Fast and Robust Bootstrap[1]

Manas Mishra* Shubha Sankar Banerjee* Rachita Mondal*

*Department of Mathematics and Statistics
Indian Institute of Technology, Kanpur

April 16, 2022

[1] Main Reference: Salibian-Barrera and Zamar (2002)

Contents

1. Introduction
2. Formulation of the Problem
3. Description of Methodology
4. Asymptotic Properties of Fast Bootstrap
5. Robustness Properties of Fast Bootstrap
6. Simulation Study
7. Data Analysis
8. References

Motivation

- Numerical stability
- Computational cost

Formulation of the Problem

- **Data:** $(y_i, \mathrm{z}_i')', (y_2, \mathrm{z}_2')', \ldots (y_n, \mathrm{z}_n')'$
- $(y_1, \mathrm{z}_1')' \overset{iid}{\sim} H$, let $x_i = (1, z_i')' \in \mathbb{R}^p$, and the model is,

$$y_i = x_i'\beta_0 + \sigma_0\varepsilon_i, \;\; i = 1, 2, ..., n \tag{1}$$

- **Ideal Situation:** y_i and z_i are independently distributed for all i.
  $y_i \sim F_0, z_i \sim G_0, (y_i, z_i')' \sim H_0$.
- $H \in \mathscr{H}_\varepsilon = \{H = (1-\varepsilon)H_0 + \varepsilon H^*\}$,where H^* is an arbitrary and unspecified function and $0 \leq \varepsilon < 1/2$.

MM-Estimation

- We consider MM-estimation which is based on two loss functions ρ_0 and ρ_1, say.
- If $\hat{\beta}_n$ is the MM-estimate of β, then it satisfies the following equations,

$$\frac{1}{n}\sum_{i=1}^{n}\rho_1^{'}\left(\frac{y_i - x_i'\hat{\beta}_n}{\hat{\sigma}_n}\right)x_i = 0 \tag{2}$$

- $\hat{\sigma}_n$ is scale S-estimate which minimizes the following equation,

$$\frac{1}{n}\sum_{i=1}^{n}\rho_0\left(\frac{y_i - x_i'\beta}{\hat{\sigma}_n(\beta)}\right) = b \tag{3}$$

- $\tilde{\beta}_n$ is the associated S-regression estimate.

Fast Bootstrap

- $\hat{\beta}_n$ can be represented as a solution of fixed point equations:

$$\hat{\beta}_n = f_n(\hat{\beta}_n)$$

- Here f_n depends on the observed data $\{(y_i, x_i), i = 1, \ldots, n\}$ and for given data f_n is given by,

$$f_n(\hat{\beta}_n) = \left[\sum_{i=1}^{n} w_i(\hat{\beta}_n) x_i x_i' \right]^{-1} \sum_{i=1}^{n} w_i(\hat{\beta}_n) x_i y_i \qquad (4)$$

$w_i(\hat{\beta}_n) = \frac{\rho_1'(r_i/\hat{\sigma}_n)}{r_i}$, where $r_i = y_i - \hat{\beta}_n' x_i$.

- Given a bootstrap sample $\{(y_i^*, x_i^*), i = 1, \ldots, n\}$ the recalculated estimates $\hat{\beta}_n^b$ solves, $\hat{\beta}_n^b = f_n^*(\hat{\beta}_n^b)$.

Fast Bootstrap

- f_n^* has the same form of f_n, except it is based on the bootstrap samples and the corresponding weights are, given by,

$$w_i^*(\hat{\beta}_n^b) = \frac{\rho_1'(r_i^b/\hat{\sigma}_n)}{r_i}, \quad \text{where } r_i^b = y_i^* - \hat{\beta}_n'^b x_i^*.$$

- Instead of computing $\hat{\beta}_n^b$ we consider, $\hat{\beta}_n^* = f_n^*(\hat{\beta}_n)$, i.e. in f_n^* we use the weights as,

$$w_i^*(\hat{\beta}_n) = \frac{\rho_1'(r_i^*/\hat{\sigma}_n)}{r_i}, \quad \text{where } r_i^* = y_i^* - \hat{\beta}_n' x_i^*.$$

- It can be shown that $\hat{\sigma}_n$ has a weighted average representation and so it is possible to define $\hat{\sigma}_n^*$ for the bootstrap samples similarly.
- $\hat{\beta}_n^*$ may not reflect true variability of $\hat{\beta}_n$, on applying correction factor our final estimate is:

$$\hat{\beta}_n^{R*} - \hat{\beta}_n = M_n(\hat{\beta}_n^* - \hat{\beta}_n) + d_n(\hat{\sigma}_n^* - \hat{\sigma}_n)$$

Asymptotic Properties of Fast Bootstrap

- Now that we have put the forward the methodology, we focus on the asymptotic properties of Fast bootstrap estimates.
- The next theorem will show that the asymptotic distribution of fast bootstrap is the same as that of MM-regression estimator.
- We proceed with stating a few regularity conditions on the form of ρ_0 and ρ_1 defined earlier.

