# Metoda maximální věrohodnosti pro volbu vyhlazovacích parametrů jádrových odhadů podmíněné hustoty

Kateřina Konečná, Ivana Horová

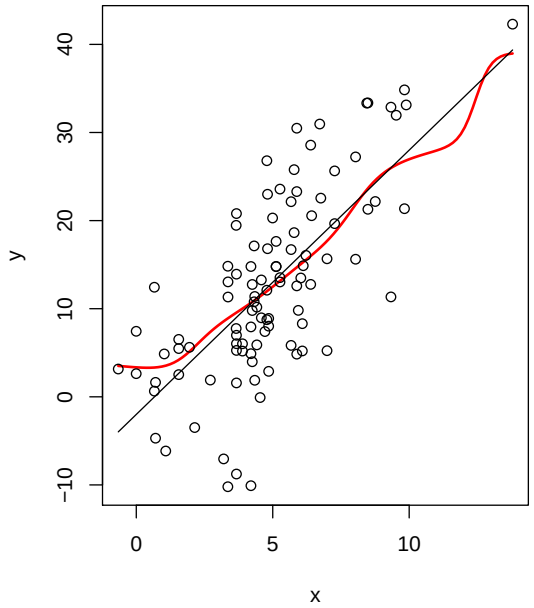
*Ústav matematiky a statistiky, Masarykova univerzita, Brno*

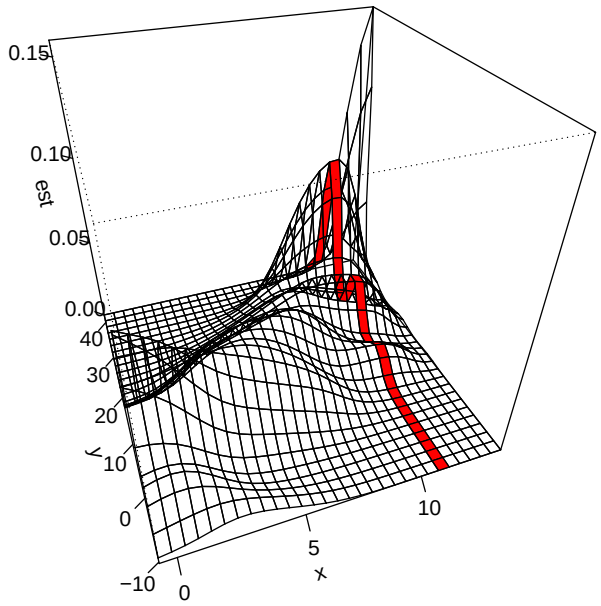
ROBUST 2016
Loučná nad Desnou, Kurzovní
11. – 16. září 2016

