Statistický seminář

Krejčová, Vojtek

Two-sample gradual change analysis

22. dubna 2025

Předpoklady

Předpoklady.

(A1) Pozorování Y_{jik}, $j \in \{1, 2\}$, $k \in \{1, \ldots, n_{ji}\}$ jsou získaná v čase i, $i \in \{1, \ldots, n\}$.

(A2) Všechna pozorování jsou nezávislá.

(A3) $\mathbb{E}(\overline{Y}_{1i} - \overline{Y}_{2i}) = \mu + \delta((i - k_0)/n)_+, i \in \{1, \ldots, n\}$, kde μ, δ jsou neznámé parametry a $k_0 = n\theta_0$ pro nějaké $\theta_0 \in (0, 1)$.

(A4) $\mathrm{Var}(Y_{jik}) = \sigma_{ji}^2 > 0$, $j \in \{1, 2\}$, $i \in \{1, \ldots, n\}$, $k \in \{1, \ldots, n_{ji}\}$.

Předpoklady

Předpoklady.

(A1) Pozorování Y_{jik}, $j \in \{1,2\}$, $k \in \{1,\dots,n_{ji}\}$ jsou získaná v čase i, $i \in \{1,\dots,n\}$.

(A2) Všechna pozorování jsou nezávislá.

(A3) $\mathbb{E}(\overline{Y}_{1i} - \overline{Y}_{2i}) = \mu + \delta((i-k_0)/n)_+, i \in \{1,\dots,n\}$, kde μ, δ jsou neznámé parametry a $k_0 = n\theta_0$ pro nějaké $\theta_0 \in (0,1)$.

(A4) $\mathrm{Var}(Y_{jik}) = \sigma_{ji}^2 > 0$, $j \in \{1,2\}$, $i \in \{1,\dots,n\}$, $k \in \{1,\dots,n_{ji}\}$.

Předpoklad homoskedasticity.

(A4*) $\mathrm{Var}(\overline{Y}_{1i} - \overline{Y}_{2i}) = \sigma^2/m$, $i \in \{1,\dots,n\}$, kde $\sigma^2 > 0$ je neznámý parametr a m může záviset na n.

Homoskedastický případ - odhady

$$\hat{k}_\mu = \arg \max_{k \in \{1,\ldots,n\}} \left\{ \frac{\left[\sum_{i=1}^n (x_{ik} - \bar{x}_k)(\bar{Y}_{1i} - \bar{Y}_{2i})\right]^2}{\sum_{i=1}^n (x_{ik} - \bar{x}_k)^2} \right\}$$

$$\hat{\delta}_k = \frac{\sum_{i=1}^n (x_{i\hat{k}} - \bar{x}_{\hat{k}})(\bar{Y}_{1i} - \bar{Y}_{2i})}{\sum_{i=1}^n (x_{i\hat{k}} - \bar{x}_{\hat{k}})^2}$$

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n (\bar{Y}_{1i} - \bar{Y}_{2i}) - \hat{\delta}_\mu \bar{x}_{\hat{k}}$$

$$\hat{k}_\mu = \arg \max_{k \in \{1,\ldots,n\}} \left\{ \frac{\left[\sum_{i=1}^n (x_{ik} - \bar{x}_k)(\bar{Y}_{1i} - \bar{Y}_{2i})\right]^2}{\sum_{i=1}^n (x_{ik} - \bar{x}_k)^2} \right\}$$

$$\hat{\delta}_k = \frac{\sum_{i=1}^n (x_{i\hat{k}} - \bar{x}_{\hat{k}})(\bar{Y}_{1i} - \bar{Y}_{2i})}{\sum_{i=1}^n (x_{i\hat{k}} - \bar{x}_{\hat{k}})^2}$$

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n (\bar{Y}_{1i} - \bar{Y}_{2i}) - \hat{\delta}_\mu \bar{x}_{\hat{k}}$$

$$\hat{k}_0 = \arg \max_{k \in \{1,\ldots,n\}} \left\{ \frac{\left[\sum_{i=1}^n x_{ik}(\bar{Y}_{1i} - \bar{Y}_{2i})\right]^2}{\sum_{i=1}^n x_{ik}^2} \right\}$$

$$\hat{\delta}_0 = \frac{\sum_{i=1}^n x_{i\hat{k}}(\bar{Y}_{1i} - \bar{Y}_{2i})}{\sum_{i=1}^n x_{i\hat{k}}^2}$$

