Test Monte Carlo su approcci Monte Carlo: Least Median of Squares come caso di studio

Dario Comanducci

Elaborato per il Corso di Statistical Learning

Master in Data Science and Statistical Learning
Università degli Studi di Firenze

22 Luglio 2024

Least Median of Squares Algoritmo di regressione *robusta* resistente agli *outliers*, basato su campionamento Monte Carlo

$$\hat{\beta} = \arg\min_{\beta} \text{median}_n \left(y_n - (\beta_0 + \beta_1 x_{1n} + \cdots + \beta_q x_{qn}) \right)^2 \quad n = 1 \ldots N$$

Outliers Un *outlier* è un'osservazione che appare deviare marcatamente dagli altri membri del campione in cui si verifica.

Robustezza Uno stimatore o un procedimento statistico è *robusto* se la sua capacità di stimare correttamente i parametri rimane valida nonostante alcune delle assunzioni che giustificano il metodo impiegato vengano meno

Least Median of Squares Algoritmo di regressione *robusta* resistente agli *outliers*, basato su campionamento Monte Carlo

$$\hat{\beta} = \arg\min_{\beta} \text{median}_n \left(y_n - (\beta_0 + \beta_1 x_{1n} + \cdots + \beta_q x_{qn}) \right)^2 \quad n = 1 \ldots N$$

Outliers Un *outlier* è un'osservazione che appare deviare marcatamente dagli altri membri del campione in cui si verifica.

Robustezza Uno stimatore o un procedimento statistico è *robusto* se la sua capacità di stimare correttamente i parametri rimane valida nonostante alcune delle assunzioni che giustificano il metodo impiegato vengano meno

Least Median of Squares Algoritmo di regressione *robusta* resistente agli *outliers*, basato su campionamento Monte Carlo

$$\hat{\beta} = \arg\min_{\beta} \text{median}_n \left(y_n - (\beta_0 + \beta_1 x_{1n} + \cdots + \beta_q x_{qn}) \right)^2 \quad n = 1 \ldots N$$

Outliers Un *outlier* è un'osservazione che appare deviare marcatamente dagli altri membri del campione in cui si verifica.

Robustezza Uno stimatore o un procedimento statistico è *robusto* se la sua capacità di stimare correttamente i parametri rimane valida nonostante alcune delle assunzioni che giustificano il metodo impiegato vengano meno

Regressione semplice ai minimi quadrati

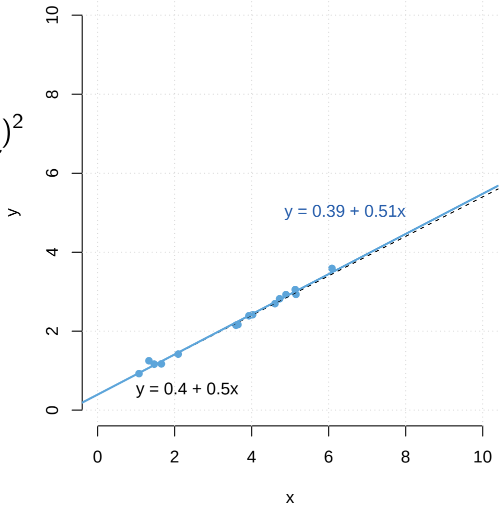
$y = \beta_0 + \beta_1 x$

N coppie di punti (x_n, y_n), $n = 1 \dots N$ con
$y_n = \beta_0 + \beta_1 x_n + \varepsilon_n, \;\; \varepsilon \sim \mathrm{N}(0, \sigma_e^2)$

$$\hat{\beta} = [\hat{\beta}_0 \;\; \hat{\beta}_1]^\top = \arg\min_{\beta} \sum_{n=1}^{N} (\underbrace{y_n - (\beta_0 + \beta_1 x_n)}_{r_n})^2$$

$$\hat{\beta} = \arg\min_{\beta} \frac{\sum_n r_n^2}{N} = \arg\min_{\beta} \mathrm{mean}_n\{r_n^2\}$$

```r
set.seed(111); b0 = 0.4; b1 = 0.5
N = 15; x = sort(runif(N, 1, 8))
sigma_e = 0.1; e = rnorm(N,0,sigma_e)
y = b0 + b1*x + e
ls = lm(y ~ x)
```



Test Monte Carlo su LMedS

D. Comanducci

La mappa d'errore
Intermezzo matematico

$$y = \beta_0 + \beta_1 x \iff \beta_1 x - y + \beta_0 = 0$$

$$\mathbf{l} = \begin{bmatrix} a \\ b \\ c \end{bmatrix} \propto \begin{bmatrix} \beta_1 \\ -1 \\ \beta_0 \end{bmatrix} \Rightarrow [x\ y\ 1]^\top \mathbf{l} = 0$$

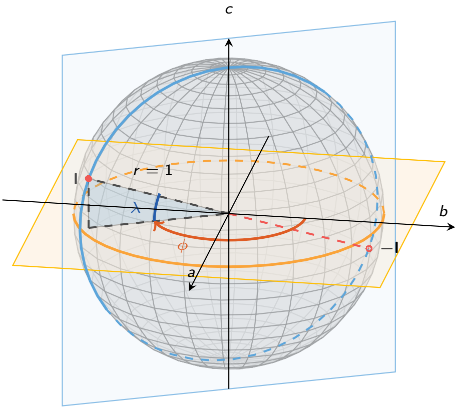
senza perdere di generalità, $\|\mathbf{l}\| = 1$