# Obsah

# Motivace a cíle práce

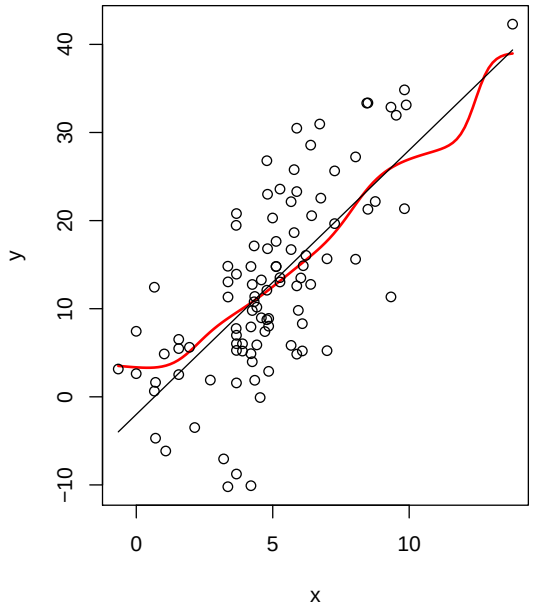
- motivace

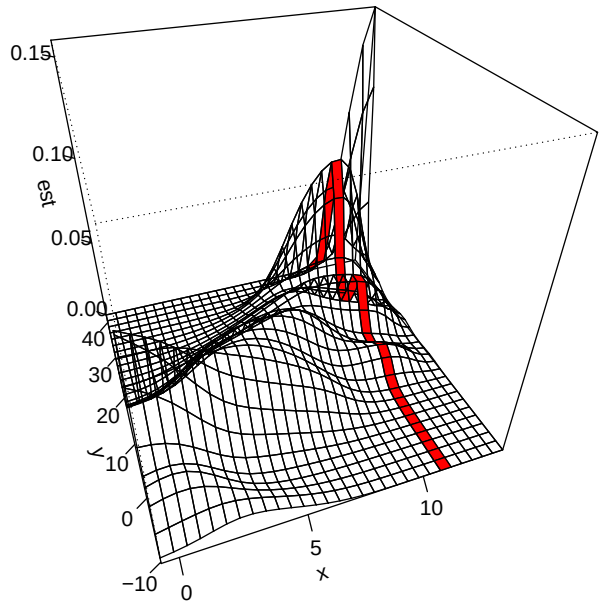


# Motivace a cíle práce

- motivace



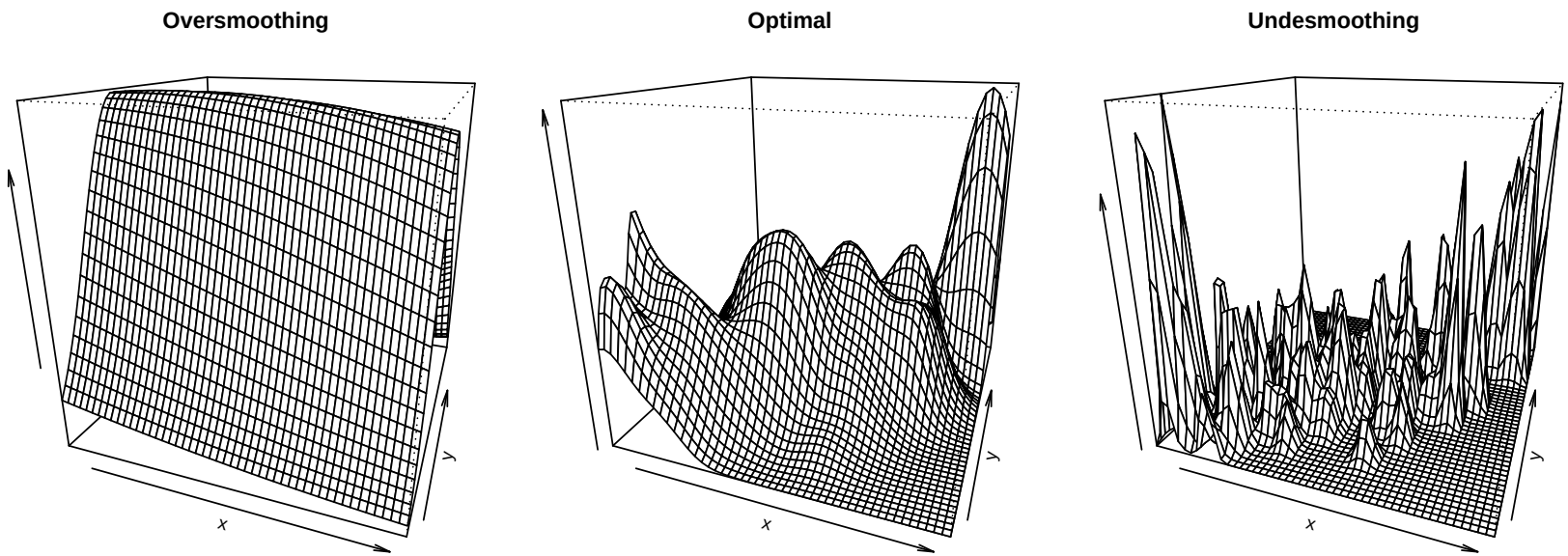

- cíle práce
  - statistické vlastnosti LL odhadu
  - odvození teoretické šířky vyhlazovacích parametrů
  - metody pro odhad šířky vyhlazovacích parametrů
  - simulační studie

# Jádrovové odhady

- jádrová funkce

# Jádrovové odhady

- jádrová funkce
- vyhlazovací parametry



# Druhy jádrových odhadů podmíněné hustoty

Jádrový odhad podmíněné hustoty

$$\hat{f}(y|x) = \frac{1}{h_y} \sum_{i=1}^{n} w_i(x) K\left(\frac{y - Y_i}{h_y}\right),$$

- Nadaraya-Watsonovy váhy

$$w_i(x) = \frac{K\left(\frac{x-X_i}{h_x}\right)}{\sum_{j=1}^{n} K\left(\frac{x-X_j}{h_x}\right)},$$

