Detekce strukturálních změn v regresi

Terézia Hatalová Tomáš Kremla

15. dubna 2025

Model

Předpokládejme regresní model:

$$Y_{in} = \boldsymbol{x}_{in}^\top \boldsymbol{\beta} + \boldsymbol{x}_{in}^\top \boldsymbol{\delta}_n \mathbb{I}\{i > m_n\} + \varepsilon_i, \quad i = 1, \ldots, n,$$

kde

- $m_n \leq n$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ a $\boldsymbol{\delta}_n = (\delta_{1n}, \ldots, \delta_{pn})^\top$ jsou neznámé parametry,
- $\boldsymbol{x}_{in} = (x_{i1,n}, \ldots, x_{ip,n})^\top$, $x_{i1,n} = 1$ jsou známé fixní (nenáhodné) hodnoty regresorů, ε_i jsou i.i.d. centrované náhodné veličiny s nenulovým rozptylem σ^2 a $\mathbb{E}|\varepsilon_i|^{2+\Delta_1} < \infty$ pro nějaké $\Delta_1 > 0$.

Parametr m_n nazveme bod změny (change point). Naším cílem je testovat:

$$H_0 : m_n = n \quad \text{proti} \quad H_1 : m_n < n.$$

Asymptotické chování za H_0

Předpokládejme:

- $\sum_{i=1}^{n} x_{ij,n} = 0$ pro $j = 2, \ldots, p$
- Existuje pozitivně definitní matice $\mathbf{C}$ typu $p \times p$ taková, že pro libovolnou posloupnost $\{l_n\}$, $\lim_{n\to\infty} l_n = \infty$, $l_n \leq n$ platí

$$\left\| \frac{1}{l_n} \left(\mathbf{C}_{k+l_n n} - \mathbf{C}_{kn} \right) - \mathbf{C} \right\| = o\left((\log l_n)^{-1} \right)$$

  stejnoměrně pro $1 \leq k \leq n - l_n$.

- Pro $n \to \infty$ platí

$$\max_{1 \leq k \leq n} \left\{ \frac{1}{k} \sum_{i=1}^{k} \|x_{in}\|^4 + \frac{1}{n-k} \sum_{i=k+1}^{n} \|x_{in}\|^4 \right\} = O(1).$$

Permutační testy

Využijeme faktu, že $(\varepsilon_1, \ldots, \varepsilon_n)$ jsou iid. Pro náhodnou permutaci $R = (R_1, \ldots, R_n)$ množiny $(1, \ldots, n)$, která je nezávislá na $(\varepsilon_1, \ldots, \varepsilon_n)$, platí

$$(\varepsilon_1, \ldots, \varepsilon_n) \stackrel{d}{=} (\varepsilon_{R_1}, \ldots, \varepsilon_{R_n}).$$

Permutační testy

Využijeme faktu, že $(\varepsilon_1, \ldots, \varepsilon_n)$ jsou iid. Pro náhodnou permutaci $R = (R_1, \ldots, R_n)$ množiny $(1, \ldots, n)$, která je nezávislá na $(\varepsilon_1, \ldots, \varepsilon_n)$, platí

$$(\varepsilon_1, \ldots, \varepsilon_n) \stackrel{d}{=} (\varepsilon_{R_1}, \ldots, \varepsilon_{R_n}).$$

Kritické hodnoty příslušné testové statistiky potom aproximujeme pomocí kvantilů podmíněné distribuce této statistiky. Tj např. pro statistiku T_n zamítáme pokud

$$T_n > d_n(1 - \alpha, \mathbf{Y}),$$

kde $d_n(1 - \alpha, \mathbf{Y})$ je $(1 - \alpha)$-kvantil rozdělení $T_n(R)|\mathbf{Y}$.

Permutační testy

Teoreticky lze kvantily $T_n(R)|\mathbf{Y}$ spočíst přesně. $n!$ je ale v praxi často moc velké. Proto vyhodnotíme $T_n(R)$ jen na B permutacích, pro B dostatečně velké (Monte-Carlo - principle). $d_n(1-\alpha, \mathbf{Y})$ potom aproximujeme empirickými kvantily.