Per $r = 1$,

$$a = r \cos \lambda \sin \phi$$

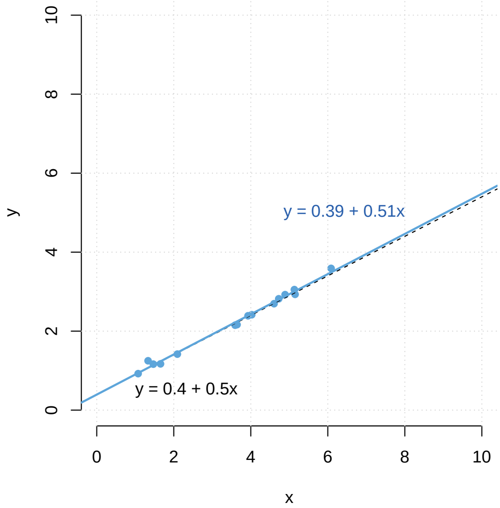
$$b = r \cos \lambda \cos \phi$$

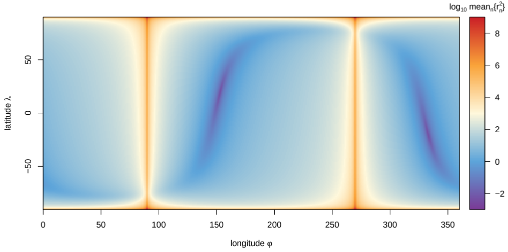
$$c = r \sin \lambda$$



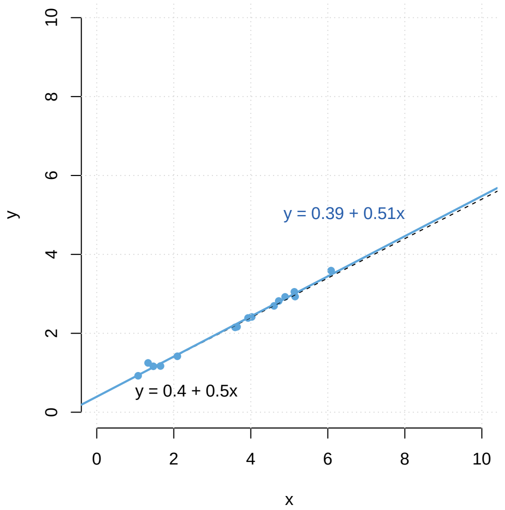
La mappa d'errore

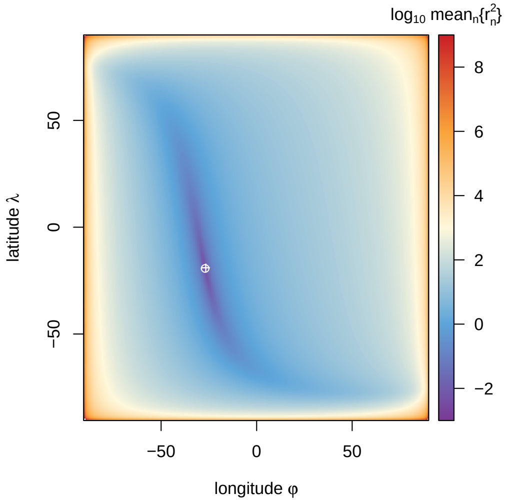
Least Mean of Squares





La mappa d'errore
Least Mean of Squares





Test Monte Carlo su LMedS

D. Comanducci

Introduzione

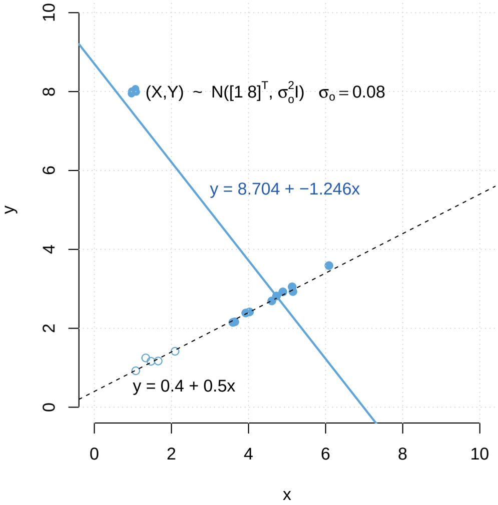
Least Median of Squares + RLS

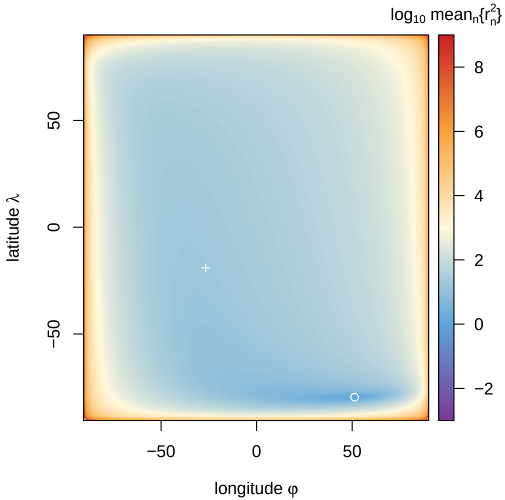
Monte Carlo

Chiusura

La mappa d'errore

Least Mean of Squares

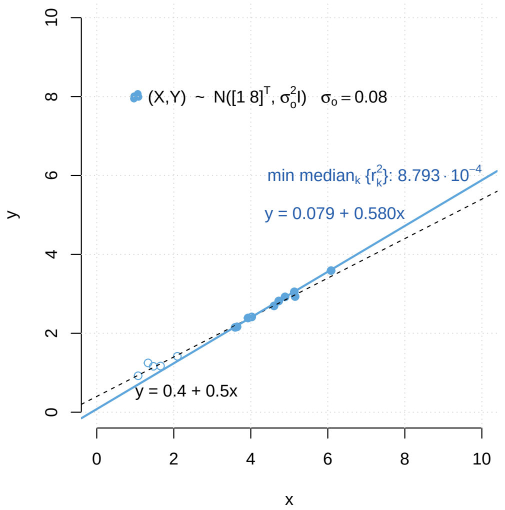


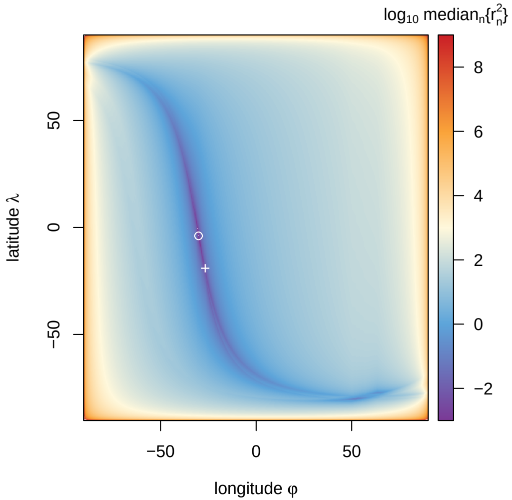


Test Monte Carlo su LMedS

D. Comanducci

La mappa d'errore
Least Median of Squares





Test Monte Carlo su LMedS

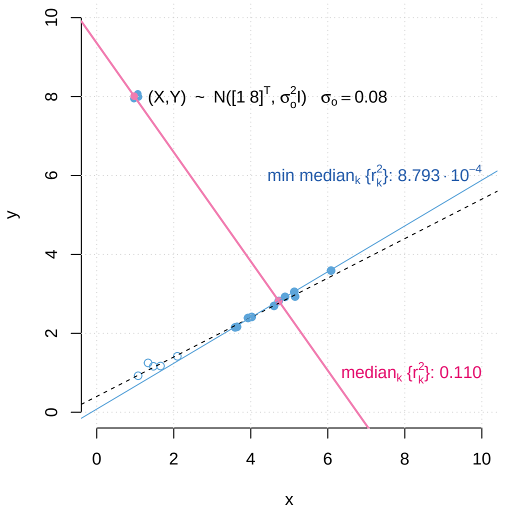
D. Comanducci

Introduzione

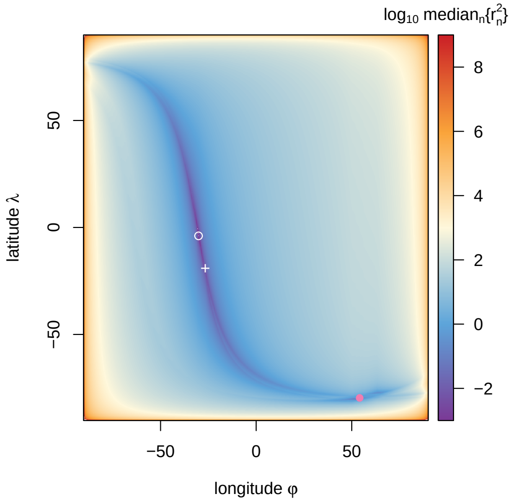
Least Median of Squares + RLS

Monte Carlo

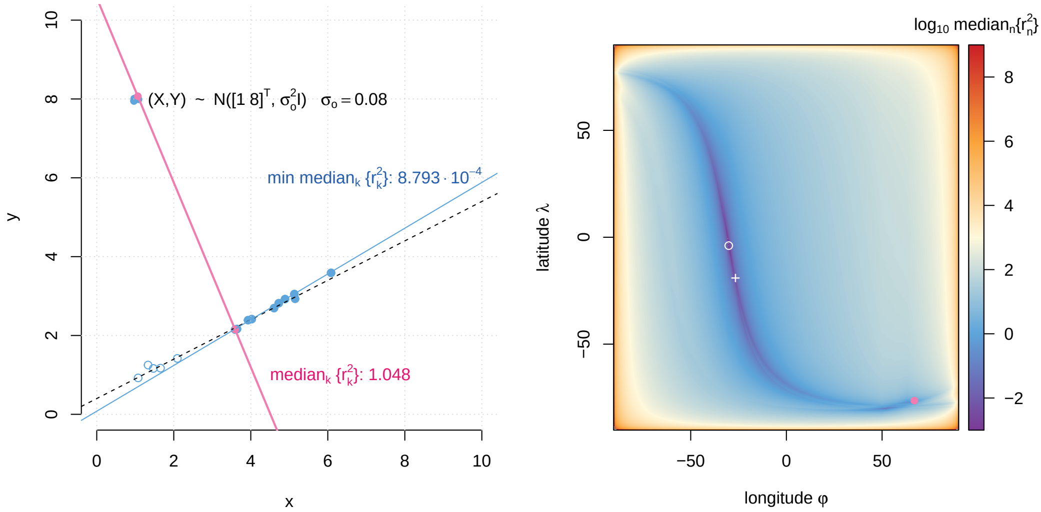
Chiusura



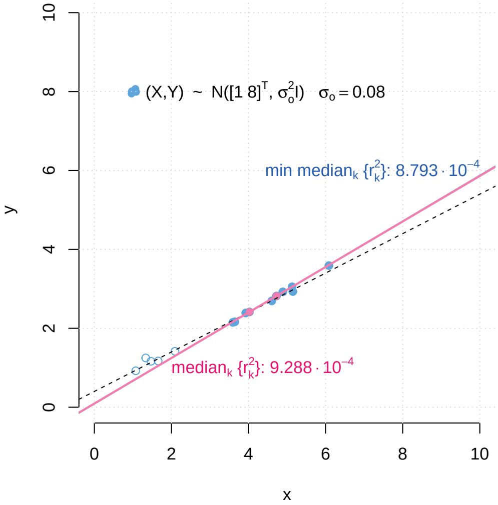
(X,Y) ~ N([1 8]^T, σ_o^2 I)   σ_o = 0.08
min median_k {r_k^2}: 8.793 · 10^−4
median_k {r_k^2}: 0.110
y
x
0
2
4
6
8
10

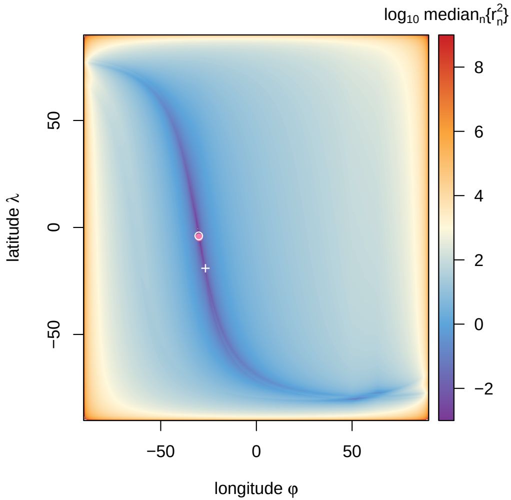


log10 median_n{r_n^2}
latitude λ
longitude φ
50
0
−50
8
6
4
2
0
−2



(X,Y) ~ N([1 8]^T, σ_o^2 I)   σ_o = 0.08
min median_k {r_k^2}: 8.793 · 10^−4
median_k {r_k^2}: 1.048
x
y
log10 median_n{r_n^2}
latitude λ
longitude φ


(X,Y) ~ N([1 8]^T, σ_o^2 I)   σ_o = 0.08
