Testy a odhady založené na minimalizaci vzdálenosti

Radim Navrátil

Ústav matematiky a statistiky
Přírodovědecká fakulta MU, Brno

Robust
16. září 2016

O čem budu povídat

1. Odhad metodou nejmenších čtverců v regresi

2. Robustní odhady v regresi
   - Příklady robustních metod
   - Pořadové testy a odhady v regresi

3. Testy a odhady založené na pořadích minimalizující vzdálenost
   - Testy minimalizující vzdálenost
   - Testy minimalizující vzdálenost - zobecnění
   - Odhady minimalizující vzdálenost
   - Odhady minimalizující vzdálenost - zobecnění

Model regresní přímky

$$Y_i = \beta_0 + \beta x_i + e_i, \quad i = 1, \ldots, n$$

- $e_1, \ldots, e_n$ jsou iid s nějakou (obecně neznámou) distribuční funkcí F a hustotou f.
- $x_1, \ldots, x_n$ jsou regresory (pevné či náhodné).

Cíle:

- Odhad parametru β.
- Test hypotézy: $\mathbf{H}_0 : \beta = 0$ proti alternativě $\mathbf{K} : \beta \neq 0$.

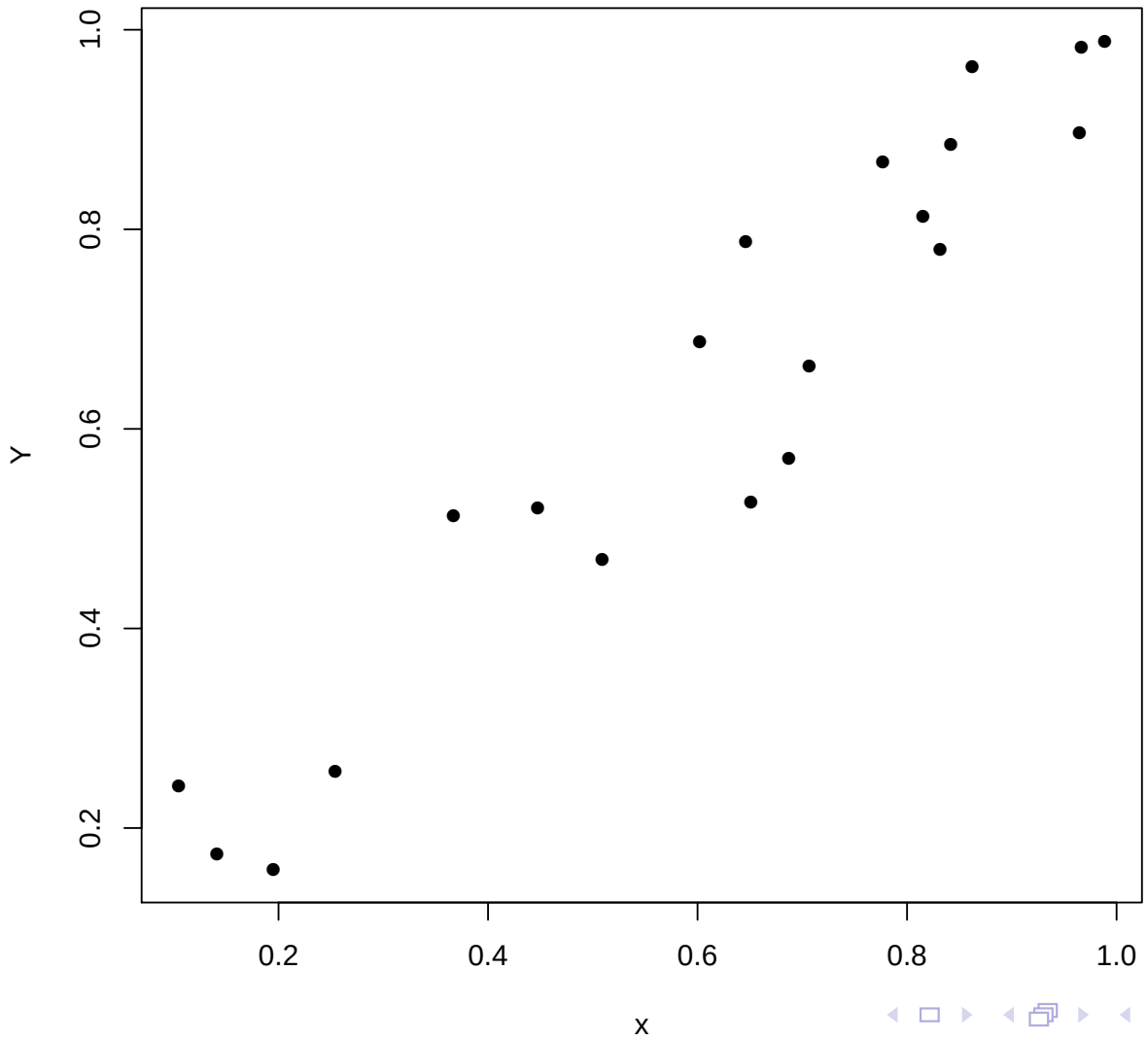
Odhad metodou nejmenších čtverců

- $(\widehat{\beta}_0, \widehat{\beta})' = \operatorname{argmin}\{\sum_{i=1}^n (Y_i - \beta_0 - \beta x_i)^2 : (\beta_0, \beta)' \in \mathbb{R}^2\}$.
- $\widehat{\beta} = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}$.
- Statistika $T = \frac{\widehat{\beta} - \beta}{\widehat{\sigma}} \sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}$ má za předpokladu normality chyb modelu t-rozdělení s $(n-2)$ stupni volnosti.

