Bootstraping Least Weighted Squares

Jiří Skuhrovec[1]

Abstract. The paper uses bootstrap approach to analyze reliability of Least Trimmed Squares, and Least Weighted Squares robust estimators. In Monte Carlo simulation it measures performance of bootstrap confidence intervals. It also proposes use of further examination of bootstrap population for detecting character and severity of contamination present in data. The main contribution of paper lays in developing basic statistical inference tools for the two estimators, for which standard asymptotic approach following from classical theory could not be used - which limited their practical usability.

Keywords: Least Weighted Squares, Least Trimmed Squares, Bootstrap, Monte Carlo, Robust statistics

JEL classification: C50
AMS classification: 93E24,62F40

1 Introduction

1.1 Outline

This paper is divided as follows: this section proposes problem of data contamination which might be considered a typical setting for the robust regressions. Section 2 then formally introduces LWS estimate as a tool for solving given problem and reviews its key theoretical properties. The final Section 4 then presents a brief description of bootstrap application on LWS. The paper concludes by discussing interpretation of bootstrap confidence intervals, and methodology comparison with recent work of other scholars.

1.2 Base model

Consider problem of regression on dataset $[X, Y]$, where Y_i is variable randomly drawn from two populations:

$$Y_i = \begin{cases} X_i\beta^T + \varepsilon_i & \text{with probability } (1-c) \\ X_i\beta^C + \varepsilon_i & \text{with probability } c. \end{cases} \tag{1}$$

Here X is $(n \times p)$ matrix of explanatory variables, Y is $(n \times 1)$ vector of response variables, ε is $(n \times 1)$ vector of independent identically distributed (*iid*) disturbances with $\varepsilon_i \sim N(0, \sigma^2)\ \forall i$, and $c \in \langle 0, \frac{1}{2} \rangle$ is known fixed probability, which we will call contamination probability. Both β^T and β^C are unknown $(p \times 1)$ vectors, where $\beta^T \neq \beta^C$. For convenience, we will denote distribution functions of the first and second populations Y^T and Y^C respectively, denoting "true" and "contaminating" data. Note that random variables Y_i generated by the model are *iid*, with distribution function given by probabilistic combination of distributions of Y^T and Y^C.

This work focuses on problem of estimating β^T parameter of "true" population. Estimating β^C is not topic of this paper. In fact, we will treat the data, as if general form of Y^C model was not known. This means, that Y^C needs not to be necessarily linearly dependent on X, and estimate of its generating model parameters (such as some $\hat{\beta}^C$) can not be obtained. We will however use exactly (1) for generating

[1]Charles University, Institute of Economic Studies, Opletalova 26, 110 00, Prague, Czech Republic, e-mail: jskuhrovec@gmail.com
Reserach was supported by Grant Agency of the Czech Republic under the project 402/09/0557.

data in our simulations, as a convenient and understandable form of contamination, which also poses sufficiently serious problem for regression methods, since it closely imitates linear model, which is what the methods are trying to find. For discussion about effect of contamination, we will need following definition.

Definition 1 (Level of contamination)**.** For any dataset generated by (1), the level of contamination is computed as follows (here denotes set cardinality) :

$$\dot{c} = \frac{\#\left\{i : Y_i = X_i\beta^C + \varepsilon_i\right\}}{n}. \tag{2}$$

Remark 1 ($\dot{c}$ distribution)**.** In (1), the level of contamination $\dot{c}$ is a random variable with binomial distribution

$$\dot{c} \sim \frac{B(n, c)}{n}. \tag{3}$$

The (1) is both reasonably simple and general framework for testing robust methods. Its advantages of the model lie in (realistic?) stochastic level of contamination $\dot{c}$, and in the fact that contamination has linear form - which poses serious challenge to estimators that look for *linear* patterns.

2 Least Weighted Squares

In this section we will define Least Weighted Squares(LWS) estimate, and briefly go through some of its known properties. For better understanding, differences with more known OLS and LWS estimates (and their minimized functions) are examined. The proposed definition is equivalent to original one given in [5], however notation has been adjusted for our needs.Before we can move on to defining LWS, we will need two more auxiliary definitions.

Definition 2 (Solution space)**.** Denote by $\mathbb{H}^n$ set of all sequences $\{h_i\}_{i=1}^n$ such that $\forall i : h_i \in \mathbb{N}, 1 \leq h_i \leq n, j \neq i \Rightarrow h_j \neq h_i$.

