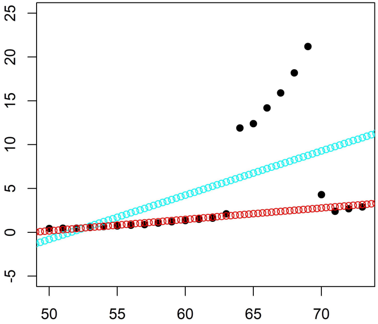
Least Weighted Absolute Value Estimator with an Application to Investment Data

Petra Vidnerová, Jan Kalina

The Czech Academy of Sciences, Institute of Computer Science, Prague

Least squares vs. robust regression



- Model $Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + e_i, \quad i = 1, \ldots, n$
- $\text{var}\, e_1 = \cdots = \text{var}\, e_n = \sigma^2$ (=nuisance parameter)
- $$X = \begin{bmatrix} X_{11} & X_{12} & \ldots & X_{1p} \\ X_{21} & X_{22} & \ldots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \ldots & X_{np} \end{bmatrix}$$
- Least squares

$$\min \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_{i1} - \cdots - b_p X_{ip})^2 \quad \text{over } (b_0, \ldots, b_p)^T \in \mathbb{R}^{p+1}$$

$$b^{LS} = (X^T X)^{-1} X^T Y$$

$$\text{var}\, b^{LS} = \sigma^2 (X^T X)^{-1}$$

Least weighted squares estimator (LWS)

- Linear regression model
  $Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + e_i, \quad i = 1, \ldots, n.$
- Residuals for a fixed value of $\mathbf{b} = (b_0, b_1, \ldots, b_p)^T \in \mathbb{R}^{p+1}$:

$$u_i(\mathbf{b}) = Y_i - b_0 - b_1 X_{i1} - \cdots - b_p X_{ip}, \quad i = 1, \ldots, n.$$

- We arrange squared residuals in ascending order:

$$u^2_{(1)}(\mathbf{b}) \leq u^2_{(2)}(\mathbf{b}) \leq \cdots \leq u^2_{(n)}(\mathbf{b}).$$

- Weight function $\psi : [0, 1] \to [0, 1]$
- The **least weighted squares** (LWS) estimator

$$\mathbf{b}^{LWS} = (b_0^{LWS}, b_1^{LWS}, \ldots, b_p^{LWS})^T = \arg\min_{b \in \mathbb{R}^{p+1}} \sum_{k=1}^{n} \psi\left(\frac{k-1}{n}\right) u^2_{(i)}(b)$$

- Appealing properties

- Víšek J.Á. (2011): Consistency of the least weighted squares under heteroscedasticity. Kybernetika **47** (2), 179 – 206.
- Kalina J., Tichavský J. (2020): On robust estimation of error variance in (highly) robust regression. Measurement Science Review **20** (1), 6 – 14.

- LWS-A: linear weight function

$$\psi(t) = 1 - t, \quad t \in [0,1]$$

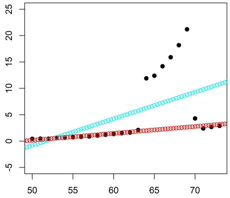
- LWS-B: logistic weight function

$$\psi(t) = \frac{1 + \exp\{-s/2\}}{1 + \exp\{s(t - \frac{1}{2})\}}, \quad t \in [0,1], \quad s > 0 \text{ (fixed)}$$

- LWS-C: trimmed linear weights for a fixed $\tau \in [1/2, 1)$

$$\psi(t) = \left(1 - \frac{t}{\tau}\right) \cdot \mathbb{1}[t < \tau], \quad t \in [0,1]$$

where $\mathbb{1}[.]$ denotes an indicator function.



- Least trimmed squares (**LTS**) estimator

$$\mathbf{b}^{LTS} = (b_0^{LTS}, b_1^{LTS}, \ldots, b_p^{LTS})^T = \arg\min_{b \in \mathbb{R}^{p+1}} \sum_{i=1}^{h} u_{(i)}^2(b)$$

  - High robustness, low efficiency

- Least trimmed absolute values (**LTA**) estimator

$$\arg\min_{b \in \mathbb{R}^p} \sum_{i=1}^{h} |u(b)|_{(i)}, \tag{1}$$

where

$$|u(b)|_{(1)} \leq |u(b)|_{(2)} \leq \cdots \leq |u(b)|_{(n)}, \tag{2}$$

  - Trimmed version of the regression median (L_1 estimator).