- lokálně-lineární váhy

$$w_i(x) = \frac{\left(\hat{s}_2(x) - \hat{s}_1(x)(x - X_i)\right) K_{h_x}(x - X_i)}{n\left(\hat{s}_0(x)\hat{s}_2(x) - \hat{s}_1^2(x)\right)},$$

kde $\hat{s}_r(x) = \frac{1}{n} \sum_i (x - X_i)^r K_{h_x}(x - X_i)$

# Metody pro odhad šířky vyhlazovacích parametrů

- Metoda křížového ověřování

$$CV(h_x, h_y) = \frac{1}{n}\sum_{i=1}^{n}\int \hat{f}^2_{-i,LL}(y|X_i)\,\mathrm{d}y - \frac{2}{n}\sum_{i=1}^{n}\hat{f}_{-i,LL}(y_i|X_i),$$

- Metoda maximální věrohodnosti

$$\mathcal{L}\left(y|x; h_x, h_y\right) = \prod_{j=1}^{n}\hat{f}_{-j,LL}\left(y_j|X_j\right).$$

# Míra kvality odhadu

- Odhad globální chyby

$$\widehat{\mathrm{ISE}}\left\{\hat{f}_{LL}(\cdot|\cdot)\right\} = \frac{\Delta}{n}\sum_{j=1}^{N}\sum_{i=1}^{n}\left(\hat{f}_{LL}(y_j|X_i) - f(y_j|X_i)\right)^2$$

$\mathbf{y} = (y_1, \ldots, y_N)$ - vektor ekvidistantních hodnot na nosiči $Y$
$\Delta$ - vzdálenost mezi sousedními hodnotami vektoru $\mathbf{y}$.

# Míra kvality odhadu

- Odhad globální chyby

$$\widehat{\mathrm{ISE}}\left\{\hat{f}_{LL}(\cdot|\cdot)\right\} = \frac{\Delta}{n}\sum_{j=1}^{N}\sum_{i=1}^{n}\left(\hat{f}_{LL}(y_j|X_i) - f(y_j|X_i)\right)^2$$

$\mathbf{y} = (y_1, \ldots, y_N)$ - vektor ekvidistantních hodnot na nosiči $Y$
$\Delta$ - vzdálenost mezi sousedními hodnotami vektoru $\mathbf{y}$.

- Logaritmus absolutní, resp. relativní chyby

$$\mathrm{E}_1 = \ln\left|\widehat{\mathrm{ISE}} - \mathrm{ISE}_{opt}\right|, \text{ resp. } \mathrm{E}_2 = \ln\left|\frac{\widehat{\mathrm{ISE}} - \mathrm{ISE}_{opt}}{\mathrm{ISE}_{opt}}\right|,$$

$\mathrm{ISE}_{opt}$ - odhad $\mathrm{ISE}$ pro optimální šířku vyhlazovacích parametrů.

# Dosažené výsledky práce I

- odvození rozptylu a vychýlení LL odhadu

$$\mathrm{AV}\left\{\hat{f}_{LL}(y|x)\right\} = \frac{1}{n^2 h_x^2 h_y} \cdot \frac{R(K)G^2(K)}{\beta_2(K)} \cdot \frac{f(y|x)}{h^2(x)},$$

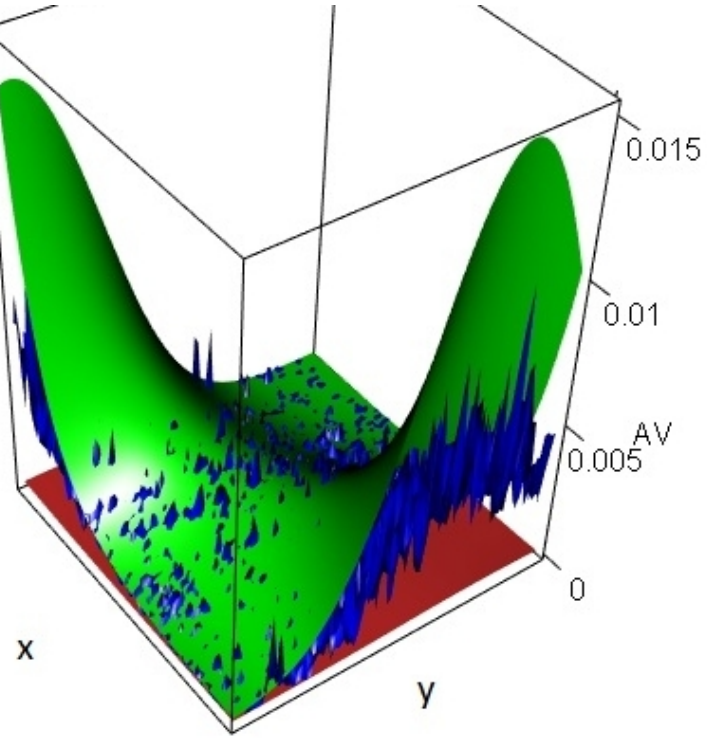
$$\mathrm{AB}\left\{\hat{f}_{LL}(y|x)\right\} = \frac{1}{2}\beta_2(K)h_x^2\left(\frac{\partial^2 f(y|x)}{\partial x^2}\right)^2 + \frac{1}{2}\beta_2(K)h_y^2\left(\frac{\partial^2 f(y|x)}{\partial y^2}\right)^2$$

