Cluster-Based Bounded Influence Regression

David E. Lawrence
Pfizer, Inc

Jeffrey B. Birch, Yajuan Chen
Virginia Tech

Abstract

A regression methodology is introduced that obtains competitive, robust, efficient, high breakdown regression parameter estimates as well as providing an informative summary regarding possible multiple outlier structure. The proposed method blends a cluster analysis phase with a controlled bounded influence regression phase, thereby referred to as *cluster-based bounded influence regression*, or CBI. Representing the data space via a special set of anchor points, a collection of point-addition OLS regression estimators forms the basis of a metric used in defining the similarity between any two observations. Cluster analysis then yields a main cluster "half-set" of observations, with the remaining observations comprising one or more minor clusters. An initial regression estimator arises from the main cluster, with a group-additive DFFITS argument used to carefully activate the minor clusters through a bounded influence regression frame work. CBI achieves a 50% breakdown point, is regression equivariant, scale and affine equivariant and distributionally is asymptotically normal. Case studies and Monte Carlo results demonstrate the performance advantage of CBI over other popular robust regression procedures regarding coefficient stability, scale estimation and standard errors. The dendrogram of the clustering process and the weight plot are graphical displays available for multivariate outlier detection. Overall, the proposed methodology represents advancement in the field of robust regression, offering a distinct philosophical view point towards data analysis and the marriage of estimation with diagnostic summary.

1. Introduction

The detection of observations not conforming to a given statistical model is a common goal of the data analyst. Many methods have been proposed to aid in the detection of such nonconforming observations or "outliers". For example, in a recent paper by Fan *et al.*[1], a hierarchical clustering method was employed that greatly improves the ability of certain multivariate control chart techniques at detecting the presence of multivariate outliers. Detecting

unusual observations in the multiple regression setting is a far more complicated process however and many techniques have been introduced (see section 2) for this purpose. As in the Fan *et al.*[1] paper, the use of clustering methodology can improve the ability of a technique to identify unusual data points in the multiple regression setting. The use of clustering to improve the properties of the bounded-influence regression method is demonstrated in this paper.

To illustrate the difficulty at detecting unusual data point, the simple example below gives the comparison of the proposed method to several existing robust procedures when the data has more than one high leverage point or "hip". The data set has 11 observations with observations 1-8 generated from the linear model

$$y_i = 100 - 4x_i + \varepsilon_i$$

where $\varepsilon_i \sim N(\mu = 0, \sigma^2 = 25)$, and with the regressor variable generated via $x_i \sim U[10,20]$. Observations 9-11 were arbitrary added to reflect a mild influence point and two hips, respectively.

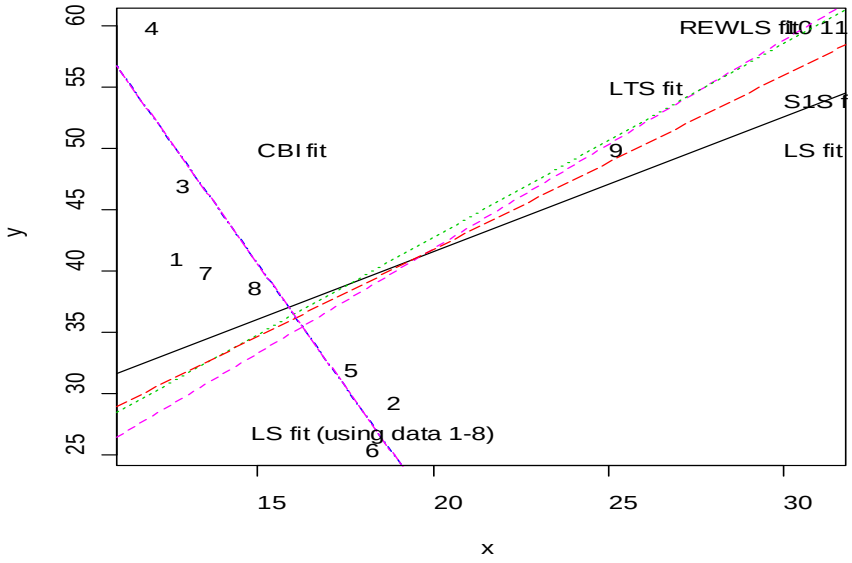

Figure 1.1: The fitted line of the different robust methods

The data are plotted in Figure 1.1 where the outlier (9) and the two hips (10, 11) are clearly seen. Regarding the collection of fits also displayed in Figure 1.1, only the proposed method (CBI) detects the correct trend of the uncontaminated data. Each of the other estimators was

dramatically misled by the joint influence of these three arbitrary points, resulting in a positive slope estimate when the true underlying slope is negative.

In building a linear regression model, a single unusual observation can dramatically influence ordinary least squares (OLS) estimation. With OLS, a single low leverage outlier can have a dramatic effect on the estimation of the general trend, especially concerning the intercept. However, a single high influence point, or hip, can have a dramatic effect on any or all parameter estimates. And, as illustrated in the example, the joint influence of several hips can have an even greater deleterious impact on parameter estimates. These coefficients and their standard errors, along with predictions, diagnostics, hypothesis tests, and other numerical measures can each become very misleading without a thorough exploratory data analysis accompanying it.

This research focuses on the study of robust, high breakdown linear regression modeling. As this discipline is extremely computationally intensive, much of the published work in this area has occurred since the early 1980's. Of course, some ideas were proposed much earlier, but generally limited in actual application. Methods such as M regression (Huber and Ronchetti[2]), and bounded influence (BI) (Huber and Ronchetti[2]) regression work well in the presence of low leverage outliers and at most one hip respectively. However, they are unable to combat a small percentage of outliers. Least median of squares (LMS) (Rousseeuw[3]) regression and least trimmed squares (LTS) (Ruppert and Carroll[4]) regression, on the other hand, are examples of high breakdown estimators as they possess the ability to provide parameter estimates with as much as 50% of the data being contaminated. Poor efficiency and numerical/computational sensitivity with large datasets has typically led to their primary use as an initial estimator feeding into other robust procedures such as M or BI estimators. Examples include Mallows 1-step (M1S) regression (Simpson *et al.*[5]) and Schweppe's 1-step (S1S) regression (Coakley and Hettmansperger[6]), which are one-step adjustments of LTS that increase efficiency versus the LTS estimator. However, two virtually identical LTS estimates may yield dramatically different M1S (or S1S) estimators (Lawrence[7]), thereby illustrating a potential negative issue with repeated sampling based methods. Another high breakdown one-step estimation method is due to Gervini and Yohai[8]. Their robust and efficient weighted least square estimate (REWLS) procedure attains full asymptotic efficiency with the assumption of normally distributed random

errors. However, according to the Monte Carlo study in Section 3.1, the REWLS, on the average, fails to correctly identify the good and bad high leverage points when the error term is not ideally normally distributed.

In this paper, CBI is introduced as a new regression methodology that obtains competitive, robust, efficient, high breakdown regression parameter estimates. Additionally, this method provides an informative summary regarding possible multiple outlier structure.

2. Review of Selected Robust Regression Methods

As the basis for linear regression analysis, the statistical model is restricted to be of the form

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon_i,$$

with the response variable, y_i, being explained as a linear function of the k regressor variables, $x_{ji}, j = 1,2 \ldots k,$ plus a random error component, ε_i, for each of the n observations, $i = 1,2 \ldots n$.

