Kvantilová regrese a odhad Paretova indexu

Jan Dienstbier

dienstbi@karlin.mff.cuni.cz

KPMS, MFF-UK, Praha

V příspěvku je načrtnut jeden z možných přístupů k odhadu Paretova indexu v regresním případě, kdy jsou náhodné veličiny stochasticky závislé na vektoru nezávislých proměnných. Metoda využívá kvantilové regrese.

Analýza extrémních událostí

Nechť jsou $X_1, \ldots, X_n$ pro $n \in N$ nezávislé stejně rozdělené náhodné veličiny se spojitou distribuční funkcí. Podle **Fisher-Tippetovy věty** lze pak vlastnosti rozdělení chvostů X_i (tj. i vlastnosti maxim $X_{n,n} := \max\{X_1, \ldots, X_n\}$) asymptoticky popsat jediným parametrem γ – tzv. **Paretovým indexem**. Mluvíme o třech **sférách přitažlivosti** neboli o rozděleních

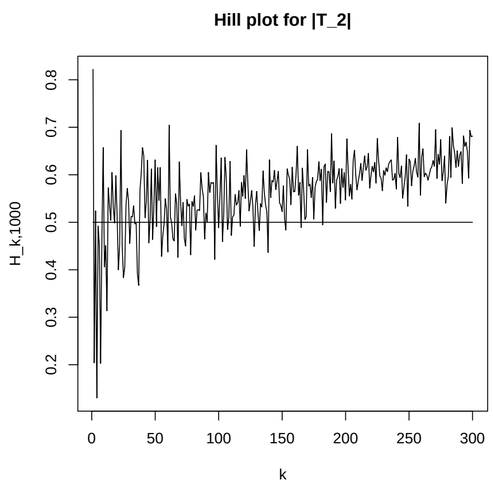
- **Fréchet-Paretova typu** pro $\gamma > 0$, tj. s těžkými chvosty,
- **Gumbelova typu** pro $\gamma = 0$, tj. s exponenciálními chvosty,
- **Weibullova typu** pro $\gamma < 0$, tj. s omezenými chvosty.

V literatuře byla popsána celá řada odhadů γ – viz Beirlant et al. (2004).

Př. Hillův odhad

$$H_{k,n} = \frac{1}{k} \sum_{i=0}^{k-1} \log X_{n-i,n} - \log X_{n-k,n}.$$

Na obrázku vidíme **závislost Hillova odhadu na** k, tj. volbě podílu dat použitých pro výpočet. Data byla získána simulací z absolutních hodnot Tukeyho rozdělení o dvou stupních volnosti ($|T_2|$). Skutečná hodnota $\gamma = 0.5$ je indikována čarou. Lze vypozorovat klesající rozptyl odhadu s rostoucím k a naopak rostoucí vychýlení. Příklad demonstruje **důležitost volby "optimálního" podílu dat**, problém, který se tak či onak objeuje u většiny odhadů extrémního indexu γ.



Regresní případ

Uvažujme nyní regresní případ $Y_i = \beta^\top \mathbf{x}_i + \mathbf{u}_i,\ i = 1, \ldots, n$, kde $\mathbf{u}_i$ jsou nezávislé náhodné veličiny. Vraťme se nyní zpět a podívejme se, co nám brání použít odhady odvozené pro stejně rozdělené náhodné veličiny.

Motivace:

$$\frac{1}{\log 2} \log \left\{ \frac{Q\left(1 - 1/4y\right) - Q\left(1 - 1/2y\right)}{Q\left(1 - 1/2y\right) - Q\left(1 - 1/y\right)} \right\} \overset{y \to \infty}{\longrightarrow} \gamma,$$

kde $Q(y) = F^{\leftarrow}(y) = \inf\{x : F(x) \geq y\}$.

$$\downarrow \quad Q(1 - 1/y) \approx X_{n-k+1,n} \text{ (!)}$$

Pickandsův odhad

$$\hat{\gamma}_{P,k} := \frac{1}{\log 2} \log \left(\frac{X_{n-\lceil k/4 \rceil + 1,n} - X_{n-\lceil k/2 \rceil + 1,n}}{X_{n-\lceil k/2 \rceil + 1,n} - X_{n-k+1,n}} \right)$$

Je tedy nutné odhadnout $Q(1 - 1/y)$ něčím "lepším" než je $X_{n-k+1,n}$.