# Dosažené výsledky práce I

- odvození rozptylu a vychýlení LL odhadu

$$\mathrm{AV}\left\{\hat{f}_{LL}(y|x)\right\} = \frac{1}{n^2 h_x^2 h_y} \cdot \frac{R(K)G^2(K)}{\beta_2(K)} \cdot \frac{f(y|x)}{h^2(x)},$$

$$\mathrm{AB}\left\{\hat{f}_{LL}(y|x)\right\} = \frac{1}{2}\beta_2(K)h_x^2\left(\frac{\partial^2 f(y|x)}{\partial x^2}\right)^2 + \frac{1}{2}\beta_2(K)h_y^2\left(\frac{\partial^2 f(y|x)}{\partial y^2}\right)^2$$



# Dosažené výsledky práce II

$$\mathrm{AMSE}\left\{\hat{f}_{LL}(y|x)\right\} = \mathrm{AV}\left\{\hat{f}_{LL}(y|x)\right\} + \mathrm{ASB}\left\{\hat{f}_{LL}(y|x)\right\}$$

# Dosažené výsledky práce II

$$\mathrm{AMSE}\left\{\hat{f}_{LL}(y|x)\right\} = \mathrm{AV}\left\{\hat{f}_{LL}(y|x)\right\} + \mathrm{ASB}\left\{\hat{f}_{LL}(y|x)\right\}$$

$$\downarrow$$

$$\mathrm{AMISE}\left\{\hat{f}_{LL}(\cdot|\cdot)\right\} = \iint \mathrm{AMSE}\left\{\hat{f}_{LL}(y|x)\right\} h(x)\, \mathrm{d}x\, \mathrm{d}y$$

# Dosažené výsledky práce II

$$\mathrm{AMSE}\left\{\hat{f}_{LL}(y|x)\right\} = \mathrm{AV}\left\{\hat{f}_{LL}(y|x)\right\} + \mathrm{ASB}\left\{\hat{f}_{LL}(y|x)\right\}$$

$$\downarrow$$

$$\mathrm{AMISE}\left\{\hat{f}_{LL}(\cdot|\cdot)\right\} = \iint \mathrm{AMSE}\left\{\hat{f}_{LL}(y|x)\right\} h(x)\,\mathrm{d}x\,\mathrm{d}y$$

$$\downarrow$$

optimální šířka vyhlazovacích parametrů: $\mathrm{AMISE}\left\{\hat{f}_{LL}(\cdot|\cdot)\right\} \to \min$

$$h_x^* = \left(\frac{c_1}{2n^2} \cdot \frac{1}{2c_4 z^{5/2} + c_5 z^{3/2}}\right)^{1/7}$$
$$h_y^* = h_x \sqrt{z},$$

