Estimating the Distribution of Linear Regression Estimates using Fast and Robust Bootstrap

Submitted by:

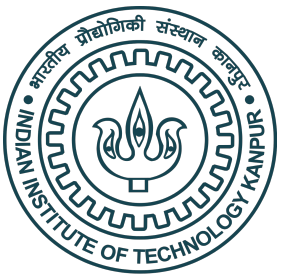
Manas Mishra [*†]

Rachita Mondal [‡†]

Shubha Sankar Banerjee [§†]

Supervised by:

Dr. Dootika Vats [†]



April 16, 2022

[*]Department of Mathematics & Statistics, Indian Institute of Kanpur, India.
[†]201340, M.Sc. Statistics (Final year).
[‡]201374, M.Sc. Statistics (Final year).
[§]201416, M.Sc. Statistics (Final year).

Abstract

In this report we discuss the method of Fast Bootstrap to obtain an estimate of the distribution of robust regression estimates. The weighted average representation of MM-estimates has been very crucial to the formulation of our problem. This method is computationally less costly as for each bootstrap sample we do not run non-convex optimization algorithm. Rather, we only solve a system of linear equations. Robustness is achieved by using weights as a decreasing function of absolute value of the residuals. The breakdown point of the quantile estimates from this method is higher than classical bootstrap estimates. We illustrate the method using a simulation study and also by performing data analysis in two different data sets.

Contents

1 Introduction

Bootstrap (Efron (1979)) is a popular approach to estimate the sampling distribution and the standard error of the robust estimates. The standard error can also be estimated using their asymptotic variances. However, only central normal models are considered while studying the asymptotic behavior of these estimates. But Normality need not hold true specially in those scenarios when robust estimation is recommended. When the distribution of errors is symmetric, the estimates of the regression coefficients and those of the scale of the errors are asymptotically independently distributed. Since the outliers need not be balanced in both sides of the regression line, many data set with outliers fails to satisfy the symmetry condition of errors. On the other hand, relaxing this condition may lead to difficult asymptotic calculation. We will focus on MM-estimators (Yohai (1987)) which is calculated with an initial S-estimate (Rousseeuw and Yohai (1984)). Using this method it is possible to have robust estimates.
Usual bootstrap may lead to the following problems,

- **Numerical Instability:** When there is outliers in the data set, bootstrap samples can have higher proportion of outliers than that in the original sample. Then bootstrap distribution will be a poor estimator of the distribution of the regression parameter estimates.
- **Computational Cost**: For high-dimensional problem, running a non-convex optimization algorithm for each bootstrap sample can be computationally expensive.
- **Recalculating residual scale estimates:** A large amount of computational cost is involved in recalculating robust scale estimates for each bootstrap sample. But if we do not calculate the scale estimates, the resulting distribution may not converge to a desired asymptotic distribution.

Hence, our basic idea is to use "Fast Bootstrap" which uses a reweighted representation of the estimates. this method is computationally simple and it is able to provide robust estimates. The rest of the paper is organized as foloows: In Section (2) we discuss the structure of the model and also briefly discuss the method of MM-estimation. In Section (3) the Fast Bootstrap methodology has been described in details. Section (4) gives us the asymptotic result for Fast

Bootstrap and Section (5) gives an idea of the robustness of Fast Bootstrap. Later in Section (6) we perform simulation to validate the method, also we perform real data analysis in Section (7). Finally, we conclude in Section (8).

2 Definition and Notation

Let us consider the regression setup with random explanatory variables.

Suppose that $(y_1, \mathbf{z}_1')', (y_2, \mathbf{z}_2')', \ldots (y_n, \mathbf{z}_n')'$ are n random vectors which are independent to each other, also they follow the common distribution function H. To include the intercept term we consider, $x_i = (1, z_i')' \in \mathbb{R}^p$.. Linear regression model is given by,

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

Assume that, y_i and z_i are independently distributed for all i. Here $y_i \sim F_0, z_i \sim G_0, (y_i, z_i')' \sim H_0$. Also, suppose that the distribution function F_0 is specified and it is symmetric.

For considering the occurrence of outliers and other deviations from the classical model, we suppose that the real distribution of the data is H, which belongs to the contamination neighborhood,

$$\mathscr{H}_\varepsilon = \{H = (1-\varepsilon)H_0 + \varepsilon H^*\} \tag{2}$$

,where H^* is an arbitrary and unspecified function and $0 \leq \varepsilon < 1/2$.

For estimating the model parameter we consider MM-estimation. This method is based on two loss functions ρ_0 and ρ_1, say. Here ρ_0 determines the breakdown point whereas, ρ_1 determines the efficiency of the estimates. If $\hat{\beta}_n$ is the MM-estimate of β, then it satifies the following equations,

$$\frac{1}{n}\sum_{i=1}^{n} \rho_1\left(\frac{y_i - x_i'\hat{\beta}_n}{\hat{\sigma}_n}\right) x_i = 0 \tag{3}$$

, where $\hat{\sigma}_n$ is scale S-estimate (Rousseeuw and Yohai (1984)). $\hat{\sigma}_n$ minimizes the following equation,

$$\frac{1}{n}\sum_{i=1}^{n}\rho_0\left(\frac{y_i - x_i'\hat{\beta}_n}{\hat{\sigma}_n(\beta)}\right) = b \tag{4}$$

Yohai (1987) has studied the asymptotic properties of the MM-estimate under $H = H_0$. But this assumption of central parametric model does not hold while using highly robust MM-estimates. In the next section we will introduce Fast Bootstrap.