Some conditions

MM-estimates are based on two loss function $\rho_0 : \mathbb{R} \to \mathbb{R}_+$ and $\rho_1 : \mathbb{R} \to \mathbb{R}_+$ (defined earlier) which determine the breakdown point and the efficiency of the estimate. They satisfy the following conditions:

C1 $\forall u \in \mathbb{R}$, $\rho_0(-u) = \rho_0(u)$ and $\rho_1(u) = \rho_1(-u)$;

C2 $\rho_0(0) = 0 = \rho_1(0)$;

C3 ρ_0 and ρ_1 are continuously differentiable functions;

C4 $\sup_x \rho_0(x) = \sup_x \rho_1(x) = 1$;

C5 If $\rho_0(u) < 1$ and $0 \leq v < u$, then $\rho_0(v) < \rho_0(u)$. Same condition holds for ρ_1.

Some established results

Salibian-Barrera and Zamar (2002) proved that that $\hat{\beta}_n$ (MM-regression estimator), $\hat{\sigma}_n$ (S-scale estimator) and $\tilde{\beta}_n$ (S-regression estimator) are consistent (weakly) for true values $\beta,\ \sigma,\ \&\ \tilde{\beta}$ where,

$$\mathbb{E}[\rho_1'((Y - X'\beta)/\sigma)] = 0$$
$$\mathbb{E}[\rho_0((Y - X'\tilde{\beta})/\sigma)] = b$$
$$\mathbb{E}[\rho_0'((Y - X'\tilde{\beta})/\sigma)] = 0$$

This result is essential in stating the first main theorem of this topic.

Convergence of Fast Robust Distribution

Theorem

If ρ_0 and ρ_1 satisfies the conditions (C1-C5) and have continuous third order derivatives, then given the consistency of $\hat{\beta}_n$, $\hat{\sigma}_n$ and $\tilde{\beta}$, and under a few regularity conditions, almost all sample sequences $\sqrt{n}(\hat{\beta}_n^{R*} - \hat{\beta}_n)$ converges weakly, as n goes to infinity, to the same limit distribution as $\sqrt{n}(\hat{\beta}_n - \beta)$.

Robustness of Fast Bootstrap

- We now focus on the robustness properties of our fast bootstrap.
- Let q_t be the t^{th} upper quantile of a statistics $\hat{\theta}_n$ i.e. q_t satisfies

$$P[\hat{\theta}_n > q_t] = t$$

- Singh (1998) defines upper breakdown point of a quantile estimate $\hat{q}_t$ as the minimum proportion of asymmetric contamination that can drive it over any finite bound.
- An estimator based on bootstrap sample can potentially break down if the expected proportion of bootstrap samples that contain more outliers than the breakdown point of the estimate (say τ^*) to be more than t.

Breakdown point of the fast bootstrap quantiles

Theorem

Let $(y_1, x_1')', \ldots, (y_n, x_n')' \in \mathbb{R}^{p+1}$ be the random sample following linear model. Assume that the explanatory variables $x_1, \ldots, x_n$ in $\mathbb{R}^p$ are in general position. Let $\hat{\beta}_n$ be an MM-regression estimate and let ε^* be its breakdown point. Then the breakdown point of the t^{th} fast bootstrap quantile estimate of the regression parameters β_j, $j = 1, \ldots, p$ is given by $\min(\varepsilon^*, \varepsilon_R)$, where ε_R satisfies

$$\varepsilon_R = \inf\{\delta \in [0,1] : P[\texttt{Binomial}(n,\delta) > n - p] \geq t\}$$

Singh (1998) obtained the upper breakdown point of bootstrap estimate $\hat{q}_t$ of q_t:

$$\varepsilon_C = \inf\{\delta \in [0,1] : P[\texttt{Binomial}(n,\delta) \geq [\varepsilon^* n]] \geq t\}$$

If $n > 2p$, then $[\varepsilon^* n] \leq [n/2] < n - p$. Thus we can clearly see that $\varepsilon_C < \varepsilon_R$.

Simulation Study

Data Description

- Generated the data $y_i = \beta_0 + \beta_1 x_i + \varepsilon,\ i = 1, \ldots, n$ for $n = 30$ and 100.

- $x_i \sim Normal(0,1),\ \beta_0 = 5$ and $\beta_1 = 5$.