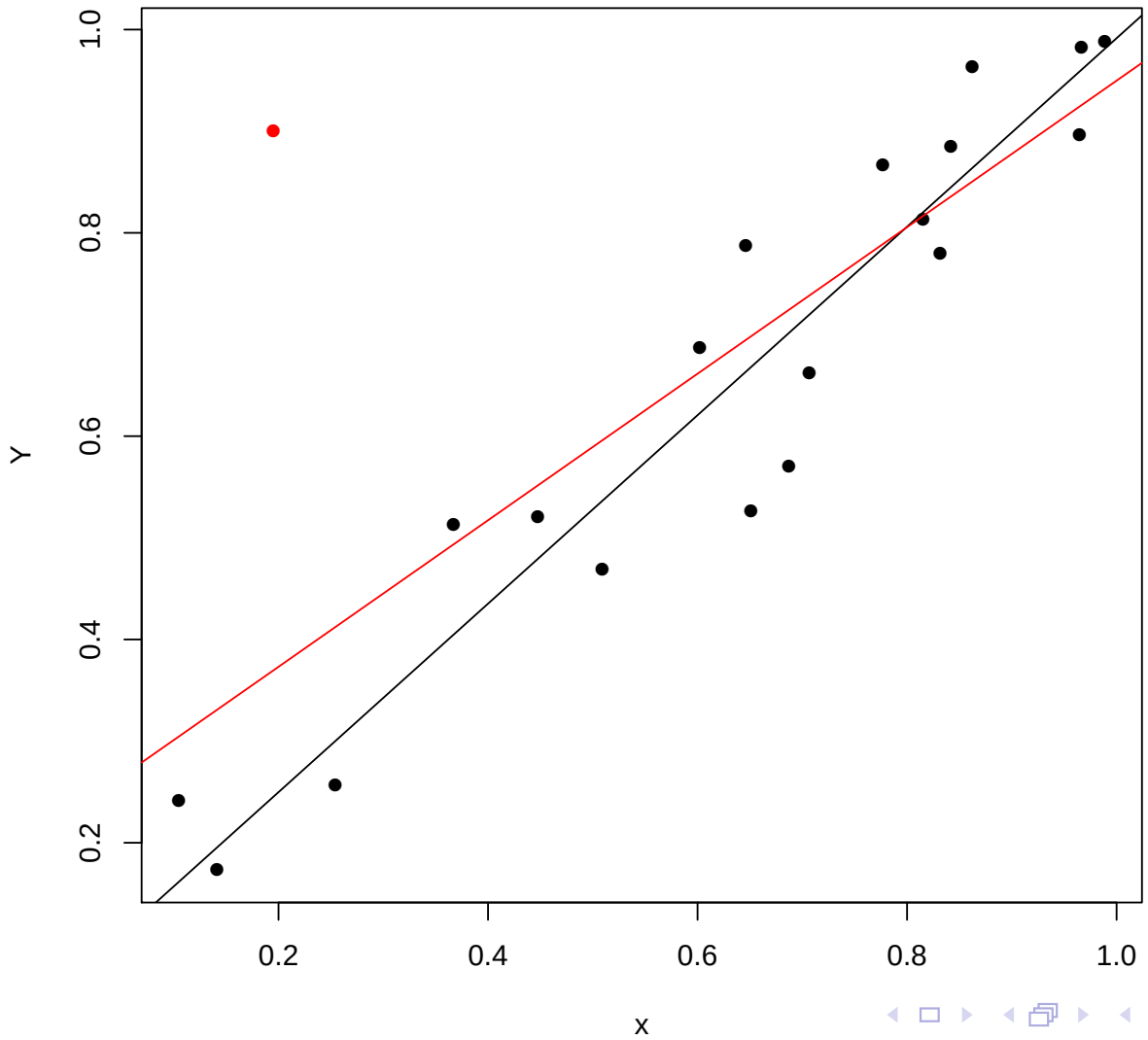
Nevýhody:

- Citlivé na porušení předpokladu normality (přítomnost outlierů).
- Citlivé na přítomnost vlivných pozorování.