Given the computational nature of the proposed method, clarity in notation becomes quite important and, therefore, this paper offers sufficient detail. The linear model also can be written matrix form as

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \text{ , where } \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I}),$$

or elementwise as

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1k} \\ 1 & x_{21} & x_{22} & \ldots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

There are $p = k + 1$ unknown parameters that form the $p \times 1$ parameter vector $\boldsymbol{\beta}$, which is to be estimated by the $p \times 1$ vector $\widehat{\boldsymbol{\beta}}$**.** This subsequently yields the estimated fits as $\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$.

Further, the $n \times 1$ vector of residuals is computed as $\boldsymbol{r} = \boldsymbol{y} - \widehat{\boldsymbol{y}}$, with r_i representing the residual for the i^{th} observation. Also, define $\boldsymbol{Z}$ as the $n \times k$ matrix containing only the $k$regressor variables, with $\boldsymbol{Z}_y$ representing the $n \times p$ matrix formed by augmenting the vector $\boldsymbol{y}$ to $\boldsymbol{Z}$. To accommodate reference to individual observations, let the i^{th} row of $\boldsymbol{X}$ be denoted by

the $1 \times p$ row vector $\boldsymbol{x}_i^T$ and the $1 \times k$ row vector $\boldsymbol{z}_i^T$ denote the i^{th} row of $\boldsymbol{Z}$. When the response variable is included, the notation for i^{th} row of $\boldsymbol{Z}_y$ is $\boldsymbol{z}_{y,i}^T$.

Consider the objective function

$$\min_{\forall \boldsymbol{b}} \sum_{i=1}^{n} r_i^2,$$

for the OLS estimator, which may be written as

$$\min_{\forall \boldsymbol{b}} \sum_{i=1}^{n} \rho(r_i),$$

with $\rho(t) = t^2$. In robust regression, the function ρ can be selected to either down weight or bound any argument rising from unusual observations. This becomes the basis for M regression (Huber and Ronchetti[2]) which has the objective function

$$\min_{\forall \boldsymbol{b}} \sum_{i=1}^{n} \rho\left(\frac{y_i - \boldsymbol{x}_i^T \boldsymbol{b}}{\hat{\sigma}}\right),$$

Where the ρ-function is chosen to be bounded and odd-symmetric, $\boldsymbol{b}$ represents an arbitrary point in the p-dimension estimation space, and where $\hat{\sigma}$ is some appropriately chosen estimate of σ. The choice for $\hat{\sigma}$ is generally limited to robust measures of scale. One such estimator that is frequently used is the median absolute deviation (MAD), where

$$MAD = 1.4826 \operatorname*{med}_{\forall i} \left| r_i - \operatorname*{med}_{\forall i} r_i \right|.$$

Taking derivatives with respect to $\boldsymbol{b}$ leads to solving p "altered normal equations",

$$\sum_{i=1}^{n} \psi\left(\frac{y_i - \boldsymbol{x}_i^T \boldsymbol{b}}{\hat{\sigma}}\right) \boldsymbol{x}_i = 0,$$

where $\psi(t) = \frac{d\rho(t)}{dt}$ and $\widehat{\boldsymbol{\beta}}_M$ is the solution for $\boldsymbol{b}$. These altered normal equations form a system of nonlinear equations that may be solved by a number of popular numerical methods including (1) Newton-Raphson and (2) iteratively reweighted least squares (IRLS), the later used in this paper. At convergence, IRLS produces the M regression parameter estimator

$$\widehat{\boldsymbol{\beta}}_M = (\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{y},$$

where $\boldsymbol{W}$ is the $n \times n$ diagonal "weight matrix", with diagonal elements denoted as w_i. Each weight, w_i, determines how much emphasis the regression will place on a particular observation.

A large weight (near 1) should indicate a good observation. An outlier or a hip, on the other hand, should get a reduced weight or perhaps even a zero weight. In M regression the i^{th} weight is calculated as $w_i = \frac{\psi(r_i/\hat{\sigma})}{r_i/\hat{\sigma}}$, a function of the i^{th} residual. Typically, the larger is the residual, the smaller is the weight.

A single hip will "pull" the fitted M regression line toward it to make the corresponding residual small, thus that weight will be large. This means that M regression can be dominated by a single hip. One solution to this problem is to use bounded influence (BI) regression. Here, the name refers to "bounding" the influence that the point $\boldsymbol{x}_i^T$ has in the regressor-space. One altered normal equation form, called the Schweppe form (Staudte[9]), is written as

$$\sum_{i=1}^{n} u(\boldsymbol{x}_i)\psi\left(\frac{y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}}{\hat{\sigma}u(\boldsymbol{x}_i)}\right)\boldsymbol{x}_i = 0.$$

Here, $u(\boldsymbol{x}_i)$ is chosen so that the effect of a large $\boldsymbol{x}_i^T$ is reduced if $(y_i, \boldsymbol{x}_i^T)$ is a hip. One choice is to have $u(\boldsymbol{x}_i) = \pi_i = \frac{1-h_{ii}}{\sqrt{h_{ii}}}$, where h_{ii} is the i^{th} diagonal element of the so-called hat matrix, $\boldsymbol{H}$, with $\boldsymbol{H} = \boldsymbol{X}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}$. The π_i value is referred to as the BI weight. The BI regression estimator can be obtained in exactly the same manner as the M-estimator via IRLS, as

$$\hat{\boldsymbol{\beta}}_{BI} = (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{y}.$$

However, the i^{th} weight now has the form $w_i = \psi\left(\frac{r_i^*}{\pi_i}\right)/\frac{r_i^*}{\pi_i}$, where r_i^* is the scaled residual $r_i/\hat{\sigma}$. Specifically, the BI weight depends on both the residual and the location of $\boldsymbol{x}_i^T$ in the regressor-space. More details involving the choice of ψ function may be found in Appendix A-1.

While M and BI estimators provide an improvement over OLS if the data has an outlier or hip, respectively, they cannot provide protection against data with even modest amounts of contamination. Ruppert and Carroll[4] introduced LTS to combat this situation, defining the objective function as

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

representing the sum of the h smallest squared residuals where h is generally taken to be $[(n + p + 1)/2]$, with [.] denoting the greatest integer function. Since this objective function is

not differentiable, no closed-form expression exists for the LTS estimator. However, algorithms are available that give the exact LTS estimator for the location model, the exact LTS estimator for the regression model based on small data sets, and a relatively accurate LTS estimator for large data sets. The algorithmic details may be found in Rousseeuw and Van Driessen[10]. Historically, methods like LTS (and its predecessor LMS) had involved repeated sampling computational methods incorporating probabilistic arguments.

One problem with high breakdown estimators such as LTS is poor efficiency due to large variability associated with estimated coefficients. The remedy for this poor efficiency is to use the LTS estimator, or another high breakdown estimator, as an initial estimator $\widehat{\boldsymbol{\beta}}_0$, with the generalized M estimator form to obtain a one-step generalized M estimator. The S1S estimator is one such estimator and results from solving the "altered" normal equations

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

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

$$\widehat{\boldsymbol{\beta}}_{S1S} = \widehat{\boldsymbol{\beta}}_0 + (\boldsymbol{X^T B X})^{-1}\boldsymbol{X^T W}\psi(\underline{\boldsymbol{r}})\hat{\sigma}_0.$$

Further computational details regarding S1S are given in Appendix A-2.

