Testy pre regresné kvantily založené na metóde sedlového bodu

Radka Sabolová KPMS MFF UK sabolova@karlin.mff.cuni.cz

Abstrakt

Metóda sedlového bodu poskytuje presné aproximácie hustoty odhadov aj pre malé rozsahy výberov. V príspevku sa budeme zaoberať testovou štatistikou, ktorá je založená na výraze v exponente v aproximácii hustoty pre M-odhady. Vďaka tvaru funkcie ψ pre regresné kvantily získame explicitný vzorec pre testovú štatistiku. Odvodené testy sú asymptoticky ekvivalentné s klasickými testami založenými na vierohodnosti, no ich relatívna chyba je iba rádu $O(n^{-1})$. Správanie navrhnutých parametrických i neparametrických testov pri rôznych rozsahoch výberov a rozdeleniach bude ilustrované v simulačnej štúdii.

Regresné kvantily

Nech $Y_1, \ldots, Y_n$ sú pozorovania spĺňajúce model

$$Y_i = \mathbf{X}_i^T \boldsymbol{\beta} + u_i, \qquad i = 1, \ldots, n,$$

kde $\boldsymbol{\beta} \in \mathbb{R}^p$, $u_i \sim \frac{1}{\sigma} g(.)$ a $(Y_i, \mathbf{X}_i)$ sú i.i.d. s hustotou $\frac{1}{\sigma} g\left(\frac{y_i - \mathbf{x}_i^T \boldsymbol{\beta}}{\sigma}\right) k(\mathbf{x}_i)$, kde $k(\mathbf{x}_i)$ je hustota $\mathbf{X}_i$.
Regresný kvantil je riešením minimalizačnej úlohy

$$\boldsymbol{\beta}_\alpha = \arg\min \left\{ \mathsf{E} \rho_\alpha (Y - \boldsymbol{X}\mathbf{t}) : \mathbf{t} \in \mathbb{R}^p \right\},$$

kde

$$\rho_\alpha(x) = |x| \left\{ (1-\alpha) \, \mathrm{I}\,[x < 0] + \alpha \, \mathrm{I}\,[x > 0] \right\}.$$

Výberový regresný kvantil je definovaný ako riešenie úlohy

$$\hat{\boldsymbol{\beta}}_\alpha = \arg\min_{\boldsymbol{\beta}_\alpha} \sum_{i=1}^n \rho_\alpha (Y_i - \mathbf{x}_i^\top t)$$

a je konzistentným odhadom vektora

$$(\beta_1 + G^{-1}(\alpha), \beta_2, \ldots, \beta_p).$$

Testujeme jednoduchú hypotézu

$$H_0 : \boldsymbol{\beta}_\alpha = \boldsymbol{\beta}_{\alpha 0}.$$

Použijeme postup pre M-odhady uvedený v [1].

Parametrický test

Ak je rozdelenie $(Y_i, \mathbf{X}_i)$ známe, je potrebné vypočítať kumulatívnu vytvárajúcu funkciu pre skóry $\psi(Y_i, \boldsymbol{\beta})$

$$K_\psi(\boldsymbol{\lambda}, \boldsymbol{\beta}_\alpha) = \log \int \left\{ e^{\alpha \boldsymbol{\lambda}^T \mathbf{x}_i} k(\mathbf{x}_i) \left(e^{-\boldsymbol{\lambda}^T \mathbf{x}_i} G\left(\frac{\mathbf{x}_i^T (\boldsymbol{\beta}_\alpha - \boldsymbol{\beta})}{\sigma} \right) + 1 - G\left(\frac{\mathbf{x}_i^T (\boldsymbol{\beta}_\alpha - \boldsymbol{\beta})}{\sigma} \right) \right) \right\} d\mathbf{x}_i.$$

Zderivovaním $K_\psi(\boldsymbol{\lambda}, \boldsymbol{\beta}_\alpha)$ a položením rovno nule získame funkcie sedlového bodu

$$\boldsymbol{\lambda}(\hat{\boldsymbol{\beta}}_\alpha)^T \mathbf{x}_i = -\log \left\{ \frac{\alpha}{1-\alpha} \frac{1 - G\left(\frac{\mathbf{x}_i^T (\hat{\boldsymbol{\beta}}_\alpha - \boldsymbol{\beta})}{\sigma}\right)}{G\left(\frac{\mathbf{x}_i^T (\hat{\boldsymbol{\beta}}_\alpha - \boldsymbol{\beta})}{\sigma}\right)} \right\}, \qquad i = 1, \ldots, n,$$