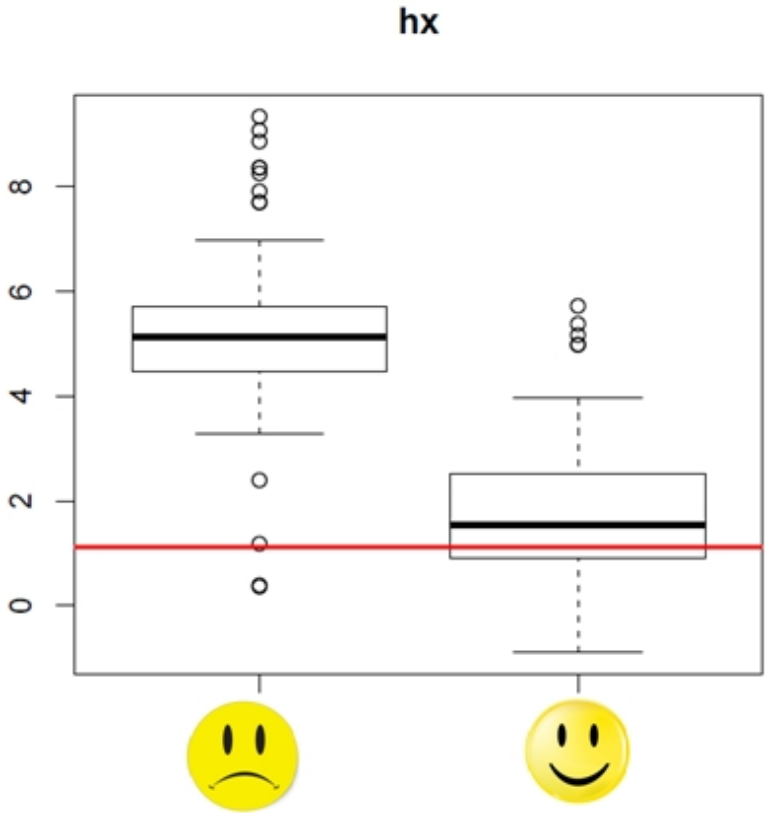
kde $z = \frac{\sqrt{c_5^2 + 32 * c_3 * c_4} - c_5}{8c_4}$

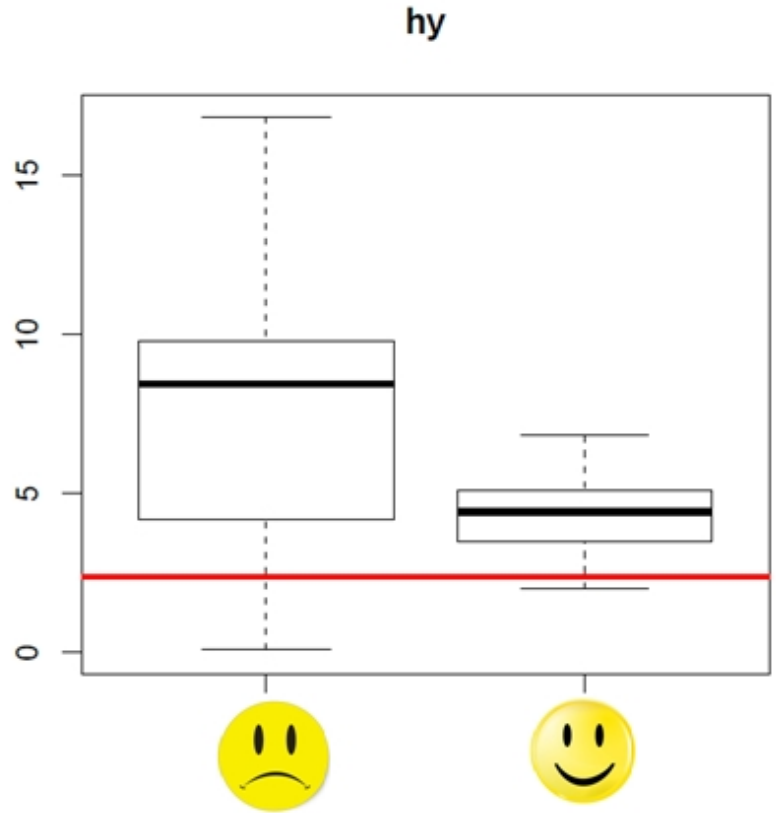
# Ověření metod na simulační studii

- Simulovaná data:
  $X_i \sim N(5, 3^2)$, $\varepsilon_i \sim N(0, 8^2)$,
  $Y_i = 3X_i - 2 + \varepsilon_i$, $i = 1, \ldots, n$.
- Počet simulovaných dat: $n = 100$.
- Počet opakování simulace: 100.