Kvantilová regrese

Myšlenkou lineární kvantilové regrese je nalezení řešení minimalizační úlohy

$$\min_{b \in \mathcal{R}^p} \sum_{i=1}^{n} \rho_\tau \left(y_i - \mathbf{x}_i^\top b \right),$$

kde ρ_τ označuje ztrátovou funkci

$$\rho_\tau(u) := u \cdot \left(\tau - I\left(u < 0\right)\right).$$

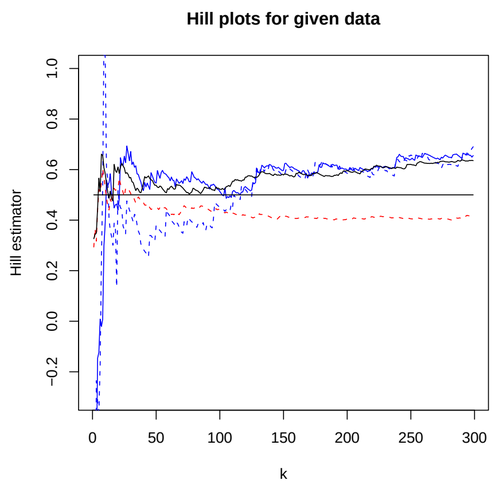
Za τ**-tý regresní kvantil** se pak označuje $\hat{Q}_Y\left(\tau | x_i\right) := x_i^\top \hat{b}\left(\tau | Y, \mathbf{x}\right)$.

Přirozeně se tak nabízí nasledující modifikace stávajících odhadů, použitelná nejen pro Pickandsův ale např. i pro Hillův odhad.

1 Obecný model	2 Model se stejně rozdělenými chybami
$\mathbf{u}_i \sim F(\theta(\mathbf{x}_i); z)$	$\mathbf{u}_i \sim F(\theta; z)$
$\downarrow$	$\downarrow$
$Q(\tau) \approx \hat{Q}_Y\left(\tau | x_i\right)$	$Q(\tau) \approx \hat{b}_0\left(\tau | Y, \mathbf{x}\right)$

Simulační studie

- metodu demonstrujeme pro model se stejně rozdělenými chybami
- jiné simulace však ukázaly její funkčnost i pro obecnější model – ten by ale vyžadoval více prostoru pro diskuzi



První obrázek ukazuje graf modifikovaného Hillova odhadu v závislosti na k, tj. pro různá $\tau = [k/n]$. Byla opět simulována náhodná čísla z $|T_2|$, tentokrát však s malým lineárním posunutím $Y_i = 0.002X_i + |T_2|, X_i = 1, \ldots, 1000$. Modře je vyznačen modifikovaný Hillův odhad využívající kvantilové regrese s lineárním trendem, čerchovaně pak modifikovaný Hillův odhad založený na kvadratickém trendu, který je v této situaci zjevně neadekvátní (zbytečně široký). Pro srovnání je zakreslen obvyklý Hillův odhad – červeně a Hillův odhad aplikovaný na data, z nichž byl zpětně odstraněn lineární trend – černě.

Na dalším obrázku znovu vidíme grafy kvantilovou regresí modifikovaných odhadů – Hillův odhad (modře) a Momentový odhad (šedě). Tentokrát byla generována data z Paretova $Pa(4)$ rozdělení a přičtena ke kvadratickému trendu $Y_i = 1/20X_i^2 + 2X_i + |Pa(4)|$, kde $X_i = \{1, \ldots, 40, 1, \ldots, 40, \ldots, 1, \ldots, 40\}$ a $i = 1, \ldots, 1200$. Skutečná hodnota $\gamma = 0.25$ je vyznačena obvyklým způsobem černou čarou.

Výhled do budoucna

Prezentovaná metoda byla načrtnuta pouze v hrubých obrysech. Řada otázek tak zůstává otevřena. Například konzistence odhadů byla dosud dokázána pouze pro v praxi nepoužitelný Pickandsův odhad – viz Chernozhukov (2000). Tento důkaz však nelze použít pro jiné odhady. Východiskem by snad mohl být funkcionální přístup k odhadu γ popsaný v Drees (1998). Zcela není dořešena ani vlastní konstrukce odhadů např. důležitá problematika jejich invariance na posunutí, otázka τ versus $\frac{k}{n}$ atp. Pro aplikaci do praxe pak schází implementace do statistického softwaru (např. do `R`).

Literatura

[1] Beirlant J. et al. (2004). *Statistics of Extremes*, John Wiley, Chichester.

[2] Drees H. (1998). *On Smooth Statistical Tail Functionals*, Scandinavian Journal of Statistics **25**, 187−210.

[3] Chernozhukov V. (2000). *Conditional Extremes and Near-Extremes: Concepts, Estimation, and Economic Applications*, Stanford Phd.D. Dissertation, `http://www.mit.edu/~vchern/ced.ps`

[4] Koenker R. (2005). *Quantile Regression*, Cambridge University Press, Cambridge