- The errors are generated from F_ε with,

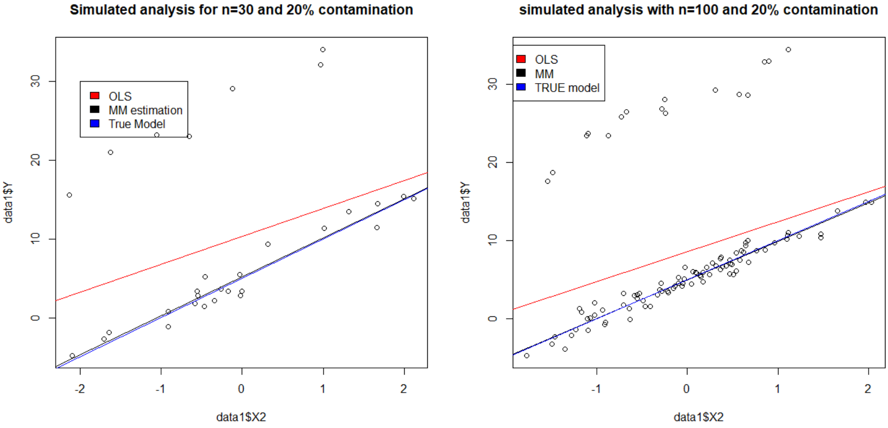
$$F_\varepsilon(x) = (1-\varepsilon)\Phi(x) + \varepsilon F_u(x)$$

  Φ is the CDF of $Normal(0,1)$andF_u is the CDF of $Uniform(20,25)$

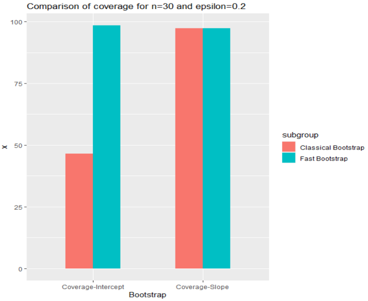
- Considered $\varepsilon = 0.0, 0.20$, i.e. considered 0% and 20% contamination in the error distribution.

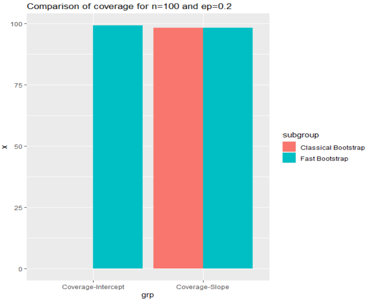
- Generated 1000 datasets from the above distribution and built 99% confidence intervals for the parameters