Homoskedastický případ - asymptotické rozdělení

Věta 1

Nechť platí (A1)-(A3) a (A4*). Potom platí

$$\sqrt{nm}\frac{\delta}{\sigma}\sqrt{\frac{\theta_0(1-\theta_0)}{1+3\theta_0}}\frac{\hat{k}_\mu - k_0}{n} \xrightarrow{D} N(0,1),$$

$$\sqrt{nm}\frac{1}{\sigma}\sqrt{\frac{(1-\theta_0)^3(1+3\theta_0)}{12}}(\hat{\delta}_\mu - \delta) \xrightarrow{D} N(0,1).$$

Homoskedastický případ - asymptotické rozdělení

Značení

$$A_k = \frac{\left(\sum_{i=1}^n (x_{ik} - \overline{x}_k)\epsilon_i\right)^2}{\sum_{i=1}^n (x_{ik} - \overline{x}_k)^2} - \frac{\left(\sum_{i=1}^n (x_{ik_0} - \overline{x}_{k_0})\epsilon_i\right)^2}{\sum_{i=1}^n (x_{ik_0} - \overline{x}_{k_0})^2},$$

$$B_k = \frac{\left(\sum_{i=1}^n (x_{ik} - \overline{x}_k)\epsilon_i\right)\left(\sum_{i=1}^n (x_{ik} - \overline{x}_k)x_{ik_0}\right)}{\sum_{i=1}^n (x_{ik} - \overline{x}_k)^2} - \sum_{i=1}^n (x_{ik_0} - \overline{x}_{k_0})\epsilon_i,$$

$$C_k = \frac{\left(\sum_{i=1}^n (x_{ik} - \overline{x}_k)x_{ik_0}\right)^2}{\sum_{i=1}^n (x_{ik} - \overline{x}_k)^2} - \sum_{i=1}^n (x_{ik_0} - \overline{x}_{k_0})^2.$$

Posloupnost $\{r_n\}$ splňuje, pro $n \to \infty$,

$$r_n \to \infty, \quad \frac{|\delta|\sqrt{n}}{r_n\sqrt{\log\log n}} \to \infty.$$

$$Z_n = \sum_{i=k_0+1}^n (\epsilon_i - \overline{\epsilon}_n)x_{ik_0} - \frac{n\theta(1-\theta)^2}{2\sum_{i=1}^n (x_{ik_0} - \overline{x}_{k_0})^2}\sum_{i=1}^n (\epsilon_i - \overline{\epsilon}_n)x_{ik_0}.$$

Lemma

Nechť platí (A1)-(A3) a (A4*). Potom platí

$$C_k = -\frac{(k_0 - k)^2}{n}\frac{\theta(1-\theta)}{1+3\theta}\left(1 + o\left(\frac{k_0 - k}{n}\right)\right)),$$

$$\max_{|k_0 - k| \le r_n|\delta|^{-1}\sqrt{n}} \left\{ \frac{\sqrt{n}}{(k_0 - k)|\delta|}|A_k| \right\} = o_p(1),$$

$$\max_{|k_0 - k| \le r_n|\delta|^{-1}\sqrt{n}} \left\{ \frac{B_k\sqrt{n}}{k_0 - k} - Z_n\frac{1}{\sqrt{n}} \right\} = o_p(1),$$

$$\frac{1}{n}\sum_{i=1}^{n}(x_{ik_0} - \overline{x}_{k_0})^2 = \frac{(1-\theta)^3}{3} - \frac{(1-\theta)^4}{4} + O(n^{-1}).$$

Homoskedastický případ - asymptotické rozdělení

Věta 2

Nechť platí (A1)-(A3) a (A4*) a $\mu = 0$. Potom platí

$$\sqrt{nm}\frac{\delta}{\sigma}\sqrt{\frac{1-\theta_0}{4}}\frac{\hat{k}_0 - k_0}{n} \xrightarrow{D} N(0,1),$$

$$\sqrt{nm}\frac{1}{\sigma}\sqrt{\frac{(1-\theta_0)^3}{3}}(\hat{\delta}_0 - \delta) \xrightarrow{D} N(0,1).$$

Heteroskedastický případ - odhady

$$\hat{k}_0(\tau^2) = \arg \max_{k \in (1,n)} \left\{ \frac{\left(\sum_{i=1}^{n} x_{ik}(\bar{Y}_{1i} - \bar{Y}_{2i})/\tau_i^2\right)^2}{\sum_{i=1}^{n} x_{ik}^2/\tau_i^2} \right\}$$

