Chapter 4
High Breakdown Regression Procedures

Introduction

While M and BI estimators provide an improvement over OLS if the data has an outlier or high influence point, they cannot provide protection against data with large amounts of contamination. As seen in Example 2.3 (Section 2.5.3), clusters of bad data may overwhelm these methods, which leads to poor estimators. Thus, the discussion next turns to methods with high breakdown points (see also Markatou and He (1994)). They have the ability to sift through as much as 50% of the data being contaminated and still provide decent estimators for explaining the general trend.

§4.1 Least Median of Squares (LMS)

All of the methods mentioned above have an objective function that involves a Σ operator. Changing the function inside this summation operator, as in M or BI regression, had limited success. Rousseeuw (1984) suggested replacing the summation by the median, paralleling the fact that the sample median is more robust than the sample mean in location estimation. This leads to Least Median of Squares regression (LMS), with the objective function being

$$\min_{\forall \mathbf{b}} \ \operatorname*{med}_{\forall i} \ (y_i - \mathbf{x}_i'\mathbf{b})^2 .$$

The resulting estimator has a 50% breakdown point, but converges at the slow rate of $O\left(n^{-1/3}\right)$ making its asymptotic efficiency against normal errors zero (Rousseeuw and Leroy, 1987, page 179). In the location model there exists a closed-form algorithm to calculate LMS. Originally, no exact algorithm was available to calculate LMS in the regression setting (excluding, of course, degenerate cases). Stromberg (1993) then provided an exact LMS algorithm, but requires that all C_{p+1}^{n} subsets be investigated. This being a daunting task, the approach generally taken is to *approximate* LMS by one of the handful of available subsampling algorithms.

Consider a subsample of size *p*, the number of unknown coefficients. This is referred to as an elemental set because assuming that the reduced **X** matrix, i.e. the matrix formed by using only the *p* rows of **X** that correspond to the *p* subsampled observations, has full rank, an exact fit can be obtained from these points and the objective function can then be evaluated.

One subsampling algorithm (Rousseeuw and Leroy, 1987, page 197) simply draws an elemental set and evaluates the objective function. This is then repeated a large number of times, and the final LMS estimate corresponds to the estimate that had the smallest observed objective function. This estimate retains the high breakdown and convergence properties of the theoretical LMS (Rousseeuw and Bassett, 1991). However, even if an exhaustive search of all C_{p}^{n} elemental sets is performed, the resulting estimator is generally not the true LMS estimator, but rather only an estimate of it (just as in the case of MVE estimation via this method). Furthermore, the calculations involved to obtain the estimated regression coefficients are of order $O\left(n^{p+1}\right)$ (Rousseeuw, 1993). Instead, the number of elemental subsets selected can be based on a probabilistic argument of how likely it is to obtain an elemental set that contains only good observations. The drawback is that since this probability is less than 1, the algorithm may break down in its calculation of a high breakdown estimator. This defeats the purpose of the high breakdown philosophy. In any event, the number of elemental subsets needed is roughly $3 \cdot 2^{p}$ and $4.6 \cdot 2^{p}$ for 95% and 99% probabilities, respectively, of obtaining at least one purely good elemental set. To obtain the regression coefficients, the order of calculation would then be these numbers of subsets multiplied by *n*, respectively (Rousseeuw, 1993).

Of course, while the probabilistic argument provides for the analysis of a strictly good elemental subset with high probability, there is no guarantee that any of those randomly obtained good elemental subsets reflects the general trend by itself. Therefore, the resulting estimator can be potentially very misleading.

A second algorithm was introduced by Rousseeuw (1993) to reduce the order of computations required and eliminate the problem of having an algorithm breakdown due to the probability of obtaining a purely good elemental set being less than one. Basically, the data is randomly assigned into blocks of size $2p-2$. Any extra points are disbursed as evenly as possible. Then, within each block, all possible subsets of size p are evaluated with the objective function. Again, the final LMS estimate corresponds to the estimate that had the smallest observed objective function.