None of the above estimators achieve full efficiency at the normal distribution while simultaneously maintaining a breakdown bound close to 50%. Gervini and Yohai[8] proposed an adaptive one-step estimation method that attains full asymptotic efficiency at the normal error distribution while at the same time has a high breakdown bound and small maximum bias. Their method, referred to as the REWLS estimator, is a weighted LS estimator computed from an initial high breakdown estimate $\widehat{\boldsymbol{\beta}}_0$, and a robust scale estimate $\hat{\sigma}_0$ such as MAD. However, rather than deleting those observations whose absolute scaled residuals are greater than a given value, the procedure will keep a number N of observations, corresponding to the smallest values of the absolute scaled residual $rs_i = \frac{|r_i(\widehat{\boldsymbol{\beta}}_0)|}{\hat{\sigma}_0}, i = 1, \ldots n$. The N has the property that in large samples under normality it will have $N/n \to 1$, which means a vanishing fraction of observations will be deleted and full efficiency will be attained (Maronna *et al.*[11]). Computational details regarding the choice of N may be found in Appendix A-3.

The REWLS estimator can be obtained as

$$\widehat{\boldsymbol{\beta}}_{REWLS} = \begin{cases} \widehat{\boldsymbol{\beta}}_0 + (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{y} & if\ \hat{\sigma}_0 > 0 \\ \widehat{\boldsymbol{\beta}}_0 & if\ \hat{\sigma}_0 = 0 \end{cases},$$

where W is the diagonal matrix with

$$w_i = \begin{cases} 1 & if\ rs_i \leq rs_N \\ 0 & otherwise \end{cases} .$$

3. Proposed method

The proposed regression methodology offers a new philosophical approach to the robust regression arena and consists of two primary phases, the cluster phase and the regression phase. First, an initial high-breakdown regression estimator is produced via a sophisticated clustering algorithm. Second, refinement of this initial regression estimator is investigated and possibly implemented under a carefully structured use of BI regression. The rationale behind this second phase is to allow for a possible improvement in efficiency, especially when the level of data contamination does not come close to approaching 50%. The proposed method has been named cluster based bounded influence regression, or CBI for short, to reflect the nature of its two phases computation process.

The cluster phase begins with high-breakdown location and scale estimation of the p dimensional regressor-response space. A special set of points, referred to as the set of anchor points, is computed that together represent the general trend of the data. Each observation is then characterized by the OLS regression fit that would occur if this individual observation is augmented to the anchor points. High breakdown location and scale estimation of this set of n OLS coefficients provides the foundation for the construction of the similarity matrix (technically, a distance matrix). The desire for a tight, compact sphere of similar coefficients exhibiting a common trend description is the basis for the selection of complete linkage hierarchical clustering (Lawrence[7]) as the default method and clustering is performed until an initial main cluster of at least $[(n + p + 1)/2]$ observations are formed. Two aspects worth mentioning are that (1) the OLS sensitivity to a single point is being exploited to our advantage in evaluating the data, and (2) the anchor points serve to alleviate repeated sampling (as required

by other 50% breakdown point estimators such as LTS) and the use of minimal sized elemental subsets that must be in general position (i.e. no singularity issues).

A simple OLS fit to this main cluster is used as the basis for the possible adjustment of the anchor set metric to more directly relate to the general trend. A revised similarity matrix is constructed, with a second cluster analysis yielding a revised, final main cluster and g minor clusters. The determination of this cluster classification structure completes the cluster phase.

To begin the regression phase, the initial CBI estimator is simply the OLS estimate of the main cluster observations. A high breakdown scale estimate is then computed. High breakdown BI leverage weights are computed from the regressor-space only. Using only the main cluster, a BI regression updates the initial CBI estimator. To this point, the minor clusters have not been utilized in the computation of the CBI regression estimator and their observations are said to be inactive. The activation process for these remaining observations has two primary stages. First, a $DFFITS_{+I}^2$ statistic is computed for each of the minor clusters, where $I = 1,2, \dots, g$. A candidate minor cluster is one such that $DFFITS_{+I}^2 < \delta$ for the cutoff value δ. Then, a single $DFFITS_{+J}^2$ statistic, denoted by J, is computed for the union of all candidate minor clusters. If $DFFITS_{+J}^2$ is "small enough", then the final CBI estimator is determined from this activation process (provided at least one minor cluster observation obtained a nonzero weight). Otherwise, the minor clusters do not play an active role (i.e. all observations possess a zero weight) and there is no further update to the current CBI regression estimator. A final CBI scale estimate is computed once the final CBI regression estimator has been determined.

The detailed algorithm consisting of ten interrelated steps for the CBI estimator is presented below. Steps 1 through 3 represent the cluster phase and steps 4 through 10 represent the regression phase. Notation is introduced as needed.

Step1

Perform minimum volume ellipsoid, MVE, estimation (see Rousseeuw and Leroy[12]) of $\boldsymbol{Z_y}$; determine the $(2p+1) \times p$ anchor point matrix, $\boldsymbol{\Omega}$. These points include $\boldsymbol{MVE}_1(\boldsymbol{Z}_y)$, the MVE location vector for $\boldsymbol{Z_y}$, and the end points of the ellipsoid of

constant distance $\chi^2_{0.975,p}$ from $\boldsymbol{MVE}_1(\boldsymbol{Z}_y)$ based on the $\boldsymbol{MVE}_2(\boldsymbol{Z}_y)$ metric, the MVE scale matrix estimator for $\boldsymbol{Z_y}$, the pair of end points is determinate by the expression $\boldsymbol{MVE}_1(\boldsymbol{Z}_y) \pm \sqrt{\lambda_i \chi^2_{0.975,p}}\, \boldsymbol{e}_i$, where λ_i and $\boldsymbol{e}_i$ is the i^{th} eigenvalue and eigenvector of $\boldsymbol{MVE}_2(\boldsymbol{Z}_y)$, respectively.

Step 2

Determine the $n \times p$ base regression estimator matrix $\boldsymbol{B}$. The i^{th} row of $\boldsymbol{B}$, denoted by the $1 \times p$ vector $\boldsymbol{b}_i$, is defined as the estimator that results from an OLS regression analysis of the set of anchor points supplemented by the addition of the i^{th} observation in the dataset. Perform an MVE estimation of $\boldsymbol{B}$, treating each row of $\boldsymbol{B}$ as an observation in p dimensions.

Step 3

Using $\boldsymbol{MVE}_2(\boldsymbol{B})$ as the distance metric, compute a $n \times n$ similarity matrix $\boldsymbol{S}$ whose elements are defined to be

$$s_{ij} = \left(\boldsymbol{b}_i - \boldsymbol{b}_j\right)^T (\boldsymbol{MVE}_2(\boldsymbol{B}))^{-1} \left(\boldsymbol{b}_i - \boldsymbol{b}_j\right).$$

Perform a cluster analysis on the dataset given the similarity matrix S and using complete linkage to obtain the tightest cluster of $\boldsymbol{b}_i$ vectors. The initial main cluster, C_0, is defined at the first instance of which a single cluster consists of at least $h = [(n + p + 1)/2]$ observations. The remaining observations fall into one of g minor clusters that are labeled as $C_1, C_2, \ldots, C_g$.

