Методы понижения дисперсии

① Ложная регрессия

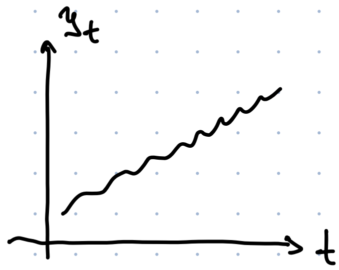
y_t – потребление эл. в Тайланде 1980 – 2020

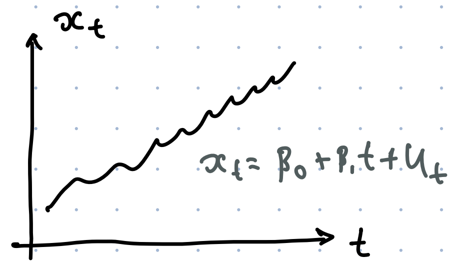
x_t – рекорды по прыжкам в высоту

$y_t = w_0 + w_1 x_t$ $\qquad w_1 > 0$

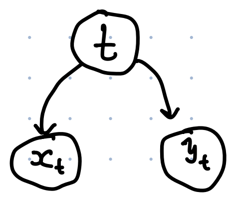
$H_0: w_1 = 0$ $\qquad$ значимая? Да!

$H_A: w_1 \neq 0$





В данных тренд!



Лечение проблемы:

а) $y'_t = y_t - (\alpha_0 + \alpha_1 \cdot t)$

$x'_t = x_t - (\beta_0 + \beta_1 \cdot t)$

$y'_t = w_0 + w_1 x'_t$

$H_0: w_1 = 0$ $\qquad$ Коэф. будет незначим

$H_A: w_1 \neq 0$

б) $\Delta x_t = x_t - x_{t-1}$ $\qquad$ Исслед. в приростах

$\Delta y_t = y_t - y_{t-1}$

в) $y_t = w_0 + w_1 \cdot x_t + w_2 \cdot t$ $\qquad$ Контрольная переменная

② Т. Фриша-Во-Ловела

$y_i \quad x_i \quad z_i$ $\qquad \boxed{Cov(x_i, z_i) = 0}$

$$y_i = \hat{w}_0 + \hat{w}_1 x_i + \hat{w}_2 z_i$$

① $y_i = \hat{\alpha}_0 + \hat{\alpha}_1 x_i$ $\qquad$ ② $y_i' = \hat{\gamma}_0 + \hat{\gamma}_2 z_i$

$y_i' = y_i - (\hat{\alpha}_0 + \hat{\alpha}_1 x_i)$

$$\Rightarrow \boxed{\hat{\gamma}_2 = \hat{w}_2}$$

③ АБ через линейную модель.

$y_1^c, \dots, y_{n_c}^c$ тут нет изм. μ_c σ_c^2

$y_1^t, \dots, y_{n_t}^t$ есть изм. μ_t σ_t^2

$H_0: \mu_c = \mu_t$

$H_A: \mu_c \neq \mu_t$

$\quad \gtrless$

y_i – метрика

$d_i = \begin{cases} 1, & \text{если набл. } i \text{ в treatment группе} \\ 0, & \text{иначе} \end{cases}$

$y_i = \hat{w}_0 + \hat{w}_1 d_i$ $\qquad d_i = 0 \quad \hat{y}_i = \hat{w}_0$

$\qquad d_i = 1 \quad \hat{y}_i = \hat{w}_0 + \hat{w}_1$

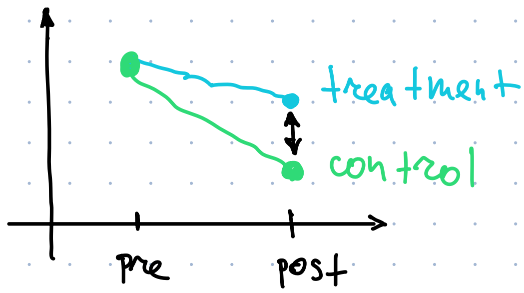
$H_0: w_1 = 0$

$H_A: w_1 \neq 0$

$\quad \gtrless$

④ DnD (difference in difference)

	Pre	Post
cont		
treat		



$$\hat{dd} = \left(\bar{y}^{t}_{post} - \bar{y}^{t}_{pre}\right) - \left(\bar{y}^{c}_{post} - \bar{y}^{c}_{pre}\right)$$

$H_0: dd = 0$

$H_A: dd \neq 0$ ($\gtrless$)

Это всё опять линейная регрессия

y_i - метрика

$$d_i = \begin{cases} 1 - \text{treatment} \\ 0 - \text{control} \end{cases}$$

$$z_i = \begin{cases} 1 - \text{post} \\ 0 - \text{pre} \end{cases}$$

$$y_i = w_0 + w_1 d_i + w_2 z_i + w_3 d_i z_i + \varepsilon_i$$

$$\hat{w}_0 = \bar{y}_{control,\, pre}$$

$$\hat{w}_1 = \bar{y}_{treat,\, pre} - \bar{y}_{contr.,\, pre}$$

$$\hat{w}_2 = \bar{y}_{contr.,\, post} - \bar{y}_{contr,\, pre}$$

$$\hat{w}_3 = \hat{dd} = \left(\bar{y}^{t}_{post} - \bar{y}^{t}_{pre}\right) - \left(\bar{y}^{c}_{post} - \bar{y}^{c}_{pre}\right)$$