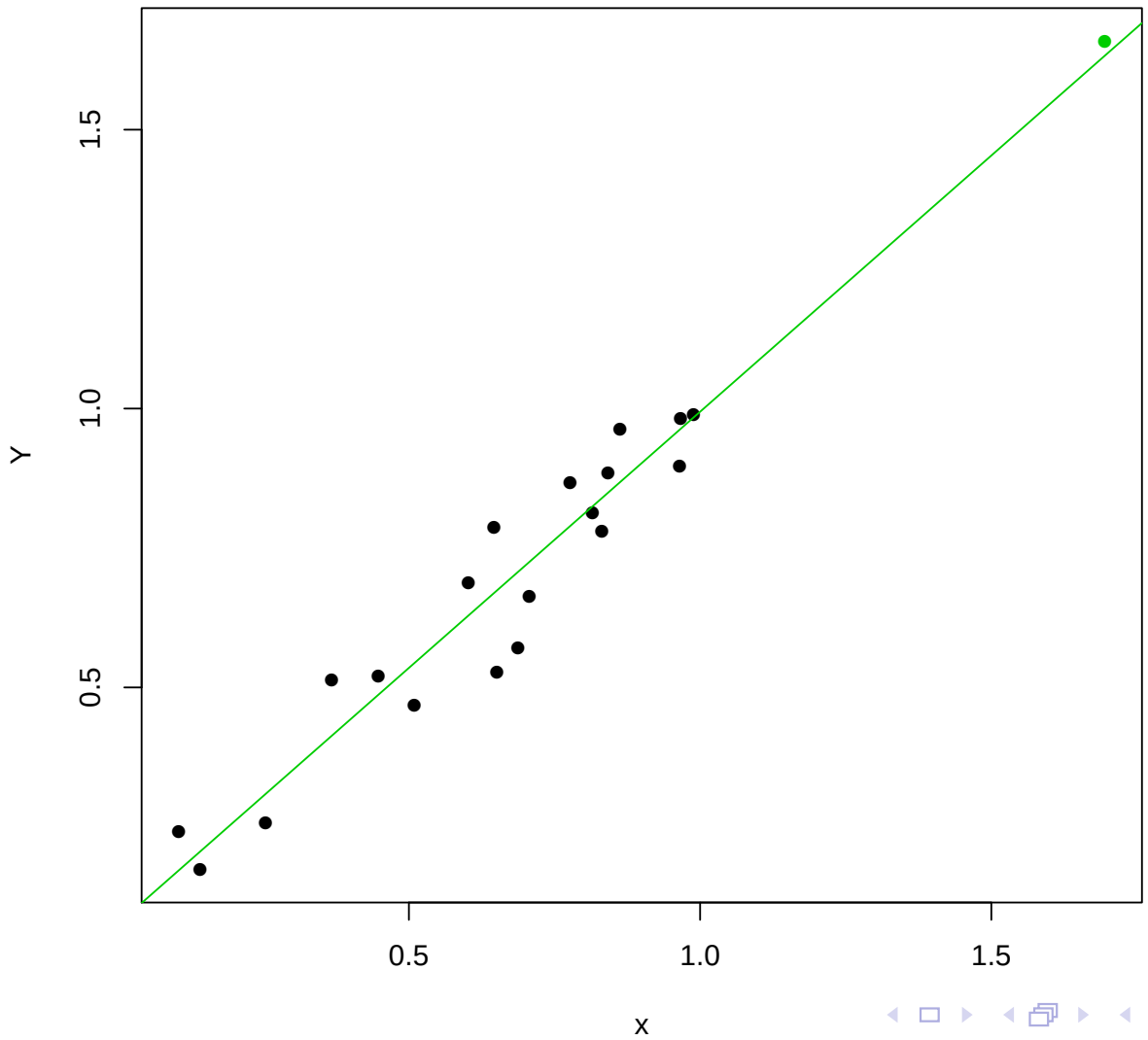
Příklad - regresní přímka s $\beta = 1$



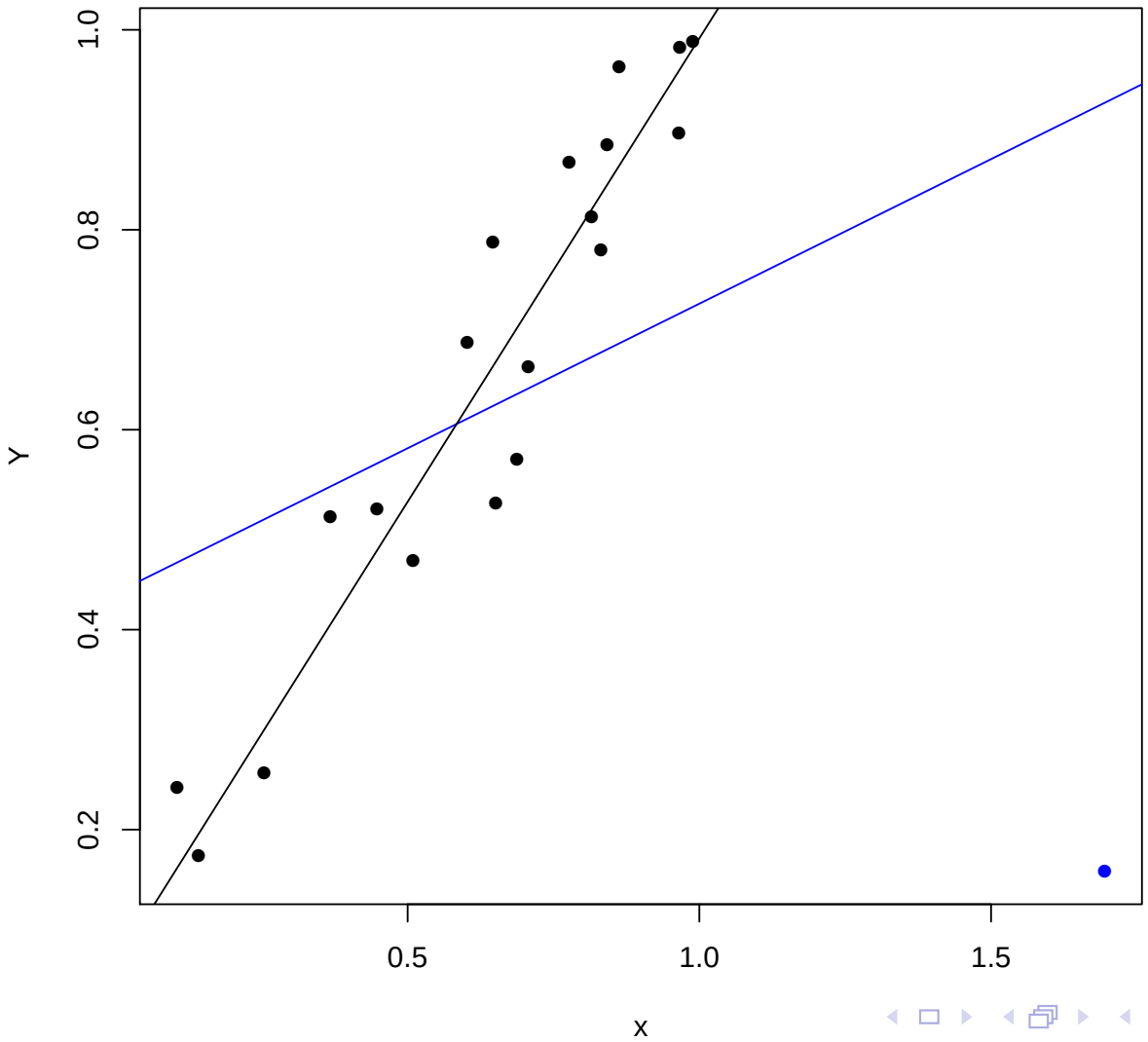
Příklad - regresní přímka s $\beta = 1$ + outlier



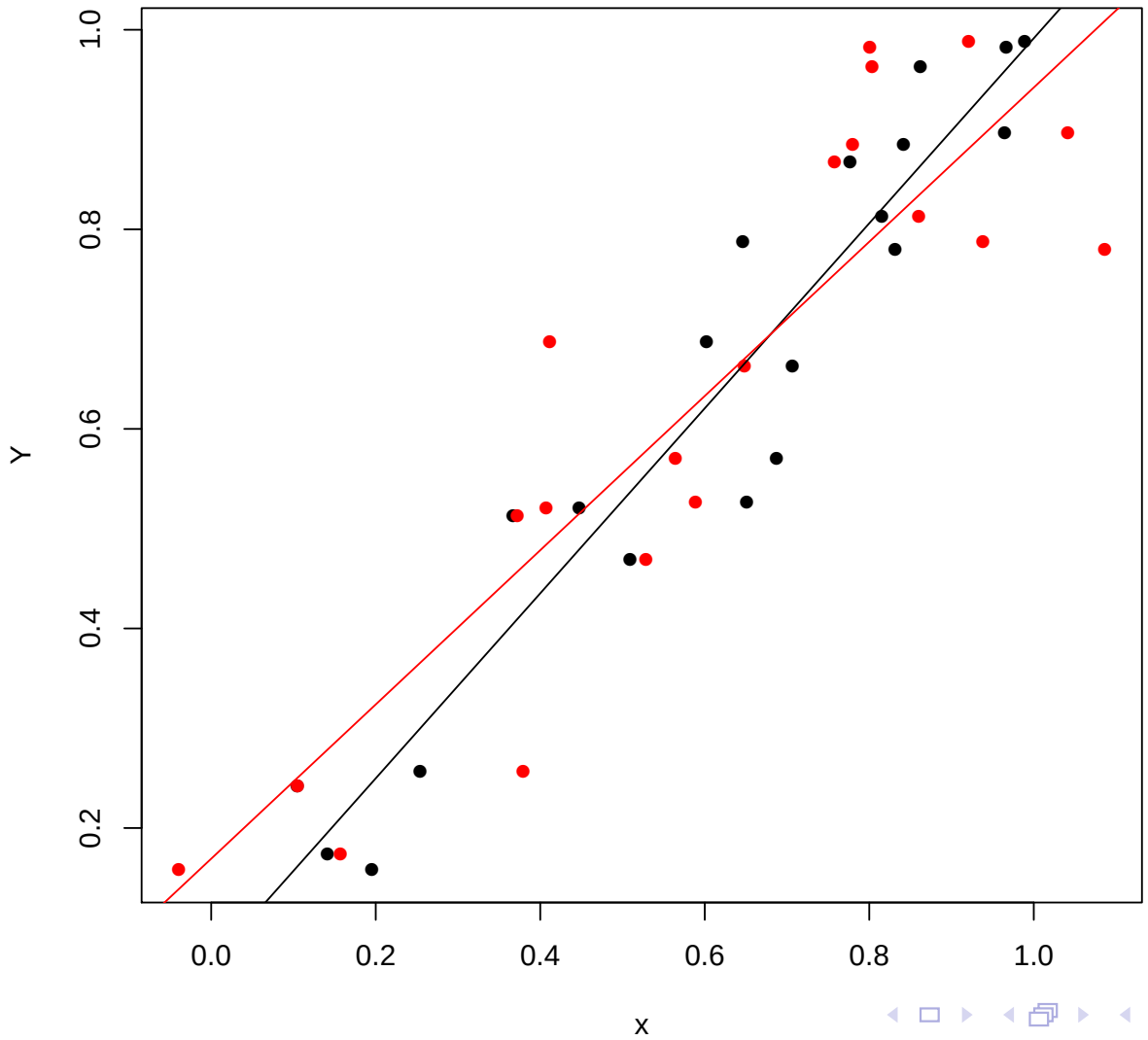
Příklad - regresní přímka s $\beta = 1 +$ leverage point