Step 4

Compute the OLS estimate $\widehat{\boldsymbol{\beta}}_0$ using the data points in C_0. A preliminary estimate of scale, $\hat{\sigma}_0$, is defined to be the MAD of all n residuals $\boldsymbol{r}\left(\widehat{\boldsymbol{\beta}}_0\right)$ where

$$r_i\left(\widehat{\boldsymbol{\beta}}_0\right) = y_i - \boldsymbol{x}_i{}^T \widehat{\boldsymbol{\beta}}_{\boldsymbol{0}}.$$

Determine the set of observations, H, such that

$$H = \left\{i: \left|r_i\left(\widehat{\boldsymbol{\beta}}_0\right)\right| \leq \hat{\sigma}_0 4.685 \sqrt{2pn}/(n - 2p)\right\}.$$

Step 5

Using the data points in H, compute the $p \times 1$ mean vector $\boldsymbol{m}_H(\boldsymbol{Z})$, of the regressor data in H,and $p \times p$ covariance matrix $\boldsymbol{V}_H(\boldsymbol{Z})$, using standard moments estimators, of the

regressor data in H, define the $p \times 1$ robust regressor distance vector $\boldsymbol{d}$ containing the p elements

$$d_i = (\boldsymbol{z}_i - \boldsymbol{m}_H(\boldsymbol{Z}))^T (\boldsymbol{V}_H(\boldsymbol{Z}))^{-1} (\boldsymbol{z}_i - \boldsymbol{m}_H(\boldsymbol{Z})).$$

Step 6

Mimic step 1 to step 3 by replacing the MVE statistics with the weighted mean and covariance estimates for the data to get the new initial main cluster, C_0 , and g minor clusters C_1, C_2 ..C_g.The weight for the i^{th} data point is define as

$$w_i = \begin{cases} 1, & i \in H \\ 0, & i \notin H \end{cases},$$

Compute the initial CBI estimator, $\widehat{\boldsymbol{\beta}}_1$, using WLS and subsequently updated the scale estimate $\hat{\sigma}_1$ as MAD of all n new residuals.

Step 7

Determine the $h \times 1$ BI leverage weight vector, $\boldsymbol{\pi}$, whose elements are defined as

$$\pi_i = \begin{cases} 1, & i \in C_0 \\ \min(1, \dfrac{\chi^2_{0.975,p-1}}{d_i}), & i \notin C_0 \end{cases}$$

Perform BI regression using only the main cluster, C_0, to obtain, at convergence of IRLS, the estimate $\widehat{\boldsymbol{\beta}}_2$.

Step 8

Let I represent any minor cluster and m_I be the size of I, and let $\boldsymbol{\pi}_{(C_0,C_I)}$ be the sub-vector set of $\boldsymbol{\pi}$ that corresponds only to the C_0 and C_I observations. Perform the BI regression with these new data points and leverage weight vector $\boldsymbol{\pi}_{(C_0,C_I)}$to obtain the estimate $\widehat{\boldsymbol{\beta}}_{+I}$ at convergence. A $DFFITS^2_{+I}$ statistic is then computed via

$$DFFITS^2_{+I} = \frac{\sum_{i=1}^{n} \left(\hat{y}_{i,+I}(\widehat{\boldsymbol{\beta}}_{+I}) - \hat{y}_i(\widehat{\boldsymbol{\beta}}_2) \right)^2}{m_I \hat{\sigma}_1^2},$$

where $\hat{y}_{i,+I}(\widehat{\boldsymbol{\beta}}_{+I})$ represent fits when using both C_0 and C_I observations and $\hat{y}_i(\widehat{\boldsymbol{\beta}}_2)$ represents fits when using just C_0 observations. This statistic is computed for each of the g minor clusters.

Step 9

Define the scalar δ to represent the maximum allowable $DFFITS_{+I}^2$ statistic. Then, let J represent the union of all activation candidate minor sets, i.e.

$$J = \bigcup_{\forall I} C_I \,|(DFFITS_{+I}^2 \leq \delta \text{ and } \exists_{\mathrm{i}\in\mathrm{I}}|\mathrm{w_i} > 0).$$

Provided that $J \neq \emptyset$, then with $\widehat{\boldsymbol{\beta}}_2$, $\hat{\sigma}_1^2$ and $\boldsymbol{\pi}_{(C_0,C_J)}$ as inputs to obtain the BI regression estimate $\widehat{\boldsymbol{\beta}}_{+J}$ and $DFFITS_{+J}^2$. The default value of δ is 4.

Step 10

$$\widehat{\boldsymbol{\beta}}_{CBI} = \begin{cases} \widehat{\boldsymbol{\beta}}_{+J}, & if\left(DFFITS_{+J}^2 \leq \delta \text{ and } \exists_{\mathrm{j}\in\mathrm{J}}\left|\mathrm{w_j} > 0\right)\right|J \neq \emptyset \\ \widehat{\boldsymbol{\beta}}_2, & otherwise \end{cases}$$

The CBI scalar estimator is then updated as the MAD of new residuals. The final CBI weights for the individual observations are simply the observations weights at convergence of BI regression used to compute $\widehat{\boldsymbol{\beta}}_{CBI}$.

Three scale estimators are provided by the CBI procedure, specifically $\hat{\sigma}^2{}_{CBI}$, $\hat{v}^2{}_{CBI}$ and $\hat{v}_w{}^2{}_{CBI}$. $\hat{\sigma}_{CBI}$ is the MAD of the CBI residuals. Given the CBI scale estimate $\hat{\sigma}_{CBI}$, the BI leverage weight vector $\boldsymbol{\pi}$, and $\widehat{\boldsymbol{\beta}}_{CBI}$, a robust mean square error that mimics the robust ANOVA scale estimate introduced by Birch[13] is found via

$$\hat{v}^2{}_{CBI} = \frac{\frac{n^2}{n-p}\hat{\sigma}^2{}_{CBI}\sum_{i=1}^{n}\psi^2\left(\frac{r_i(\widehat{\boldsymbol{\beta}}_{\mathrm{CBI}})}{\pi_i\hat{\sigma}_{CBI}}\right)}{\sum_{i=1}^{n}\psi'\left(\frac{r_i(\widehat{\boldsymbol{\beta}}_{\mathrm{CBI}})}{\pi_i\hat{\sigma}_{CBI}}\right)}.$$

Using the effective sample size, $n_w = \sum_{i=1}^{n} w_i$ (Birch[14]), a modified version of the robust analysis of variance scale estimate then becomes $\hat{v}_w{}^2$

$$\hat{v}_w{}^2{}_{CBI} = \frac{\frac{n_w{}^2}{n_w-p}\hat{\sigma}^2{}_{CBI}\sum_{i=1}^{n}\psi^2\left(\frac{r_i(\widehat{\boldsymbol{\beta}}_{\mathrm{CBI}})}{\pi_i\hat{\sigma}_{CBI}}\right)}{\sum_{i=1}^{n}\psi'\left(\frac{r_i(\widehat{\boldsymbol{\beta}}_{\mathrm{CBI}})}{\pi_i\hat{\sigma}_{CBI}}\right)}.$$

Once the CBI estimate is obtained, the BI based analysis of variance methods of Birch[13] and Birch and Agard[15] can be used to perform inference on any single parameter or any subset of parameters.