$H_0: w_3 = 0$

$H_A: w_3 \neq 0$ ($\gtrless$)

⑤ CUPED (2013) Alex Deng

Controlled-experiment using pre-experiment data

$$Z = \frac{\bar{y}^t - \bar{y}^c}{se(\bar{y}^t - \bar{y}^c)} \qquad n_c \quad n_t$$

CUPED хочет ↓ se
не увеличивая размер
выборки

$$\Delta \in \left[\bar{y}^t - \bar{y}^c - 1.96 \cdot se(\ldots) ; \ \bar{y}^t - \bar{y}^c + 1.96 \cdot se(\ldots) \right]$$

$$\Delta = \mathbb{E}(\bar{y}^t) - \mathbb{E}(\bar{y}^c)$$

y Найдём предиктор x, который не влияет
на разбиение control/treatment

$$\hat{y} = y - \theta \cdot (x - \mathbb{E}(x)) \qquad \hat{\theta} = \frac{s\mathrm{Cov}(x, y)}{s\mathrm{Var}(x)}$$

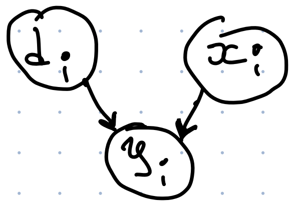
$$\left[\bar{\hat{y}}^t - \bar{\hat{y}}^c - 1.96 \cdot se(\ldots) ; \ \bar{\hat{y}}^t - \bar{\hat{y}}^c + 1.96 \cdot se(\ldots) \right] \downarrow$$

$$\mathbb{E}(\hat{y}) = \mathbb{E}(y) - \theta \cdot (\mathbb{E}(x) - \mathbb{E}(x)) = \mathbb{E}(y)$$

т.е. оценка $\hat{\Delta} = \bar{\hat{y}}^t - \bar{\hat{y}}^c$ всё ещё несмещ.

$$\mathrm{Var}(\hat{y}) = \mathrm{Var}(y) + \mathrm{Var}(x) \cdot \theta^2 - 2\theta \cdot \mathrm{Cov}(y, x)$$

$d_i \in \{0, 1\}$ рандомно нез. от x



Не CUPED

$$y_i = w_0 + w_1 d_i + \varepsilon_i \qquad \left[\begin{array}{l} \varepsilon_i \sim \text{гомоск.} \\ \varepsilon_i \sim \text{гетер.} \end{array}\right.$$

МНК $\hat{w}_1$ CI $H_0: w_1 = 0$

$H_A: w_1 \neq 0$

Добавим щепотку элегантности

$$y_i = w_0 + w_1 d_i + \underbrace{f(x_i)}_{\text{контрольные пер.}} + \varepsilon_i$$

$$+ \, w_2 x_i$$

МНК $\hat{w}_1$ CI $H_0: w_1 = 0$

$H_A: w_1 \neq 0$

итоговое se для $\hat{w}_1$ будет меньше

CUPED

Выжмем максимум из предпосылки нез. d_i и $f(x_i)$

$$Var(\hat{w} \mid X) = (X^T X)^{-1} \cdot \sigma^2 \qquad X = \begin{bmatrix} 1 & d_1 & x_1 \\ 1 & d_2 & x_2 \\ \vdots & \vdots & \vdots \\ 1 & d_n & x_n \end{bmatrix}$$

$$\hat{y}_i = \hat{w}_0 \cdot 1 + \hat{w}_1 d_i + \hat{w}_2 x_i$$

$$\mathrm{Var}(\hat{w} \mid X) = \left(\frac{X^T X}{n}\right)^{-1} \cdot \frac{\sigma^2}{n}$$

$$\frac{X^T X}{n} = \begin{bmatrix} \frac{\sum 1}{n} & \frac{\sum d_i}{n} & \frac{\sum x_i}{n} \\ \frac{\sum d_i}{n} & \boxed{\frac{\sum d_i}{n}} & \boxed{\frac{\sum d_i x_i}{n}} \\ \frac{\sum x_i}{n} & \boxed{\frac{\sum d_i x_i}{n}} & \boxed{\frac{\sum x_i^2}{n}} \end{bmatrix}$$

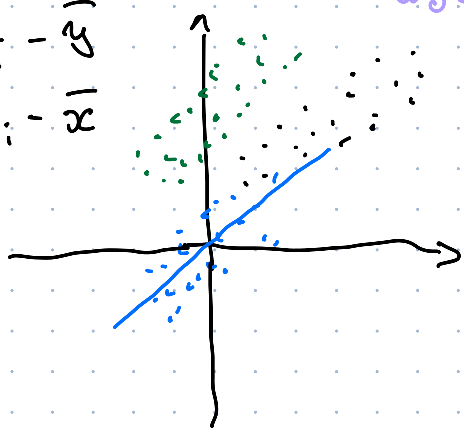
$\mathrm{Cov}(d_i, x_i) = 0$

При больших n в матрице $\mathrm{Var}(\hat{w} \mid X)$ мы знаем ковар.