min median_k {r_k^2}: 8.793 · 10^−4
median_k {r_k^2}: 9.288 · 10^−4
y
x
0
2
4
6
8
10


log10 median_n{r_n^2}
latitude λ
longitude φ
50
0
−50
8
6
4
2
0
−2

Algoritmo LMedS

Input: tentativi T; $\mathcal{D} = \{(y_n, \mathbf{x}_n)\}$, $y_n \in \mathbb{R}$, $\mathbf{x}_n \in \mathbb{R}^q$

1. Poni $t := 1$; med $:= +\infty$; $\hat{\beta} = [\,]$
2. Seleziona casualmente un sottoinsieme minimale $\mathcal{D}_q^{(t)} = \{(y_i, \mathbf{x}_i)\} \subset \mathcal{D}$ di $p = q+1$ dati per cui $\beta^{(t)}$ è t.c. $y_i - (\beta_0^{(t)} + \beta_1^{(t)} x_{1i} + \ldots + \beta_q^{(t)} x_{qi}) = 0$
3. Calcola $r_n := y_n - ([1\ \mathbf{x}_k^\top]\beta^{(t)})\ \forall (y_n, \mathbf{x}_n) \in \mathcal{D}$
4. Se med $> \text{med}_n r_n^2$
   $\hat{\beta} := \beta^{(t)}$; med $:= \text{med}_n r_n^2$
5. $t := t + 1$
6. Se $t < T$, vai a 2
7. Restituisci $\hat{\beta}$ e med

		T	
p	τ_N	$N \le \tau_N$	$N > \tau_N$
2	50	$c_{N,p}$	1000
3	22	$c_{N,p}$	1500
4	17	$c_{N,p}$	2000
5	15	$c_{N,p}$	2500
6	14	$c_{N,p}$	3000
≥ 7		3000	3000

$$N = \|\mathcal{D}\|$$

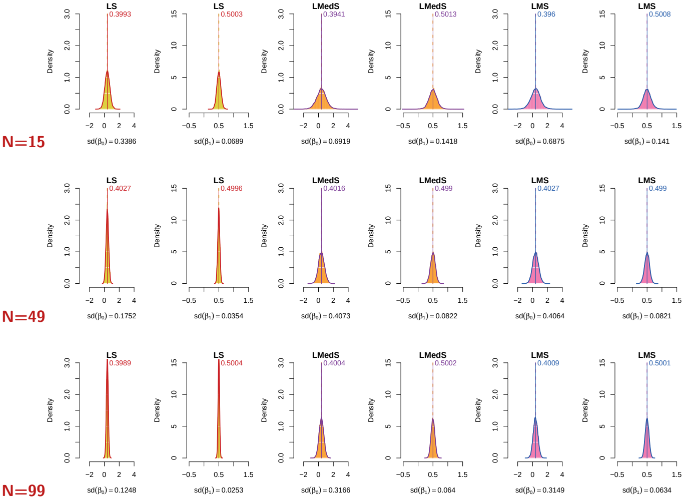
$$c_{N,p} = \binom{N}{p} = \frac{N!}{p!(N-p)!}$$

Least Median of Squares vs. Least Squares

E in assenza di outliers?

```
nsim=10000
for N %in% c(15,49,99){
# DGP:
b0 = 0.4; b1 = 0.5
x = runif(N, 1, 8)
sigma_e = 0.5
e = rnorm(N,0,sigma_e)
y = b0 + b1*x + e
...}
```

- LS: Least Squares
- LMedS: Least Median of Squares
- LMS: lmsreg(...)