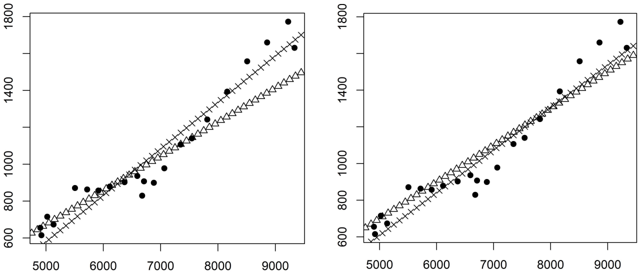
- Least weighted absolute value (**LWA**) estimator

$$\arg\min_{b \in \mathbb{R}^p} \sum_{i=1}^{n} w_i |u(b)|_{(i)}. \tag{3}$$

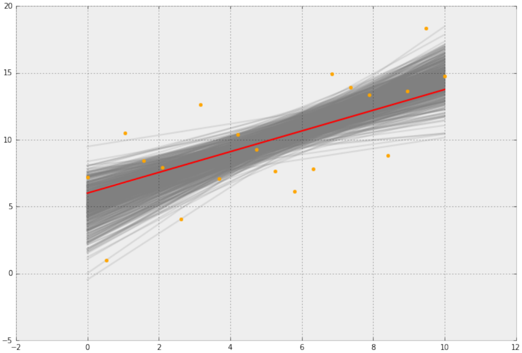
  - Implicitly weighted regression median

- Dataset of U.S. investments with $n = 22$ yearly values (in 10^9 USD)
- X: Gross domestic product
- Y: Gross private domestic investments

Results of four robust regression estimators in the investment dataset.
Left: results of the LTS fit (triangles) and LWS fit (crosses).
Right: results of the LTA fit (triangles) and LWA fit (crosses).



Principles of bootstrap



Nonparametric bootstrap for var b^{LWS}

- Data rows $(X_{i1}, \dots, X_{ip}, Y_i)$, $i = 1, \dots, n$
- $S > 0$
- Compute the least weighted squares estimator $\hat{\beta}_{LWS}$ of β in the model $Y \sim X$
- FOR $s = 1$ to S
  - Generate n new bootstrap data rows

    $$({}_{(s)}X^*_{j1}, \dots, {}_{(s)}X^*_{jp}, {}_{(s)}Y^*_j), \quad j = 1, \dots, n,$$

    by sampling with replacement from data rows $(X_{i1}, \dots, X_{ip}, Y_i)$, $i = 1, \dots, n$
  - Consider a linear regression model in the form

    $${}_{(s)}Y^*_j = {}_{(s)}\gamma_0 + {}_{(s)}\gamma_{1(s)}X^*_{j1} + \cdots + {}_{(s)}\gamma_{p(s)}X^*_{jp} + {}_{(s)}v_j, \quad j = 1, \dots, n \quad (4)$$

  - Estimate ${}_{(s)}\gamma = ({}_{(s)}\gamma_0, {}_{(s)}\gamma_1, \dots, {}_{(s)}\gamma_p)^T$ in (1) by the LWS
  - Store the estimate from the previous step as ${}_{(s)}\hat{\gamma}_{LWS}$
- Compute the empirical covariance matrix from values ${}_{(s)}\hat{\gamma}_{LWS}$, $r = 1, \dots, R$

The classical and robust estimates of the intercept and slope are accompanied by nonparametric bootstrap estimates of standard deviances (s_0 and s_1) and covariances (s_{01}). MSE denotes the mean square error evaluated within a leave-one-out cross validation.

Estimator	Intercept	Slope	s_0	s_1	s_{01}	MSE
Least squares	−582	0.239	108.9	0.016	−1.67	10 948
LTS	−375	0.207	742.0	0.106	−5.74	16 489
LWS	−601	0.242	207.2	0.031	−2.40	12 033
LTA	−312	0.204	721.6	0.112	−5.58	16 207
LWA	−551	0.232	224.8	0.030	−2.49	12 251

Values of five different loss functions computed for five estimators over the investment dataset. This reveals the tightness of the algorithms for computing the individual robust regression estimators.

	Loss function				
Estimator	$\sum_{i=1}^n u_i^2$	$\sum_{i=1}^h u_{(i)}^2$	$\sum_{i=1}^n w_i u_{(i)}^2$	$\sum_{i=1}^h \lvert u\rvert_{(i)}$	$\sum_{i=1}^n w_i \lvert u\rvert_{(i)}$
LS	198 796	80 834	4225	995	51.7
LTS	245 484	61 298	4019	835	45.4
LWS	223 132	63 661	3914	844	45.0
LTA	247 037	62 597	4004	791	46.2
LWA	220 925	64 076	3985	826	41.3

Conclusions

- LTS popular
- LWS more promising but little known
- LTA with a small number of applications
- Novel proposal of LWA
  - Reliable algorithm
  - More flexible than LTA
  - Performance similar to LWS
- Future research