, odpovídající situaci, kdy neznáme $\rho_1, \ldots, \rho_m$, je při $n \to \infty$ asymptoticky ekvivalentní optimálnímu t-testu, a tedy je asymptoticky optimální.

Další otázka, která s použitím testu $\psi_V^{(m)}$ vzniká, je, jakou utrpíme ztrátu, jestliže použijeme modifikovaného testu $\psi_V^{(m)}$ v případě, že veličiny $X_1, \ldots, X_n$ jsou skutečně závislé a kdy je tedy optimální klasický t-test (3.1). Intuitivně je

zřejmé, že chceme-li pak testem $\psi_V^{(m)}$ dosáhnout stejné asymptotické síly jako t-testem, musíme vzít více pozorování.

Z předcházejících úvah vyplývá, že síla testu $\psi_V^{(m)}$ proti alternativě μ je rovna

$$(3.14)\qquad \beta_V^{(m)}(\mu) = 1 - \Phi\left(u_\alpha - n^{1/2}\mu\left[\sigma^2\left(1+2\sum_{k=1}^{m}\rho_k\right)\right]^{-1/2}\right) + o(1).$$

V případě nezávislosti je $\rho_k = 0$, $k=1,\ldots,m$, a tedy síla (3.14) je s přesností řádu $o(1)$ shodná se silou $\beta_t(\mu)$ klasického t-testu ψ_t. To znamená, že v případě nezávislosti je relativní asymptotická vydatnost e testu $\psi_V^{(m)}$ vzhledem k t-testu rovna 1. Jinak řečeno, pokud ψ_t požaduje n pozorování k dosažení určité síly a $\psi_V^{(m)}$ požaduje k_n pozorování k dosažení stejné síly, platí $k_n/n \to 1$ při $n\to\infty$.

Hodges a Lehmann (1970) srovnávali statistické postupy pomocí tzv. deficience $d_n = k_n - n$, tj. dodatečného počtu pozprování, který potřebuje méně vydatný postup. Albers (1978a) studoval deficienci $d_n(V_m,t)$ testu $\psi_V^{(m)}$ vzhledem k t-testu a odvodil

$$(3.15)\qquad d_n(V_m,t) = m\,u_\alpha^2 + O(n^{-1}),$$

což lze interpretovat tak, že chráníme-li se proti možnému vlivu m-závislosti použitím modifikovaného testu $\psi_V^{(m)}$, potřebujeme k tomu asymptoticky o $m.u_\alpha^2$ více pozorování více než t-test.

4. Modifikace M-odhadu parametru polohy při m-závislých pozorováních (Portnoy 1977)

Nechť $X_1, X_2, \ldots$ je posloupnost m-závislých náhodných veličin se společnou marginální distribuční funkcí $F(x-\theta)$ takovou, že $F(x) + F(-x) = 1$, $x \in R^1$. Nechť $\hat\theta_n = \hat\theta_n(X_1,\ldots,X_n)$ je libovolný konsistentní odhad θ založený na $X_1,\ldots,X_n$ (např. $\hat\theta_n = \bar X_n$). Pak M-odhad $T_n = T_n(X_1,\ldots,X_n)$ parametru θ definujeme jako řešení rovnice

$$(4.1)\qquad \sum_{i=1}^{n}\psi(X_i - t) = 0 \quad \text{vzhledem k } t$$

nejbližší $\hat{\theta}_n$ [a větší, jestliže jsou 2 kořeny stejně vzdálené od $\hat{\theta}_n$; položíme $T_n=0$, jestliže neexistuje kořen rovnice (4.1)] ; ψ je vhodně zvolená funkce, jejíž vlastnosti dále upřesníme.

Následující věta udává asymptotické rozdělení pravděpodobnosti posloupnosti $\sqrt{n}(T_n-\theta)$ za předpokladu m-závislosti.

VĚTA 4.1. Nechť ψ je spojitá ohraničená funkce, která má stejnoměrně spojitou a ohraničenou derivaci ψ' na doplňku libovolného okolí uzavřené množiny D Lebesgueovy míry 0 takové, že $0 \notin D$. Nechť dále platí

$$(4.2) \qquad \int_{R^1} \psi(x)\,dF(x-t) \quad \text{je ryze rostoucí v } t \text{ v okolí } t=0;$$

$$(4.3) \qquad \int \psi(x)\,dF(x) = 0; \qquad \gamma = \int \psi'(x)\,dF(x) > 0.$$

Nechť F má spojitou derivaci v okolí D. Pak $\sqrt{n}(T_n-\theta)$ má při $n\to\infty$ asymptoticky normální rozdělení $N(0, \sigma_0^2/\gamma^2)$, kde

$$(4.4) \qquad \sigma_0^2 = \lim_{n\to\infty} \frac{1}{n} \operatorname{var}_0\left(\sum_{i=1}^{n} \psi(X_i)\right).$$

Důkaz: viz Portnoy (1977).

