Change point (bod zlomu)

Julie Váňová

12. dubna 2023

Statistický seminář 1

Motivace k problému

Uvažujeme regresní model

$$Y_{in} = \mathbf{x}_{in}^{\mathsf{T}}\beta + \mathbf{x}_{in}^{\mathsf{T}}\delta_n \cdot 1_{[i>m_n]} + \epsilon_i, \qquad i = 1, ..., n.$$

- $\mathbf{x}_{in} = (\mathbf{x}_{i1,n}, ..., \mathbf{x}_{ip,n})^{\mathsf{T}}$, $i = 1, ..., n$, - předem dané body
- $\epsilon_1, ..., \epsilon_n$ i.i.d. - chybové členy
- $\beta = (\beta_1, ..., \beta_p)^{\mathsf{T}}$ a $\delta_n = (\delta_{1n}, ..., \delta_{pn})^{\mathsf{T}} \neq \mathbf{0}$ - (neznámé) parametry
- $m_n \leq n$ - neznámý parametr (change point)

Motivace k problému

- chceme testovat hypotézu

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

- zajímá nás odhad m_n

Testové statistiky

- budeme uvažovat následující testové statistiky

$$T_n = \max_{p<k<n-p} \frac{S_{kn}^{\mathsf{T}} C_{kn}^{-1} C_{nn} C_{kn}^{o-1} S_{kn}}{\hat{\sigma}_n^2}$$

$$T_n(q) = \sup_{0<t<1} \frac{S_{\lfloor (n+1)t \rfloor n}^{\mathsf{T}} C_{nn}^{-1} S_{\lfloor (n+1)t \rfloor n}}{q^2(t)\hat{\sigma}_n^2}$$

$$T_n^* = \max_{1\leq k<n} \sqrt{\frac{n}{k(n-k)}} \cdot \frac{|S_{kn}^*|}{\hat{\sigma}_n}$$

$$T_n^*(q) = \sup_{0<t<1} \frac{|S_{\lfloor (n+1)t \rfloor n}^*|}{\sqrt{n} q(t) \hat{\sigma}_n}$$

Odhad rozptylu

- $\hat{\sigma}_n^2$ je odhad σ^2 splňující

$$\hat{\sigma}_n^2 - \sigma^2 = o_p((\log(\log n))^{-1/2}), \quad n \to \infty$$

-

$$\hat{\sigma}_n^2 = \frac{1}{n-p} \min_{p<k<n-p} \left(\sum_{i=1}^{k} (Y_{in} - \mathbf{x}_{in}^{\mathsf{T}} \hat{\beta}_k)^2 + \sum_{i=k+1}^{n} (Y_{in} - \mathbf{x}_{in}^{\mathsf{T}} \hat{\beta}_k^0)^2 \right),$$

kde $\hat{\beta}_k$ a $\hat{\beta}_k^0$ jsou LSE založené na $Y_1, ..., Y_k$ a $Y_{k+1}, ..., Y_n$

- dá se dokázat, že

$$\hat{\sigma}_n^2 = \frac{1}{n-p} \left(\sum_{i=1}^{n} (Y_{in} - \mathbf{x}_{in}^{\mathsf{T}} \hat{\beta}_n)^2 - \max_{p<k<n-p} (S_{kn}^{\mathsf{T}} C_{kn}^{-1} C_{nn} C_{kn}^{o-1} S_{kn}) \right)$$

- u T_n^* a $T_n^*(q)$ se rovněž používá

$$\tilde{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^{n} (Y_{in} - \mathbf{x}_{in}^{\mathsf{T}} \hat{\beta}_n)^2$$

Limitní chování T_n a T_n^*

Věta 1:

- Nechť platí předpoklady (A1)-(A3) a (B1). Potom za H_0

$$\lim_{n\to\infty} \mathrm{P}\big(a(\log n)\sqrt{T_n} \le t + b_p(\log n)\big) = \exp\left(-2e^{-t}\right),$$

$$\lim_{n\to\infty} \mathrm{P}\big(a(\log n)\, T_n^* \le t + b_1(\log n)\big) = \exp\left(-2e^{-t}\right),$$

kde

$$a(y) = \sqrt{2\log y}, \quad b_p(y) = 2\log y + \frac{p}{2}\log\log y - \log(\Gamma(\frac{p}{2})), \quad y > 1,$$

a $\Gamma(p) = \int_0^\infty t^{p-1}\mathrm{e}^{-t}\, dt.$

Asymptotické rozdělení $T_n(q)$ a $T_n^*(q)$

Věta 2:

- Nechť platí předpoklady (A1)-(A3), (B1) a (C1). Potom za H_0

$$\sqrt{T_n(q)} \xrightarrow[n\to\infty]{D} \sup_{0<t<1} \frac{\sqrt{\sum_{i=1}^{p} B_i^2(t)}}{q(t)},$$

$$T_n^*(q) \xrightarrow[n\to\infty]{D} \sup_{0<t<1} \frac{|B_1(t)|}{q(t)},$$

kde $\{B_i(t); t \in (0,1)\}_{i=1}^{p}$ jsou nezávislé Brownovy mosty.