Heteroskedastický případ - odhady

$$\hat{k}_0(\tau^2) = \arg \max_{k \in (1,n)} \left\{ \frac{\left(\sum_{i=1}^{n} x_{ik}(\bar{Y}_{1i} - \bar{Y}_{2i})/\tau_i^2\right)^2}{\sum_{i=1}^{n} x_{ik}^2/\tau_i^2} \right\}$$

$$\hat{k}_0(\hat{\tau}^2) = \arg \max_{k \in (1,n)} \left\{ \frac{\left(\sum_{i=1}^{n} x_{ik}(\bar{Y}_{1i} - \bar{Y}_{2i})/\hat{\tau}_i^2\right)^2}{\sum_{i=1}^{n} x_{ik}^2/\hat{\tau}_i^2} \right\} = \arg \max_{k \in (1,n)} T_{2,\hat{\tau}^2}(k)$$

Bootstrap algoritmus

- $\widehat{D}_i = \widehat{\delta}_0(\frac{i-\widehat{k}}{n})_+$
- $b \in \{1, \ldots, B\}$ $D^*_{ib} = \widehat{D}_i + \widehat{\tau}_i \varepsilon_{ib}$, kde $\varepsilon_{ib} \sim N(0, 1)$
- Výpočet $\widehat{k}^*_b$ z $D^*_{1b}, \ldots, D^*_{nb}$
- Výpočet kvantilů q^*_α z $\widehat{k}^*_1 - \widehat{k}, \ldots, \widehat{k}^*_B - \widehat{k}$

Tvrzení 4.1.

Mějme $H_0 : k_0 \geq k_1$ a $H_1 : k_0 < k_1$, pak pomocí náhodné veličiny $K \sim \text{dist.}(\widehat{k} - k_0)$ definujeme p-hodnota $P(K < \widehat{k} - k_0)$, a jako její aproximaci $\sum_{b=1}^{B} I(\widehat{k_b^*} - \widehat{k} < \widehat{k} - k_1)$.

Simulační studie

Algoritmus

1. Vyberme si n a $\theta_0 = k_0/n$.
2. Vyberme σ^2_{1i}. σ^2_{2i}, n_{1i} a n_{2i}.
3. Vypočtěme $\tau_i^2 = \frac{\sigma^2_{1i}}{n_{1i}} + \frac{\sigma^2_{2i}}{n_{2i}}$.
4. $s \in \{1, \dots, S\}$, pro $\forall i$ vygenerujeme $D_i = \bar{Y}_{1i} - \bar{Y}_{2i} \sim N((i - k_0)_+, \tau_i^2)$.
5. Vygenerujeme $\widehat{\tau}_i^2 \sim \frac{\sigma^2_{1i}\chi^2_{n_{1i}-1}}{n_{1i}(n_{1i}-1)} + \frac{\sigma^2_{2i}\chi^2_{n_{2i}-1}}{n_{2i}(n_{2i}-1)}$.
6. Vypočteme $\widehat{k}_0^{(s)}$ a k tomu interval spolehlivosti, strannost (bais), a MSE.

Homoskedastické simulace

		θ_0	$\widehat{\sigma}^2_{\text{pooled}}$				$\widehat{\sigma}^2_{ji}$			
			$\widehat{k}_\mu$	$\widehat{k}_0$	$\widehat{k}_\mu^{\text{corr}}$	$\widehat{k}_0(\widehat{\tau}^2)$	$\widehat{k}_\mu$	$\widehat{k}_0$	$\widehat{k}_\mu^{\text{corr}}$	$\widehat{k}_0(\widehat{\tau}^2)$
$n = 10$	$n_{ji} = 10$	0.1	88.8	95.2	93.3	94.0	89.5	93.9	93.7	94.5
		0.2	92.4	95.9	92.8	94.9	91.7	94.8	94.3	95.6
		0.4	94.1	92.3	92.9	91.9	95.5	92.4	93.3	92.0
		0.6	92.8	93.2	92.6	92.4	93.9	89.9	92.2	90.2
		0.8	90.4	90.5	90.0	90.8	89.6	87.7	89.2	89.1
		0.9	78.1	78.3	79.2	78.4	80.1	74.8	76.4	76.0
	$n_{ji} = 20$	0.1	93.9	92.0	94.4	92.1	95.1	92.8	94.6	93.0
		0.2	96.3	92.8	95.6	92.4	95.4	92.7	95.8	94.7
		0.4	93.5	92.0	92.3	91.1	92.7	91.6	92.0	90.8
		0.6	87.6	90.1	90.1	89.8	89.3	87.1	88.9	88.4
		0.8	88.8	89.0	89.7	87.8	89.9	86.9	87.2	88.4
		0.9	72.1	70.4	74.9	70.1	72.4	70.9	72.6	70.3
$n = 20$	$n_{ji} = 10$	0.1	96.5	94.3	94.0	94.9	94.9	93.5	95.8	93.1
		0.2	97.1	94.1	95.0	93.7	96.9	93.0	95.0	93.6
		0.4	94.0	93.7	93.8	93.9	94.4	92.1	93.0	92.8
		0.6	93.2	90.9	92.7	92.7	91.9	92.1	91.7	91.8
		0.8	94.8	95.6	94.3	95.3	93.8	94.5	92.5	93.7
		0.9	84.1	84.3	84.8	84.0	83.1	81.8	84.4	80.9
	$n_{ji} = 20$	0.1	97.3	95.0	94.4	94.9	97.0	93.5	93.4	95.3
		0.2	95.1	94.3	94.1	94.3	93.5	93.9	94.1	94.0
		0.4	93.0	93.1	93.1	92.9	93.6	93.6	93.1	94.7
		0.6	91.9	90.7	92.8	92.8	91.7	93.6	92.0	91.4
		0.8	93.2	91.9	91.3	90.7	91.8	92.5	91.5	89.3
		0.9	79.5	81.4	83.4	79.3	82.3	80.4	82.4	82.4