Dále budeme uvažovat speciální model m-závislosti, ve kterém pozorování $X_1,\ldots,X_n$ mají tvar

$$(4.5) \qquad X_i = \theta + Y_i + \rho Y_{i-1} + \rho Y_{i+1}; \quad i=1,\ldots,n,$$

kde $Y_1,\ldots,Y_n$ jsou nezávislé náhodné veličiny se společnou distribuční funkcí G a hustotou g takovou, že $g(x)=g(-x)$, $x \in R^1$ [dodefinujme $Y_0=Y_n$, $Y_{n+1}=Y_1$] ; θ je parametr polohy a ρ je další parametr, $|\rho| < 1$. Pak $X_1,\ldots,X_n$ jsou m-závislé, kde $m=2$ a mají stejné rozdělení pravděpodobnosti, jehož distribuční funkci označíme F.

Z věty 4.1 vyplývá, že M-odhad $T_n=T_n(X_1,\ldots,X_n)$ má asymptoticky normální rozdělení. Přesněji, platí věta

VĚTA 4.2. Nechť náhodné veličiny $X_1,\ldots,X_n$ vyhovují modelu (4.5), kde $Y_1,\ldots,Y_n$ jsou nezávislé náhodné veličiny se společnou hustotou g, $g(-x)=g(x)$ $\forall x$, která má konečný druhý mo-

ment a jejíž charakteristická funkce splňuje podmínku

(4.6) $$\int u^2 |\varphi_Y(u)| du < \infty .$$

Nechť T_n je M-odhad definovaný ve (4.1), kde funkce ψ je absolutně spojitá a má ohraničenou derivaci ψ'. Pak

(4.7) $$\sqrt{n}(T_n - \theta) \xrightarrow{\mathcal{D}} N(0, \sigma^2), \quad \text{kde}$$

(4.8) $$\sigma^2 = \frac{\int \psi^2 \, dG}{\left(\int \psi' \, dG\right)^2} + 4\rho \, \frac{\int y\psi(y) \, dG(y)}{\int \psi' \, dG} + O(\rho^2).$$

Důkaz : viz Portnoy (1977).

Poznámka. První člen pravé strany (4.8) je roven asymptotickému rozptylu M-odhadu, založeného na veličinách $Y_1, \ldots, Y_n$. Pozorování jsou však $X_1, \ldots, X_n$.

Označme

(4.9) $$V(g, \psi) = \frac{E\,\psi^2(Y)}{(E\,\psi'(Y))^2} + 4\rho \, \frac{E[Y.\psi(Y)]}{E(\psi'(Y))} ;$$

asymptotický rozptyl $\sqrt{n}(T_n - \theta)$ je pro malá $|\rho|$ přibližně roven $V(g, \psi)$. Předpokládáme, že neznáme přesně rozdělení G; pouze víme, že $G \in \mathcal{F}_1$, kde

(4.10) $$\mathcal{F}_1 = \left\{ G = (1-\varepsilon)\Phi + \varepsilon H; \; H \text{ symetrická spojitá distribuční funkce} \right\}$$

a Φ je distribuční funkce normálního rozdělení $N(0,1)$; $0 \le \varepsilon < 1$; [model kontaminovaného normálního rozdělení]. Za těchto podmínek hledáme M-odhad odpovídající funkci ψ_0 a rozdělení $g_0 \in \mathcal{F}_1$ (nejméně příznivé rozdělení) tak, aby asymptotický rozptyl $V(g_0, \psi_0)$ tvořil sedlový bod funkce $V(g, \psi)$, tj. aby platilo

(4.11) $$V(g, \psi_0) \le V(g_0, \psi_0) \le V(g_0, \psi)$$

pro vš. $g \in \mathcal{F}_1$ a vš. ψ .

Pro případ nezávislých pozorování (tj. $\varrho=0$) je tento problém vyřešen v Huberově práci (1964). V tomto případě řešení odpovídá funkci

$$(4.12)\quad \psi(x)=\begin{cases} x & \dots\ |x|\le k \\ k.\mathrm{sign}\ x & \dots\ |x|>k \end{cases}$$

kde k závisí na ε podle vztahu

$$(4.13)\quad 2\Phi(k)-1+\frac{2}{k}\Phi'(k)=\frac{1}{1-\varepsilon}\ .$$

Portnoy (1977) řešil analogický problém pro model (4.5). Při zenedbání členů řádu $O(\varrho^2)$ (tj. pro malé hodnoty $|\varrho|$) dospěl k M-odhadu odpovídajícímu funkci

$$(4.14)\quad \psi_0(x)=\begin{cases} x & \dots |x|\le k \\ k.\mathrm{sign}\ x-\dfrac{2\varrho a}{1+2\varrho(1-a)}(x-k\,\mathrm{sign}\,x) & \dots |x|>k, \end{cases}$$

kde $a=(1-\varepsilon)(2\Phi(k)-1)$ a k je dáno vztahem (4.13); příslušné nejméně příznivé rozdělení g_0 je stejné jako v nezávislém případě a je dáno vztahem

$$(4.15)\quad -\frac{g_0'(x)}{g_0(x)}=\begin{cases} x & \dots |x|\le k \\ k.\mathrm{sign}\ x & \dots |x|>k\ . \end{cases}$$