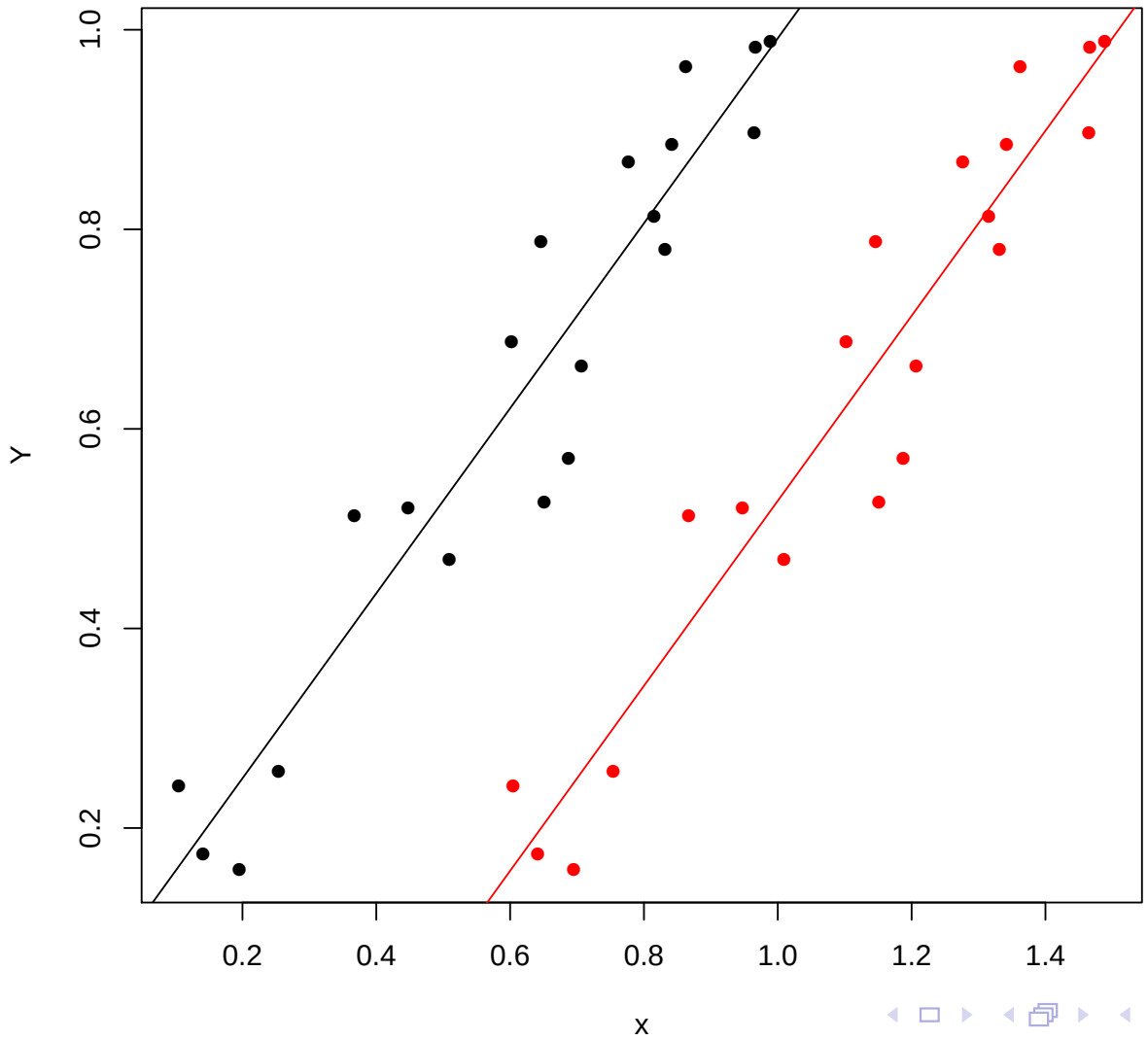
Příklad - regresní přímka s $\beta = 1 +$ outlier leverage point



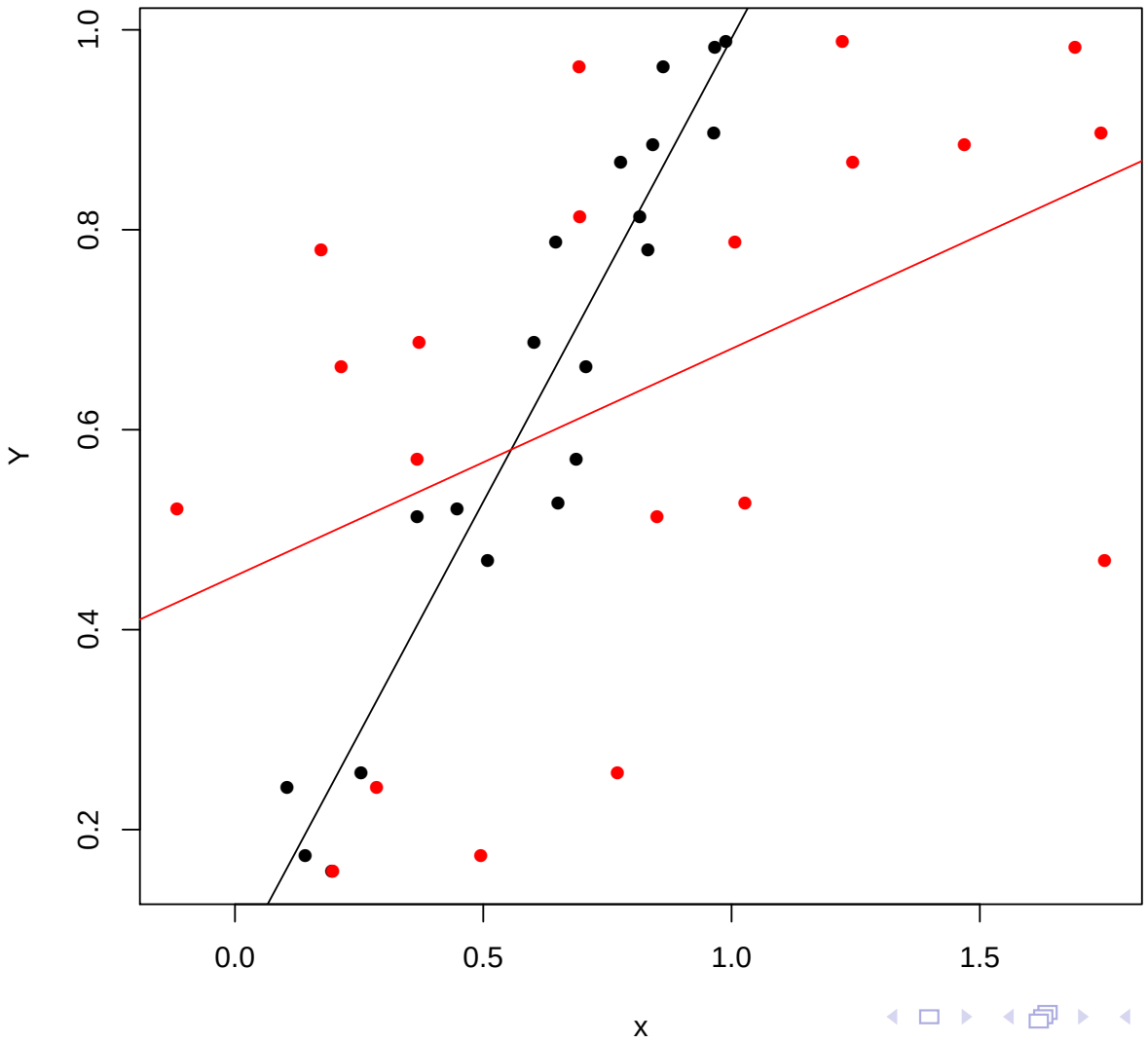
Příklad - regresní přímka s $\beta = 1 +$ chyby měření



Příklad - regresní přímka s $\beta = 1 +$ chyby měření



Příklad - regresní přímka s $\beta = 1 +$ chyby měření



Robustní odhady v regresi

- Huber (1981) - M-odhady
- Siegel (1982) - Least Median of Squares
- Rousseeuw (1983) - Least Trimmed Squares
- Koenker, Bassett (1978) - Kvantilová regrese
- Jaeckel (1972) - R-odhady

R-odhady v regresi

- Označme $\mathcal{D}_n(b) = \sum_{i=1}^{n} (Y_i - x_i b)\, \varphi\left(\frac{R_i(b)}{n+1}\right)$.
- $R_i(b)$ je pořadí $(Y_i - x_i b)$ mezi $(Y_1 - x_1 b), \ldots, (Y_n - x_n b)$.
- $\varphi : (0,1) \mapsto \mathbb{R}$ je nekonstantní, neklesající, integrovatelná se čtvercem, $\varphi(1-t) = -\varphi(t),\ 0 < t < 1$.
- Jaeckel (1972) definoval odhad parametru β jako $\widehat{\beta}_n = \arg\min\{\mathcal{D}_n(b) : b \in \mathbb{R}\}$.
- Dále označme $S_n(b) = \sum_{i=1}^{n} (x_i - \bar{x})\, \varphi\left(\frac{R_i(b)}{n+1}\right)$
- Jurečková (1971) definovala odhad parametru β jako $\widehat{\beta}_n = \frac{1}{2}\sup\{b \in \mathbb{R} : S_n(b) > 0\} + \frac{1}{2}\inf\{b \in \mathbb{R} : S_n(b) < 0\}$.

