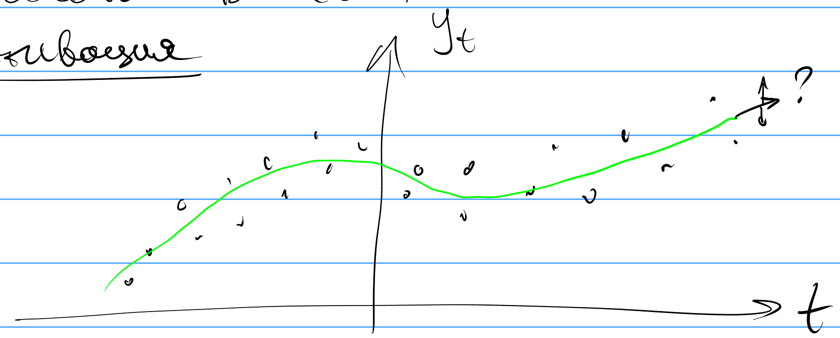
Привет

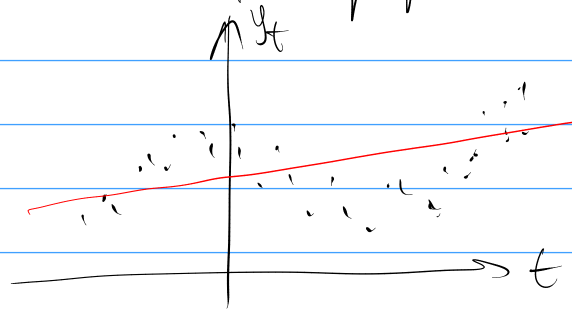
2021 - 12 - 0~~7~~3

- LOESS = LOcal regrESSion
- немного Байеса.

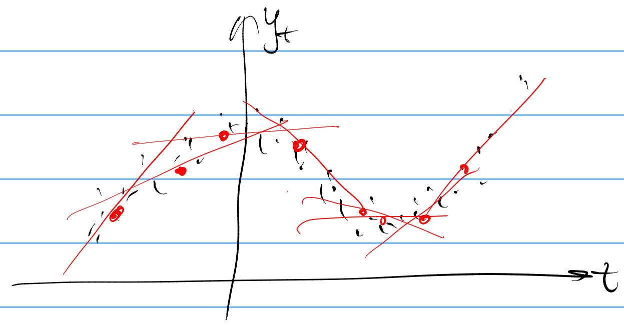
мотивация



обычная регрессия



LOESS



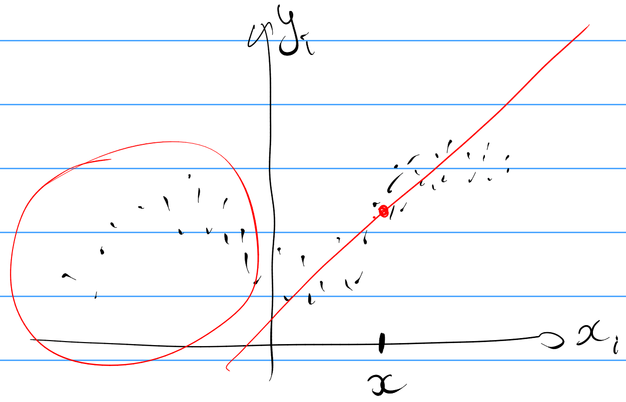
МНК $\sum \left(y_i - (\hat\beta_1 + \hat\beta_2 \cdot x_i)\right)^2 \to \min_{\hat\beta_1, \hat\beta_2}$

$\hat\beta_1$ $\hat\beta_2$

МНК локально

Хотим линию, которая хорошо приближает точки рядом с заданной т-ой x.

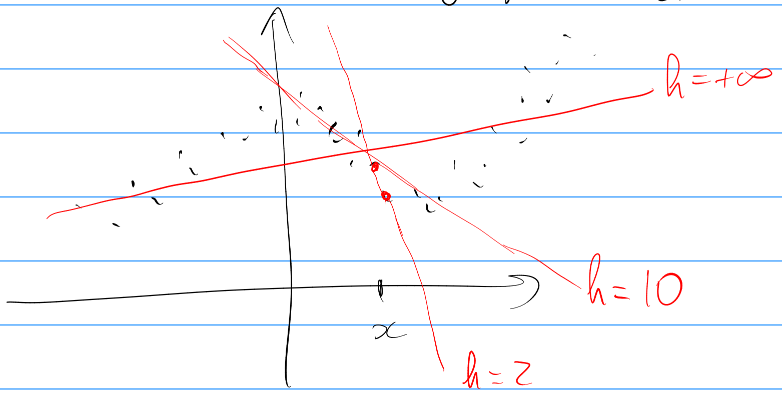
на вход $\Rightarrow$ $\hat\beta_1(x)$ $\hat\beta_2(x)$



Придаём малый вес точкам x_i далеко от x.