Jestliže $x\to\infty$, funkce $\psi_0(x)$ lineárně klesá k $-\infty$. Na druhé straně, $P_0(\psi_0(X)<0)=E_0X^2.O(\varrho^2)$. Proto Portnoy navrhuje následující modifikaci funkce ψ_0, vhodnou při $E_0X^2<\infty$: usekneme-li funkci ψ_0 v bodech, kde protíná osu x, dostaneme

$$(4.16)\quad \psi_1(x)=\begin{cases} x & |x|\le k \\ k.\mathrm{sign}\ x & k<|x|\le k' \\ 0 & k'<|x|, \end{cases}$$

kde k vyhovuje (4.13), $a=(1-\varepsilon)(2\Phi(k)-1)$ a

$$(4.17)\quad k'=k\,\frac{1+2\varrho a}{2\varrho a}\ .$$

Poznámka. Funkce ψ_1 je shodná s funkcí, kterou navrhl Hampel (viz např. Andrews a kol.(1972)).

Ze známého vztahu mezi M-odhady a L-odhady (lineární kombinace pořádkových statistik - viz např. Jaeckel (1971)) můžeme odvodit L-odhad, který také poskytuje minimaximální řešení problému s přesností do členů řádu $O(|\varrho|)$. Uvažujme L-odhad tvaru

$$(4.18) \qquad T_n^* = \sum_{i=1}^{n} c_i X^{(i)}, \quad \text{kde} \quad c_i = J\left(\frac{i}{n+1}\right), \; i=1,\ldots,n;$$

volíme-li funkci $J(t)$, $0<t<1$ tvaru

$$(4.19) \quad J(t) = \begin{cases} -2\varrho & \ldots\ 0<t\le\alpha,\ 1-\alpha\le t<1 \\ \dfrac{1+4\alpha\varrho}{1-2\alpha} & \ldots\ \alpha\le t\le 1-\alpha, \end{cases}$$

pak při vhodné volbě konstanty α, $0<\alpha<\frac{1}{2}$, dostáváme L-odhad s asymptoticky minimaximálním rozptylem v modelu kontaminovaného normálního rozdělení. Všimněme si, že koeficienty u krajních pořádkových statistik jsou záporné; při $\varrho = 0$ dostáváme useknutý průměr.

LITERATURA :

[1] Albers,W.(1978a). Testing the mean of a normal population under dependence. Ann.Statist.6,1337-1344.

[2] Albers,W.(1978b). One-sample rank tests under autoregressive dependence. Ann.Statist.6,836-845.

[3] Andrews,D.F.,Bickel,P.J., Hampel,F.R., Huber,P.J., Rogers,W.H., Tukey,J.W. (1972). Robust Estimates of Location. Survey and Advances.Princeton University Press.

[4] Gastwirth,J.L. and Rubin,H.(1971). Effect of dependence on the level of some one-sample tests. J.Amer.Statist.Assoc. 66, 816-820.

[5] Gastwirth,J.L. and Rubin,H.(1975a). Asymptotic distribution theory of empiric cdf for mixing processes. Ann.Statist. 3, 809-824.

[6] Gastwirth,J.L. and Rubin,H.(1975b). The behavior of robust estimators on dependent data. Ann.Statist. 3, 1070-1100.

[7] Gastwirth,J.L., Rubin,H. and Wolf,S.S.(1976). The effect of autoregressive dependence on a nonparametric test. IEEE Trans.Information Theory IT-13, 311-313.

[8] Hodges,J.L.Jr. and Lehmann,E.L.(1970). Deficiency. Ann.Statist. 41, 783-801.

[9] Huber,P.J.(1964). Robust estimation of a location parameter. Ann.Math.Statist.35, 73-101.

[10] Jaeckel,L.A.(1971). Robust estimates of location Symmetry and asymmetry contamination. Ann.Math. Statist. 42, 1020-1034.

[11] Koul,H.L.(1977). Behavior of robust estimators in the regression model with dependent errors. Ann.Statist. 5, 681_699.

[12] Lehmann,E.L.(1966). Some concepts of dependence. Ann.Statist.37, 1137-1152.

[13] Modestino,J.W.(1969). Nonparametric and adaptive detecting on dependent data. Technical Report 27, Dept.of Electrical Engineering, Princeton Univ.

[14] Ramachandramurty,P.V. and Rao,M.S.(1971). Some problems of robust estimation against dependence. Sankhyā A 33, 193-200.

[15] Scheffé,H.(1959). The analysis of Variance. Wiley, N.York.

[16] Serfling,R.J.(1968). The Wilcoxon two-sample statistic on strongly mixing processes. Ann.Math.Statist. 39, 1202-1209.