Příklad - Ráztoka

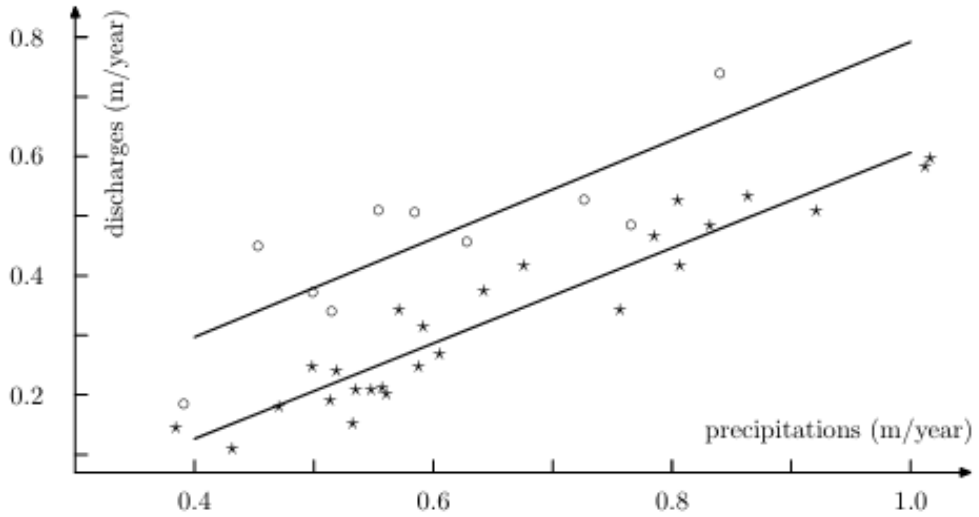

Figure 1. Malá Ráztoka: Data and model.

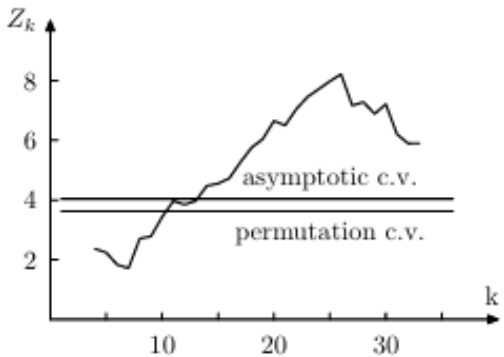

Figure 2a. Statistics Z_k.

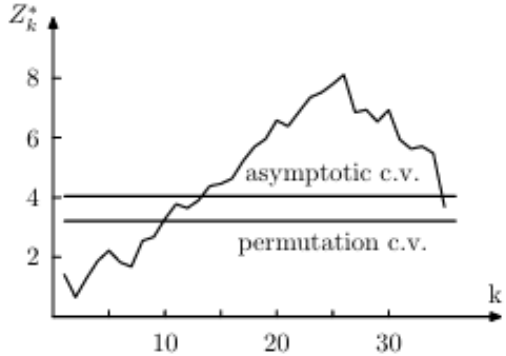

Figure 2b. Statistics Z_k^*.

Simulace

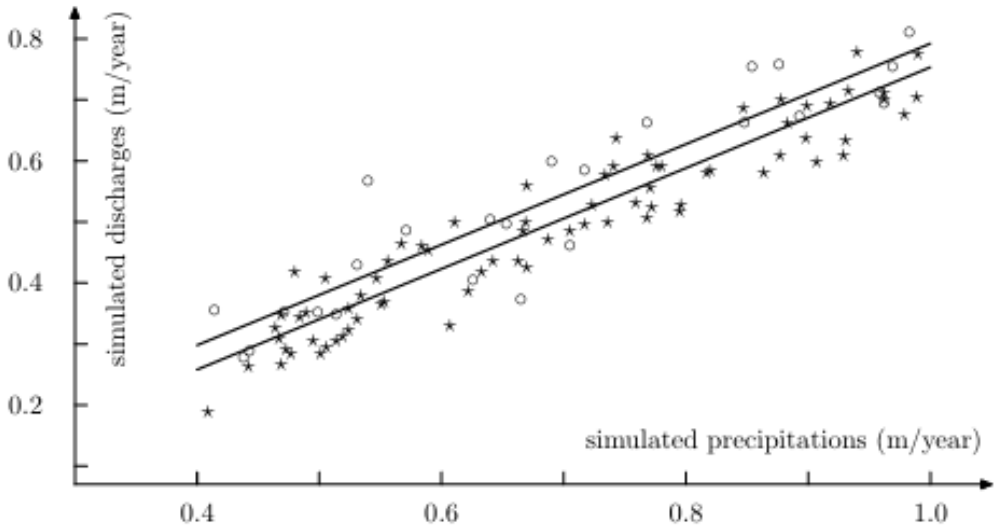

Figure 3. Simulated data and model.

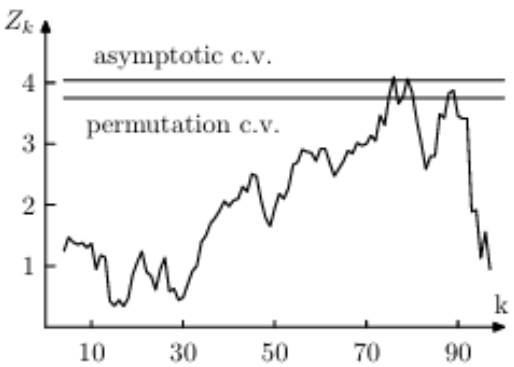

Figure 4a. Statistics Z_k.