Pořadové testy v regresi

- $S_n = n^{-1/2} S_n(0) = \sum_{i=1}^n (x_i - \bar{x}) \varphi\left(\frac{R_i}{n+1}\right)$.
- R_i je pořadí Y_i mezi $Y_1, \ldots, Y_n$.
- $\varphi : (0,1) \mapsto \mathbb{R}$ je nekonstantní, neklesající, integrovatelná se čtvercem.
- $A^2(\varphi) = \int_0^1 (\varphi(t) - \overline{\varphi})^2 dt, \quad \overline{\varphi} = \int_0^1 \varphi(t) dt.$
- $Q_n = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$.
- **Testová statistika:** $T_n^2 = \frac{S_n^2}{n A^2(\varphi) Q_n}$.
- T_n^2 má za platnosti nulové hypotézy asymptoticky χ^2 rozdělení s 1 stupněm volnosti.

Pořadové testy v regresi - příklady

- $\varphi(u) = u$ Wilcoxonův test
- $\varphi(u) = \text{sign}\,(2u - 1)$ mediánový (znaménkový) test
- $\varphi(u) = \Phi^{-1}(u)$ van der Waerdenův test

Testy minimalizující vzdálenost

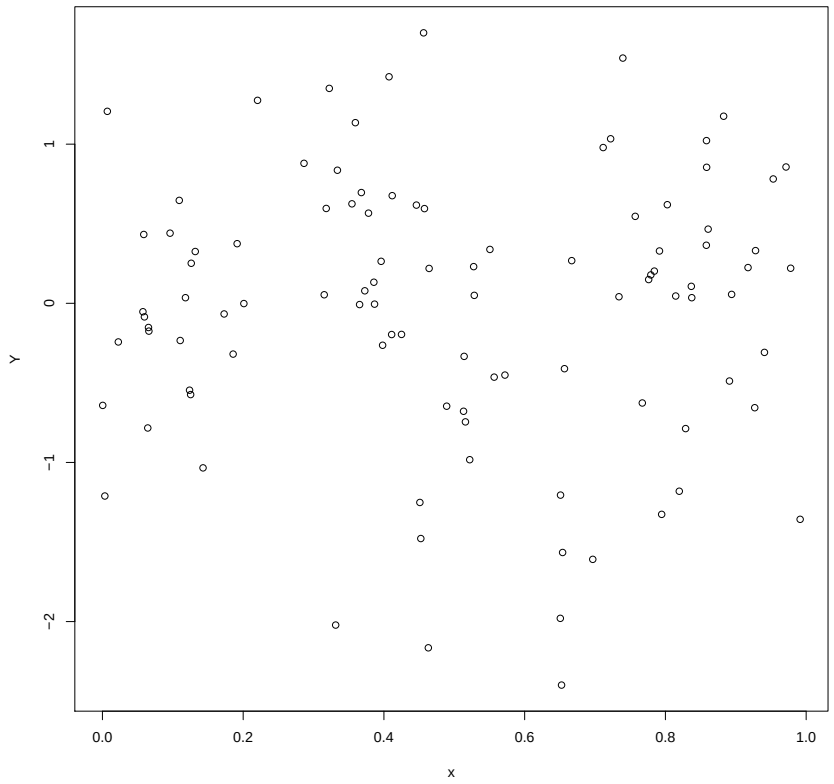
- Model regresní přímky: $Y_i = \beta_0 + x_i\beta + e_i, \quad i = 1, \ldots, n.$
- Motivace: Koul (2002) uvažoval odhady parametru β minimalizující jisté vzdálenosti a ukázal, že v některých situacích mají lepší vydatnost než R-odhady.
-

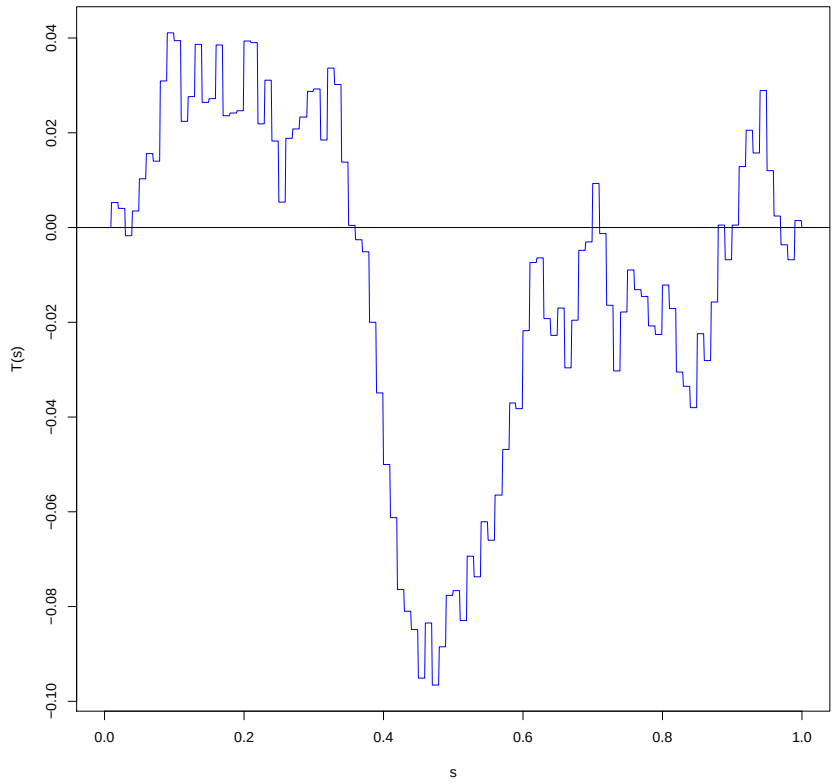
$$T_{g,n}(s) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} g(x_i)\mathbb{I}\{R_i \leq ns\}, \quad 0 \leq s \leq 1,$$