Many theoretical properties of the CBI estimator have been studied and proved by Lawrence[7]. For example, it has been demonstrated that the CBI regression estimator belongs to the family of high breakdown regression estimators; with a breakdown point approaching 50% as $n \to \infty$. It was further showed that the CBI estimator is asymptotically normally distributed. That is,

$$\sqrt{n}\left((\widehat{\boldsymbol{\beta}}_{CBI} - \boldsymbol{\beta}\right) \xrightarrow{Law} N[\mathbf{0}, \boldsymbol{M}^{-1}\boldsymbol{Q}\boldsymbol{M}^{-1}],$$

where the $\boldsymbol{M}$ and $\boldsymbol{Q}$ is defined as

$$\boldsymbol{M} = E_F\left[\left(w + \frac{d\omega(\boldsymbol{x}, r)}{dr} r\right)\boldsymbol{x}\boldsymbol{x}^T\right],$$

$$\boldsymbol{Q} = E_F[\omega^2(\boldsymbol{x}, r) r^2 \boldsymbol{x}\boldsymbol{x}^T],$$

$$\mathrm{w} = \omega(\boldsymbol{x}, r).$$

The function $\omega(\boldsymbol{x}, r)$, the weight function is nonnegative, bounded and measurable in $(\boldsymbol{x}, r)$. The CBI regression estimator has also been shown to achieve regression equivariance, scale equivariance and affine equivariance properties (see Rousseeuw and Leroy[12] for definitions of these equivariance properties). These equivariance properties also impact the following Monte Carlo simulation study by the fact that the values defined for the regression coefficients and scale will not impact the final Monte Carlo results; i.e., these values are themselves arbitrary and meaningless. Overall, the theoretical foundation for the CBI methodology strongly supports its inclusion in the class of high breakdown regression estimators.

Reflection on the development of the CBI algorithm yields an interesting and diverse discussion onto itself. Motivation initially stemmed from an interest in how iteration breaks down M and BI estimators and a curiosity about joint influence diagnostics in general. The joint influence aspect itself led to the inclusion of some sort of clustering mechanism to identify these various subgroups of problematic observations. Many forms of the initial similarity matrix construct were considered, including one based on the altered hat matrix. Further, initial strategies were more spatially oriented and were utilizing single-linkage clustering to take advantage of the chaining property that is often considered a detrimental property of the method but could track a regression trend under this alternative use. In fact, such a CBI version was proposed early in its development (Lawrence[7]).

A major breakthrough in the development of the CBI algorithm occurred with the introduction of the anchor set. Ironically, this thought arose during development of a closed-form computation method for a multivariate C_p statistic in a completely different research area. However, it was clear that this anchor set could alleviate the random subsampling with elemental sets issues (faced with the leading high-breakdown estimators) as it was large enough to fit the regression model without any singularity issues. Further, it had a direct implementation into the clustering framework. The exploitation of the OLS breakdown property would form the basis of this new paradigm. Common regression estimates would indicate common trends (either general trend or common deviant trend that would reflect joint influence) and, very importantly, there is no spatial requirement directly involved. Joint influence can involve observations scattered across the response-regressor space. As a direct consequence, clustering moved from single-linkage to complete-linkage to more appropriately capture what are effectively similar regression estimates.

Iteration has both beneficial and detrimental aspects, so the CBI algorithm had to be robust to such negative effects. Earlier versions of CBI allowed for minor clusters to be added sequentially. From the research, it was deemed more prudent to assess them individually, then together, to avoid estimator drift due to iteration as well as to further bolster the robustness versus joint influence of several minor clusters.

Overall, while the technical and computational details of the CBI algorithm have evolved during the development process, the general philosophy and intent have remained steadfast. The goal was to take an efficient low-breakdown point method, BI regression, and improve the breakdown point while not making a huge sacrifice regarding efficiency. A more thorough discussion of the motivation of each step of the CBI algorithm may be found in chapter 5 of Lawrence[7].

3.1 Case Studies and Comparison

Two well-known datasets are used to illustrate and compare the CBI method to several other robust techniques, (1) the Pendleton and Hocking[16] (PH) data, and (2) the Hawkins *et al.*[17] (HKB) data. The PH dataset has three regressors $\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3$ and $n = 26$ observations. The parameters to be estimated are $\boldsymbol{\beta}^T = (20 \;\; 3 \;\; -2 \;\; 0)$. Three low-leverage outliers were artificially created and inserted as observations 11, 17 and 18. One hip was inserted as observation 24.

The CBI cluster phase of the PH data resulted in a main cluster of 19 observations (four more than h =15) and five minor clusters. A summary of the entire CBI regression analysis is provided as Table 3.1 and Figure 3.1.

Table 3.1: Summary of the CBI regression analysis of the PH dataset

<table>
<tr><td colspan="5">Cluster History</td></tr>
<tr><td>Step</td><td colspan="2">Clusters</td><td colspan="2">n=26</td></tr>
<tr><td>Initial</td><td colspan="2">$C_0 = \{2\ ,5\ ,7\ ,8\ ,9\ ,12:16\ ,19:23\}$</td><td colspan="2">h=15</td></tr>
<tr><td rowspan="5">Final</td><td colspan="2" rowspan="5">$C_0 = \{1::5\ ,7\ ,8,10,12:16,19,21:23,25\ ,26\}$
$C_1 = \{6\ ,9\ ,20\} C_2 = \{11\}$
$C_3 = \{17\} C_4 = \{18\}$
$C_5 = \{24\}$</td><td colspan="2">Initial OLS :</td></tr>
<tr><td>intercept</td><td>26.987</td></tr>
<tr><td>X_1</td><td>2.601</td></tr>
<tr><td>X_2</td><td>-2.108</td></tr>
<tr><td>X_3</td><td>-0.173</td></tr>
<tr><td>Minor Sets</td><td colspan="2">$\text{DFFITS}_{+I}{}^2$</td><td colspan="2">Activate</td></tr>
<tr><td>C_1</td><td colspan="2">1.9153</td><td colspan="2">YES</td></tr>
<tr><td>C_2</td><td colspan="2">0</td><td colspan="2">NO</td></tr>
<tr><td>C_3</td><td colspan="2">0</td><td colspan="2">NO</td></tr>
<tr><td>C_4</td><td colspan="2">0</td><td colspan="2">NO</td></tr>
<tr><td>C_5</td><td colspan="2">0</td><td colspan="2">NO</td></tr>
<tr><td>Candidate J</td><td colspan="2">$\text{DFFITS}_{+J}{}^2$</td><td colspan="2">Activate</td></tr>
<tr><td>C_1</td><td colspan="2">1.9153</td><td colspan="2">YES</td></tr>
<tr><td colspan="5">Parameter Estimate</td></tr>
<tr><td>Parameter</td><td>Estimate</td><td>Sd.Error</td><td>t</td><td>P -value</td></tr>
<tr><td>intercept</td><td>25.615</td><td>13.677</td><td>1.873</td><td>0.038</td></tr>
<tr><td>X_1</td><td>2.719</td><td>0.695</td><td>3.909</td><td>0.000</td></tr>
<tr><td>X_2</td><td>-2.136</td><td>0.321</td><td>-6.638</td><td>0.000</td></tr>
<tr><td>X_3</td><td>-0.194</td><td>0.441</td><td>-0.440</td><td>0.331</td></tr>
<tr><td>Scale</td><td colspan="4">$\hat{\sigma}_{CBI} = 0.516 \quad \hat{v} = 0.306 \quad \hat{v}_w = 0.254$</td></tr>
</table>

The final CBI fitted equation is

$$\hat{y}_i = 25.615 + 2.719x_{1i} - 2.136\ x_{2i} - 0.194x_{3i}.$$

It is clear (p-value = 0.331) that $\boldsymbol{x}_3$ is not significant in the presence of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, a correct decision for this case study. The intercept, $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are each statistically significant (p-values of 0.038, 0.000and 0.000, respectively) terms, as they should be. According to the CBI weight plot in Figure 3.1, four observations received zero weight, these being the three outliers and the one hip.

