

Cluster-Based Bounded Influence Regression

by

David E. Lawrence

Dissertation submitted to the faculty of
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Statistics

Committee

Dr. Jeffrey B. Birch, chair
Dr. Christine Anderson-Cook
Dr. Eric P. Smith
Dr. George R. Terrell
Dr. Keying Ye

July 17, 2003
Blacksburg, Virginia

Keywords: High-breakdown, Robust, LTS, Linear, Outlier

Copyright 2003, David E. Lawrence.

Cluster-Based Bounded Influence Regression

David E. Lawrence

Abstract

In the field of linear regression analysis, a single outlier can dramatically influence ordinary least squares estimation while low-breakdown procedures such as M regression and bounded influence regression may be unable to combat a small percentage of outliers. A high-breakdown procedure such as least trimmed squares (LTS) regression can accommodate up to 50% of the data (in the limit) being outlying with respect to the general trend. Two available one-step improvement procedures based on LTS are Mallows 1-step (M1S) regression and Schweppe 1-step (S1S) regression (the current state-of-the-art method). Issues with these methods include (1) computational approximations and sub-sampling variability, (2) dramatic coefficient sensitivity with respect to very slight differences in initial values, (3) internal instability when determining the general trend and (4) performance in low-breakdown scenarios. A new high-breakdown regression procedure is introduced that addresses these issues, plus offers an insightful summary regarding the presence and structure of multivariate outliers. This 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 “halfset” of observations, with the remaining observations becoming one or more minor clusters. An initial regression estimator arises from the main cluster, with a multiple point addition DFFITS argument used to carefully activate the minor clusters through a bounded influence regression framework. CBI achieves a 50% breakdown point, is regression equivariant, scale equivariant and affine equivariant and distributionally is asymptotically normal. Case studies and Monte Carlo studies demonstrate the performance advantage of CBI over S1S and the other high breakdown methods regarding coefficient stability, scale estimation and standard errors. A dendrogram of the clustering process is one graphical display available for multivariate outlier detection. Overall, the proposed methodology represents advancement in the field of robust regression, offering a distinct philosophical viewpoint towards data analysis and the marriage of estimation with diagnostic summary.

Table of Contents

	Page
Abstract	ii
Glossary of Acronyms	ix
Common Notation	xi
List of Tables	xxi
List of Figures	xxiv
 Chapter 1 Classical Regression Analysis	
Introduction	1
1.1 Background and Notation	4
1.2 Ordinary Least Squares	6
1.3 Terminology	6
1.4 Leverage and the Hat Matrix	9
1.4.1 Altered Hat Matrix	11
1.5 Outlier Diagnostics	11
1.5.1 Influence Diagnostics	13
1.6 Case Study: Stackloss Data	15
 Chapter 2 Robust Regression	
Introduction	18
2.1 M and Bounded Influence Regression	19
2.2 ψ -functions	22
2.3 Iterated Reweighted Least Squares	25
2.4 Case Study: Stackloss Data	26
2.5 Illustrating the Robustness Properties of OLS, M and BI Regression	29
2.5.1 Example 2.1: A Single Low Leverage Outlier	30
2.5.2 Example 2.2: A Single High Leverage Point	34
2.5.3 Example 2.3: A High Leverage Cluster	37

		Page
2.5.4	Concluding Remarks on the Previous Three Examples	39
2.6	Joint Influence Diagnostics	40
2.7	Multiple Outlier Detection Procedures	46
2.7.1	K-Clustering	47
2.7.2	Stalactite Plot	48
2.8	Chapter Summary	50
Chapter 3	Multivariate Location and Scale Estimation	
	Introduction	51
3.1	Classical Estimation	52
3.2	Outlier Resistant Methods	53
3.2.1	Coordinatewise Median	53
3.2.2	Stahel-Donoho Estimator	53
3.2.3	Transformed One-Step Weighted Dispersion Estimator	54
3.2.4	Minimum Volume Ellipsoid Estimation	55
3.2.5	Minimum Covariance Determinant Estimation	56
3.2.6	Stalactite Estimation	58
3.2.7	Hadi Forward Search	58
3.3	Chapter Summary	59
Chapter 4	High Breakdown Regression Procedures	
	Introduction	60
4.1	Least Median of Squares (LMS)	60
4.2	Least Trimmed Squares (LTS)	63
4.3	One-Step Generalized M Estimators	65
4.3.1	Mallows 1-Step Estimator	66
4.3.2	Schweppe 1-Step Estimator	67
4.4	Case Study: Stackloss Data	69