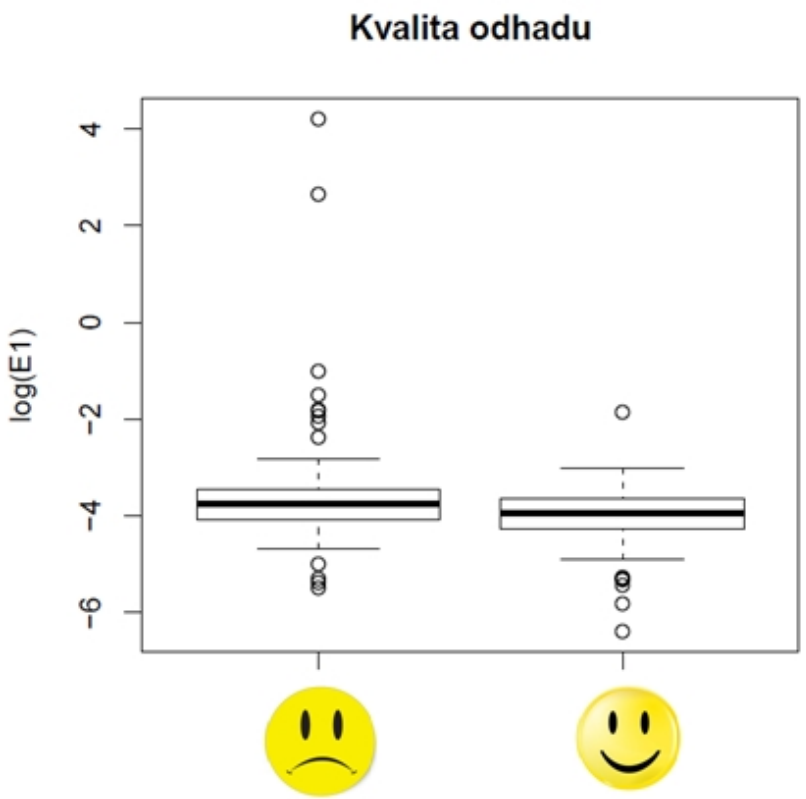
# Ověření metod na simulační studii

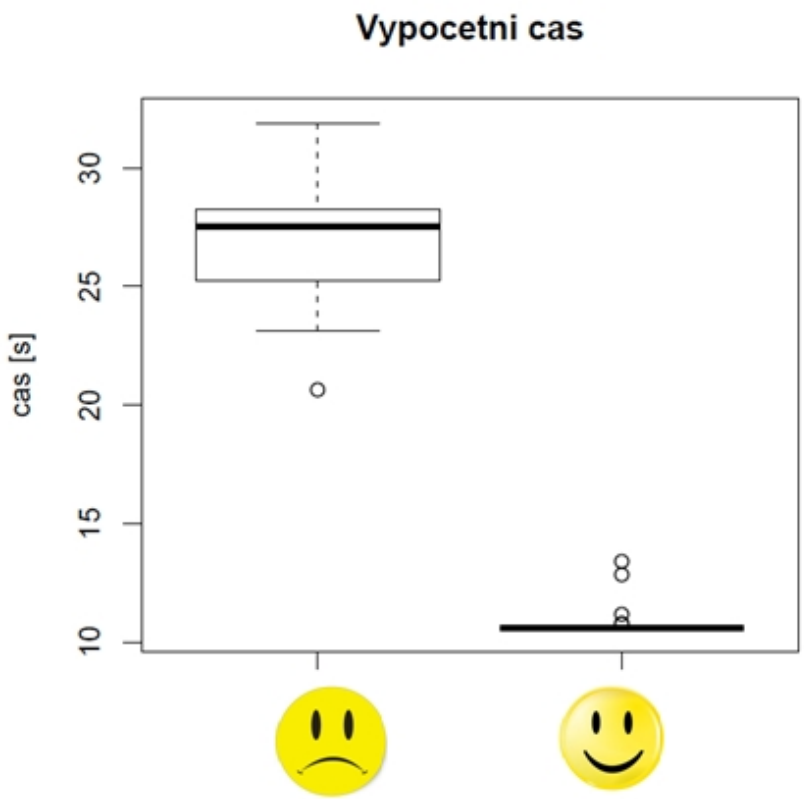
- Simulovaná data:
  $X_i \sim N(5, 3^2)$, $\varepsilon_i \sim N(0, 8^2)$,
  $Y_i = 3X_i - 2 + \varepsilon_i$, $i = 1, \ldots, n$.
- Počet simulovaných dat: $n = 100$.
- Počet opakování simulace: 100.



# Ověření metod na simulační studii



# Reference


- Bashtannyk, D. M., Hyndmann, R. J. (2001). *Bandwidth Selection for Kernel Conditional Density Estimation.* Computational Statistics & Data Analysis **36**(3), 279–298
- Fan, J., Yim, T. H., *A Crossvalidation Method for Estimating Conditional Densities*, Biometrika, 2004.
- Hansen, B. E., *Nonparametric Conditional Density Estimation*, unpublished manuscript, 2004.
- Konečná K., Horová, I., Koláček J., *Conditional Density Estimations*. In C H Skiadas. Theoretical and Applied Issues in Statistics and Demography. Athens: ISAST, 2014.