$K_h(x_i - x)$ ← весовая ф-ция

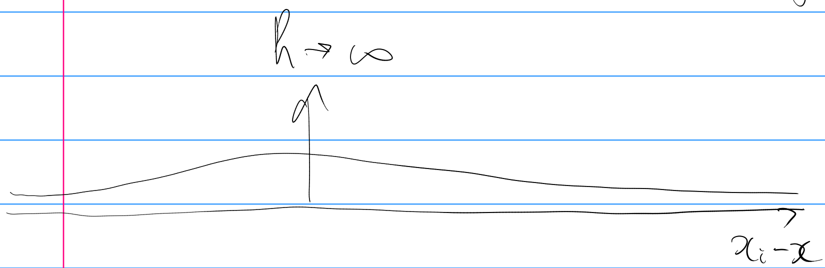
чем $\|x_i - x\|$ больше, тем $K_h(x_i - x)$ меньше обоб.: $K_h(x_0 - x)$ – присваиваем вес 1 ближайшим x_i-м к т. x в количестве h штук.

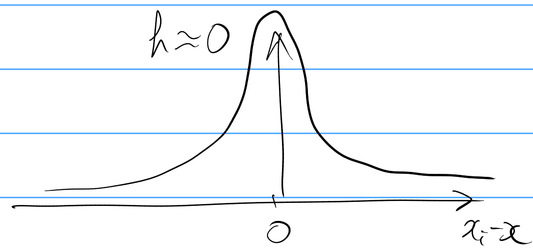


K_h – ядерная ф-ция

классика $K_h(x_i - x)$ – плотность $N(0; h^2)$

$$K_h(x_i - x) = \frac{1}{\sqrt{2\pi h^2}} \cdot \exp\left(-\frac{(x_i - x)^2}{h^2}\right)$$

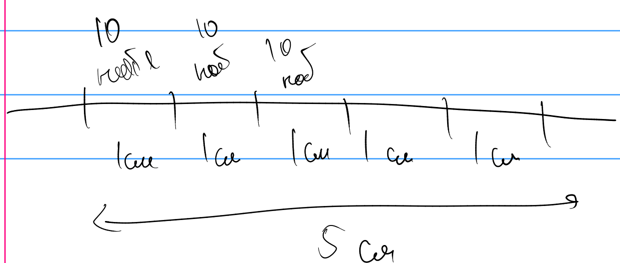


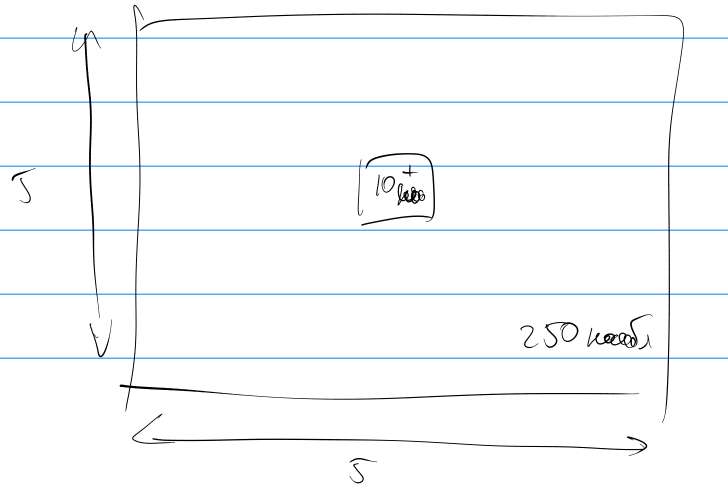


Q. Почему плохо подходит для мн-ой регрессии?

$$\min_{\hat\beta_1(x)\,\hat\beta_2(x)\,\hat\beta_3(z)} \sum_{i=1}^{n} K_h\left(\left\|\binom{x_i}{z_i} - \binom{x}{z}\right\|\right) \cdot \left(y_i - \hat\beta_1 - \hat\beta_2 x_i - \hat\beta_3 z_i\right)^2$$

50 наблюд





Как тут строить дов. интервалы?