Сделаем вид что данные центрированы и $w_0 = 0$

$y'_i = y_i - \bar{y}$

$x'_i = x_i - \bar{x}$



Отдельное упражнение

$$\mathrm{Var}(\hat{w} \mid X) = \begin{bmatrix} \frac{\sum d_i}{n} & \frac{\sum d_i x_i}{n} \\ \frac{\sum d_i x_i}{n} & \frac{\sum x_i^2}{n} \end{bmatrix}^{-1} \cdot \frac{\sigma^2}{n}$$

$$\mathrm{Var}_{cuped}(\hat{w} \mid X) = \begin{bmatrix} \cdots & 0 \\ 0 & \cdots \end{bmatrix}^{-1} \cdot \frac{\sigma^2}{n} =$$

$$= \frac{1}{\frac{\sum_i d_i \cdot \sum_i x_i^2}{n^2}} \cdot \frac{\sigma^2}{n} \cdot \begin{bmatrix} \frac{\sum_i x_i^2}{n} & 0 \\ 0 & \frac{\sum_i d_i}{n} \end{bmatrix} = \begin{bmatrix} \frac{\sigma^2}{\sum_i d_i} & \\ & \frac{\sigma^2}{\sum_i x_i^2} + \blacksquare \end{bmatrix}$$

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad - bc} \cdot \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

CUPED в виде линейной регрессии:

1. $\hat{y}_i = \hat{w}_0 + \hat{w}_1 d_i + \hat{w}_2 \cdot x_i$

   $r_i = y_i - \underline{\hat{w}_2 \cdot x_i}$

2. $\hat{r}_i = \hat{\gamma}_0 + \hat{\gamma}_1 \cdot d_i$

$\gamma_1 = \hat{w}_1$

$\hat{\gamma}_0 = \hat{w}_0$

$se(\hat{\gamma}_1)$ ●

Трюк

$se(\hat{w}_1)$ она будет не той ● $+$ $\blacksquare$

Это способ сказать питону что $Cov(d_i, x_i) = 0$

⊕ Площадка для расширения метода

А можно ли заиспользовать вместо одного x_i сразу много?

→ CUMPED

① $\hat{y}_i = \hat{w}_0 + \hat{w}_1 d_i + (\hat{w}_2 \cdot x_i + \hat{w}_3 h_i)$

$r_i = y_i - (\hat{w}_2 \cdot x_i + \hat{w}_3 h_i)$

Можно воткнуть нелинейную модель.

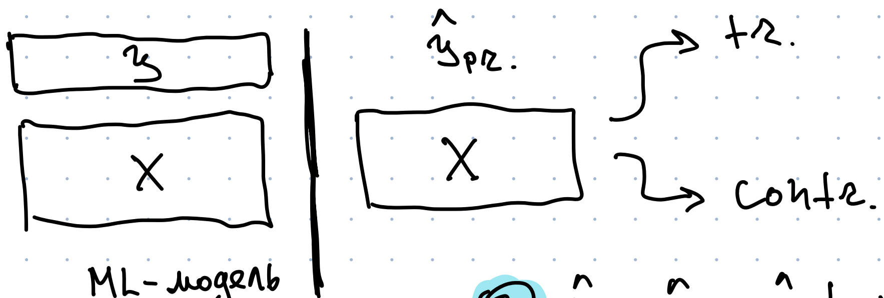
② $\hat{r}_i = \hat{\gamma}_0 + \hat{\gamma}_1 \cdot d_i$

$se(\hat{\gamma}_1)$

→ Наверное, простейший способ воткнуть что-то нелинейное – CUPAC

(Control Using Predictions as Covariate)

①



② $\hat{y}_i = \hat{w}_0 + \hat{w}_1 d_i + \hat{w}_2 \hat{y}_{pr,i}$

$r_i = y_i - \hat{w}_2 \hat{y}_{pr,i}$

③ $\hat{r}_i = \hat{\gamma}_0 + \hat{\gamma}_1 d_i$

Как гарантировать нез. x_i от d_i

y_i – лайки на видосы

$x_i = y_i^{pre}$ – лайки „вчера"

⑥ На самом деле DnD тоже частный случай CUPED

$$y_i^{post} = w_0 + w_1 \cdot d_i + \overbrace{w_2}^{= 1} \cdot y_i^{pre}$$

$$r_i = y_i^{post} - y_i^{pre}$$

$$\hat{r}_i = \hat{\gamma}_0 + \hat{\gamma}_1 \cdot d_i \qquad \hat{\gamma}_1 = \widehat{dd}$$