Remark 2. $\mathbb{H}^n$ is a set including all permutations of integers from 1 to n, hence $|\mathbb{H}^n| = n!$.

Definition 3 (Weight function)**.** By weight function we mean any function $w(x)$, that satisfies following conditions

1. $w(x)$ is function $\langle 0, 1\rangle \to \langle 0, 1\rangle$.
2. $w(x)$ is non-increasing.
3. $w(0) = 1$.

Definition 4 (Least Weighted Squares estimator)**.** We will call LWS a solution to following optimization problem. Denote:

$$r_i^2(\beta) = (Y_i - \beta X_i)^2 \qquad \text{(residual)}, \tag{4}$$

$$S^2_{\text{LWS}}(\beta, h) = \sum_{i=1}^{n} w\left(\frac{h_i - 1}{n - 1}\right) r_i^2(\beta) \qquad \text{(objective function)}, \tag{5}$$

$$\left[\hat{\beta}_{\text{LWS}}, \hat{h}\right] = \underset{\beta\in\mathbb{R}^p, h\in\mathbb{H}^n}{\arg\min}\ S^2_{\text{LWS}}(\beta, h) \qquad \text{(LWS estimate)}. \tag{6}$$

The LWS computation is quite complicated, as it requires discrete optimization over $\mathbb{H}$, and has been broadly addressed in [4]. Here we only need to know, that in reasonable time (20 s) we can compute exact LWS solution for $n < 10$, and approximate (yet quite reliable) solution for $n < 100.000$.

In [6] Víšek shows that $\hat{\beta}_{\text{LWS}}$ is asymptotically linear, $\sqrt{n}$-consistent, and asymptotically normal under conditions generalizing Gauss-Markov theorem. Here we however examine LWS behavior under different conditions - with contamination present in data. As this would hardly be feasible analytically, we provide empirical examination with use of bootstrap.

3 Application of LWS and LTS

Now we will provide definition of Least Trimmed Squares (LTS) estimate. Definition notation considerably differs from original one provided for example in [2], however these two are equivalent as shown in [7]. This definition shows, that LTS belongs to class of LWS estimators.

Definition 5 (Least Trimmed Squares)**.** We will call a LWS estimate $[\hat{\beta}, \hat{h}]$ with weights[1] $w = I(x < t)$ $t \in \left\langle \frac{1}{2}, 1 \right)$ the Least Trimmed Squares estimator $\hat{\beta}_{LTS}$ at level t.

The nature of estimate can perhaps better be understood in following form, demonstrating that objective function of LTS is a plain (not weighted) sum of residuals on chosen observations, or equivalently an S^2_{OLS} computed on chosen subset of data.

Remark 3. $\left[\hat{\beta}_{\text{LTS}}, \hat{h}\right] = \underset{\beta \in \mathbb{R}^p, h \in \mathbb{H}^n}{\arg\min} \sum_{i=1}^{nt} r^2_{h_i}(\beta)$

Example 1 (Least Trimmed Squares)**.** Consider (1) with $n = 100$ and $c = 0.4$, hence roughly 40 observations are contaminating. We might then use LTS with $t = 0.6$, so that our estimator will assign unit weight to *some* 60 observations, and no weight to others. LTS will then choose (assign $w = 1$) such subset of data, that has the lowest S^2_{OLS} (plain sum of squared residuals) among all possible subsets of size 60. The $\hat{h}$ will not be unique. However the $\hat{h}_1, .., \hat{h}_{60}$ will be any permutation of selected observations indices, and $\hat{h}_{61}, .., \hat{h}_{100}$ will be any permutation of the other indices.

Such estimation is based on idea, that some (nt)-sized subset of observations is uncontaminated, hence closely follows linear model Y^T, and will be most likely candidate for minimizing S^2_{OLS}. Since (1) implies that Y^T has *probably*[2] generated majority of the data, we might then believe that $\hat{\beta}_{\text{LTS}}$ is reasonable estimator of β^T.

To realize this, consider the most likely case when $\dot{c} = 0.4$ (exactly 40 observations come from Y^C). Now, if we are choosing subsamples of size $nt = 60$, then these subsamples will have contamination $0 \leq \dot{c} \leq \frac{2}{3}$. The subsamples with $\dot{c}$ close to 0 are most likely to minimize S^2, since they actually contain data from some linear model (Y^T).