Reweighted Least Squares (RLS)

Incremento dell'efficienza

$$\widehat{|r|} = \sqrt{\mathrm{med}_n r_n^2}$$

$$s^{(0)} = 1.4826\left(1+\frac{5}{N-p}\right)\widehat{|r|}$$

$$w_n^{(0)} = \mathbb{1}(|r_n/s_0| \leq 2.5)$$

$$\hat{\sigma} = \sqrt{\frac{\sum_n w_n^{(0)} r_n^2}{\sum_n w_n^{(0)} - p}}$$

$$w_n = \mathbb{1}(|r_n/\hat{\sigma}| \leq 2.5)$$

$$\hat{\beta} = \arg\min_{\beta} \sum_n w_n r_n^2$$

Reweighted Least Squares (RLS)

Incremento dell'efficienza

$$\widehat{|r|} = \sqrt{\mathrm{med}_n r_n^2}$$

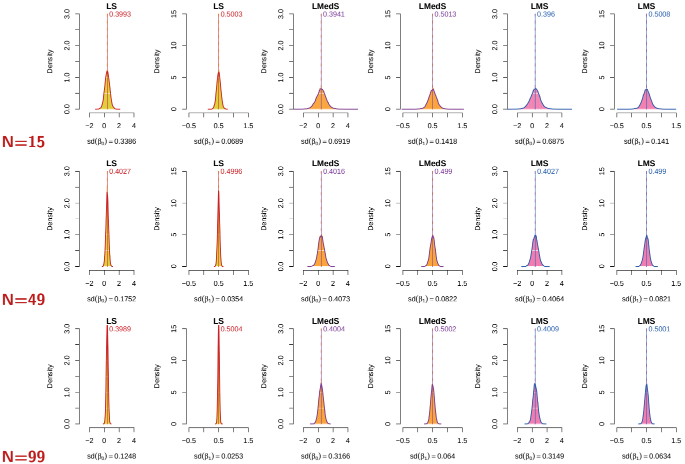
$$s^{(0)} = 1.4826\left(1+\frac{5}{N-p}\right)\widehat{|r|}$$

$$w_n^{(0)} = \mathbb{1}(|r_n/s_0| \le 2.5)$$

$$\hat{\sigma} = \sqrt{\frac{\sum_n w_n^{(0)} r_n^2}{\sum_n w_n^{(0)} - p}}$$

$$w_n = \mathbb{1}(|r_n/\hat{\sigma}| \le 2.5)$$

$$\hat{\beta} = \arg\min_{\beta} \sum_n w_n r_n^2$$



Reweighted Least Squares (RLS)

Incremento dell'efficienza

$$\widehat{|r|} = \sqrt{\text{med}_n r_n^2}$$

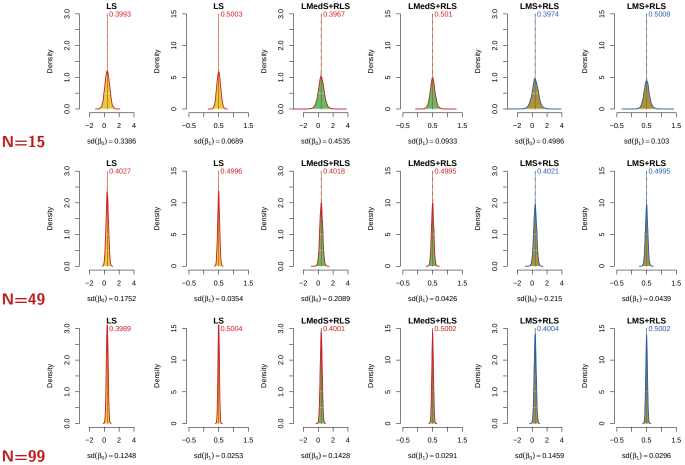
$$s^{(0)} = 1.4826\left(1 + \frac{5}{N-p}\right)\widehat{|r|}$$

$$w_n^{(0)} = \mathbb{1}(|r_n/s_0| \leq 2.5)$$

$$\hat{\sigma} = \sqrt{\frac{\sum_n w_n^{(0)} r_n^2}{\sum_n w_n^{(0)} - p}}$$

$$w_n = \mathbb{1}(|r_n/\hat{\sigma}| \leq 2.5)$$

$$\hat{\beta} = \arg\min_{\beta} \sum_n w_n r_n^2$$



Reweighted Least Squares (RLS)

Incremento dell'efficienza

$$\widehat{|r|} = \sqrt{\mathrm{med}_n r_n^2}$$

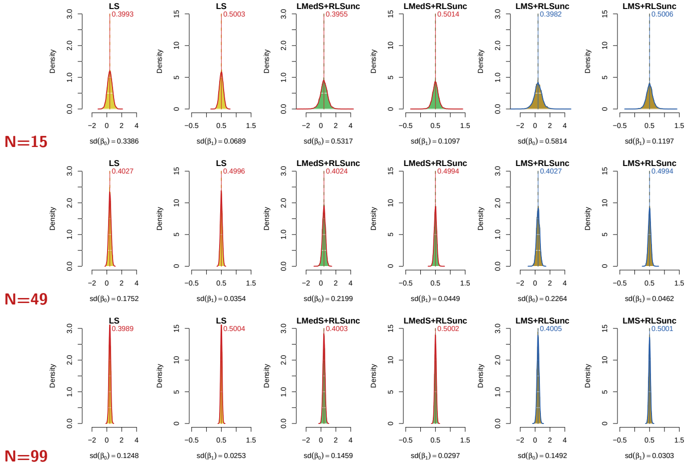
$$s^{(0)} = 1.4826 \cancel{\left(1 + \frac{5}{N-p}\right)} \widehat{|r|}$$

$$w_n^{(0)} = \mathbb{1}(|r_n/s_0| \leq 2.5)$$

$$\hat{\sigma} = \sqrt{\frac{\sum_n w_n^{(0)} r_n^2}{\sum_n w_n^{(0)} - p}}$$

$$w_n = \mathbb{1}(|r_n/\hat{\sigma}| \leq 2.5)$$

$$\hat{\beta} = \arg\min_{\beta} \sum_n w_n r_n^2$$

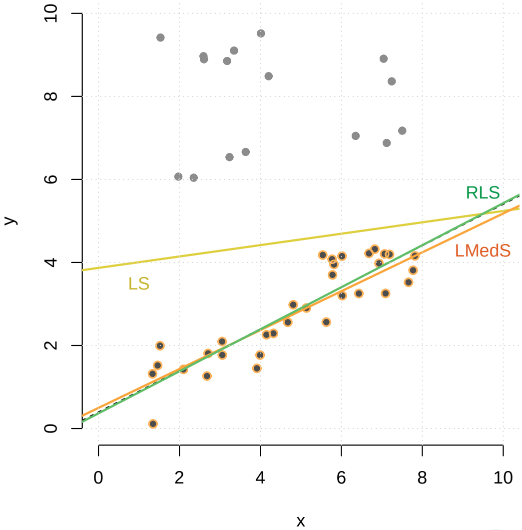


LMedS+RLS vs LS

Con outliers, scenario #1

$x \sim \mathrm{U}(a=1, b=8)$
$\varepsilon \sim \mathrm{N}(\mu=0, \sigma=0.5)$
$b_0 = 0.4, \quad b_1 = 0.5$
$y_n = b_0 + b_1 x_n + \varepsilon_n \quad (n = 1 \ldots N)$
$N \in \{15, 49, 99\}$