$$K^*_{g,n} = \int_0^1 T^2_{g,n}(s)dL(s),$$

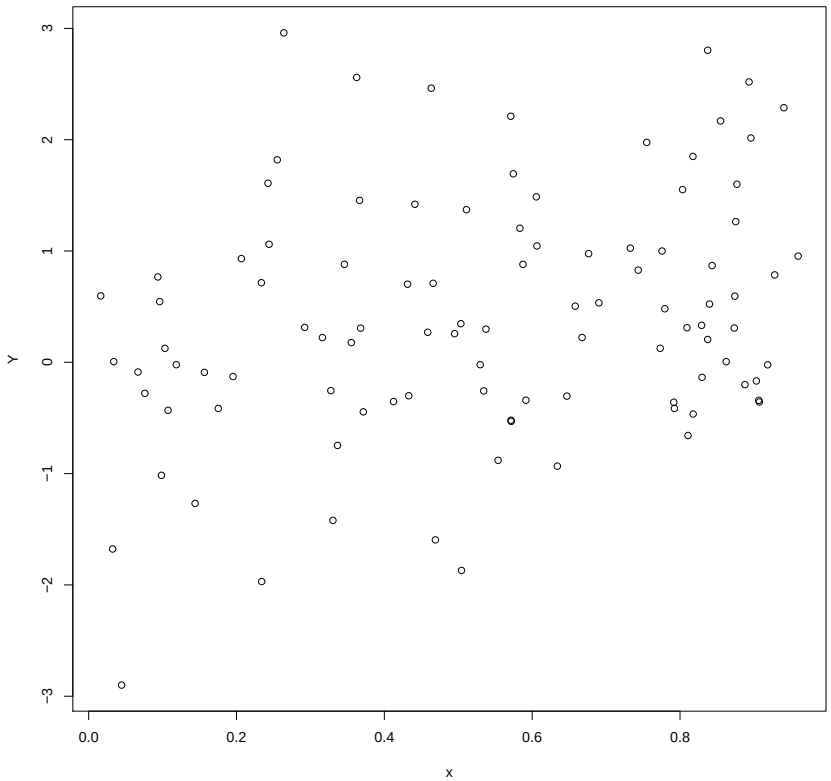
- R_i je pořadí Y_i mezi $Y_1, \ldots, Y_n$.
- L je distribuční funkce na $[0, 1]$ a g reálná (váhová) funkce taková, že $\sum_{i=1}^n g(x_i) = 0$ a $\sum_{i=1}^n g^2(x_i) = 1$.