3 Fast Bootstrap

By solving equation (3) we obtain the MM-regression estimate of β_0 as $\hat{\beta}_n$. Now, our goal is to estimate the sampling distribution of $\hat{\beta}_n$. For this purpose a computer-intensive method will be used to generate a large number of recalculated $\hat{\beta}_n^*$ based on the plug-in approach used in bootstrap as introduced by Efron (1979). The Empirical Distribution function of the recalculated statistics will be used to estimate the sampling distribution of $\hat{\beta}_n$.

Let us suppose that $\tilde{\beta}_n$ is the S-regression estimate of β_0. Let us define the residuals for each par $(y_i, x_i')'$ corresponding to $\hat{\beta}_n$ and $\tilde{\beta}_n$.

$$r_i = y_i - \hat{\beta}_n' x_i$$
$$\tilde{r}_i = y_i - \tilde{\beta}_n' x_i \quad i = 1, \ldots, n$$

Let us now define the following weights,

$$w_i = \frac{\rho_1'(r_i/\hat{\sigma}_n)}{r_i}$$
$$v_i = \frac{\hat{\sigma}_n}{nb}\frac{\rho(\tilde{r}_i/\hat{\sigma}_n)}{\tilde{r}_i}, \quad i = 1, \ldots n \tag{5}$$

Based on the weights in (5) $\hat{\beta}_n$ and $\hat{sigma}_n$ can be represented as weighted average form as follows,

$$\begin{aligned}\hat{\beta}_n &= \left[\sum_{i=1}^{n} w_i x_i x_i'\right]^{-1} \sum_{i=1}^{n} w_i x_i y_i \\ \hat{\sigma}_n &= \sum_{i=1}^{n} v_i (y_i - \tilde{\beta}_n' x_i)\end{aligned} \tag{6}$$

Suppose that, $\{(y_i^*, x_i^{*'})', i = 1, \ldots, n\}$ be a bootstrap sample from the observation. Based on the bootstrap samples we define the following forms of residuals,

$$\begin{aligned} r_i^* &= y_i^* - \hat{\beta}_n' x_i^* \\ \tilde{r}_i^* &= y_i^* - \tilde{\beta}_n' x_i^* \quad , i = 1, \ldots, n \end{aligned}$$

Using the above residuals new weights can be defined as,

$$\begin{aligned} w_i^* &= \frac{\rho_1'(r_i^*/\hat{\sigma}_n)}{r_i^*} \\ v_i^* &= \frac{\hat{\sigma}_n}{nb} \frac{\rho(\tilde{r}_i^*/\hat{\sigma}_n)}{\tilde{r}_i^*}, \quad i = 1, \ldots n \end{aligned} \tag{7}$$

Using the weights as defined in (7) the recalculated parameter estimates are given by,

$$\begin{aligned}\hat{\beta}_n^* &= \left[\sum_{i=1}^{n} w_i^* x_i^* x_i^{*'}\right]^{-1} \sum_{i=1}^{n} w_i^* x_i^* y_i^* \\ \hat{\sigma}_n^* &= \sum_{i=1}^{n} v_i^* (y_i^* - \tilde{\beta}_n' x_i^*)\end{aligned} \tag{8}$$

Note that, the estimates $\hat{\beta}_n$ and $\hat{\sigma}_n$ used in the calculating weights w_i^* and v_i^* are kept fixed. So $\hat{\beta}_n^*$ amd σ_n^* may not reflect the actual variability of the random vector $(\hat{\beta}_n', \hat{\sigma}_n)'$. For this reason

a linear correction has been imposed on the recalculated $\hat{\beta}_n^*$ and $\hat{sigma}_n^*$. Let,

$$M_n = \hat{\sigma}_n \left[\sum_{i=1}^{n} \rho_1''(r_i/\hat{\sigma}_n, x_i) x_i x_i' \right]^{-1} \sum_{i=1}^{n} w_i x_i x_i'$$
$$d_n = a_n^{-1} \left[\sum_{i=1}^{n} \rho_1''(r_i/\hat{\sigma}_n, x_i) x_i x_i' \right]^{-1} \sum_{i=1}^{n} \rho_1''(r_i/\hat{\sigma}_n, x_i) r_i x_i$$
$$a_n = \frac{1}{nb} \sum_{i=1}^{n} \frac{\tilde{r}_i}{\hat{\sigma}_n} \rho_0'(\tilde{r}_i/\hat{\sigma}_n)$$

The the recalculated $\hat{\beta}_n - \beta$ for Fast Bootstrap is given by,

$$\hat{\beta}_n^{R*} - \hat{\beta}_n = M_n(\hat{\beta}_n^* - \hat{\beta}_n) + d_n(\hat{\sigma}_n^* - \hat{\sigma}_n) \tag{9}$$

Let us now discuss, the reason behind taking the above form of $\hat{\beta}_n^{R*} - \hat{\beta}_n$ as our final estimate. note that, $\hat{\beta}_n, \hat{\sigma}_n, \tilde{\beta}_n$ satisfiy the following equations,

$$\frac{1}{n} \sum_{i=1}^{n} \rho_1' \left(\frac{r_i(\hat{\beta}_n)}{\hat{\sigma}_n} \right) x_i = 0$$
$$\frac{1}{n} \sum_{i=1}^{n} \rho_0 \left(\frac{r_i(\tilde{\beta}_n)}{\hat{\sigma}_n} \right) = b$$
$$\frac{1}{n} \sum_{i=1}^{n} \rho_0' \left(\frac{r_i(\tilde{\beta}_n)}{\hat{\sigma}_n} \right) x_i = 0$$