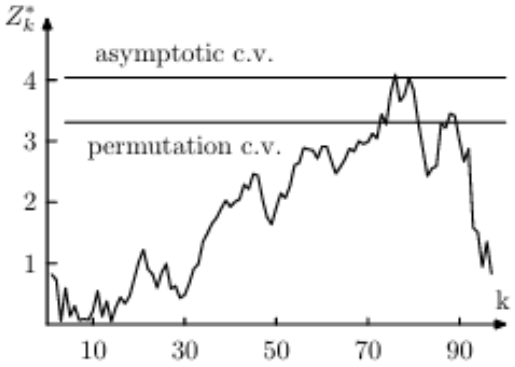

Figure 4b. Statistics Z_k^*.

Simulace

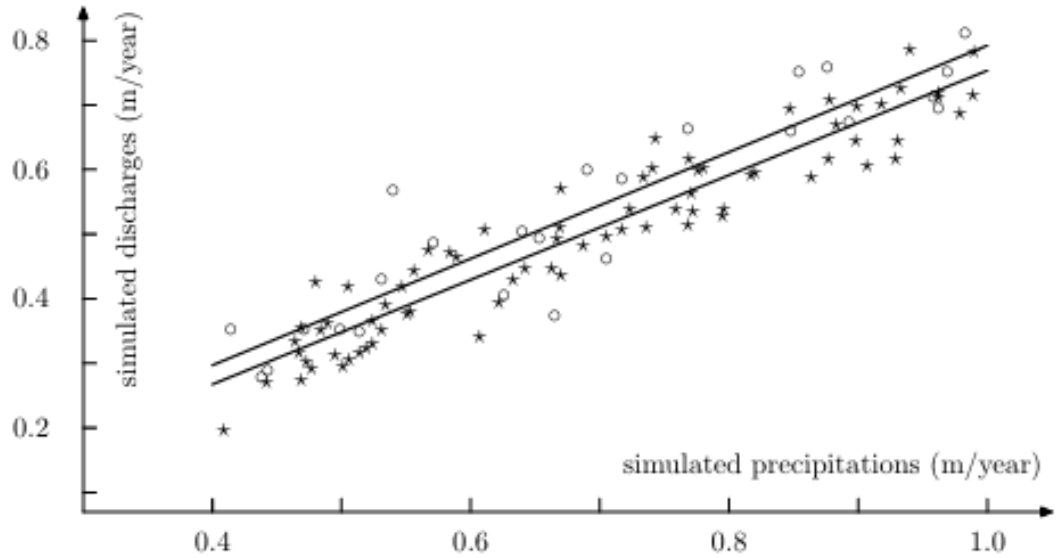

Figure 5. Simulated data and model.

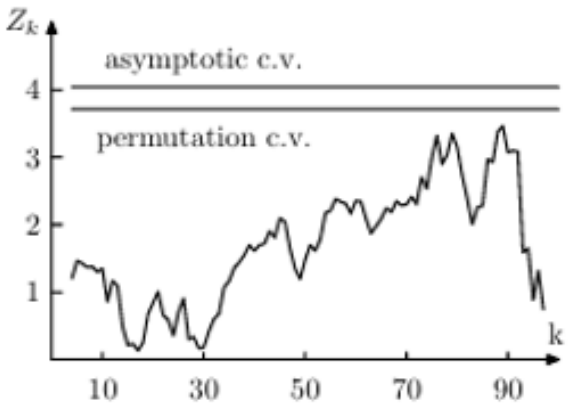

Figure 6a. Statistics Z_k.

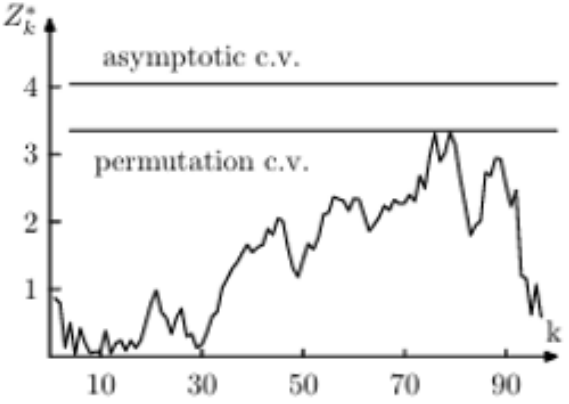

Figure 6b. Statistics Z_k^*.