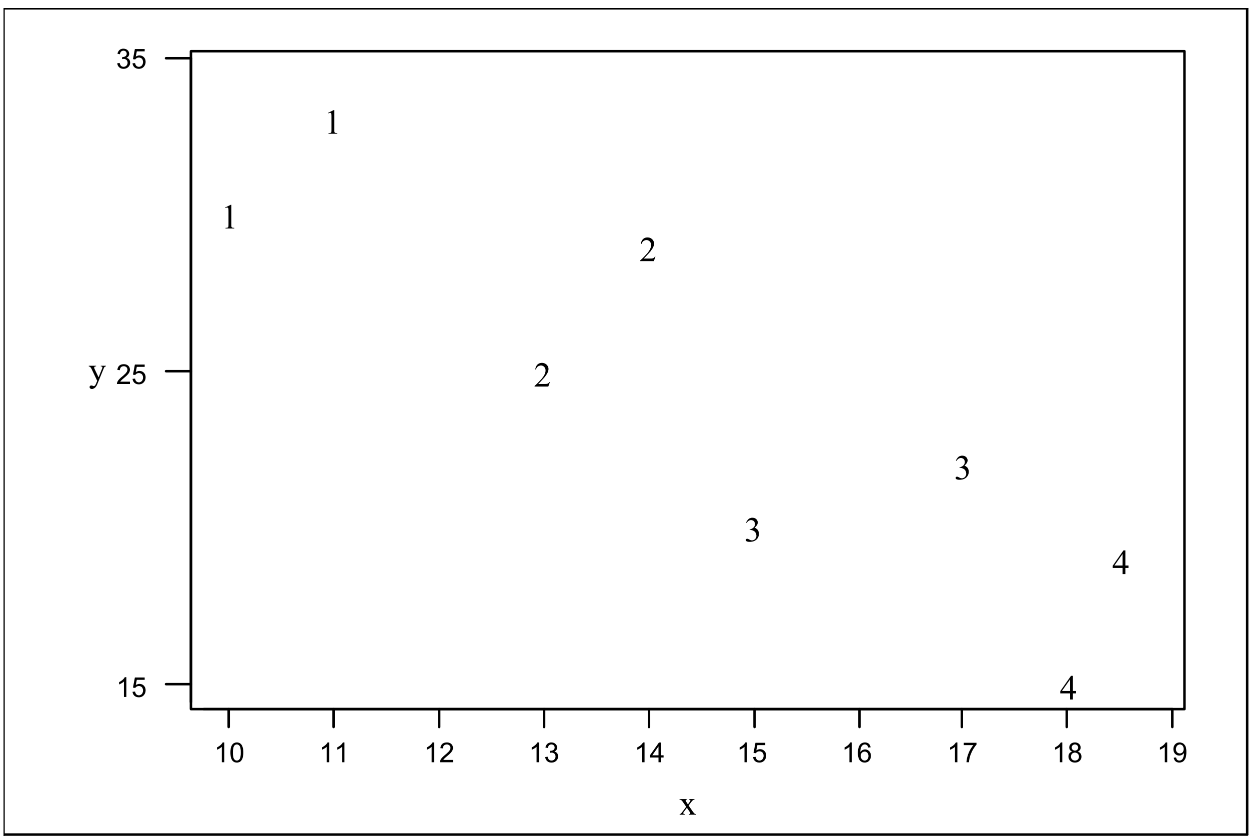

Figure 4.1: Possible configuration of the four blocks used in the second LMS subsampling algorithm, given eight observations on a scatterplot.

It is guaranteed that at least one elemental set will consist only of good observations. The advantage of this algorithm is theoretical. It achieves a high breakdown point, but the estimates obtained could be very misleading. A major problem is that there is no guarantee that any block will provide enough information concerning the general trend. To illustrate this idea, suppose that a simple linear regression (SLR) model is posed for a dataset of eight good observations. Here, the block size is $2(2)-2=2$, and there are $8/2=4$ blocks. Next, suppose that the data are shown in Figure 4.1, and are labeled as to which block each observation was randomly assigned. Even though the slope is obviously negative, all four blocks result in positive slope estimates. It seems that the algorithm relies on asymptotic combinatoric probabilities, that obtaining this type of block structure goes to zero as n gets large. If there are replications at the regressor locations, such as in a designed experiment, this scenario may not be all that uncommon.

The first algorithm will be utilized in the case studies to come. The number of randomly drawn elemental sets is generally taken to be 500 or 1000 in practice. In some case studies to follow, as many as 50,000 randomly drawn elemental subsets were used in an attempt to avoid an algorithm breakdown.