It can be easily shown that the estimates can be expressed as follows:

$$\hat{\beta}_n = A_n(\hat{\beta}_n, \hat{\sigma}_n)^{-1} \, v_n(\hat{\beta}_n, \hat{\sigma}_n)$$
$$\hat{\sigma}_n = \hat{\sigma}_n u_n(\tilde{\beta}_n, \tilde{\sigma}_n) \tag{10}$$
$$\tilde{\beta}_n = B_n(\tilde{\beta}_n, \hat{\sigma}_n)^{-1} \, w_n(\tilde{\beta}_n, \hat{\sigma}_n)$$

where,

$$A_n(\beta_1,\sigma) = \frac{1}{n}\sum_{i=1}^{n}\frac{\rho_1'(r_i/\sigma)}{r_i}x_i x_i'$$
$$v_n(\beta_1,\sigma) = \frac{1}{n}\sum_{i=1}^{n}\frac{\rho_1'(r_i/\sigma)}{r_i}y_i x_i$$
$$u_n(\beta_2,\sigma) = \sum_{i=1}^{n}\frac{\rho_0(\tilde{r}_i/\sigma)}{nb\tilde{r}_i}\tilde{r}_i$$
$$B_n(\beta_2,\sigma) = \frac{1}{n}\sum_{i=1}^{n}\frac{\rho_0'(\tilde{r}_i/\sigma)}{\tilde{r}_i}x_i x_i'$$
$$w_n(\beta_2,\sigma) = \frac{1}{n}\sum_{i=1}^{n}\frac{\rho_0'(\tilde{r}_i/\sigma)}{\tilde{r}_i}y_i x_i$$

The set of equations (10) can be written as a fixed point of a suitably chosen function. Let, $f : \mathbb{R}^{2p+1} \to \mathbb{R}^{2p+1}$ be defined for $\beta_1 \in \mathbb{R}^p$, $\beta \in \mathbb{R}$, $\sigma \in \mathbb{R}$. Then we can write,

$$f(\beta_1,\sigma,\beta_2) = \begin{pmatrix} A_n(\beta_1,\sigma)^{-1}v_n(\beta_1,\sigma) \\ \\ \sigma u_n(\beta_2,\sigma) \\ \\ B_n(\beta_2,\sigma)^{-1}w_n(\beta_2,\sigma) \end{pmatrix}$$

Although f is dependent on n, for sake of simplicity we write,

$$f(\hat{\beta}_n,\hat{\sigma}_n,\tilde{\beta}_n) = (\hat{\beta}_n,\hat{\sigma}_n,\tilde{\beta}_n)'$$

Since ρ_1, ρ_0 are differentiable, using Taylor Series expansion of f about limiting values of the estimates i.e. about $(\beta,\sigma,\tilde{\beta})$ we get,

$$\begin{pmatrix} \hat{\beta}_n \\ \hat{\sigma}_n \\ \tilde{\beta}_n \end{pmatrix} = f(\beta,\sigma,\tilde{\beta}) + \nabla f(\beta,\sigma,\tilde{\beta}) \begin{pmatrix} \hat{\beta}_n - \beta \\ \hat{\sigma}_n - \sigma \\ \tilde{\beta}_n - \tilde{\beta} \end{pmatrix} + R_n \tag{11}$$

Here R_n is the remainder term and $\nabla f(\beta, \sigma, \tilde{\beta})$ is the matrix of first order partial derivative. Since it is possible to show that $||R_n|| = o_p(1)$ equation (11) can be written as,

$$\begin{pmatrix} \hat{\beta}_n \\ \hat{\sigma}_n \\ \tilde{\beta}_n \end{pmatrix} = f(\beta, \sigma, \tilde{\beta}) + \nabla f(\beta, \sigma, \tilde{\beta}) \begin{pmatrix} \hat{\beta}_n - \beta \\ \hat{\sigma}_n - \sigma \\ \tilde{\beta}_n - \tilde{\beta} \end{pmatrix} + o_p(1)$$

If we define $\theta_n = (\hat{\beta}_n, \hat{\sigma}_n, \tilde{\beta}_n)'$ and $\theta = (\hat{\beta}, \hat{\sigma}, \tilde{\beta})'$, equation (11) becomes,

$$\sqrt{n}(\theta_n - \theta) = (\mathtt{I} - \nabla f(\theta))^{-1} \sqrt{n}(f(\theta) - \theta) + o_p(1) \tag{12}$$

Under certain conditions it can be shown that, $\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n) \sim \sqrt{n}(\hat{\theta}_n - \theta)$ and $\sqrt{n}(f(hat\theta_n) - \hat{\theta}_n) \sim \sqrt{n}(f^*(\theta) - \theta)$. Again, if we approximate $(\mathtt{I} - \nabla f(\theta))^{-1}$ by $(\mathtt{I} - \nabla f(\hat{\theta}_n))^{-1}$, we get,

$$\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n) \sim (\mathtt{I} - \nabla f(\hat{\theta}_n))^{-1} \sqrt{n}(f^*(\hat{\theta}_n) - \hat{\theta}_n)$$

Using the above form we define the corrected $\hat{\theta}_n^*$ as,

$$\hat{\theta}_n^{R*} = \hat{\theta}_n + \sim (\mathtt{I} - \nabla f(\hat{\theta}_n))^{-1} (f^*(\hat{\theta}_n) - \hat{\theta}_n)$$

From here the expression (9) arises.