$n_{\text{sims}} = 10\,000$
$p_o = 0.333$
$N_o = \mathrm{round}(p_o N)$
$n_o = \mathrm{sample}(1:N, N_o, \mathrm{repl.} = F)$
$y_{n_o} \sim \mathrm{U}(a=6, b=10)$

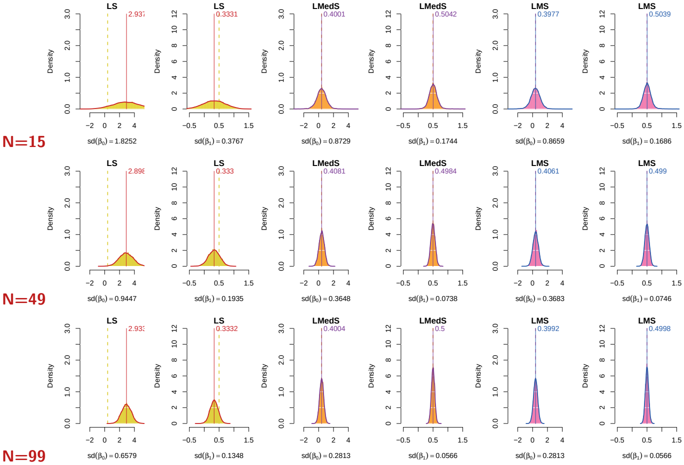


LMedS+RLS vs LS

Con outliers, scenario #1

$x \sim \mathrm{U}(a=1, b=8)$
$\varepsilon \sim \mathrm{N}(\mu=0, \sigma=0.5)$
$b_0=0.4, \quad b_1=0.5$
$y_n=b_0+b_1 x_n+\varepsilon_n \quad (n=1\ldots N)$
$N \in \{15, 49, 99\}$

$n_{\mathrm{sims}} = 10\,000$
$p_o=0.333$
$N_o=\mathrm{round}(p_o N)$
$n_o=\mathrm{sample}(1:N, N_o, \mathrm{repl.}=F)$
$y_{n_o} \sim \mathrm{U}(a=6, b=10)$



LMedS+RLS vs LS

Con outliers, scenario #1

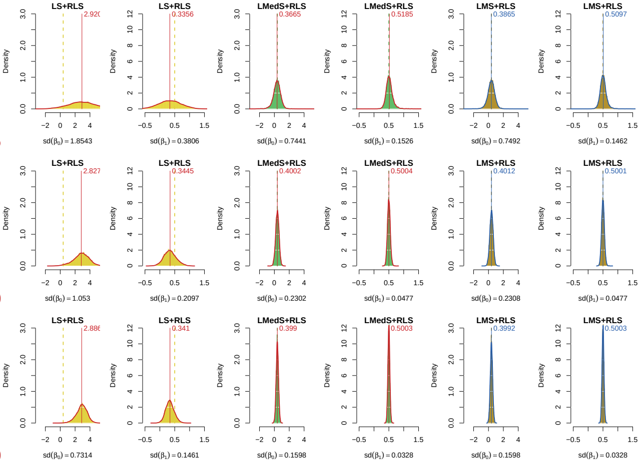
$x \sim \mathrm{U}(a=1, b=8)$
$\varepsilon \sim \mathrm{N}(\mu=0, \sigma=0.5)$
$b_0 = 0.4, \quad b_1 = 0.5$
$y_n = b_0 + b_1 x_n + \varepsilon_n \quad (n = 1 \ldots N)$
$N \in \{15, 49, 99\}$

$n_{\text{sims}} = 10\,000$
$p_o = 0.333$
$N_o = \text{round}(p_o N)$
$n_o = \text{sample}(1:N, N_o, \text{repl.} = F)$
$y_{n_o} \sim \mathrm{U}(a=6, b=10)$

N=15

N=49

N=99

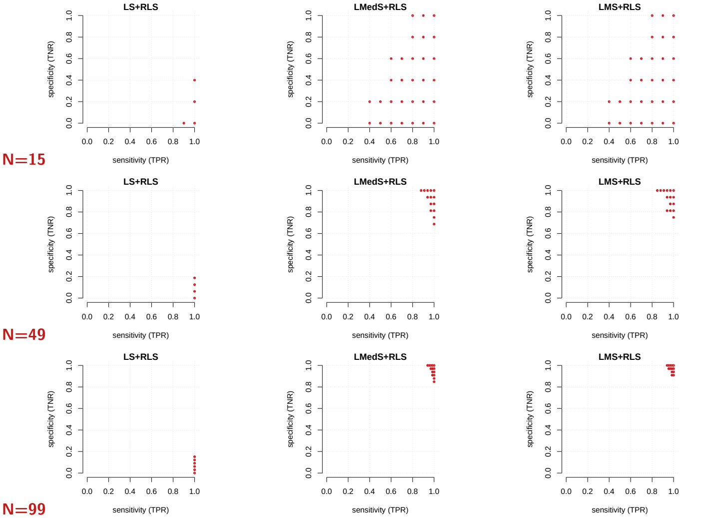


LMedS+RLS vs LS

Con outliers, scenario #1

$x \sim \mathrm{U}(a=1, b=8)$
$\varepsilon \sim \mathrm{N}(\mu=0, \sigma=0.5)$
$b_0=0.4, \quad b_1=0.5$
$y_n=b_0+b_1 x_n+\varepsilon_n \quad (n=1\ldots N)$
$N \in \{15, 49, 99\}$

$n_{\text{sims}} = 10\,000$
$p_o=0.333$
$N_o=\text{round}(p_o N)$
$n_o=\text{sample}(1:N, N_o, \text{repl.}=F)$
$y_{n_o} \sim \mathrm{U}(a=6, b=10)$

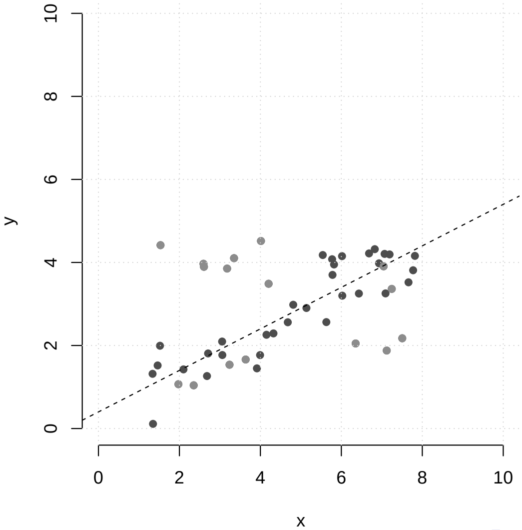


LMedS+RLS vs LS

Con outliers, scenario #2

$x \sim \mathrm{U}(a=1, b=8)$
$\varepsilon \sim \mathrm{N}(\mu=0, \sigma=0.5)$
$b_0=0.4, \quad b_1=0.5$
$y_n=b_0+b_1 x_n+\varepsilon_n \quad (n=1\ldots N)$
$N \in \{15, 49, 99\}$

$n_{\text{sims}} = 10\,000$
$p_o=0.333$
$N_o=\text{round}(p_o N)$
$n_o=\text{sample}(1:N, N_o, \text{repl.}=F)$
$y_{n_o} \sim \mathrm{U}(a=1, b=5)$