Testování založené na permutačním principu

Věta 3:

- Nechť platí předpoklady (A1)-(A3) a (B1). Potom

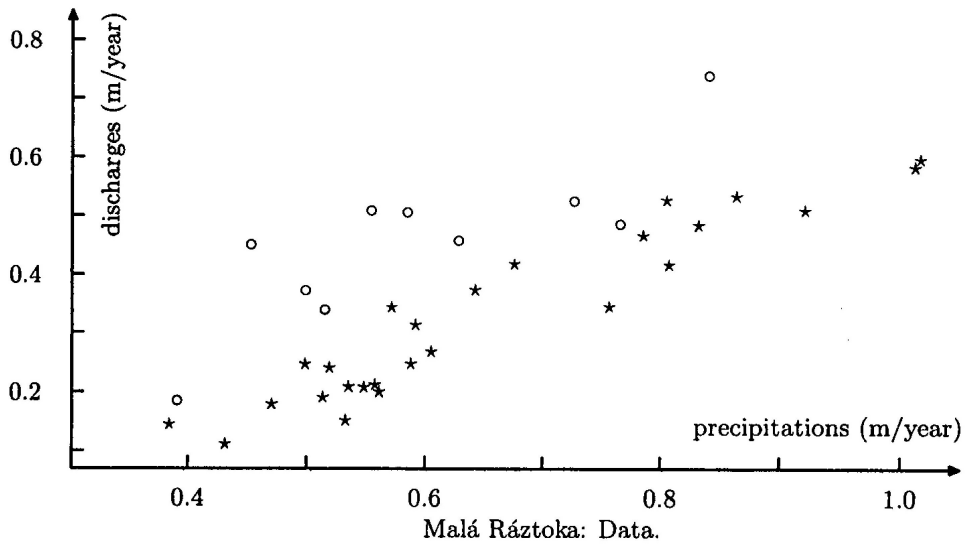
$$\lim_{n\to\infty} \mathrm{P}\left(a(\log n)\sqrt{T_n(\mathbf{R})} \leq t + b_p(\log n) | \mathbf{Y}_n \right) = \exp\left(-2e^{-t}\right)$$

v pravděpodobnosti, a

$$\lim_{n\to\infty} \mathrm{P}\left(a(\log n)\, T_n^*(\mathbf{R}) \leq t + b_1(\log n) | \mathbf{Y}_n \right) = \exp\left(-2e^{-t}\right)$$

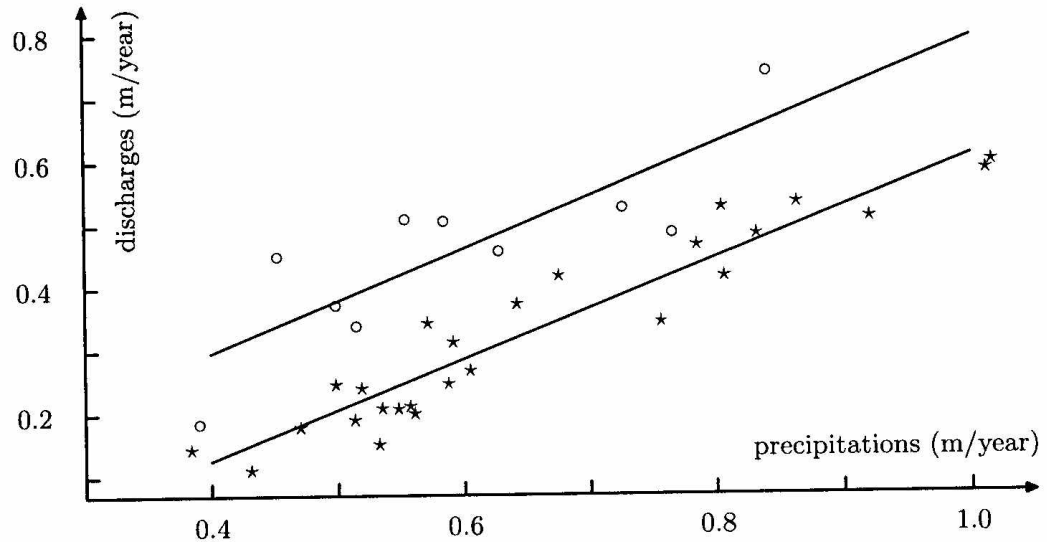
v pravděpodobnosti, kde $t \in \mathcal{R}_1$, a $a(y)$, $b_p(y)$ jsou jako ve větě 1.

Příklad

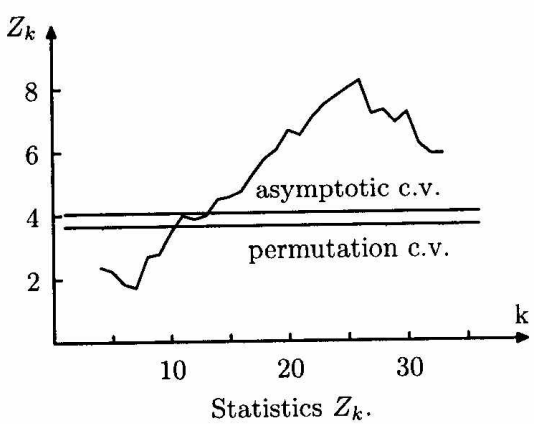


Malá Ráztoka: Data.

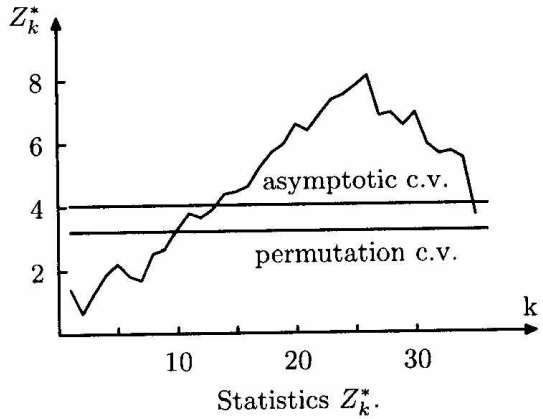
Příklad



Malá Ráztoka: Data and model.



Statistics Z_k.



Statistics Z_k^*.

Příklad - semi-simulační studie