However, with probability 0.457 (recall Remark 1), the number of contaminating points will get above 40, and it will not be possible to choose uncontaminated subset. All subsets will then contain some potential outliers (generated by Y^C) which will seriously distort the β^T estimate, like they do in standard OLS case.

Literature regarding robust estimates sometimes assesses estimators using *breakdown point*, the level of contamination at which estimator breaks down i.e. fails to be bounded in probability. Precise definition of breakdown point is complicated and can be found for example in [2]. LTS is considered very robust, because of having breakdown point $1 - t$ (up to 0.5), so that it breaks down only in case when $\dot{c} \geq 1 - t$.[3]

Now, as suggested in Example 1, setting weights to exactly fit the most likely amount of contamination is not desirable. For LTS it would be perhaps reasonable to set $t = 1 - \dot{c}$, so that LTS could assign zero weights to all contamination observations, but not to any further ones, which would only decrease estimate efficiency. In case when we know only c, the question of setting t gets more subtle. The Example 1 demonstrates, that for $c = 0.4$, the use of $\hat{\beta}_{\text{LTS}}$ with $t = 0.6$ is risky. In case when real contamination $\dot{c}$ gets above $1 - t = 0.4$, estimate gets heavily distorted by outliers. In contrast with that, if contamination falls below breakdown point $1 - t$, in particular $\dot{c} < 0.4$ nothing serious happens - only some information is lost, and our estimate is slightly less exact, than if we used $t = 1 - \dot{c}$. So, in choice of t, there is some trade-off between loss of information and risk of getting contaminated subsample. Clearly we should tend to choose some $0 < t < 1 - c$.

Problem of such trade-off in setting weight is more general. Using calculus of variations, optimal weight function $w^*(x, c)$ might possibly be derived for estimating β^T in (1), minimizing some measure such as mean squared error. Such result would certainly be useful[4], it would however again possibly need

[1] Here and furhter on, I denotes logical indicator function.

[2] With some probability $p \geq 0.5$. This follows from $c \leq 0.5$ and Remark 1

[3] Whereas OLS is breaks down for any $\dot{c} > 0$, which is perhaps the main motivation for using robust methods in practice where some errors might be present in data.

[4] Some results regarding optimal weights can be found in [3], the paper however deals with slightly different problem.

some additional assumptions regarding contamination. We will now focus on more practical goals. We now propose a general form of weight function, which is in no sense optimal, however it is both reasonably simple and versatile for practical purposes.

$$w_{a,b,t}(x) = (1 - ax)^b I(x < t) \text{ where } a \in \langle 0, 1\rangle \, b \in \mathbb{R}^+ \quad t \in \langle 0, 1\rangle \tag{7}$$

We will use this form in the rest of paper, written shortly as $w_{a,b,t}$. Note that all previously discussed forms of weight function (including LTS and OLS estimates) can be written in this form. Here t is a trim level analogous to LTS, and a, b parametrize differences in sensitivity to points close or far from regression plane.

4 Bootstraping LWS

We will now apply bootstrap approach on LWS statistic and examine the results. The percentile and ABC bootstrap methods for estimating confidence intervals (described for example in [1]) shall be used. Can we expect reasonable results from bootstrap applied on LWS statistic? In [8] Willems and Van Aelst justifiy application of bootstrap on LTS by its asymptotic normality. Since LWS is smoother generalization of LTS, which is also asymptotically normal, it is reasonable to use bootstrap here as well.

By $B \in \mathbb{N}$ times resampling observations $o = [Y, X]$, we shall obtain B bootstrap samples $\hat{o}_1, ..., \hat{o}_B$. On each of those samples we will then evaluate LWS to obtain $(p \times 1)$ vectors $\hat{\beta}_1, ..., \hat{\beta}_B$. By examining percentiles of their distribution, we will then compute confidence intervals $\left[\hat{\beta}^L, \hat{\beta}^U\right]$.

Example 2 (LWS on moderately contaminated data)**.** Consider (1), with $n = 100, p = 3, \beta^T = [1, 1, 1], \beta^C = [-1, -1, -1], \dot{c} = 0.2$ and LWS with $w_{1,1,0.7}$. We examine behavior of $\hat{\beta}_1$ and its confidence intervals with $\alpha = 0.1$. We report OLS results as benchmark - however Gauss-Markov conditions do not hold, and we therefore do not expect it to work well.