According to Theorem 4.3 of Salibian-Barrera (2000), the asymptotic behavior of $\hat{\beta}_n$ depends on $\hat{\sigma}_n$. Hence it is important to consider scale estimate $\hat{\sigma}_n$ to estimate the distribution of $\hat{\beta}_n$. Now consider the following remarks which will give an insight regarding the fastness and robustness of the method.

Remark 1. *For each bootstrap sample we do not solve equations (3) and (4) every time, rather we recalculate the quantities mentioned in (8). Also, the correction factors M_n, d_n, a_n which arise from the linear systems and the weighted averages are calculated only once. Hence, it reduces the load of computation.*

Remark 2. *ρ_1' is a redescending score function, i.e. $\rho_1'(r) = 0$, for $\mid r \mid \geq c > 0$. Since, $\hat{\beta}_n$ is estimated using ρ_1', smaller weights are applied to the outlying observations. Hence, the*

method gives stable estimates even in the presence of outliers. The extreme outliers (with corresponding residuals $\mid r_i \mid > c\hat{\sigma}_n$*) receives zero weight and so they do not affect the recalculated coefficients.The recalculated* $\hat{\sigma}_n^*$ *are not affected by the outliers since the weights* v_i *are decreasing in absolute values of the residuals.*

4 Asymptotic Properties of Fast Bootstrap

For the above proposed fast bootstrap estimator, our focus now is to derive its asymptotic distribution and to show that it is the same as that of MM-regression estimator.

We assume the following regularity conditions on ρ_0 and ρ_1".

R 1. *Both* ρ_0 *and* ρ_1 *are even functions i.e.* $\forall u \in \mathbb{R}$, $\rho_0(-u) = \rho_0(u)$ *and* $\rho_1(u) = \rho_1(-u)$*;*

R 2. $\rho_0(0) = 0 = \rho_1(0)$*;*

R 3. ρ_0 *and* ρ_1 *are continuously differentiable functions;*

R 4. $\sup_x \rho_0(x) = \sup_x \rho_1(x) = 1$*;*

R 5. *If* $\rho_0(u) < 1$ *and* $0 \leq v < u$*, then* $\rho_0(v) < \rho_0(u)$*. Same condition holds for* ρ_1*.*

Beaton and Tukey (1974) proposed a family of functions which satisfy R (1)- (R 5), denoted by ρ, where,

$$\rho(u) = \begin{cases} 3\left(\frac{u}{d}\right)^2 - 3\left(\frac{u}{d}\right)^4 + \left(\frac{u}{d}\right) & \text{if } |u| \leq d \\ 1 & \text{if } |u| > d \end{cases} \quad \text{where } d > 0 \text{ is a fixed constant} \tag{13}$$

We now state the main theorem o the convergence of Fast bootstrap distribution.

Theorem 1. *Let* ρ_1 *and* ρ_2 *be real functions satisfying (R1)-(R5). We further assume that the have continuous third derivatives. Let* $\hat{\beta}_n$ *be the MM-regression estimator,* $\hat{\sigma}_n$ *the S-scale and* $\tilde{\beta}_n$ *the associated S-regression estimator. We assume that they are consistent, i.e.,* $\hat{\beta}_n \xrightarrow{p} \beta$*,*

$\hat{\sigma}_n \xrightarrow{p} \sigma$ and $\tilde{\beta}_n \xrightarrow{p} \tilde{\beta}$, where β, σ and $\tilde{\beta}$ are the solutions of the following equations:

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

Now if the following conditions hold,

1. *The following matrices exist and are finite:*

$$\mathbb{E}[\rho_1'(r)/rXX']^{-1}, \quad \mathbb{E}[\rho_0'(r)/rXX'], \quad \mathbb{E}[\rho_1'(r)XX'] \quad \mathbb{E}[\rho_1'(r)rXX']$$
$$\mathbb{E}[\rho_0''(r)XX'] \quad \mathbb{E}[\rho_1''(r)XX'], \quad \mathbb{E}[\rho_0''(r)rX], \quad \mathbb{E}[\rho_1''(r)rX]$$

2. $\mathbb{E}[\rho_0''(r)r] \neq 0$ *and finite*

3. $\rho_1'(u)/u$, $\rho_1'(u)/u$, $(\rho_0'(u) - \rho_0''(u)u)/u^2$ *and* $(\rho_1'(u) - \rho_1''(u)u)/u^2$ *are continuous*

then almost all sample sequences $\sqrt{n}(\hat{\beta}_n^{R*} - \hat{\beta}_n)$ converges weakly, as n goes to ∞, to the same limit distribution as $\sqrt{n}(\hat{\beta}_n - \beta)$.

It is interesting to note that the Tukey's family (13) satisfies Assumption 3 stated above.

We made another assumption on the consistency of $\hat{\sigma}_n$, $\tilde{\beta}_n$ and $\hat{\beta}_n$. Salibian-Barrera (2000) found regularity conditions that suffice to prove the consistency and asymptotic distribution of these estimates for any $F \in \mathscr{H}_\varepsilon$ (2).

5 Robustness of Fast Bootstrap

Now that we have established the consistency of the fast bootstrap estimator, we now focus on their robustness properties.

For $t \in (0,1)$, let q_t be the t^{th} upper quantile of a statistic $\hat{\theta}_n$, then q_t satisfies the following equation:

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

To discuss about the robustness of any estimator, a classical metric to judge said property is using the breakdown point.