	θ_0	$n = 10$								$n = 20$			
		$\widehat{\sigma}^2_{\text{pooled}}$				$\widehat{\sigma}^2_{ji}$				$\widehat{\sigma}^2_{ji}$			
		$\widehat{k}_\mu$	$\widehat{k}_0$	$\widehat{k}_\mu^{\text{corr}}$	$\widehat{k}_0(\widehat{\tau}^2)$	$\widehat{k}_\mu$	$\widehat{k}_0$	$\widehat{k}_\mu^{\text{corr}}$	$\widehat{k}_0(\widehat{\tau}^2)$	$\widehat{k}_\mu$	$\widehat{k}_0$	$\widehat{k}_\mu^{\text{corr}}$	$\widehat{k}_0(\widehat{\tau}^2)$
H01	0.1	65.1	66.8	58.0	67.9	88.1	92.5	92.5	93.2	97.5	93.1	95.1	93.8
	0.4	69.9	68.9	67.7	70.7	96.3	94.7	95.7	92.2	94.1	93.3	94.4	95.1
	0.8	74.7	71.2	72.2	68.5	88.5	86.5	86.7	88.4	95.9	94.9	95.2	91.4
	0.9	84.1	82.8	77.9	80.5	78.6	77.4	78.8	78.1	77.5	80.4	80.8	78.9
H02	0.1	60.5	63.2	50.6	65.6	83.7	94.2	92.5	93.9	95.3	93.0	95.0	93.1
	0.4	65.9	65.4	63.9	69.6	90.6	88.4	91.0	93.0	93.6	92.5	92.8	93.2
	0.8	69.4	69.5	74.1	72.9	90.4	89.2	91.0	86.3	89.6	88.4	90.7	90.4
	0.9	80.0	78.2	77.2	83.7	76.8	71.9	75.6	75.8	82.6	84.2	82.3	81.9
H10	0.1	90.0	93.9	92.4	94.0	88.8	94.6	93.4	92.8	92.6	94.0	93.1	95.3
	0.4	97.1	99.6	99.7	99.7	92.7	91.5	94.3	91.1	94.6	94.9	94.9	93.3
	0.8	89.0	93.6	93.3	92.2	91.6	92.5	90.7	89.0	95.3	91.4	94.8	90.5
	0.9	94.1	87.5	87.5	68.5	92.4	88.8	86.1	73.4	87.5	86.6	89.1	82.3
H11	0.1	65.6	65.4	55.6	69.9	89.9	93.1	91.8	93.1	90.4	95.5	91.7	94.2
	0.4	71.0	67.2	67.4	70.8	92.4	94.8	93.2	91.9	93.7	92.4	93.1	93.3
	0.8	79.0	78.7	76.1	71.0	92.2	84.3	90.7	89.8	96.8	95.9	96.4	89.3
	0.9	95.5	92.5	84.6	73.1	93.1	90.3	87.5	66.2	88.4	86.8	86.3	80.1
H12	0.1	58.9	63.9	49.9	64.0	88.3	95.6	92.0	94.2	88.8	94.4	92.4	93.1
	0.4	70.3	68.3	65.3	69.5	90.0	87.1	88.6	91.5	94.5	93.0	94.7	92.9
	0.8	79.6	77.5	78.8	72.2	91.1	90.1	92.7	87.9	93.9	88.5	91.3	88.8
	0.9	93.2	88.3	81.7	78.0	91.3	85.2	85.7	74.3	88.6	87.3	90.9	82.2
H20	0.1	80.9	92.4	91.0	99.4	82.1	91.5	88.0	98.5	97.0	97.2	98.9	99.3
	0.4	93.8	94.6	93.2	99.8	93.0	89.6	89.9	93.7	97.4	95.4	96.6	99.2
	0.8	78.0	75.5	79.1	74.1	78.2	77.4	76.2	73.8	93.2	90.6	91.1	94.1
	0.9	90.6	86.2	84.8	78.3	90.0	86.5	83.3	77.5	79.0	78.4	83.7	74.4
H21	0.1	58.8	63.8	48.8	66.9	76.6	88.7	80.7	94.5	87.8	87.4	87.0	93.3
	0.4	60.6	62.8	61.9	65.2	83.9	82.0	83.3	88.4	85.0	85.4	84.2	90.8
	0.8	73.3	70.7	68.0	69.3	79.9	79.4	76.9	79.8	84.7	84.8	84.3	88.1
	0.9	93.3	88.7	81.8	82.5	87.6	85.4	81.3	78.9	81.1	81.9	79.7	77.4
H22	0.1	49.8	58.4	39.4	69.1	61.6	84.4	73.7	94.8	79.0	83.2	80.4	93.0
	0.4	57.1	61.2	56.0	72.5	74.1	68.2	75.6	90.2	78.0	81.9	81.0	93.3
	0.8	72.6	67.5	72.1	67.6	82.6	79.0	74.9	73.5	76.7	75.4	75.5	89.4
	0.9	92.2	86.6	79.2	79.9	90.3	83.7	83.2	82.1	83.9	78.7	81.9	76.8