Table 1: Bootstrap confidence intervals, with moderate contamination ($\dot{c} = 0.2$)

statistic	$\hat{\beta}_1$	$\hat{\beta}_1^L$	$\hat{\beta}_1^U$
OLS	1.31	0.79	1.84
LWS (PRC)	1.01	0.71	1.43
LWS (BCA)	1.01	0.68	1.40

As expected, OLS estimate and confidence intervals went off (in case of β_0 and β_2 confidence interval did not even contain β^T. On contrary LWS estimates with both confidence intervals seem quite accurate. Note, that the used weights were still not fully appropriate[5].

Figure 1 shows the full distribution of bootstrap population. One-peaked distribution which is close to normal and has median near actual LWS estimate is a good indicator, that estimated parameter is reliable. Thus we may now conclude that we found LWS method reliable with moderate contamination and both percentile and BCA method were performing equally and fairly accurately.

Example 3 (LWS on heavily contaminated data)**.** Again use (1), with $n = 100, p = 3, \beta^T = [1, 1, 1], \beta^C = [-1, -1, -1]$, and examine $\hat{\beta}_1$ estimate on population with $\dot{c} = 0.45$. The robustness of weight function was this time increased to $w_{1,1,0.5}$, again to slightly overshoot contamination level.

Although the estimate itself is quite close to β_1^T, the confidence intervals are very broad, and biased to left. Apparently, they contain $\beta_1^C = -1$, which (even though did not break down the LWS estimate itself) possibly broke down many estimates made on bootstrap samples. This can nicely be seen from bootstrap histogram in Figure 2.

[5] More appropriate (and efficient) would perhaps be $w_{1,0,0.8}$ - however result with such weight would not be too interesting as the case of $\dot{c}$ knowledge is not very common in practice.

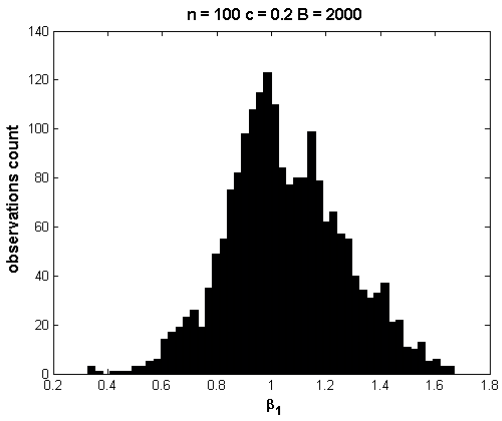

Figure 1: Histogram of $\hat{\beta}_1^1, ..., \hat{\beta}_1^B$ population

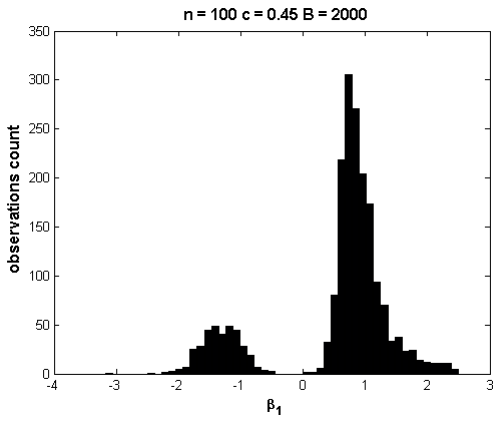

Figure 2: Histogram of $\hat{\beta}_1^1, ..., \hat{\beta}_1^B$ population

Table 2: Bootstrap confidence intervals, with heavy contamination ($\dot{c} = 0.45$)

statistic	$\hat{\beta}_1$	$\hat{\beta}_1^L$	$\hat{\beta}_1^U$
OLS	-0.37	-1	0.25
LWS (PRC)	0.81	-1.51	1.63
LWS (BCA)	0.81	-1.46	1.91

The $\hat{\beta}$ distribution is far from being normal. In contrary, it nicely reflects structure of our data - majority of which has been generated using $\beta_1^T = 1$ and minority using $\beta_1^C = -1$. The information that we can gain from histogram could hardly be obtained while using confidence intervals only.