Definition 1 (Breakdown Point (Singh (1998))). *The upper breakdown point of a quantile estimate $\hat{q}_t$ is the minimum proportion of asymmetric contamination than can drive it over any finite bound.*

We discuss two scenarios where quantile estimates based on fast bootstrap can breakdown.

- When the proportion of outliers in the original data is larger than the breakdown point of the estimate. The estimates can be unreliable along with any conclusion which could be derived from them.
- Let τ^* be the expected proportion of bootstrap samples that contain more outliers than the breakdown point of the estimate which implies that we expect $\tau^* \times 100\%$ of the recalculated $\hat{\beta}_n^*$'s to be unreliable. The estimate $\hat{q}_t$ can be affected severely if $\tau^* > t$.

As used in classical theory, breakdown point is related to the geometrical characteristics of the data. In similar way, these characteristics affect the breakdown point of the fast bootstrap quantile estimates. We define the concept of General position (Rousseeuw and Leroy (1987)) to progress with our discussion.

Definition 2 (General Position). *We say k points in $\mathbb{R}^p$ are in general position if no subset of size $p+1$ of them determines an affine subspace of dimension p. In other words, for every subset $x_{i_1},\ldots,x_{i_{p+1}}$, $1 \le i_j \le k$, $i_j \neq i_l$ if $j \neq l$, there are no vector $v_0 \in \mathbb{R}^p \backslash \{0\}$ and scalar $\alpha \in \mathbb{R}$ such that,*

$$x_{i_j}' v_0 = \alpha, \text{ for } j = 1,\ldots,p+1$$

If explanatory variables are in general position, then the condition ensures that the estimates we are dealing with are bounded.

We now state the main result of this theorem which establishes the breakdown point of the quantile estimators based on fast bootstrap.

Theorem 2 (Breakdown point of fast bootstrap quantiles for regression model). *Let $(y_1, x_1')', \ldots, (y_n, x_n')' \in \mathbb{R}^{p+1}$ be the random sample following linear model (1). Assume that the explanatory*

variables $x_1, \ldots, x_n$ in $\mathbb{R}^p$ are in general position (See Def. 2). 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 $\beta_j,\ j = 1, \ldots, p$ is given by $\min(\varepsilon^*, \varepsilon_R$, *where ε_R satisfies*

$$\varepsilon_R = \inf\{\delta \in [0,1] : P[Binomial(n, 1-\delta) < p] \geq t\} \tag{14}$$

the above equation is equivalent to:

$$\varepsilon_R = \inf\{\delta \in [0,1] : P[Binomial(n, \delta) \geq n - p] \geq t\} \tag{15}$$

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

$$\varepsilon_C = \inf\{\delta \in [0,1] : P[Binomial(n, \delta) \geq [\varepsilon^* n]] \geq t\} \tag{16}$$

where $[x]$ denotes the smallest integer larger than or equal to x and ε^* is the breakdown point of the estimate being bootstrapped. Since $[\varepsilon^* n] \leq [n/2] < n - 1$ for $n > 3$, we see from the two above equations that $\varepsilon_C < \varepsilon_R$.

An intuitive discussion on the above theorem can be made as follows: It can be shown that given a bootstrap sample of size n, if without loss of generality, the first k observations are "good" and the remaining $n-k$ are arbitrary outliers, we get bounded estimate $\hat{\beta}^*$, the value of which can only be modified by a finite amount (amount being depending on the k first observations, and not the others). Then, considering all possible bootstrap samples which contain at least p points that are not outliers, we find a bound that depends only on the original dataset. To drive the t^{th} fast bootstrap quantile estimate above any bound, we need to have atleast $t\%$ of the bootstrap samples containing less than p "good" points. That proportion, say δ of outliers in the original sample should then satisfy the condition that $P[Binomial(n, 1-\delta) < p] \geq t$.

In other words, ε_R is that proportion of the original sample which needs to be contaminated

by asymmetric values to blow up the fast bootstrap quantile estimate over any finite bound.

6 Simulation Study

After establishing the theoretical properties,in this section we show that the Fast Bootstrap actually works in practical scenario by using it on simulated data.Here we report the coverage and length of the confidence interval of the parameters β in the true model.We will also report the computation time to show the fastness of this method.

We considered sample sizes n=30 and 100 with p=2 explanatory variables.The independent variables included an intercept x_1:1 and $x_2 \sim$ N(0,1).The errors were generated as follows:

Algorithm 1 Generation of errors

START

1. Generate random number r from U(0,1)
2. If $r \leq 1-\varepsilon$ generate e_i from N(0,1) else go to step 3
3. Generate e_i from U(20,25)
4. Repeat step 1,2 and 3 until n errors are generated

STOP

We used $\varepsilon = 0.00$ and 0.20 and here we report the results of all the 4 cases. We used $\beta_0 = 5$ and $\beta_1 = 5$ and then generated $Y = \beta_0 + \beta_1 X_1 + \varepsilon$. We have generated 1000 datasets from the above distribution and built 99% confidence intervals for the parameters.

We used MM-regression estimates obtained with $\psi = \rho_{4.685}$ in Tukey's family. The S-scale was obtained with $\rho_{1.54764}$ also in Tukey's family. This choice yields estimates with simultaneous 50% breakdown point and 95% efficiency when the data are normally distributed.

From Figure 1 we can see that Fast and robust bootstrap gives better coverage and shorter length whenever there is contamination in the data.The reason for this seems to be that the empirical asymptotic variance formula is numerically unstable (especially for contaminated datasets). In case of no contamination it works as well as the OLS regression bootstrap but it still takes less time than classical bootstrap as shown in the Figure 2.