	Nr. of observations (n_{ji})		
	$n_{ji} = m$	$m\{1 + 3I(i \text{ odd})\}/2$	$m\{1 + 3I(i > n/2)\}/2$
σ_{ji} constant ($\sigma_{ji} = \sigma$)		H01	H02
$\sigma_{ji} = \sigma(1 + 2I(i > k_0))$	H10	H11	H12
$\sigma_{ji} = \sigma(1 + 2I(i \text{ even}))$	H20	H21	H22

Age cat.	girls $\overline{Y}_1$ ($\hat{\sigma}_1$)	girls n_1	boys $\overline{Y}_2$ ($\hat{\sigma}_2$)	boys n_2	p-values: t-test	p-values: Bonferroni	p-values: BH	p-values: $\hat{k}_0(\hat{\tau}^2)$	p-values: $\hat{k}_0^{bc}(\hat{\tau}^2)$	Age
6–7	1.89 (0.17)	33	1.87 (0.18)	19	0.780	1.000	0.780	1.000	1.000	6
7–8	2.00 (0.21)	43	1.98 (0.20)	38	0.646	1.000	0.763	1.000	1.000	7
8–9	2.01 (0.21)	33	2.06 (0.21)	38	0.369	1.000	0.479	1.000	1.000	8
9–10	2.06 (0.18)	42	2.14 (0.18)	29	0.081.	1.000	0.117	0.999	0.997	9
10–11	2.19 (0.22)	42	2.17 (0.19)	45	0.713	1.000	0.773	0.861	0.846	10
11–12	2.23 (0.15)	30	2.31 (0.23)	37	0.062.	0.800	0.100	0.113	0.117	11
12–13	2.26 (0.13)	41	2.35 (0.23)	40	0.047*	0.615	0.088.	0.003**	0.003**	12
13–14	2.30 (0.22)	32	2.53 (0.21)	36	0.000***	0.001***	0.000***	0.000***	0.000***	13
14–15	2.28 (0.23)	31	2.66 (0.19)	20	0.000***	0.000***	0.000***	0.000***	0.000***	14
15–16	2.37 (0.17)	29	2.72 (0.22)	26	0.000***	0.000***	0.000***	0.000***	0.000***	15
16–17	2.33 (0.19)	17	2.83 (0.28)	9	0.001***	0.006**	0.001**	0.000***	0.000***	16
17–18	2.35 (0.18)	25	2.76 (0.16)	13	0.000***	0.000***	0.000***	0.000***	0.000***	17
18–19	2.33 (0.17)	34	2.87 (0.10)	14	0.000***	0.000***	0.000***	0.000***	0.000***	18

Jumping speed