LMedS+RLS vs LS

Con outliers, scenario #2

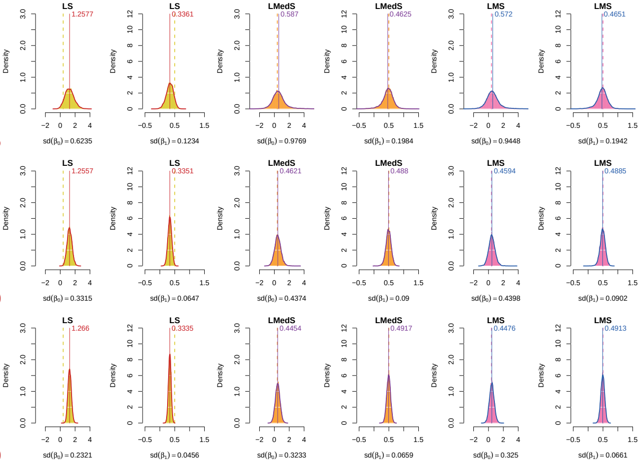
$x \sim \mathrm{U}(a=1, b=8)$
$\varepsilon \sim \mathrm{N}(\mu=0, \sigma=0.5)$
$b_0=0.4, \quad b_1=0.5$
$y_n = b_0 + b_1 x_n + \varepsilon_n \quad (n=1\ldots N)$
$N \in \{15, 49, 99\}$

$n_{\mathrm{sims}} = 10\,000$
$p_o = 0.333$
$N_o = \mathrm{round}(p_o N)$
$n_o = \mathrm{sample}(1:N, N_o, \mathrm{repl.}=F)$
$y_{n_o} \sim \mathrm{U}(a=1, b=5)$

N=15

N=49

N=99

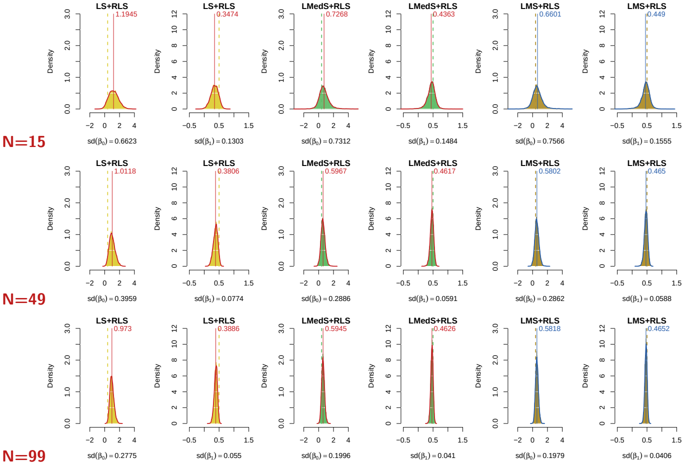


LMedS+RLS vs LS

Con outliers, scenario #2

$x \sim \mathrm{U}(a=1, b=8)$
$\varepsilon \sim \mathrm{N}(\mu=0, \sigma=0.5)$
$b_0=0.4, \quad b_1=0.5$
$y_n=b_0+b_1 x_n+\varepsilon_n \quad (n=1\ldots N)$
$N \in \{15, 49, 99\}$

$n_{\text{sims}} = 10\,000$
$p_o=0.333$
$N_o=\text{round}(p_o N)$
$n_o=\text{sample}(1:N, N_o, \text{repl.}=F)$
$y_{n_o} \sim \mathrm{U}(a=1, b=5)$



LMedS+RLS vs LS

Con outliers, scenario #2

$x \sim \mathrm{U}(a=1, b=8)$

$\varepsilon \sim \mathrm{N}(\mu=0, \sigma=0.5)$

$b_0 = 0.4, \quad b_1 = 0.5$

$y_n = b_0 + b_1 x_n + \varepsilon_n \quad (n = 1 \ldots N)$

$N \in \{15, 49, 99\}$

$n_{\text{sims}} = 10\,000$

$p_o = 0.333$

$N_o = \text{round}(p_o N)$

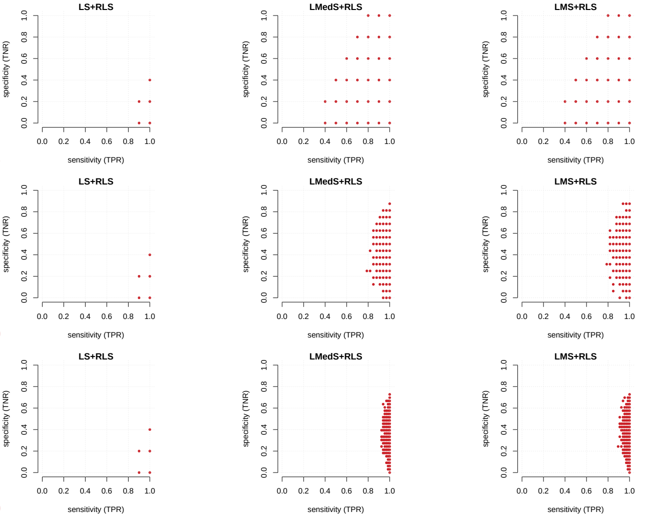
$n_o = \text{sample}(1:N, N_o, \text{repl.} = F)$

$y_{n_o} \sim \mathrm{U}(a=1, b=5)$

N=15

N=49

N=99

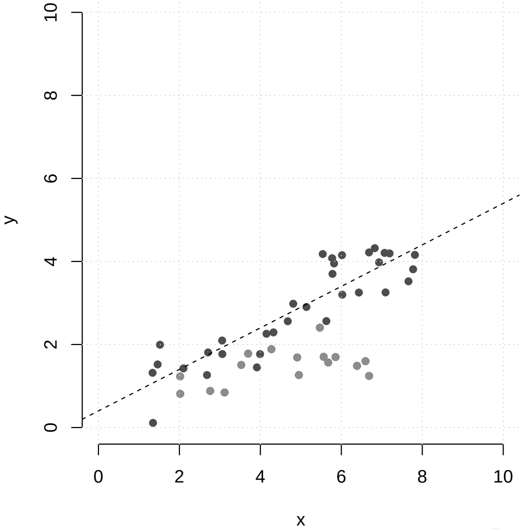


LMedS+RLS vs LS

Con outliers, scenario #3

$x \sim \mathrm{U}(a = 1, b = 8)$
$\varepsilon \sim \mathrm{N}(\mu = 0, \sigma = 0.5)$
$b_0 = 0.4, \quad b_1 = 0.5$
$y_n = b_0 + b_1 x_n + \varepsilon_n \quad (n = 1 \ldots N)$
$N \in \{15, 49, 99\}$

$n_{\text{sims}} = 10\,000$
$p_o = 0.333$
$N_o = \text{round}(p_o N)$
$n_o = \text{sample}(1 : N, N_o, \text{repl.} = F)$
$y_{n_o} \sim \mathrm{N}(\mu = 1.5, \sigma = 0.4)$