Additionally, the LMS estimator that is obtained from a random subsampling algorithm can be modestly improved by adjusting the intercept. By viewing the residuals as a univariate sample, the exact LMS algorithm for location estimation can be performed. The updated intercept is found as the "location LMS" of the residuals from the current LMS regression estimator. This procedure is guaranteed to reduce (or not change) the LMS objective function, and it also eliminates the condition of always having at least p residuals being zero (because of the exact fit). This intercept adjustment procedure is incorporated into all LMS calculations.

§4.2 Least Trimmed Squares (LTS)

Recall that one drawback to LMS is that it possesses a very slow convergence rate of $O\left(n^{-1/3}\right)$. Rousseeuw (1984) introduced Least Trimmed Squares (LTS) to remedy this situation.

It also possesses a 50% breakdown point, but converges at the faster rate of $O\left(n^{-1/2}\right)$. Here, the objective function is

$$\min_{\forall \mathbf{b}} \sum_{i=1}^{h} r_{[i]}^{2},$$

which represents the sum of the h smallest squared residuals. As mentioned before (in Chapter 3), h is generally taken to be $\left[(n+p+1)/2\right]$. The problem is that no closed-form algorithm exists, except for the location model, to construct the true LTS estimator. Instead, the algorithms stated previously for approximating LMS can also be used to approximate LTS simply by changing the objective function being evaluated for each elemental set.

The intercept adjustment step is also available for the LTS estimator. Simply replace the current intercept with the "location LTS" estimate of the residuals from the current LTS regression estimator. This update is performed in all LTS calculations. Agullo (2001) offers more discussion regarding LTS algorithms.

Both LMS and LTS are inefficient estimators. In fact, for the location model, LMS has a 1.39% asymptotic efficiency versus the sample mean under normal errors. LTS is only slightly better, having a 7.14% asymptotic efficiency versus the sample mean under normal errors. Therefore, several methods employ LTS, which converges more rapidly than LMS, as an initial estimator and perform some improvement calculation. These are referred to as one-step estimators. The idea is to utilize the high breakdown properties of these initial estimators, but improve on their lack of efficiency. However, this introduces another problem. The improvement step generally will require weights for the observations. Thus, robust weights are required to retain the desired high breakdown properties. This leads back to the material of Chapter 3. As mentioned in Section 3.3, a popular choice among the robust weighting schemes is the MVE-based Mallows weight.

The remainder of Chapter 4 will introduce two competing one-step estimators and provide an overall high breakdown regression analysis of the stackloss case study.

§4.3 One-Step Generalized M Estimators

The previous high breakdown estimators, LMS and LTS, attack the regression problem by changing the objective function to an expression that leads to improved breakdown point capabilities. Because of poor efficiency and numerical sensitivity due to both the random subsampling process as well as to small internal movements of the data (these topics are discussed later in Chapter 8), other high breakdown regression techniques have been developed. One remedy for the poor efficiency is to incorporate a high breakdown initial estimator with the generalized M estimator to obtain the one-step generalized M estimator. The objective function has the same form as the bounded influence estimator of Section 2.1, but with robust leverage weights. The solution is no longer found via the IRLS procedure, but instead through a one step Taylor series expansion of the objective function. This estimator can be written as the sum of two terms: the initial estimator and the one-step improvement calculation. LTS has become the initial estimator of choice for many one-step improvement algorithms. It has a high breakdown point and converges more rapidly than LMS. The GM-estimators inherit the high breakdown point of LTS, but improve on the efficiency aspect.

In the following discussion on one-step GM estimators, it is understood that

(1) the initial estimator, $\hat{\boldsymbol{\beta}}_0$, is LTS,

(2) the residuals from the initial fit are denoted by $r_i(\hat{\boldsymbol{\beta}}_0) = y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_0$,

(3) a diagonal matrix $\mathbf{W} = diag(w_i(\mathbf{x}_i))$ of robust Mallows weights, with

$$w_i = \min\left[1, \left[\frac{\chi^2_{.95,p-1}}{RD_i^2}\right]\right],$$

is calculated using MVE estimates (based solely on the regressor space), and

(4) The robust scale estimate, $\hat{\sigma}_0$, is based on the LMS estimate (Rousseeuw and Leroy, 1987, pg. 202), and is found as

$$\hat{\sigma}_0 = 1.4826 \cdot \left(1 + \frac{5}{n-p}\right) \underset{\forall i}{med} \; |r_i(\hat{\boldsymbol{\beta}}_0)| \; .$$