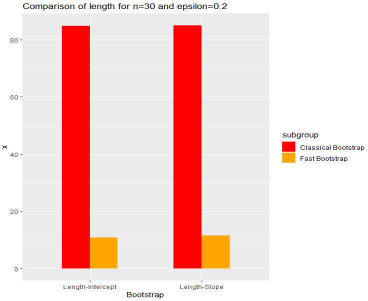
Robustness regression fits

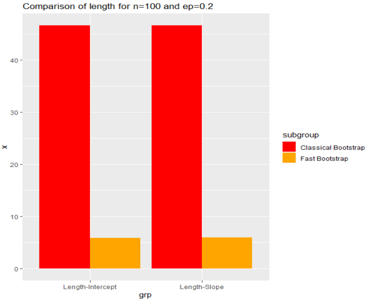


Numerical stability results

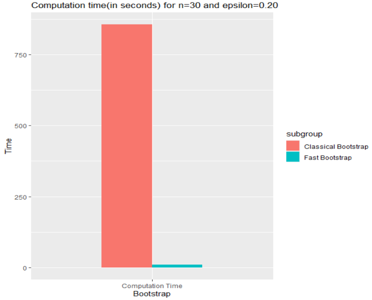


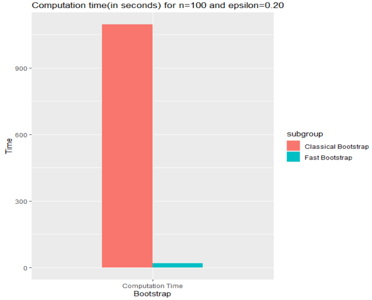


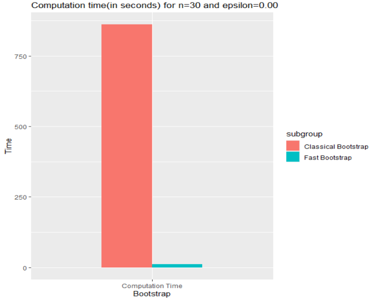


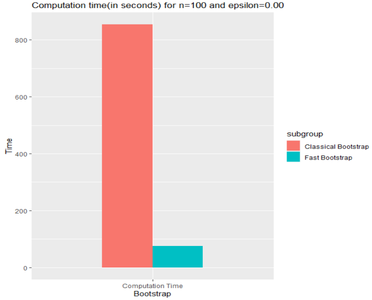


Computational cost results

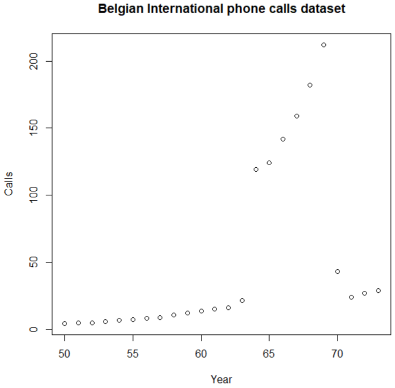






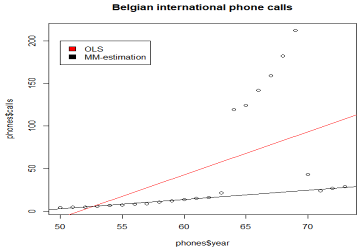


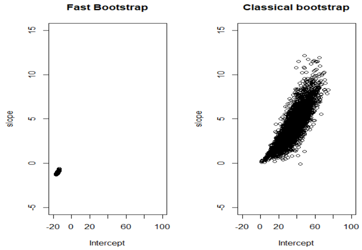
1)Belgian International Phone Calls[2]
Using 10000 fast bootstrap calculations we estimate the distribution of robust regression estimates and compare results with classical bootstrap method.

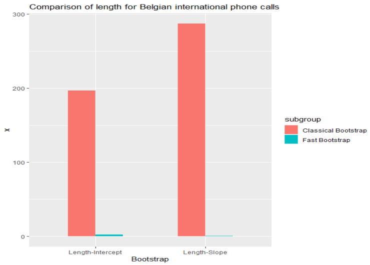


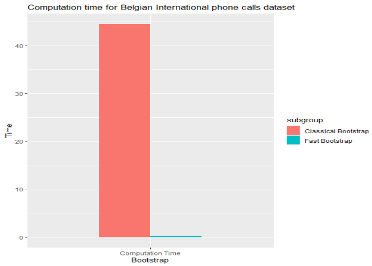
[2] MASS Package in R

Data Analysis-II



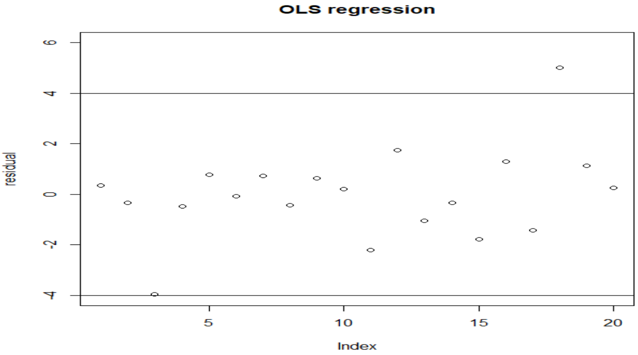






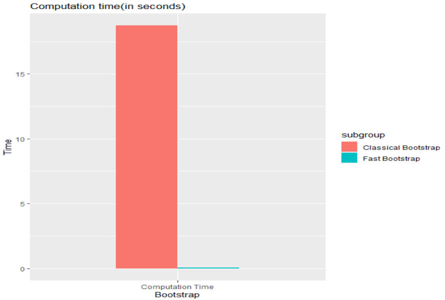
2)Verbal test score data[3]

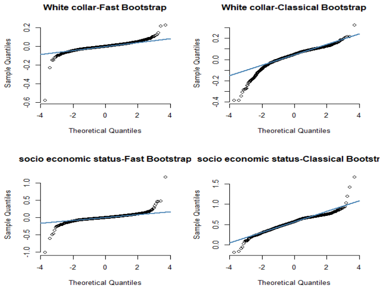
The data consist of verbal mean test scores from 20 schools.There are 5 explanatory variables.The plot of residuals below confirms presence of outliers.

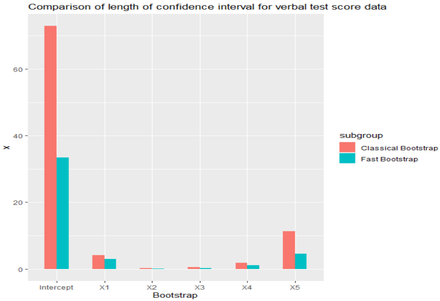


[3]Coleman et. al (1966)

Results on Verbal test score data







Reference

Salibian-Barrera, M. and Zamar, R. H. (2002). Bootstrapping robust estimates of regression. *Annals of Statistics*, pages 556–582.

Singh, K. (1998). Breakdown theory for bootstrap quantiles. *The Annals of Statistics*, 26(5):1719 – 1732.