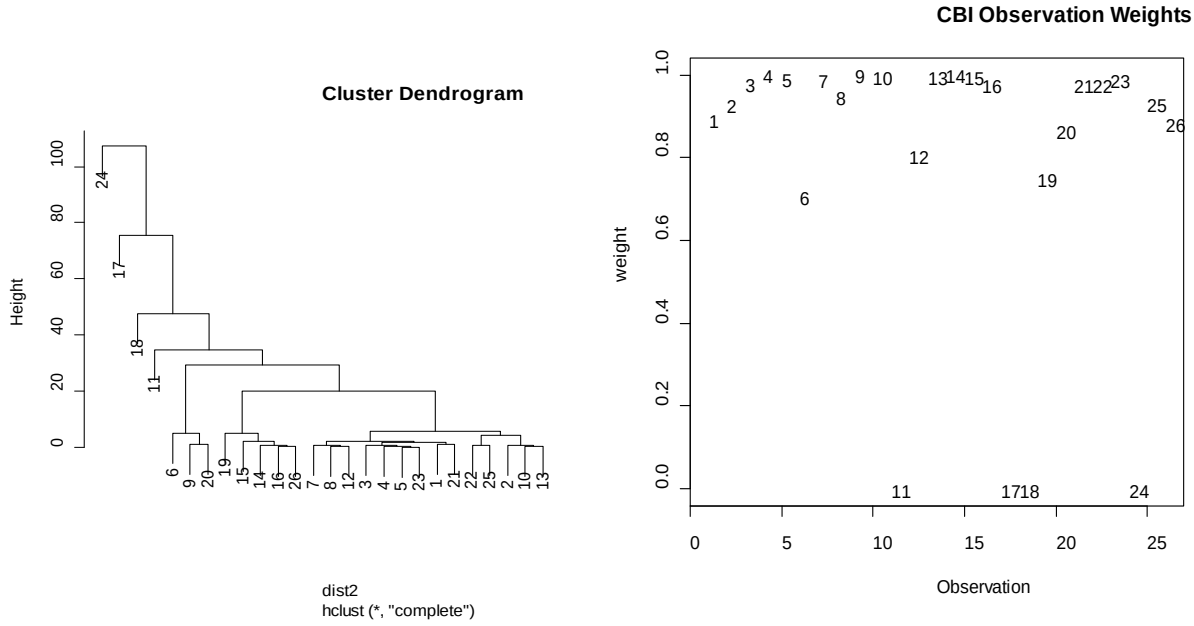

Figure 3.1: Cluster dendrogram and final observation weights of PH dataset

Other competing regression methods are applied to the PH dataset and the corresponding estimates are given by Table 3.2.

Table 3.2: Robust analysis of parameter estimate summary of PH dataset

Parameter	**OLS**	**LTS**	**S1S**	**REWLS**	**BI**	**CBI**	**OLS without outliers**
Intercept	8.205	10.961	40.96	8.931	17.954	25.615	24.270
X_1	3.560	3.384	1.974	3.523	3.120	2.719	2.791
X_2	-1.640	-1.712	-2.538	-1.697	-1.971	-2.136	-2.112
X_3	0.334	0.483	-0.781	0.4337	0.052	-0.196	-0.156

The estimated coefficients resulting from the different estimation methods described in Section 1 for the PH data reveal some interesting results, especially as they relate to the CBI algorithm. First, it is seen that the BI estimator has coefficient estimates very close to the true parameter vector. The CBI estimator began with estimates based on the final main cluster and then improved upon them through the minor cluster activation process. It is interesting to note that the estimated coefficients using the OLS method for the 22 good observations is nearly identical to those obtained by the CBI method. Thus, the CBI estimator is actually closer to the observed trend of the data than is the BI estimator.

We note that the PH data had no troublesome jointly influential observations. Consider next the HBK data which has a cluster of ten hips (as observations 1 through 10) and another cluster of four good high leverage points (observations 11 through 14). Since the true parameters were not reported by Hawkins *et al.*[17], the goal in analyzing this dataset was to ascertain the ability of the robust methods to distinguish between the outliers and the non-outliers occurring at the high leverage points.

The CBI method applied to the HBK data resulted in a weight for each observation (Figure 3.2). Figure 3.2 shows that the first ten observations received zero weight, the ideal case. The four good leverage observations, on the other hand, all have weights greater than zero, as they should be; especially the observations 11, 12 and 14 received very high weights each close to 1.

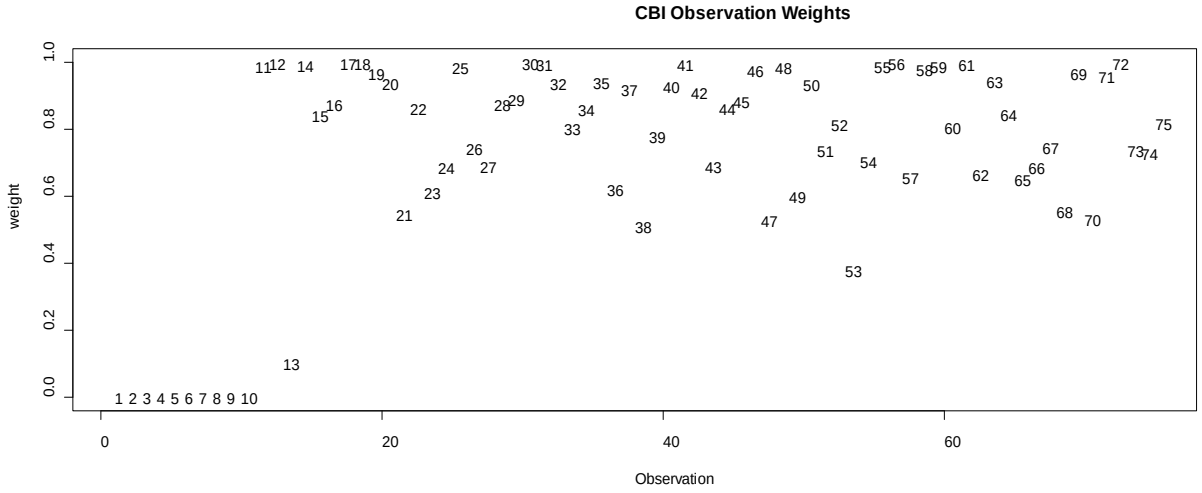

Figure 3.2: The final CBI regression observation weights of HBK dataset

The final CBI fitted equation is

$$\hat{y}_i = -0.224 + 0.097\ x_{1i} + 0.045x_{2i} - 0.064x_{3i}.$$

A summary of the CBI regression analysis is provided as Table 3.3.

Table 3.3: CBI analysis of parameter estimate summaryof HBK dataset

Parameter Estimate				
Parameter	Estimate	Sd.Error	t	P -value
intercept	-0.224	0.169	-1.326	0.190
X_1	0.097	0.107	0.901	0.371
X_2	0.045	0.061	0.736	0.464
X_3	-0.064	0.055	-1.165	0.249
Scale	$\hat{\sigma}_{CBI} = 0.867 \quad \hat{v} = 0.890\ \hat{v}_w = 0.646$			

A comparison of the CBI results to other competing regression methods is given in Table 3.4. It is seen that the REWLS estimate provide the same result as the OLS estimate without hips, this result is not surprising because the REWLS method took advantage of the fact that the hips in this case have larger residuals as determined by its initial LTS estimate. The CBI estimates, while not identical, are very close to the OLS estimates based on the good data points.