§4.3.1 Mallows 1-Step Estimator

The Mallows-1-step (M1S) estimator, a Generalized M estimator, was introduced by Simpson, Ruppert, and Carroll (1992). The focus of the 1-step improvement is to incorporate a leverage control term and an outlier control term in the estimation of $\hat{\boldsymbol{\beta}}$. The residuals from the initial estimate are utilized, with the M1S estimator being the solution to the altered normal equations

$$\sum_{i=1}^{n} w_i \psi\left(\frac{r_i(\hat{\boldsymbol{\beta}}_0)}{\hat{\sigma}_0}\right)\mathbf{x}_i = \mathbf{0}.$$

Outliers are controlled by the ψ-function downweighting large scaled residuals. For our discussion the Huber ψ-function is used. In addition, with the w_i's being robust Mallows weights, high leverage points get downweighted due to this term. Using Newton-Raphson to solve the altered normal equations, the form of the estimator is

$$\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_0 + \mathbf{H}_0^{-1}\mathbf{g}_0,$$

where

$$\mathbf{g}_0 = \hat{\sigma}_0 \sum_{i=1}^{n} \psi\left(\frac{r_i(\hat{\boldsymbol{\beta}}_0)}{\hat{\sigma}_0}\right) w_i \mathbf{x}_i$$

and

$$\mathbf{H}_0 = \sum_{i=1}^{n} w_i \mathbf{x}_i \mathbf{x}_i' \psi^{(1)}\left(\frac{r_i(\hat{\boldsymbol{\beta}}_0)}{\hat{\sigma}_0}\right),$$

with $\psi^{(1)}$ being the first derivative of ψ, i.e. $\psi^{(1)}(u) = \dfrac{\partial \psi(u)}{\partial u}$. This can be simplified when written in matrix notation as

$$\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_0 + (\mathbf{X}'\mathbf{B}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\boldsymbol{\psi}\,\hat{\sigma}_0.$$

Here, $\boldsymbol{\psi}$ is an $n \times 1$ vector, $\mathbf{W}$ is the $n \times n$ weight matrix defined earlier in Section 4.3, and $\mathbf{B}$ is the $n \times n$ diagonal matrix such that $\mathbf{B} = diag\left(w_i \psi^{(1)}\left(\frac{r_i(\hat{\boldsymbol{\beta}}_0)}{\hat{\sigma}_0}\right)\right)$. Using the Huber ψ-function, the diagonal elements of $\mathbf{B}$ become

$$b_{ii} = \begin{cases} w_i, & if \left| r_i(\hat{\beta}_0) \right| \le c\,\hat{\sigma}_0, \\ 0, & otherwise. \end{cases}$$

The $\boldsymbol{\psi}$ vector elements are calculated as

$$\psi_i = \begin{cases} -c, & if \;\; r_i(\hat{\boldsymbol{\beta}}_0) < -c\,\hat{\sigma}_0, \\ \dfrac{r_i(\hat{\boldsymbol{\beta}}_0)}{\hat{\sigma}_0}, & if \;\; \left| r_i(\hat{\boldsymbol{\beta}}_0) \right| \le c\,\hat{\sigma}_0, \\ c, & if \;\; r_i(\hat{\boldsymbol{\beta}}_0) > c\,\hat{\sigma}_0. \end{cases}$$

To further the analysis beyond estimation, standard errors are needed for each of the coefficients. If the $p \times p$ matrix $\mathbf{M}_0$ is defined as

$$\mathbf{M}_0 = \hat{\sigma}_0^2 \sum_{i=1}^{n} w_i^2\, \mathbf{x}_i \mathbf{x}_i' \psi^2\left(\frac{r_i(\hat{\boldsymbol{\beta}}_0)}{\hat{\sigma}_0} \right),$$

then the estimated (asymptotic) covariance matrix for the parameter estimates is given by

$$Cov(\hat{\boldsymbol{\beta}}) = \mathbf{H}_0^{-1}\mathbf{M}_0\mathbf{H}_0^{-1}.$$

By defining the matrix $\mathbf{V} = diag\left(w_i \psi\left(\dfrac{r_i(\hat{\boldsymbol{\beta}}_0)}{\hat{\sigma}_0} \right) \right)$, the estimated covariance matrix can be written in matrix form as

$$Cov(\hat{\boldsymbol{\beta}}) = \hat{\sigma}_0^2 (\mathbf{X}'\mathbf{B}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{V}^2\mathbf{X})(\mathbf{X}'\mathbf{B}\mathbf{X})^{-1}.$$

Thus, standard errors for the M1S coefficients are determined by the square root of the diagonal elements of this estimated covariance matrix.