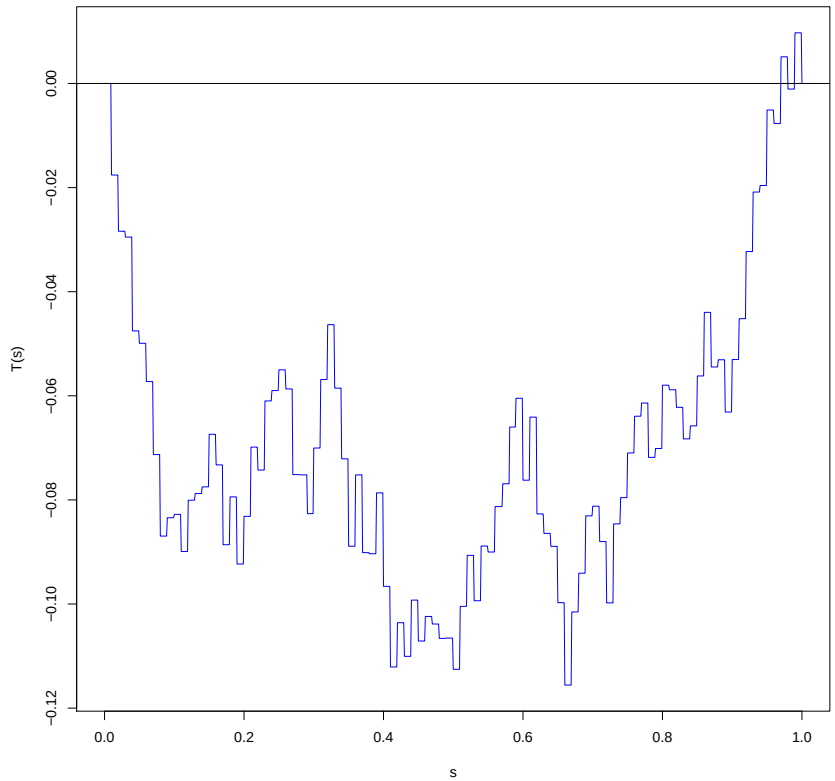
Regresní model s $\beta = 0$



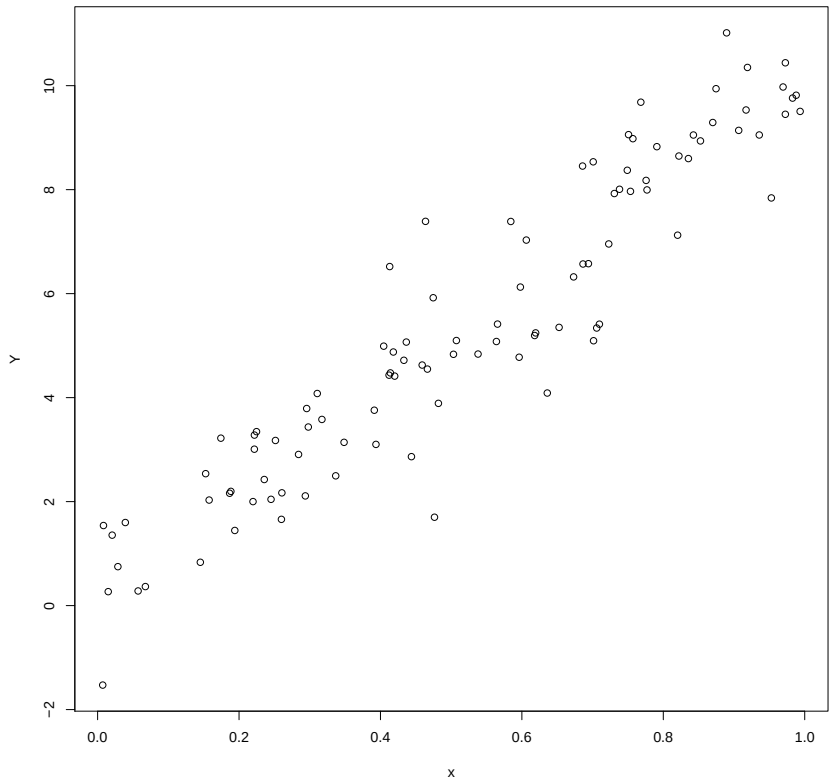


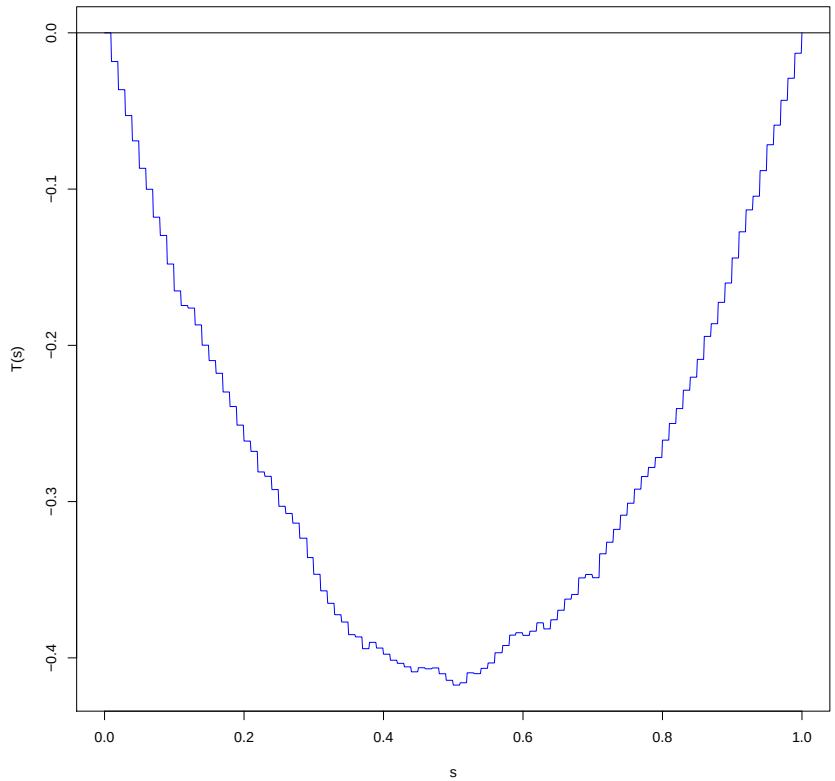
Regresní model s $\beta = 1$





Regresní model s $\beta = 10$





Testová statistika

- $K^*_{g,n} = \int_0^1 T^2_{g,n}(s) dL(s)$.
- $K^*_{g,n} = -\frac{1}{2n} \sum_{i=1}^n \sum_{j=1}^n g(x_i) g(x_j) \left| L\left(\frac{R_i}{n} -\right) - L\left(\frac{R_j}{n} -\right) \right|$.
- Rozdělení $K^*_{g,n}$ za hypotézy nezávisí na F (permutační princip).
- Za mírných předpokladů a za platnosti hypotézy

$$K^*_{g,n} \xrightarrow{d} Y_L, \quad \text{kde} \quad Y_L = \int_0^1 B^2(s) dL(s),$$

$B(s)$ je Brownův můstek v $\mathcal{C}[0,1]$.

Testová statistika II

- Vyšetřováno také asymptotické rozdělení za lokální alternativy.
- Volba funkce g. Největší síla pro $g(x_i) = \frac{x_i - \bar{x}}{\sqrt{\sum_{j=1}^{n}(x_j - \bar{x})^2}}$.
- Volba funkce L.
- Test funguje i za přítomnosti chyb v měření. Jejich přítomnost pouze snižuje sílu testu.

Simulace

e_i	$\mathcal{N}\left(0, \frac{3}{2}\right)$			$Log\left(0, \frac{3}{\sqrt{2}\pi}\right)$			$Lap\left(0, \frac{\sqrt{3}}{2}\right)$			$t(6)$		
β	MD	W	t	MD	W	t	MD	W	t	MD	W	t
0	5.15	4.41	4.85	5.34	4.67	5.25	5.03	4.50	4.82	4.90	4.20	4.61
0.1	30.0	29.3	32.3	33.4	32.2	32.6	43.6	40.0	34.6	35.8	34.3	33.9
−0.1	29.2	29.7	31.5	34.0	33.1	34.0	44.2	40.3	34.9	35.0	33.4	34.2
0.2	81.2	81.9	85.4	84.9	84.5	84.7	90.5	88.8	85.0	87.8	87.1	85.5
−0.2	80.6	81.6	85.1	84.9	84.7	85.3	90.5	88.9	84.8	86.8	86.1	84.4

Tabulka: Četnost zamítnutí hypotézy $\mathbf{H}_0 : \beta = 0$ (v procentech) testu minimalizujícího vzdálenost (MD), Wilcoxonova testu pro regresi (W) a klasického t-testu pro regresi (t); $n = 30$, $\alpha = 0.05$.

Simulace

e_i	$\mathcal{N}\left(0, \frac{3}{2}\right)$			$Log\left(0, \frac{3}{\sqrt{2}\pi}\right)$			$Lap\left(0, \frac{\sqrt{3}}{2}\right)$			$t(6)$		
β	MD	W	t	MD	W	t	MD	W	t	MD	W	t
0	5.15	4.41	4.85	5.34	4.67	5.25	5.03	4.50	4.82	4.90	4.20	4.61
0.1	30.0	29.3	32.3	33.4	32.2	32.6	43.6	40.0	34.6	35.8	34.3	33.9
−0.1	29.2	29.7	31.5	34.0	33.1	34.0	44.2	40.3	34.9	35.0	33.4	34.2
0.2	81.2	81.9	85.4	84.9	84.5	84.7	90.5	88.8	85.0	87.8	87.1	85.5
−0.2	80.6	81.6	85.1	84.9	84.7	85.3	90.5	88.9	84.8	86.8	86.1	84.4

Tabulka: Četnost zamítnutí hypotézy $\mathbf{H}_0 : \beta = 0$ (v procentech) testu minimalizujícího vzdálenost (MD), Wilcoxonova testu pro regresi (W) a klasického t-testu pro regresi (t); $n = 30$, $\alpha = 0.05$.

Simulace

e_i	$\mathcal{N}\left(0, \frac{3}{2}\right)$			$Log\left(0, \frac{3}{\sqrt{2}\pi}\right)$			$Lap\left(0, \frac{\sqrt{3}}{2}\right)$			$t(6)$		
β	MD	W	t	MD	W	t	MD	W	t	MD	W	t
0	5.15	4.41	4.85	5.34	4.67	5.25	5.03	4.50	4.82	4.90	4.20	4.61
0.1	30.0	29.3	32.3	33.4	32.2	32.6	43.6	40.0	34.6	35.8	34.3	33.9
−0.1	29.2	29.7	31.5	34.0	33.1	34.0	44.2	40.3	34.9	35.0	33.4	34.2
0.2	81.2	81.9	85.4	84.9	84.5	84.7	90.5	88.8	85.0	87.8	87.1	85.5
−0.2	80.6	81.6	85.1	84.9	84.7	85.3	90.5	88.9	84.8	86.8	86.1	84.4

Tabulka: Četnost zamítnutí hypotézy $\mathbf{H}_0 : \beta = 0$ (v procentech) testu minimalizujícího vzdálenost (MD), Wilcoxonova testu pro regresi (W) a klasického t-testu pro regresi (t); $n = 30$, $\alpha = 0.05$.

Proč uvažovat L^2 normu?

- Původní statistika: $K^*_{g,n} = \int_0^1 T^2_{g,n}(s) dL(s)$.
- $K^{p*}_{g,n} = \int_0^1 |T_{g,n}(s)|^p dL(s), \quad$ pro $p \in [1, \infty)$.
- $K^{p*}_{g,n} = \max\{|T_{g,n}(s)| : s \in (0,1)\} \quad$ pro $p = \infty$.
- Rozdělení $K^{p*}_{g,n}$ za hypotézy nezávisí na F (permutační princip).
- Za mírných předpokladů a za platnosti hypotézy

$$K^{p*}_{g,n} \xrightarrow{d} Y^p_L, \quad \text{kde} \quad Y^p_L = \int_0^1 |B(s)|^p dL(s),$$

$B(s)$ je Brownův můstek v $\mathcal{C}[0,1]$.

Simulace

e_i	$\mathcal{N}\left(0, \frac{3}{2}\right)$			$Log\left(0, \frac{3}{\sqrt{2}\pi}\right)$			$Lap\left(0, \frac{\sqrt{3}}{2}\right)$			$t(6)$		
β	L^2	L^1	L^∞	L^2	L^1	L^∞	L^2	L^1	L^∞	L^2	L^1	L^∞
0	5.15	5.03	2.73	5.34	5.31	2.61	5.03	5.18	2.71	4.90	4.97	2.64
0.1	30.0	31.0	17.8	33.4	34.1	20.6	43.6	43.0	30.6	35.8	36.6	22.4
−0.1	29.2	30.2	17.2	34.0	34.7	21.4	44.2	43.4	30.4	35.0	35.8	21.6
0.2	81.2	82.6	62.4	84.9	85.7	68.1	90.5	90.3	78.9	87.8	88.5	72.2
−0.2	80.6	82.6	62.2	84.9	86.0	68.8	90.5	90.4	78.9	86.8	87.4	71.2

Tabulka: Četnost zamítnutí hypotézy $\mathbf{H}_0 : \beta = 0$ (v procentech) testu minimalizujícího vzdálenost založeného na L^2, L^1 a L^∞ normě; $n = 30$, $\alpha = 0.05$.