ktoré sa následne dosadia do štatistiky

$$h(\hat{\boldsymbol{\beta}}_\alpha) = \sup_\lambda \{ -K_\psi(\lambda, t) \} = -K_\psi(\boldsymbol{\lambda}(\hat{\boldsymbol{\beta}}_\alpha), \hat{\boldsymbol{\beta}}_\alpha)$$

a za H_0 platí

$$-2n \log \int \left(\frac{G\left(\frac{\mathbf{x}_i^T (\hat{\boldsymbol{\beta}}_\alpha - \boldsymbol{\beta})}{\sigma}\right)}{\alpha} \right)^\alpha \left(\frac{1 - G\left(\frac{\mathbf{x}_i^T (\hat{\boldsymbol{\beta}}_\alpha - \boldsymbol{\beta})}{\sigma}\right)}{1-\alpha} \right)^{1-\alpha} k(\mathbf{x}_i) \, d\mathbf{x}_i \xrightarrow[n \to \infty]{\mathcal{D}} \chi_p^2.$$

Neparametrický test

Ak je rozdelenie $(Y_i, \mathbf{X}_i)$ neznáme, je potrebné pracovať s reziduami r_i a ich empirickou distribučnou funkciou (viď [2])

$$r_i = y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_\alpha, \qquad I_i = \mathrm{I}[r_i < 0], \qquad i = 1, \ldots, n,$$

$$I_{ij}(\boldsymbol{\beta}) = \mathrm{I}[r_i + \mathbf{x}_j^T (\hat{\boldsymbol{\beta}}_\alpha - \boldsymbol{\beta}) < 0], \qquad i, j = 1, \ldots, n,$$

$$G_n^j(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^n I_{ij}(\boldsymbol{\beta}) \qquad i, j = 1, \ldots, n.$$

Prvým krokom bude nájdenie emprirických distribučných funkcií, ktoré spĺňajú hypotézu a zároveň minimalizujú Kullback-Leiblerovu vzdialenosť od $(1/n, \ldots, 1/n)$. Pre regresné kvantily majú tieto váhy tvar

$$w_{ij} = \frac{\left(\frac{\alpha}{1-\alpha} \frac{1 - G_n^j(\boldsymbol{\beta}_{\alpha 0})}{G_n^j(\boldsymbol{\beta}_{\alpha 0})} \right)^{I_{ij}(\boldsymbol{\beta}_{\alpha 0})}}{\sum_{k=1}^n \left(\frac{\alpha}{1-\alpha} \frac{1 - G_n^j(\boldsymbol{\beta}_{\alpha 0})}{G_n^j(\boldsymbol{\beta}_{\alpha 0})} \right)^{I_{kj}(\boldsymbol{\beta}_{\alpha 0})}}, \; i, j = 1, \ldots, n.$$

Potom

$$K(\boldsymbol{\lambda}(\hat{\boldsymbol{\beta}}_\alpha), \hat{\boldsymbol{\beta}}_\alpha) = \frac{1}{n} \sum_{j=1}^n K^j(\boldsymbol{\lambda}(\hat{\boldsymbol{\beta}}_\alpha), \hat{\boldsymbol{\beta}}_\alpha)$$