Table 1: Coverage of 99% confidence interval

n,ε		Fast and robust bootstrap	classical bootstrap
n=30 ε=0.00	β_o	0.978	0.978
	β_1	0.974	0.978
n=30 ε=0.20	β_o	0.984	0.464
	β_1	0.974	0.979
n=100 ε=0.00	β_o	0.982	0.992
	β_1	0.987	0.988
n=100 ε=0.20	β_o	0.991	0.00
	β_1	0.985	0.989

Table 2: Length of 99% confidence interval

n,ε		Fast and robust bootstrap	classical bootstrap
n=30 ε=0.00	β_o	0.967	0.928
	β_1	1.012	0.928
n=30 ε=0.20	β_o	1.076	8.490
	β_1	1.146	8.490
n=100 ε=0.00	β_o	0.526	0.511
	β_1	0.536	0.511
n=100 ε=0.20	β_o	0.581	4.667
	β_1	0.590	4.667

Figure 1: Coverage and length of 99% confidence interval obtained by Fast and robust bootstrap

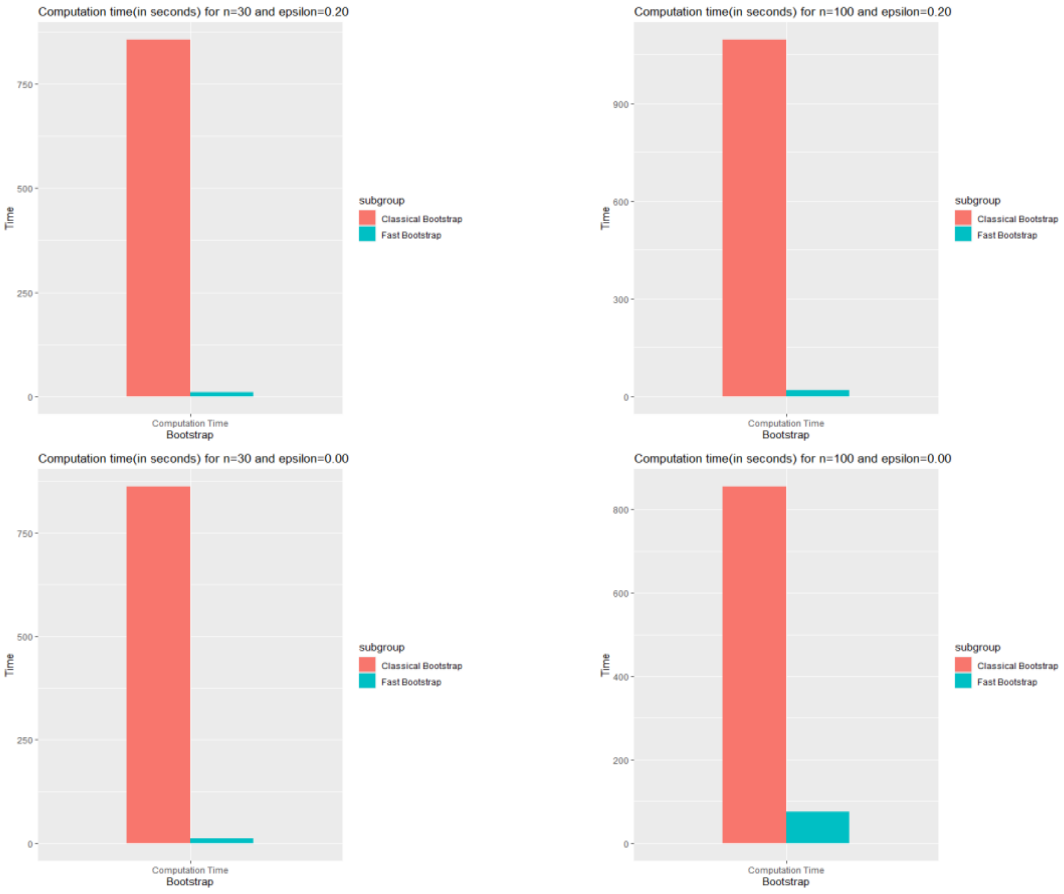


Figure 2: Comparison of computational cost

In the beginning we stated that how classical bootstrap deals with problem of numerical stability and computational cost. As Figure (1) shows how the Fast and Robust Bootstrap gives

numerically more stable estimates similarly Figure (2) clearly shows how the computational cost of Fast and Robust Bootstrap is much less when compared to Classical Bootstrap

7 Data Analysis

We now illustrate the stability of the inference based on the fast bootstrap on a simple and a multiple linear regression analysis. In both cases we compare the inference obtained using the bootstrap and the fast bootstrap. These examples simultaneously illustrate the serious effect of the outliers on the inference derived from the bootstrap and the robustness of the fast bootstrap.

7.1 Belgian International phone calls data set

This data consist of the number of international phone calls (in tens of millions) originated in Belgium between 1950 and 1973. From 1964 to 1969 the observations were mistakenly recorded.Instead of the number of calls, their total duration in minutes was registered.The linear regression model used to fit the data is

$$Calls = \alpha_0 + \beta_0 Year + \varepsilon,$$

where α_0 and β_0 are the parameters of interest and the errors are assumed to be independent and identically distributed with mean 0 and unknown but constant variance.Figure (3) displays the data with the robust and least squares fits. To obtain confidence intervals for the regression parameters β, we use the bootstrap and fast bootstrap methods.