Table 3.4: Robust analysis of parameter estimate summary of HBK dataset

Parameter	OLS	LTS	S1S	REWLS	BI	CBI	OLS without hips
Intercept	-0.388	-0.612	-0.004	-0.180	-0.934	-0.224	-0.180
X_1	0.239	0.255	0.041	0.081	0.144	0.097	0.081
X_2	-0.335	0.048	0.021	0.039	0.192	0.045	0.039
X_3	0.383	-0.106	-0.082	-0.051	0.184	-0.064	-0.051

The CBI estimates are close to the trend of the data for both case studies and the weight plots also show that it can correctly identify the outliers and hips for the case studies considered here. Results from a small Monte Carlo study are presented in the next section to further evaluate the ability if the competing regression methods to detect multiple outliers, especially those occurring at high leverage points.

4. Monte Carlo Study

In this Monte Carlo study, the simulated dataset utilized the original regressor values of the HBK dataset, but generated a new response vector while maintaining observations 1 through 10 as a high influence cluster. Specifically, the $n = 75$ observations were generated by the linear model

$$y_i = \begin{cases} \varepsilon_i, & i \in (1:10) \\ 0.2 - 0.15x_{1i} + 0.1x_{3i} + \varepsilon_i, & i \notin (1:10) \end{cases}.$$

With the random errors generated from the following distributions

$$\varepsilon_i \sim \begin{cases} N(\mu = 10, \sigma^2 = 0.385^2), & i \in (1:10) \\ N(\mu = 0, \sigma^2 = 0.5^2), & i \notin (1:10) \end{cases}.$$

The results of this Monte Carlo study are provided in Table 4.1. Here, the parameters to be estimated are $\boldsymbol{\beta}^T = (0.2 - 0.15 \;\; 0.1)$ and $\sigma_\varepsilon{}^2 = 0.25$. The number of Monte Carlo repetitions was 2000.

Table 4.1: Simulation results for Monte Carlo study (The crossed cells are not applicable)

	OLS	**LTS**	**S1S**	**REWLS**	**BI**	**CBI**	
$\widehat{\mathrm{E}}[\widehat{\boldsymbol{\beta}}]$	0.029	−0.104	0.209	−0.105	−0.424	0.218	
	−0.019	−0.122	−0.156	−0.125	−0.090	−0.147	
	−0.307	0.069	−0.005	0.063	0.119	−0.006	
	0.456	0.188	0.105	0.189	0.295	0.093	
$\widehat{E}[\hat{\sigma}^2]$	3.478	0.363		0.312	0.358	0.329	$\widehat{E}[\hat{v}_w{}^2]$
$\widehat{E}[\hat{v}^2]$					0.016	0.469	0.276
$\widehat{\mathrm{E}}[\mathrm{se}[\widehat{\boldsymbol{\beta}}]]$	0.345		0.211	0.114	0.024	0.177	0.141
	0.217		0.075	0.069	0.015	0.095	0.076
	0.128		0.073	0.053	0.013	0.073	0.057
	0.107		0.074	0.042	0.009	0.065	0.051
$\widehat{\boldsymbol{\beta}}$	0.689	2.333	2.316	1.848	0.619	1.903	
	−0.310	−1.168	−0.570	−1.077	−0.735	−0.759	
	0.379	1.165	1.746	0.7	−0.115	1.144	
	0.127	0.665	0.298	0.630	0.080	0.189	
Range	0.401	0.977	1.066	0.660	0.354	0.642	
	−0.223	−0.597	−0.693	−0.372	−0.251	−0.466	
	0.179	0.329	0.373	0.373	0.102	0.176	
	0.078	0.192	0.094	0.111	0.080	0.090	
Min	0.235	0.901	0.869	0.556	0.264	0.637	
	−0.423	−0.400	−0.473	−0.137	0.141	−0.307	
	−0.188	0.501	0.396	0.419	0.405	0.329	
	0.046	0.183	0.102	0.125	0.048	0.083	
Max	0.202	1.019	0.945	0.504	0.265	0.543	
	0.364	−0.313	−0.549	−0.029	−0.405	−0.198	
	0.566	0.706	0.396	0.475	0.140	0.345	
IQR	0.039	0.193	0.113	0.193	0.005	0.073	

According to the characteristics of the estimators in Table 4.1, it is seen that CBI estimator had overall better performance. For example, consider $\widehat{E}[\widehat{\boldsymbol{\beta}}]$, the simulated expected coefficient

vector for each estimation method. We see that S1S and CBI were similar, with little exhibited bias. LTS and REWLS, on the other hand, were very close to each other, demonstrated a moderate bias. OLS and BI were severely biased as expected. All simulated scale estimates, $\hat{E}[\hat{\sigma}^2]$, overestimated, on the average, the true scale parameter of 0.25. On the other hand, the simulated robust scale estimate, $\hat{E}[\hat{v}^2]$ for the BI procedure severely underestimated the scale parameter. This led to the smallest expected standard errors of the BI coefficients, results based on the average of the simulated coefficient standard errors using the average square root of the diagonal value of $\hat{v}^2 * (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}$ matrix. Among the scale estimates, the robust scale estimate based on the effective sample size, ${\hat{v}_w}^2$, for the CBI procedure had the smallest bias, on the average.

Between CBI and S1S, the CBI coefficients had the smaller standard error and were more stable, both in terms of the observed range as well as with respect to the IQR. The REWLS improved the stability of LTS and had smaller standard error for its coefficients. Both OLS and BI exhibited very tight distributions for each of the four coefficients was of little consequence given the extreme bias that was exhibited.

The average observation weights are denoted as $\overline{\boldsymbol{w}}$, and the standardized average weight $\overline{ws}_i$ is defined as

$$\overline{ws}_i = \frac{\overline{w}_i - Min(\overline{\boldsymbol{w}})}{Max(\overline{\boldsymbol{w}}) - Min(\overline{\boldsymbol{w}})}.$$

Table 4.2: Standardized average weight for observations 1-14

Observation	REWLS	BI	CBI	Observation	REWLS	BI	CBI
1	0.014	0.971	0.002	8	0.096	0.981	0.000
2	0.144	0.991	0.002	9	0.131	0.985	0.004
3	0.159	0.991	0.005	10	0.186	0.992	0.006
4	0.000	0.964	0.001	11	0.243	0.000	0.614
5	0.107	0.987	0.005	12	0.243	0.000	0.533
6	0.132	0.986	0.002	13	0.248	0.000	0.677
7	0.152	0.990	0.003	14	0.231	0.000	0.646

Considering the result in Table 4.2, the CBI, on average, was more likely to identify the hips. For example, it gave almost 0 weights on the average to all the hips and weights greater than 0.5 to all the good leverage points. The REWLS, ended with the low weights to all the bad and good leverage points. The BI, on the other hand, mistakenly attributed the weights, provided very high weight for the first ten bad leverage points and 0 weights for the four good leverage points.

5. Conclusions

The proposed CBI methodology is a comprehensive regression analysis procedure. The goal is to be competitive with methods such as LTS (Ruppert and Carroll[4]), S1S (Coakley and Hettmansperger[6]) and REWLS (Gervini and Yohai[8]) when the data is highly contaminated but also be able to compete with the efficient M and BI regression methods (Huber and Ronchetti[2]) when the data has few or no problematic observations. Specifically, the first case study shows that the CBI outperformed the other high breakdown procedures under the low contamination situation. The Monte Carlo study, on the other hand, shows that the CBI is one of the two procedures (S1S and CBI) that provide unbiased regression coefficients. Between the unbiased procedures, the CBI has the smaller standard errors of the regression coefficients and has more stable of the coefficient estimates.