$$w_i = \underbrace{K_h(x_i - x)} \leftarrow \boxed{y_i \text{ не входит}}$$

$$\sum w_i \cdot \left(y_i - (\hat\beta_1 \cdot 1 + \hat\beta_2 \cdot x_i)\right)^2 =$$

$$= \sum \left(\underline{y_i \sqrt{w_i}} - (\hat\beta_1 \cdot \sqrt{w_i} + \hat\beta_2 \cdot \sqrt{w_i}\, x_i)\right)^2$$

$$\sum \left(\tilde y_i - (\hat\beta_1\, a_i + \hat\beta_2 \cdot b_i)\right)^2$$

$\tilde y_i$	a_i	b_i
~	~	~

бутстрап / se_{HC} / se

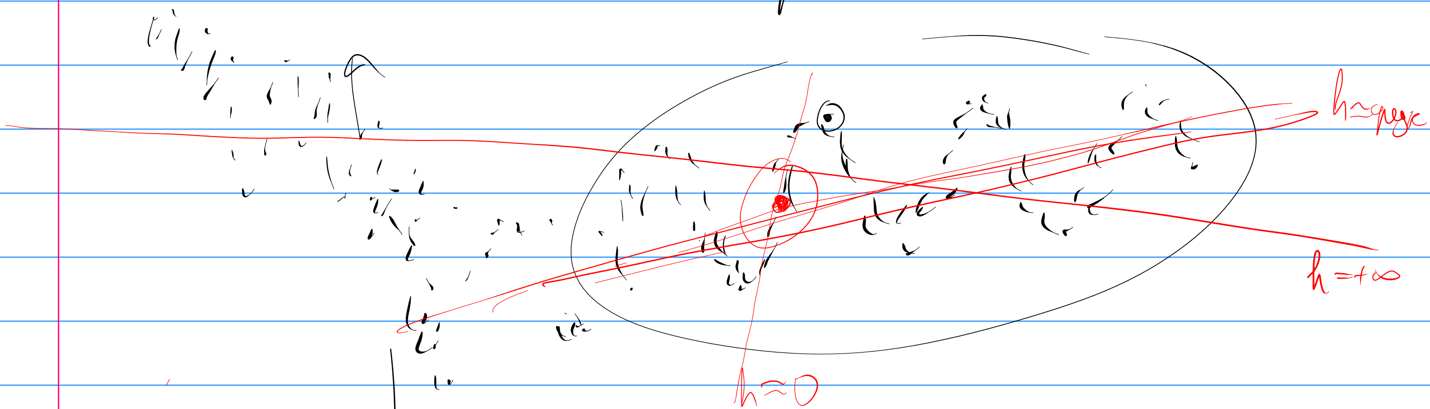
95% и-вал накрывает $E(\tilde y_i \mid a_i, b_i)$

$\Downarrow$

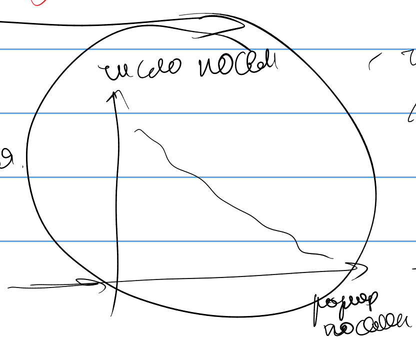
95% интервал и-вал накроет $E(y_i \mid x_i)$

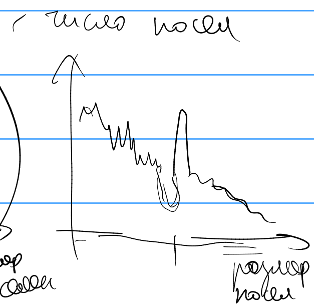
- Как выбрать h?
- Может быть несколько верных h.

$h \in \{0.1, 1, \infty\}$

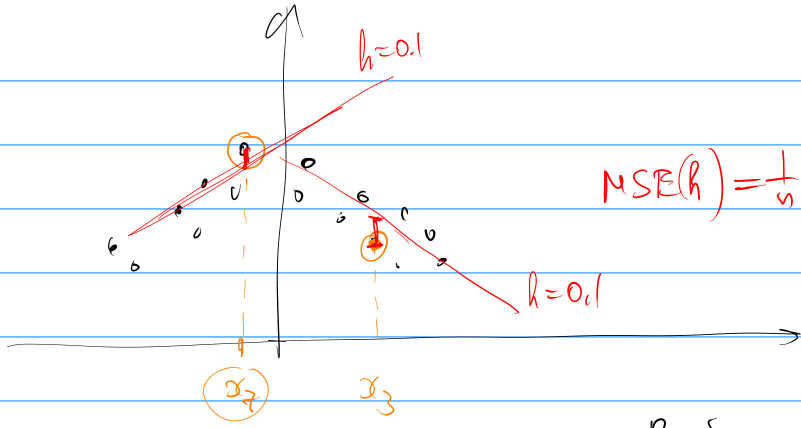


- h_{cv} кросс валидация





$h \in \{0.1;\ 1;\ +\infty\}$



$$MSE(h) = \frac{1}{n}\sum \left(y_i - \hat{y}_i^{cv}(h)\right)^2$$

* Выбираем одно h для всех точек
* $\hat\beta_1(x)$ и $\hat\beta_2(x)$ — их столько, в скольких точках делаем прогноз.