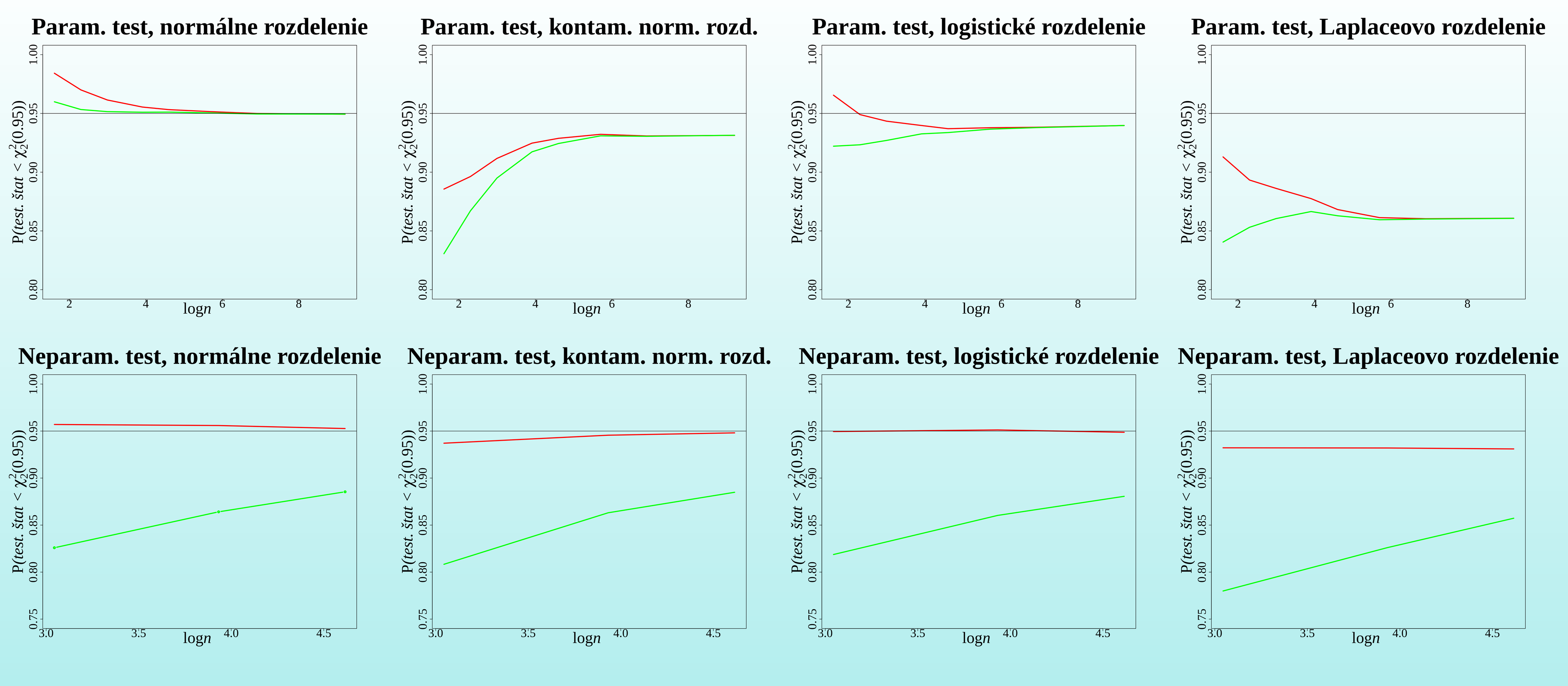
$$K^j(\boldsymbol{\lambda}(\hat{\boldsymbol{\beta}}_\alpha), \hat{\boldsymbol{\beta}}_\alpha) = \log \left\{ \left(\frac{\sum_{i=1}^n w_{ij} I_i}{\alpha} \right)^\alpha \left(\frac{\sum_{i=1}^n w_{ij} (1 - I_i)}{1-\alpha} \right)^{1-\alpha} \right\}$$

a za H_0 platí

$$-2 \sum_{j=1}^n \log \left\{ \left(\frac{\sum_{i=1}^n w_{ij} I_i}{\alpha} \right)^\alpha \left(\frac{\sum_{i=1}^n w_{ij} (1 - I_i)}{1-\alpha} \right)^{1-\alpha} \right\} \xrightarrow[n \to \infty]{\mathcal{D}} \chi_p^2.$$

Simulácie

Navrhnuté testy pre regresné kvantily (v grafoch červenou farbou) porovnáme s Waldovým testom (zelenou). Zvolili sme $\mathbf{x}_i^T = (1, \frac{i-1}{n})$. Nasledujúce obrázky porovnávajú rýchlosť konvergencie rozdelenia testových štatistík ku χ_2^2 rozdeleniu za platnosti H_0 a v parametrickom prípade ilustrujú robustnosť testov. Použité rozdelenia boli: normálne, kontaminované normálne, Laplaceovo, logistické.
Pri použití parametrických testov sa predpokladalo normálne rozdelenie u_i, je zrejmé, že navrhnutý test je najmä pre menšie rozsahy výberov robustnejší než Waldov test. Pri neparametrických testoch je aproximácia testovej štatistiky χ^2 rozdelením pre testy založené na metóde sedlového bodu omnoho presnejšia než pre Waldov test. Vidíme, že táto aproximácia funguje veľmi dobre už aj pre malé výbery.



Literatúra

[1] Ronchetti E, Robinson J, Young GA (2003): Saddlepoint Approximations and Tests Based on Multivariate M-estimates, The Annals of Statistics 31: 1154–1169
[2] Ronchetti E, Welsh AH (1994): Empirical Saddlepoint Approximations for Multivariate M-estimators, Journal of the Royal Statistical Society 56, no.2: 313–326

Podpora

Tento výstup vznikol v rámci projektu SVV 265 315.