LMedS+RLS vs LS

Con outliers, scenario #3

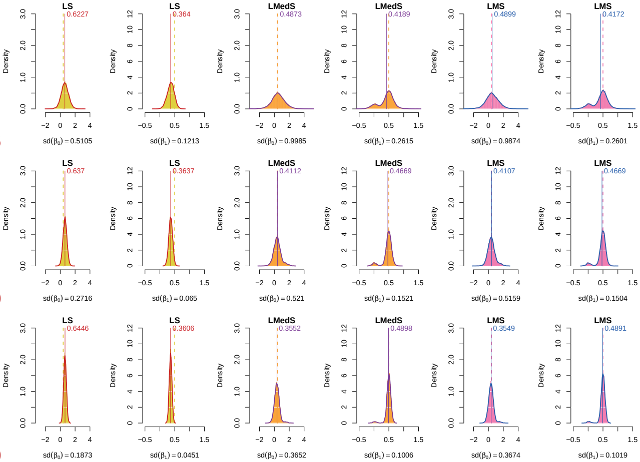
$x \sim \mathrm{U}(a=1, b=8)$
$\varepsilon \sim \mathrm{N}(\mu=0, \sigma=0.5)$
$b_0=0.4, \quad b_1=0.5$
$y_n=b_0+b_1x_n+\varepsilon_n \quad (n=1\ldots N)$
$N \in \{15, 49, 99\}$

$n_{\text{sims}} = 10\,000$
$p_o=0.333$
$N_o=\mathrm{round}(p_o N)$
$n_o=\mathrm{sample}(1:N, N_o, \mathrm{repl.}=F)$
$y_{n_o} \sim \mathrm{N}(\mu=1.5, \sigma=0.4)$

N=15

N=49

N=99



LMedS+RLS vs LS

Con outliers, scenario #3

$x \sim \mathrm{U}(a=1, b=8)$

$\varepsilon \sim \mathrm{N}(\mu=0, \sigma=0.5)$

$b_0 = 0.4, \quad b_1 = 0.5$

$y_n = b_0 + b_1 x_n + \varepsilon_n \quad (n = 1 \ldots N)$

$N \in \{15, 49, 99\}$

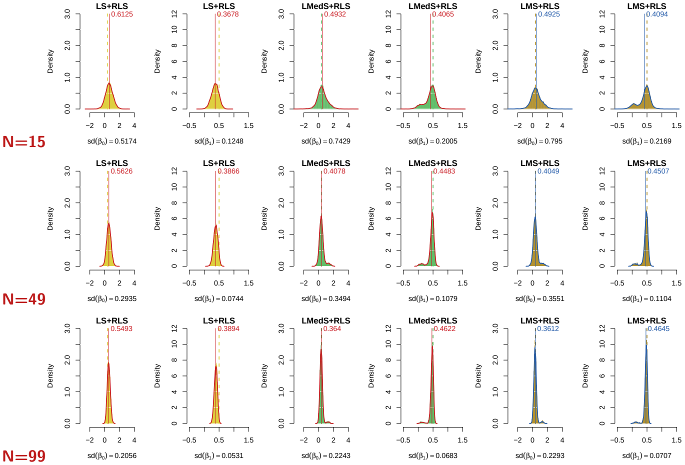
$n_{\text{sims}} = 10\,000$

$p_o = 0.333$

$N_o = \text{round}(p_o N)$

$n_o = \text{sample}(1:N, N_o, \text{repl.} = F)$

$y_{n_o} \sim \mathrm{N}(\mu=1.5, \sigma=0.4)$

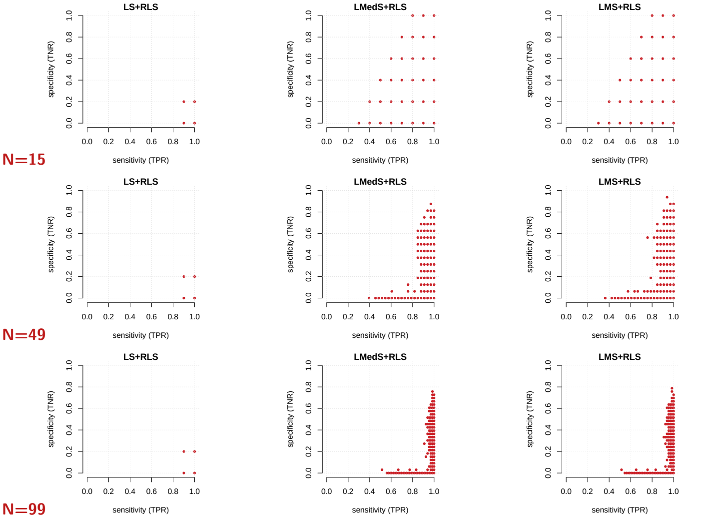


LMedS+RLS vs LS

Con outliers, scenario #3

$x \sim \mathrm{U}(a=1, b=8)$
$\varepsilon \sim \mathrm{N}(\mu=0, \sigma=0.5)$
$b_0=0.4, \quad b_1=0.5$
$y_n=b_0+b_1 x_n+\varepsilon_n \quad (n=1\ldots N)$
$N \in \{15, 49, 99\}$

$n_{\text{sims}} = 10\,000$
$p_o=0.333$
$N_o=\text{round}(p_o N)$
$n_o=\text{sample}(1:N, N_o, \text{repl.}=F)$
$y_{n_o} \sim \mathrm{N}(\mu=1.5, \sigma=0.4)$

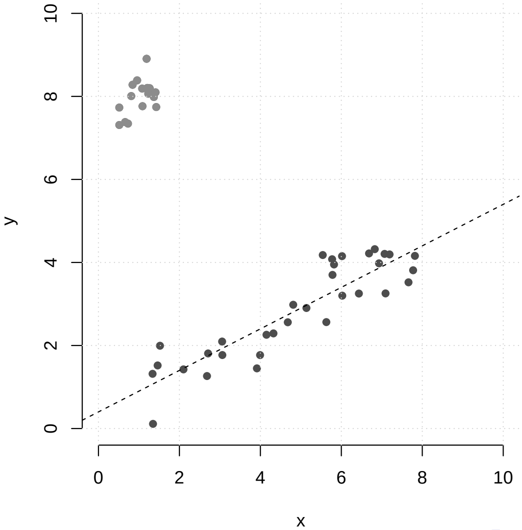


LMedS+RLS vs LS

Con outliers, scenario #4

$x \sim \mathrm{U}(a=1, b=8)$
$\varepsilon \sim \mathrm{N}(\mu=0, \sigma=0.5)$
$b_0=0.4, \quad b_1=0.5$
$y_n=b_0+b_1 x_n+\varepsilon_n \quad (n=1\ldots N)$
$N \in \{15, 49, 99\}$

$n_{\text{sims}} = 10\,000$
$p_o=0.333$
$N_o=\text{round}(p_o N)$
$n_o=\text{sample}(1:N, N_o, \text{repl.}=F)$
$x_{n_o} \sim \mathrm{N}(\mu=1.5, \sigma=0.35)$
$y_{n_o} \sim \mathrm{N}(\mu=8, \sigma=0.4)$