Simulace

e_i	$\mathcal{N}\left(0, \frac{3}{2}\right)$			$Log\left(0, \frac{3}{\sqrt{2}\pi}\right)$			$Lap\left(0, \frac{\sqrt{3}}{2}\right)$			$t(6)$		
β	L^2	L^1	L^∞	L^2	L^1	L^∞	L^2	L^1	L^∞	L^2	L^1	L^∞
0	5.15	5.03	2.73	5.34	5.31	2.61	5.03	5.18	2.71	4.90	4.97	2.64
0.1	30.0	31.0	17.8	33.4	34.1	20.6	43.6	43.0	30.6	35.8	36.6	22.4
−0.1	29.2	30.2	17.2	34.0	34.7	21.4	44.2	43.4	30.4	35.0	35.8	21.6
0.2	81.2	82.6	62.4	84.9	85.7	68.1	90.5	90.3	78.9	87.8	88.5	72.2
−0.2	80.6	82.6	62.2	84.9	86.0	68.8	90.5	90.4	78.9	86.8	87.4	71.2

Tabulka: Četnost zamítnutí hypotézy $\mathbf{H}_0 : \beta = 0$ (v procentech) testu minimalizujícího vzdálenost založeného na L^2, L^1 a L^∞ normě; $n = 30$, $\alpha = 0.05$.

Simulace

e_i	$\mathcal{N}\left(0, \frac{3}{2}\right)$			$Log\left(0, \frac{3}{\sqrt{2}\pi}\right)$			$Lap\left(0, \frac{\sqrt{3}}{2}\right)$			$t(6)$		
β	L^2	L^1	L^∞	L^2	L^1	L^∞	L^2	L^1	L^∞	L^2	L^1	L^∞
0	5.15	5.03	2.73	5.34	5.31	2.61	5.03	5.18	2.71	4.90	4.97	2.64
0.1	30.0	31.0	17.8	33.4	34.1	20.6	43.6	43.0	30.6	35.8	36.6	22.4
−0.1	29.2	30.2	17.2	34.0	34.7	21.4	44.2	43.4	30.4	35.0	35.8	21.6
0.2	81.2	82.6	62.4	84.9	85.7	68.1	90.5	90.3	78.9	87.8	88.5	72.2
−0.2	80.6	82.6	62.2	84.9	86.0	68.8	90.5	90.4	78.9	86.8	87.4	71.2

Tabulka: Četnost zamítnutí hypotézy $\mathbf{H}_0 : \beta = 0$ (v procentech) testu minimalizujícího vzdálenost založeného na L^2, L^1 a L^∞ normě; $n = 30$, $\alpha = 0.05$.

Odhady minimalizující vzdálenost

- $$T_{g,n}(s,t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} g(x_i) \mathbb{I}\{R_{i,t} \leq ns\}, \quad 0 \leq s \leq 1,$$
$$K^*_{g,n}(t) = \int_0^1 T^2_{g,n}(s,t) dL(s),$$
- $R_{i,t}$ je pořadí $Y_i - x_i t$ mezi $Y_1 - x_1 t, \ldots, Y_n - x_n t$.
- L je distribuční funkce na $[0,1]$ a g reálná (váhová) funkce taková, že $\sum_{i=1}^{n} g(x_i) = 0$.
- $\widehat{\beta_{g,n}} = \text{argmin}\{K^*_{g,n}(t) : t \in \mathbb{R}\}$.

Odhady minimalizující vzdálenost - vlastnosti

- $K^*_{g,n}(t) = -\frac{1}{2n}\sum_{i=1}^{n}\sum_{j=1}^{n} g(x_i)g(x_j)\left|L\left(\frac{R_{i,t}}{n}-\right) - L\left(\frac{R_{j,t}}{n}-\right)\right|.$
- Za mírných předpokladů platí

$$\sqrt{n}(\widehat{\beta_{g,n}} - \beta) \xrightarrow{d} Y, \quad \text{kde} \quad Y \sim \mathcal{N}\left(0, \sigma^2_{f,L}\cdot\frac{(GQ)^4}{G^2Q^3}\right).$$

- $Q = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2.$
- $G = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} g^2(x_i).$
- $GQ = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} g(x_i)(x_i-\bar{x}).$

Odhady minimalizující vzdálenost - speciální případ

- Volba $g(x_i) = x_i - \bar{x}$ vede na asymptoticky vydatný odhad parametru β.
- Pro $g(x_i) = x_i - \bar{x}$ a $L(s) = s$ platí

$$\sqrt{n}(\widehat{\beta_{g,n}} - \beta) \xrightarrow{d} Y, \quad \text{kde} \quad Y \sim \mathcal{N}\left(0, \sigma_f^2 \cdot Q^{-1}\right),$$

$$\sigma_f^2 = \frac{\int\int [F(x \wedge y) - F(x)F(y)] f^2(x) f^2(y) dx dy}{\left(\int f^3(x) dx\right)^2}.$$

Srovnání ARE odhadů

F	$ARE(MD, W)$	$ARE(MD, LSE)$
Laplace	1.667	1.309
Logistic	0.988	1.034
Normal	0.957	0.914
Cauchy	1.278	∞

Tabulka: Asymptotická relativní vydatnost odhadu minimalizujícího vzdálenost (MD), R-odhadu založeného na Wilcoxonových skórech (W) a odhadu metodou nejmenších čtverců (LSE) pro různé chyby modelu.

Odhady minimalizující vzdálenost - zobecnění

- Zobecnění opět spočívá v uvažování jiné než L^2 normy.
-

$$T_{g,n}(s,t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} g(x_i) \mathbb{I}\{R_{i,t} \leq ns\}, \quad 0 \leq s \leq 1,$$

$$K_{g,n}^{p*}(t) = \int_0^1 |T_{g,n}(s,t)|^p dL(s), \quad \text{pro} \quad p \in [1, \infty].$$

- $R_{i,t}$ je pořadí $Y_i - x_i t$ mezi $Y_1 - x_1 t, \ldots, Y_n - x_n t$.
- L je distribuční funkce na $[0,1]$ a g reálná (váhová) funkce taková, že $\sum_{i=1}^{n} g(x_i) = 0$.
- $\widehat{\beta_{g,n}^p} = \text{argmin}\{K_{g,n}^{p*}(t) : t \in \mathbb{R}\}$.

Závěrečné poznámky, výhledy do budoucna

- Pro $p = 1$ příslušný odhad má v mnoha situacích vyšší vydatnost než původní odhad založený na L^2 normě.

Co by se dalo ještě udělat:

- Zaměřit se více na volbu funkce g a robustifikaci vzhledem k vlivným pozorováním.
- Volba optimální normy pro daný model.
- Popsat chování odhadů v modelu s chybami měření.
- Vícerozměrný případ.

Poděkování

Děkuji za pozornost.

Práce byla spolufinancována grantem MUNI/A/1234/2015.