§4.3.2 Schweppe 1-Step Estimator

Another generalized M estimator is the Schweppe-1-Step (S1S) estimator introduced by Coakley and Hettmansperger (1993). The focus is on the selection of an appropriate weighting scheme. The M1S estimator is modified by replacing the Mallows form of the altered normal equations with the Schweppe form of the altered normal equations. Basically, this entails adding a weight to the denominator of the ψ-function argument, which improves the efficiency of the estimator (Coakley and Hettmansperger (1993)). The altered normal equations for the S1S estimator are

$$\sum_{i=1}^{n} w_i \psi\left(\frac{r_i(\hat{\boldsymbol{\beta}}_0)}{\hat{\sigma}_0 w_i}\right)\mathbf{x}_i = \mathbf{0}.$$

These equations are of the same form as those for BI regression. The difference is that the S1S method uses the same Mallows weights that are used in the M1S method rather than the hat diagonal-based Welsch weights that are used in BI regression.

A Gauss-Newton approximation using a first-order Taylor series expansion about the initial estimate $\hat{\boldsymbol{\beta}}_0$ yields a one-step improvement of the form

$$\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_0 + (\mathbf{X}'\mathbf{B}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\boldsymbol{\psi}\,\hat{\sigma}_0 .$$

This has the same form as the M1S method, and uses the same weight matrix, **W**, but with changes in the definition of the **B** matrix and $\boldsymbol{\psi}$ vector. Now, $\mathbf{B} = diag\left(\psi^{(1)}\left(\frac{r_i(\hat{\boldsymbol{\beta}}_0)}{\hat{\sigma}_0 w_i}\right)\right)$. Using the Huber ψ-function, the diagonal elements of **B** are

$$b_{ii} = \begin{cases} 1, & if\ \left| r_i(\hat{\boldsymbol{\beta}}_0) \right| \le c\,\hat{\sigma}_0 w_i, \\ 0, & otherwise. \end{cases}$$

The $\boldsymbol{\psi}$ vector entries are calculated as

$$\psi_i = \begin{cases} -c, & if\ r_i(\hat{\boldsymbol{\beta}}_0) < -c\,\hat{\sigma}_0 w_i, \\ \dfrac{r_i(\hat{\boldsymbol{\beta}}_0)}{\hat{\sigma}_0 w_i}, & if\ \left| r_i(\hat{\boldsymbol{\beta}}_0) \right| \le c\,\hat{\sigma}_0 w_i, \\ c, & if\ r_i(\hat{\boldsymbol{\beta}}_0) > c\,\hat{\sigma}_0 w_i. \end{cases}$$

By defining the matrix $\mathbf{V} = diag\left(w_i \psi \left(\frac{r_i(\hat{\boldsymbol{\beta}}_0)}{\hat{\sigma}_0 w_i} \right) \right)$, the estimated covariance matrix can be written in matrix form as

$$Cov(\hat{\boldsymbol{\beta}}) = \hat{\sigma}_0^2 (\mathbf{X}'\mathbf{B}\mathbf{X})^{-1} (\mathbf{X}'\mathbf{V}^2\mathbf{X})(\mathbf{X}'\mathbf{B}\mathbf{X})^{-1}.$$

Standard errors for the S1S coefficients are then determined by the square root of the diagonal elements of this estimated covariance matrix.

§4.4 Case Study: Stackloss Data

In order to obtain the M1S and S1S estimates, LTS and MVE (without the 1-step improvement) are first estimated to provide a high breakdown starting point and robust weights for the improvement step. Using the repeated subsample (elemental set) algorithm with 50,000 iterations yields

$$\mathbf{MVE}_1(\mathbf{Z}) = \begin{bmatrix} 56.50 \\ 19.75 \\ 88.00 \end{bmatrix}$$

and

$$\mathbf{MVE}_2(\mathbf{Z}) = \begin{bmatrix} 99.6575 & 40.5274 & 31.8904 \\ 40.5274 & 38.2021 & 23.9178 \\ 31.8904 & 23.9178 & 82.3836 \end{bmatrix},$$