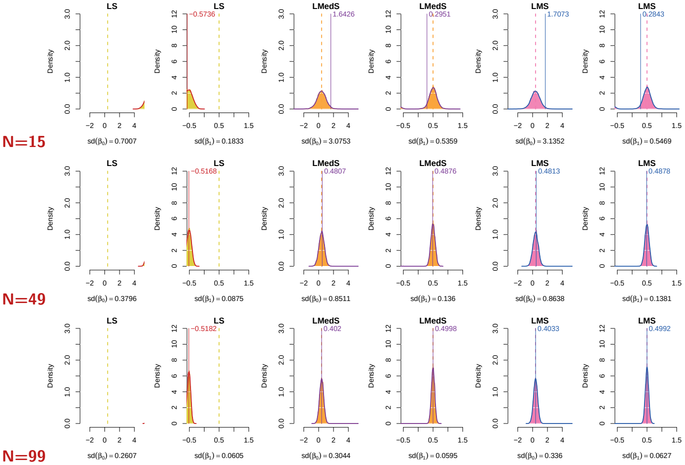


LMedS+RLS vs LS

Con outliers, scenario #4

$x \sim \mathrm{U}(a=1, b=8)$
$\varepsilon \sim \mathrm{N}(\mu=0, \sigma=0.5)$
$b_0 = 0.4, \quad b_1 = 0.5$
$y_n = b_0 + b_1 x_n + \varepsilon_n \quad (n = 1 \ldots N)$
$N \in \{15, 49, 99\}$

$n_{\text{sims}} = 10\,000$
$p_o = 0.333$
$N_o = \text{round}(p_o N)$
$n_o = \text{sample}(1 : N, N_o, \text{repl.} = F)$
$x_{n_o} \sim \mathrm{N}(\mu = 1.5, \sigma = 0.35)$
$y_{n_o} \sim \mathrm{N}(\mu = 8, \sigma = 0.4)$

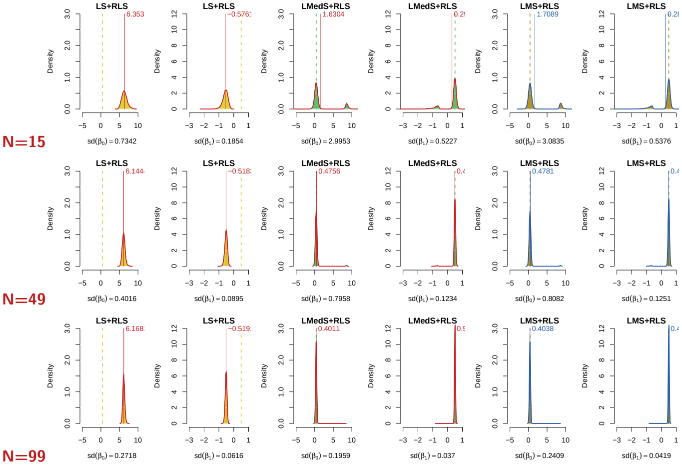


LMedS+RLS vs LS

Con outliers, scenario #4

$$x \sim \mathrm{U}(a=1, b=8)$$
$$\varepsilon \sim \mathrm{N}(\mu=0, \sigma=0.5)$$
$$b_0=0.4, \quad b_1=0.5$$
$$y_n=b_0+b_1 x_n+\varepsilon_n \quad (n=1\ldots N)$$
$$N \in \{15, 49, 99\}$$

$$n_{\text{sims}} = 10\,000$$
$$p_o=0.333$$
$$N_o=\text{round}(p_o N)$$
$$n_o=\text{sample}(1:N, N_o, \text{repl.}=F)$$
$$x_{n_o} \sim \mathrm{N}(\mu=1.5, \sigma=0.35)$$
$$y_{n_o} \sim \mathrm{N}(\mu=8, \sigma=0.4)$$

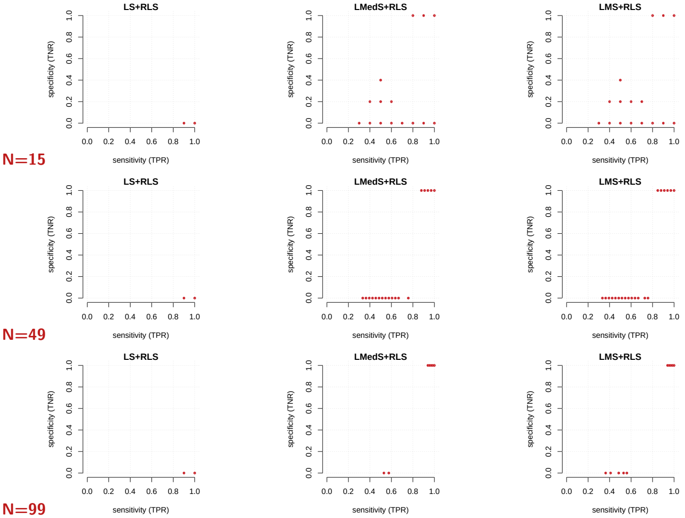


LMedS+RLS vs LS

Con outliers, scenario #4

$x \sim \mathrm{U}(a=1, b=8)$
$\varepsilon \sim \mathrm{N}(\mu=0, \sigma=0.5)$
$b_0=0.4, \quad b_1=0.5$
$y_n=b_0+b_1 x_n+\varepsilon_n \quad (n=1\ldots N)$
$N \in \{15, 49, 99\}$

$n_{\text{sims}} = 10\,000$
$p_o=0.333$
$N_o=\text{round}(p_o N)$
$n_o=\text{sample}(1:N, N_o, \text{repl.}=F)$
$x_{n_o} \sim \mathrm{N}(\mu=1.5, \sigma=0.35)$
$y_{n_o} \sim \mathrm{N}(\mu=8, \sigma=0.4)$



Conclusioni su Least Median of Squares

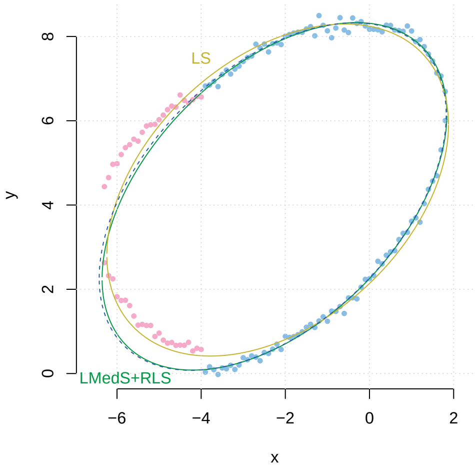
Take away

- Least Median of Squares (LMedS) è un algoritmo robusto ideato per resistere agli outliers
- Least Median of Squares (LMedS) è scarsamente efficiente
- Utile per stima iniziale da raffinare, ad esempio con Reweighted Least Squares (RLS) o M-estimators
- Test Monte Carlo applicati sul 33% di outliers hanno mostrato che la robustezza di LMedS+RLS dipende da come gli outliers sono distribuiti e dal numero di dati a disposizione
- La robustezza aumenta all'aumentare del numero di dati
- L'approccio di LMedS è generalizzabile anche al di fuori della regressione

Oltre la regressione

Esempio: stima di coniche

- $ax^2 + bxy + cy^2 + dx + ey + f = 0$
- $p = 5$ punti bastano per definire una conica
- residui: distanza geometrica punti-conica o Sampson Error



Riferimenti bibliografici

- P. J. Rousseeuw and A. Leroy. *Robust Regression and Outlier Detection*. Wiley Series in Probability and Statistics. Wiley, 1987
- J. Fox, S. Weisberg. Robust Regression in R – online appendix to *An R Companion to Applied Regression*, 3rd edition (Sage 2019)
- F. E. Grubbs, *Procedures for detecting outlying observations in samples*. Technometrics. 11(1):1–21, February 1969
- Z. Zhang. *Parameter estimation techniques: A tutorial with application to conic fitting*. Image and Vision Computing, 15(1):59–76, 1997
- R. I. Hartley and A. Zisserman. Multiple View Geometry in Computer Vision. Cambridge University Press, 2nd edition, 2004