We performed 10,000 bootstrap recalculations. Scatterplots of $\hat{\beta}_n^{R*} - \hat{\beta}_n$ for the fast bootstrap and of $\hat{\beta}_u^* - \hat{\beta}_n$ for the bootstrap are presented in Figure (4A). We clearly see that the fast bootstrap estimates are more stable. This is also reflected in the length of the confidence intervals in Figure (4B) where the length of Fast and robust bootstrap confidence interval is extremely low in comparison to that of classical bootstrap.

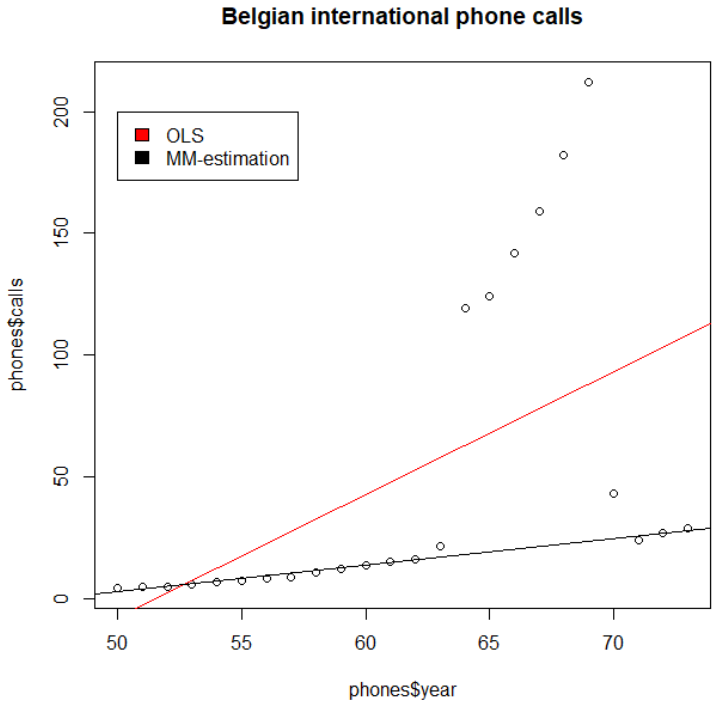
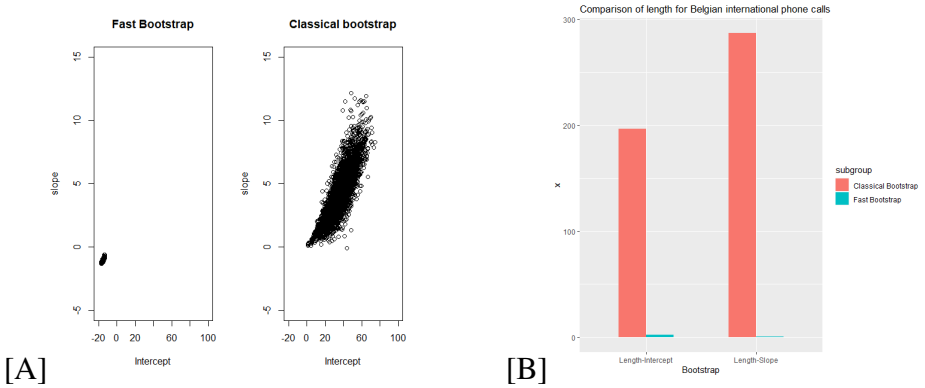


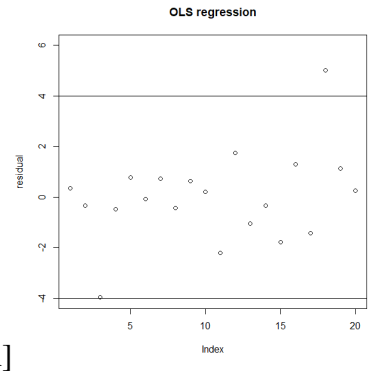
Figure 3: OLS and robust regression fits on the Belgian international phone calls

Coefficient	Fast Bootstrap	Classical Bootstrap
Intercept	(.52, 2.53)	(-438.71, -1.47)
Slope	(.99, 1.20)	(409.16, 11.89)

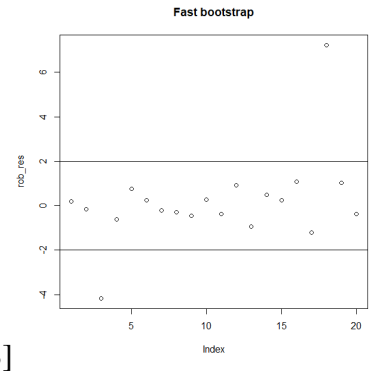


[A] [B]

Figure 4



[A]



[B]

Figure 5

7.2 Verbal Test Score Data

For drawing inferences drawing from a multiple regression analysis we chose this data set. These data contain observations drawn from 20 schools in the United States. They were first studied by Coleman (1966).The data consist of verbal mean test scores for sixth-graders drawn from 20 schools in the Mid-Atlantic and New England states. The explanatory variables are: staff salaries per pupil (Staff Salary), percent age of white-collar fathers (White Collar), socioeconomic status composite deviation (Soc. Status), mean teacher's verbal test score (Teacher Score) and mean mother's educational level (Mother Ed.). We fit a multiple linear regression model to these data to find which variables have a significant effect on the mean verbal test score of the students. We used the classical least squares fit and a 50% breakdown point and 95% efficient MM regression estimate with score functions in Tukey's family . Figure (5) contains the plot of the residuals obtained with the least squares and MM-regression estimates. From this plot it is clear that these data contain outliers and that the least squares fit is not appropriate. To determine which coefficients are significantly different from 0, we built 95% confidence intervals using both bootstrap and fast bootstrap methods to estimate the distribution of the robust MM-regression estimator. We used 5000 bootstrap samples to estimate the appropriate quantiles of the marginal distributions.The resulting confidence intervals are displayed in table given below.