as the MVE estimates defining an ellipsoid having a (minimum) volume of 224.3446. Corresponding to an objective function evaluation of 3.1797, the LTS estimator produces the fitted equation

$$\hat{y}_i = -37.031 + 0.734\, x_{1i} + 0.438\, x_{2i} + 0.000\, x_{3i}.$$

Given these preliminary calculations, along with the LTS-based scale estimate of $\hat{\sigma}_0 = 1.0793$, the high breakdown regression estimators M1S and S1S both yield the equation

$$\hat{y}_i = -40.8148 + 1.0443\, x_{1i} + 0.6805\, x_{2i} - 0.2133\, x_{3i},$$

as shown in Table 4.1, along with their respective asymptotic standard errors.

Table 4.1: High breakdown regression for stackloss data.

Parameter	LTS	M1S	M1S s.e.	S1S	S1S s.e.
Intercept	-37.031	-40.815	5.169	-40.815	5.169
x_1	0.734	1.044	0.289	1.044	0.289
x_2	0.438	0.681	0.197	0.681	0.197
x_3	0.000	-0.213	0.132	-0.213	0.132

M1S and S1S essentially differ in their respective weighting schemes, particularly in cases with high influence points. Since the stackloss data has no high influence points, just the four outliers, it is not surprising that the two 1-step methods agree.

By viewing the standard errors for the M1S and S1S estimators, it is evident that acid concentration (x_3) is not significant in the presence of air flow (x_1) and temperature (x_2). One could extend the analysis further by viewing observation weights, leverage weights, plots, etc., but this is omitted for the discussion purposes here.

§4.5 Computational Issues for High Breakdown Regression

The main goal of high breakdown regression methods like those mentioned in this chapter is to keep large quantities (up to 50% of the data) of outliers from ruining the analysis. Thus, outliers in the response and high influence points are under control, not exerting any undue influence on the regression analysis. There is, however, a major problem that still needs to be addressed. The M1S and S1S methods require both an initial estimate and a set of weights in order to proceed with the one-step improvement. The initial estimate is taken to be LTS, with the weights being robust Mallows weights based on the MVE estimator. Both have problems attached to them. Recall, LTS is generally not the solution to its objective function, but merely an estimate of the estimate. As pointed out by Hettmansperger and Sheather (1992), LMS (and LTS for that matter) is highly sensitive to small changes in the middle of the regressor space. The process of repeated subsampling can easily result in drastically different final estimates for LTS. This results from the objective function in question having many local minima at vastly different locations.

In a similar fashion, the MVE estimator is also very unstable in the sense that drastically different results are common if the repeated subsampling algorithm were itself repeated. Thus, the robust weights generated from this algorithm may become very different in another simulation. Cook and Hawkins (1990) discuss this lack of repeatability problem when trying to mimic the results of Rousseeuw and van Zomeren (1990). Of note, for a data set having 20 observations and 5 regressors, it took the authors nearly 60,000 iterations to obtain the true MVE. This indicates that determining the number of iterations needed in subsampling by probabilistic arguments such as those given by Rousseeuw and Leroy (1987) and Rousseeuw (1993) may fail to find the proper estimates. The alternative would be to incorporate the FSA approach in obtaining the MVE estimates, which is much more stable in terms of the effects of random starts. This approach would definitely increase the computational time required to perform the regression since the initial regression estimator and the robust weights are no longer calculated in a parallel fashion.

Basing a method on LTS-based initial estimate leaves the method vulnerable to an internal breakdown. To obtain decent results will require an enormous amount of calculation, not the small number of iterations (say under 1000) as currently suggested (Rousseeuw and Leroy (1987), Rousseeuw (1993)). Even so, the researcher is not guaranteed to avoid misleading results. This extends to M1S and S1S methods as well. These methods are very reliant on their initial estimates, and very different results are a disturbing reality. Agostinelli and Markatou (1998) also offer a one-step robust regression estimator into the field.

In conclusion, it is stressed that a current high breakdown estimator such as LTS may not be reproducible. Two researchers analyzing the same data by the same regression procedure may obtain vastly different results. Case studies to come in Chapter 7 show a wide disparity of values over a small number of repeated analyses, while extending the discussion to include M1S and S1S, and their inherent reproducibility issues, as well. This issue is another reason as to why taking a different approach in obtaining a high breakdown regression estimator is justified.