Further, that the user can rely on the CBI method to perform well across the spectrum of data contamination levels is an advantage, especially when the user may not be aware of the finer details of robust regression. Additionally, the CBI methodology provides valuable insight into the data structure, identifying multiple outliers or subgroups of similar observations. With a dendrogram illustrating the cluster history, a minor cluster activation summary and a final CBI regression estimator, scale estimate and observation weights, a CBI regression analysis provides an extensive amount of information in a compact tabular and graphical summary form.

All numerical results in this work were programmed using R and the programs are available from the authors upon request. The CBI algorithm is surprisingly fast. For example, the CBI algorithm for the case study required a few seconds using a moderately equipped PC.

Appendix A-1

Both the M and BI regression procedures have altered normal equations that involve a ψ function. This ψ function dictates some robustness properties of the estimator. For example, ψ determines whether to downweight or bound a large value of its argument, which generally involves a rescaled residual. One possible choice is the Huber ψ function, defined as

$$\psi(t) = \begin{cases} -c_H & t < -c_H \\ t & -c_H \le t \le c_H \\ c_H & t > c_H \end{cases},$$

where c_H is some constant (often referred to as the tuning parameter). To attain 95% efficiency for the location model under normally distributed errors, c_H is set to 1.345.

Another frequently used choice for ψ is the bisquare ψ function, defined as

$$\psi(t) = \begin{cases} t(1-(t/c_B)^2)^2, & t < |c_B| \\ 0, & t \ge |c_B| \end{cases}.$$

The tuning parameter c_B is given the value of 4.685 to achieve 95% efficiency under normally distributed errors for the location model. However, for BI regression it is generally preferred to incorporate the correction factor $c = \sqrt{2pn}/(n-2p)$ to account for the presence of π_i in the ψ function argument that is used to produce the BI estimator (Walker[18], Birch and Agard[15]). In this case, the Huber and the bisquare ψ functions use $c_H = 1.345\sqrt{2pn}/(n-2p)$ and $c_B = 4.685\ \sqrt{2pn}/(n-2p)$ respectively. More choices for the ψ function are provided by Wilcox[19].

Appendix A-2

Computationally, S1S requires two inputs: (1) $\widehat{\boldsymbol{\beta}}_0$, the initial LTS estimator, and (2) the $n \times 1$ vector of Mallows weights, whose elements, RD_i , are determined via $RD_i = (\boldsymbol{z}_i - \boldsymbol{MVE}_1(\boldsymbol{Z}))^T(\boldsymbol{MVE}_2(\boldsymbol{Z}))^{-1}(\boldsymbol{z}_i - \boldsymbol{MVE}_1(\boldsymbol{Z}))$, where $\boldsymbol{MVE}_1(\boldsymbol{Z})$ is the $k \times 1$ mean vector and $\boldsymbol{MVE}_2(\boldsymbol{Z})$ is the $k \times k$ covariance matrix based on minimum volume ellipsoid estimation (see Rousseeuw and Leroy[12]) and utilizing only the regressor space.

Regarding the 1-step improvement, $r_i(\widehat{\boldsymbol{\beta}}_0) = y_i - \boldsymbol{x}_i{}^T\widehat{\boldsymbol{\beta}}_0$ is the i^{th} initial residual, the robust scale estimate $\hat{\sigma}_0$ (Rousseeuw and Leroy[12]) is

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

The diagonal matrix $\boldsymbol{W}$ has diagonal elements

$$w_i = min\left[1, \left[\frac{\chi^2_{0.975,p-1}}{RD_i}\right]\right],$$

And the diagonal matrix $\boldsymbol{B}$ is defined as $\boldsymbol{B} = diag\left(\psi^{(1)}\left(\frac{r_i(\hat{\boldsymbol{\beta}}_0)}{\hat{\sigma}_0 w_i}\right)\right)$,where $\psi^{(1)}$ is the first derivative of ψ.

Appendix A-3

To obtain N, first define G as the distribution function of the absolute scaled errors under the normal model. That is

$$G(t) = 2\Phi(\mathrm{t}) - 1,$$

where $\Phi(\mathrm{t})$ is the cumulative standard normal distribution.

Let $rs_{(1)} < rs_{(2)} \ldots \leq rs_{(n)}$ denote the order statistic of the absolute scaled residual and define

$$i_0 = max\{i: rs_{(i)} < \eta, i = 1, \ldots n\},$$

where $\eta = G^{-1}(\gamma)$ and γ is a large value such as $\gamma = 0.95$. Define

$$q = \min_{i>i_0}\left\{\frac{i-1}{G(rs_{(i)})}\right\},$$

and

$$N = [q].$$

Essentially, this procedure would delete those observations with $rs_{(i)}$ values well above the identity line in a normal Q-Q plot of $rs_{(i)}$ against the respective quantiles of G .

References

1. Fan SK, Huang HK, Chang YJ. Robust Multivariate Control Charts for Outlier Detection Using Hierarchical Cluster Tree in SW2. *Quality and Reliability in Engineering International* 2012.
2. Huber PJ, Ronchetti E. *Robust statistics* Wiley: Hoboken, N.J., 2009.
3. Rousseeuw PJ. Least median of squares regression. *Journal of the American Statistical Association* 1984; **79:** 871-880.

4. Ruppert D, Carroll RJ. Trimmed least-squares estimation in the linear-model. *Journal of the American Statistical Association* 1980; **75:** 828-838.

5. Simpson DG, Ruppert D, Carroll RJ.On one-step gm estimates and stability of inferences in linear-regression. *Journal of the American Statistical Association* 1992; **87:** 439-450.

6. Coakley CW, Hettmansperger TP. A bounded influence, high breakdown, efficient regression estimator. *Journal of the American Statistical Association* 1993; **88:** 872-880.

7. Lawrence DE. Cluster-Based Bounded Influence Regression. Ph.D. dissertation, Department of Statistics, Virginia Tech. 2003.

8. Gervini D, Yohai VJ. A class of robust and fully efficient regression estimators. *Annals of Statistics* 2002; **30:** 583-616.

9. Staudte RGaS SJ. *Robust estimation and testing* John Wiley & Sons: New York, 1990.

10. Rousseeuw PJ, Van Driessen K. Computing lts regression for large data sets. *Data Mining and Knowledge Discovery* 2006; **12:** 29-45.

11. Maronna RA, Martin RD, Yohai VJ. *Robust statistics : Theory and methods* J. Wiley: Chichester, England, 2006.

12. Rousseeuw PJ, Leroy AM. *Robust regression and outlier detection* Wiley-Interscience: Hoboken, NJ, 2003.

13. Birch JB. Estimation and inference in multiple regression using robust weights: a unified approach, *Technical Report*, 92-2, Department of Statistics, Virginia Tech.1992;

14. Birch JB. *Exploratory and Robust Data Analysis*, pre-publication course packet, Virginia Tech. 2010;

15. Birch JB, Agard DB. Robust inference in regression - a comparative-study. *Communications in Statistics-Simulation and Computation* 1993; **22:** 217-244.

16. Pendleton OJ, Hocking RR. Diagnostic techniques in multiple linear regression using proc matrix. *SUGI* 1981**:** 195-201.

17. Hawkins DM, Bradu D, Gordon VK. Location of several outliers in multiple-regression data using elemental sets. *Technometrics* 1984; **26:** 197-208.

18. Walker E. Influence, Collinearity and Robust Estimation in Regression. Ph.D disseration, Department of Statistics, Virginai Tech.1984.

19. Wilcox RR. *Introduction to robust estimation and hypothesis testing* Elsevier/Academic Press: Amsterdam; Boston, 2005.