Coefficient	Fast Bootstrap	Classical Bootstrap
Intercept	(13.8, 47.2)	(-16.08, 56.59)
Staff salary	(-3.15, -0.17)	(-3.74, 0.374)
White collar	(0.034, 0.133)	(-0.065, 0.147)
Soc. status	(0.554, 0.780)	(0.316, 0.793)
Teacher Score	(0.605, 1.729)	(0.184, 1.931)
Mother Ed.	(-6.43, -1.84)	(-7.30, 3.92)

The only significant coefficients using the bootstrap (at the 5% level) are those of Soc. Status and Teacher Score. The confidence intervals constructed with the fast bootstrap indicate that all coefficients are significant at this level. Also note that the lengths of confidence intervals by classical bootstrap is longer than that of Fast and Robust Bootstrap.

To explore the shape of the estimates of the marginal distributions obtained with each method, we used QQ-plots of the marginal bootstrap distributions. Figure (6) contains these plots for two marginal distributions, the other marginal distributions being very similar. As expected, the marginal distributions of the bootstrap have heavier tails than those of the fast bootstrap, resulting in unduly long confidence intervals.

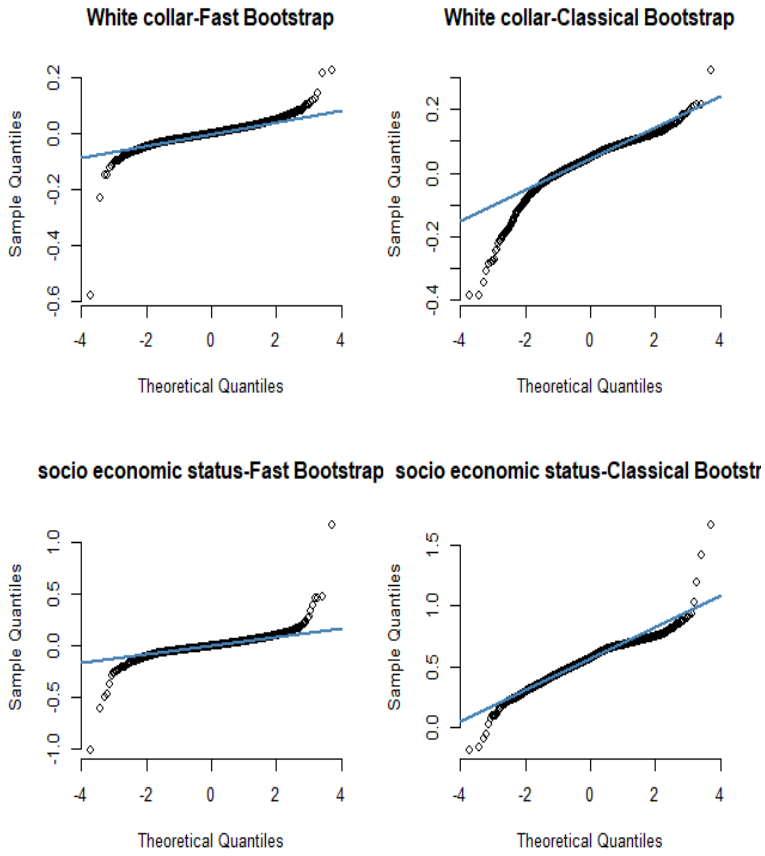


Figure 6: QQ-plots of the bootstrap and fast bootstrap marginal distributions for the verbal test score data

8 Concluding Remark

We have discussed the concept of fast bootstrap, which we use as a tool in estimation of distribution of robust regression estimates. We have discussed its asymptotic properties and robustness. It is clear that this process yields estimates faster than normal Bootstrap estimates and are much robust than the same. We have simulated data and implemented the technique to validate the concepts. We have found the it yields good coverages for the estimates confidence intervals even when used on data with significant contamination. Finally we have used the technique on two real life data-sets and obtained positive results as far as theoretical and computational properties are concerned.

Supplementary Materials

Interested readers may visit the following link to access the codes used for simulation tasks and the application on real data-sets:

https://github.com/manas16may/MTH-516A-Non-Parametric-inference

Acknowledgement

We express our gratitude towards Prof. Dootika Vats for her valuable feedback and constant guidance on this project.

References

Beaton, A. E. and Tukey, J. W. (1974). The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. *Technometrics*, 16(2):147–185.

Coleman, J. S. (1966). *Equality Of Educational Opportunity [Summary Report]*, volume 1. US Department of Health, Education, and Welfare, Office of Education.

Efron, B. (1979). The 1977 rietz lecture. *The Annals of Statistics*, 7(1):1–26.

Rousseeuw, P. and Leroy, A. (1987). Robust regression and outlier detection: Wiley interscience. *New York*.

Rousseeuw, P. and Yohai, V. (1984). Robust regression by means of s-estimators. In *Robust and nonlinear time series analysis*, pages 256–272. Springer.

Salibian-Barrera, M. (2000). *Contributions to the theory of robust inference*. Citeseer.

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

Yohai, V. J. (1987). High breakdown-point and high efficiency robust estimates for regression. *The Annals of statistics*, pages 642–656.