Confidence intervals performance Of course, no conclusions can be made, based on two examples. These were presented only to give reader some flavor of working with LWS bootstrap confidence and pose few tentative statements, rather than to prove something. Now we are about to run more thorough simulation, in which we will try to measure reliability of bootstrap confidence intervals at level $\alpha = 10\%$. Used approach directly follows from definition of confidence interval:

1. Generate $K = 1000$ populations of size n, using (1) with some β^T.
2. On each population perform bootstrap estimate of LWS confidence intervals at level α, thus obtain $[\hat{\beta}_1^L, \hat{\beta}_1^U], ..., [\hat{\beta}_K^L, \hat{\beta}_K^U]$.
3. Compute some aggregate statistics on these intervals.

In line with standard approach, we will perform two simulations, with uncontaminated and contaminated data. In the first one will demonstrate us how serious mistake we can make by plugging robust estimate into setting where it is not necessary, second will then measure performance in setting similar to Example 2. We will use two simple measures of confidence interval performance[6]:

1. $hits = \frac{1}{K}\#\{[\hat{\beta}_i^L, \hat{\beta}_i^U] : \hat{\beta}_i^L < \beta^T < \hat{\beta}_i^U\}$
2. $width = \frac{1}{K}\sum_{i=1}^{K} \hat{\beta}_i^U - \hat{\beta}_i^L$

Now consider (1) with $n = 100, p = 5$. Each of two different settings $c = 0$, and $c = 0.2$ will be used to generate 1000 random populations $[Y, X]$, on each of which we will evaluate OLS, LWS and respective four confidence interval estimators. Two aggregate measures defined above will be reported, measured separately for uncontaminated and moderately contaminated population.

Like expected, in uncontaminated case we see OLS superiority, however also LWS methods perform very well - in terms of hits they are close to optimal value $1 - \alpha = 0.9$, the PRC method has even hit β^T

[6]These are measures for location model($p = 1$). For general p-dimensional model, these are computed separately for each dimension and average values over all p values are reported.

Table 3: Bootstrap confidence intervals coverage

stat.	$\dot{c}$	LWS_{PRC}	LWS_{BCA}	OLS_{BCA}	OLS_{norm}
hits	0	0.986	0.910	0.882	0.906
	0.2	0.952	0.908	0.792	0.796
width	0	0.98	0.95	0.41	0.42
	0.2	0.74	0.72	0.79	0.80

too often. The width of both OLS-based methods was considerably lower, which follows from its higher efficiency.

To conclude, in this section we saw that LWS performance, and more notably performance of its bootstrap confidence intervals is quite good, as long as used weight function at least approximately reflects type of contamination. From two bootstrap methods (percentile and BCA) we found BCA as slightly better performing, which was expected by theory.

5 Conclusion

Major contribution of this work should be seen as practical. As a by-product of this work a Matlab toolbox has been created, able to obtain not only LWS, but also more popular LTS estimates with greater reliability and speed than presently used packages. The toolbox is freely available together with this work.[7]. As its integral part, the source codes of presented simulations are provided, so that any researcher can *exactly* replicate and verify our results, and possibly also modify codes for purposes of further research.

The lack of inference tools is one of major drawbacks of robust statistics. The potential of bootstrap to overcome such problem has been known for many years, however until recently there was not enough computational power to practically apply such approach. Toolbox distributed with this work is able to estimate confidence intervals within few seconds, making LWS method potentially more useful in statistical practice.

References

[1] Bradley Efron and Robert J. Tibshirani. *An Introduction to Boostrap*. Chapman and Hall, 1993.

[2] Frank R. Hampel, Elvezio M. Ronchetti, Peter J. Rousseeuw, and Werner A. Stahel. *Robust Statistics - The Approach Based on Influence Functions*. John Wiley and Sons, Inc., 1986.

[3] Libor Mašíček. Optimality of the least weighted squares estimator. *Kybernetika*, 40:715–734, 2004.

[4] Jiří Skuhrovec. Analysis of lws empirical properties using bootstrap. *Diploma thesis*, 2010.

[5] Jan Amos Víšek. The regression with high breakdown point. *Robust*, pages 324–356, 2000.

[6] Jan Amos Víšek. The least weighted squares ii. consistency and asymptotic normality. *Bull. of the Czech Economet. Society*, 9:1–28, 2002.

[7] Jan Amos Víšek. Consistency of the least weighted squares. *Kybernetika*, 2008(submitted).

[8] Gert Willems and Stefan Van Aelst. Fast and robust bootstrap for lts. *Computational Statistics & Data Analysis*, 48(4):703–715, April 2005.

[7] At *http://www.skuhry.com/lws* together with more comprehensive descriptive paper.