$\boxed{\text{Байес.}}$

Классич.

→ $\Theta = (\Theta_1, \dots, \Theta_k)$ — неизв. конст.

→ модель для набл-х данных $f(y|\Theta)$

$\boxed{\text{метод: ML}}$

на выходе:
→ точечные оценки $\hat\Theta_{ML}$
→ интервальные оценки $[\hat\Theta_{ML} - 1.96\cdot se(\hat\Theta_{ML});\ \hat\Theta_{ML} + 1.96\cdot se(\hat\Theta_{ML})]$

~~$P(\Theta_1 > \Theta_2 \mid$ данных$)$?~~ не отвечаем в принципе

$P(7>3)=1$ $\quad P(\text{неизв} > \text{неизв}) \in \{0,1\}$ неизв.

$P(3>7)=0$

confidence interval CI 0,95

Байесовский подход

→ $\Theta = (\Theta_1, \dots, \Theta_d)$ — случ. величины

$f(\Theta)$ априорный з-н распр.

! надо явно задать закон распр.

→ модель для набл-х данных $f(y|\Theta)$

$\boxed{\text{метод: формула усл-ой вер-сти; техн.: MCMC}}$

на выходе → апостериорный $f(\Theta|y)$

→ точечные оценки $\hat\Theta = Med(\Theta|y)$, $\hat\Theta = Mod(\Theta|y)$, $\hat\Theta = E(\Theta|y)$

→ интерв-ые оценки $P(\Theta \in [\Theta_L; \Theta_H] \mid y) = 0.94$

credible interval CI 0.94

→ $P(\Theta_1 > \Theta_2 \mid y)$ можно ответить

Задачка о Маше и медведях.

Классич-ий подход. Байесовский подход.

θ

Маша спряталась в точке θ на прямой.

3 медведя.

y_1, y_2, y_3 – где медведь ищет Машу. [нерв]

модель для данных: $(y_i|\theta) \sim N(\theta; 1)$

классич.

$\boxed{\theta \text{ – неизв. конст.}}$

Байесовский

θ – случ. величина

$\theta \sim N(0; 10)$ [выбор иссл-ля] априорное распр

$f(\theta) = \frac{1}{\sqrt{2\pi \cdot 10}} \cdot \exp\left(-\frac{1}{2}\frac{\theta^2}{10}\right)$

$$\ell(y|\theta) = \ln f(y_1|\theta) + \ln f(y_2|\theta) + \ln f(y_3|\theta) =$$

$$= \ln\frac{1}{\sqrt{2\pi}} - \frac{1}{2}(y_1-\theta)^2 + \ldots = \text{const} - \frac{1}{2}(y_1-\theta)^2 - \frac{1}{2}(y_2-\theta)^2 - \frac{1}{2}(y_3-\theta)^2$$

$$\hat{\theta}_{ML} = \underset{\theta}{\text{argmax}}\, \ell(y|\theta)$$

рбб

$$\underbrace{f(\theta|y)}_{\text{апостериорное}} = \frac{f(y,\theta)}{f(y)} = \frac{f(y|\theta)\cdot f(\theta)}{f(y)}$$

$$\ln f(\theta|y) = \underbrace{\ln f(y|\theta)}_{\ell(y|\theta)\text{ ф. правд}} + \underbrace{\ln f(\theta)}_{\text{лог. априорн}} - \underbrace{\ln f(y)}_{\text{не зависит от }\theta} =$$

забиваем на то, что не зависит от θ

$$= \underbrace{\text{const} - \frac{1}{2}(y_1-\theta)^2 - \frac{1}{2}(y_2-\theta)^2 - \frac{1}{2}(y_3-\theta)^2}_{\ln\text{ правд}} + \underbrace{\text{const} - \frac{1}{20}\cdot\theta^2}_{\ln\text{ априорн}} \underbrace{+\text{const}}_{\ln f(y)}$$

сюда и константы и y

$$= -\frac{1}{2}\left(3{,}1\,\theta^2 - 2\theta(y_1+y_2+y_3)\right) + \text{const}$$

$$= \text{выд. полн. квадрата} = -\frac{1}{2}\cdot\frac{(\theta - \text{слон})^2}{\text{еж}^2} + \text{конст.} = \ln f(\theta|y)$$

$$(\theta|y) \sim N(\text{слон}, \text{еж}^2)$$

$$\hat{\theta}_B = E(\theta|y) = \text{слон}.$$