	Page
4.5 Computational Issues for High Breakdown Regression	70
 Chapter 5 Proposed Regression Methodology	
Introduction	72
5.1 Cluster Analysis	74
5.2 The Proposed Method	77
5.3 Detailed Simple Example	83
5.4 The CBI Algorithm Philosophy	96
5.4.1 The Cluster Phase of the CBI Algorithm	97
5.4.1.1 Clustering Foundation	98
5.4.1.2 Linkage Selection	104
5.4.1.3 Clustering Stopping Rule	104
5.4.1.4 Revised Similarity Matrix	105
5.4.2 The Sequential Regression Phase of the CBI Algorithm	106
5.4.2.1 The Scale Estimate	107
5.4.2.2 Weighting Schemes	108
5.5 Case Study: Stackloss Data	110
5.6 Inferential Analysis with the CBI regression estimator	123
5.7 Chapter Summary	130
 Chapter 6 Theoretical Properties	
Introduction	132
6.1 Equivariance Properties	133
6.1.1 MVE Properties	134
6.1.2 Equivariance: OLS	143
6.1.3 Equivariance and Scale Estimation	144
6.1.4 Equivariance and the DFFITS Statistic	146
6.1.5 Equivariance and Bounded Influence Regression	148

		Page
6.1.6	Notation Used for CBI Equivariance Proofs	150
6.1.7	Regression Equivariance	152
6.1.8	Scale Equivariance	157
6.1.9	Affine Equivariance	161
6.2	Breakdown Point of the CBI Estimator	165
6.3	Asymptotic Distribution Theory	172
6.3.1	The Influence Function	173
6.3.2	Functionals in a Regression Setting	175
6.3.3	Asymptotic Distribution of the CBI Estimator	180
6.4	Chapter Summary	182
Chapter 7	Case Study Comparisons	
	Introduction	183
7.1	Case Study: The Pendleton-Hocking Data	183
7.1.1	Ordinary Least Squares Analysis	184
7.1.2	Bounded Influence Analysis	186
7.1.3	High Breakdown Regression Analysis	188
7.1.4	CBI Regression Analysis	191
7.2	Case Study: The Hawkins, Bradu and Kass Data	194
7.2.1	Ordinary Least Squares Analysis	195
7.2.2	Bounded Influence Analysis	197
7.2.3	High Breakdown Regression Analysis	198
7.2.3.1	Repeatability Issue for M1S and S1S	200
7.2.4	CBI Regression Analysis	200
7.3	Chapter Summary	203
Chapter 8	Monte Carlo	
	Introduction	204

	Page
8.1 Study #1: Uncontaminated Data with 2 Regressors	209
8.1.1 Results for Monte Carlo Study #1A (Fixed Regressor Space)	210
8.1.2 Results for Monte Carlo Study #1B (Random Regressor Space)	212
8.2 Study #2: Simple Linear Regression with a High Influence Cluster	215
8.2.1 Results for Monte Carlo Study #2A (Fixed Regressor Space)	216
8.2.2 Results for Monte Carlo Study #2B (Random Regressor Space)	218
8.3 Study #3: Pendleton-Hocking Data	220
8.4 Study #4: Hawkins-Bradru-Kass Data	223
8.5 Study #5: Random 40% Contamination with 3 Regressors	226
8.5.1 Results for Monte Carlo Study #5A (Fixed Regressor Space)	227
8.5.2 Results for Monte Carlo Study #5B (Random Regressor Space)	230
8.6 Study #6: Clustered 40% Contamination (Random Sign) with 3 Regressors...	232
8.6.1 Results for Monte Carlo Study #6A (Fixed Regressor Space)	233
8.6.2 Results for Monte Carlo Study #6B (Random Regressor Space)	235
8.7 Study #7: Clustered 40% Contamination (Single Cluster) with 3 Regressors .	237
8.7.1 Results for Monte Carlo Study #7A (Fixed Regressor Space)	238
8.7.2 Results for Monte Carlo Study #7B (Random Regressor Space)	240
8.8 Study #8: Mixed 40% Contamination with 5 regressors	243
8.8.1 Results for Monte Carlo Study #8A (Fixed Regressor Space)	244
8.8.2 Results for Monte Carlo Study #8B (Random Regressor Space)	247
8.9 Chapter Summary	249
 Chapter 9 Future Directions and Summary	
Introduction	251
9.1 Current Methodologies	251
9.1.1 Ordinary Least Squares	251
9.1.2 M Regression	252
9.1.3 BI Regression	252

		Page
9.1.4	LMS Regression	252
9.1.5	LTS Regression	252
9.1.6	MIS Regression	253
9.1.7	SIS Regression	253
9.2	CBI Regression	253
9.3	Future CBI Research	254
9.4	Conclusion	256
	References	259
	Appendix A (Datasets)	
A.1	Stackloss Data	265
A.2	Pendleton-Hocking Data	266
A.3	Hawkins-Bradru-Kass Data	267
A.4	Fixed Regressor Space Values for the Monte Carlo Studies of Chapter 8	268
	Appendix B (Detailed Algorithms)	
B.1	Stahel-Donoho Projection-based Location and Scale Estimator	275
B.2	Transformed One-step Weighted Dispersion Estimator	276
B.3	Exact MVE Estimator via the Feasible Solution Algorithm	277
B.4	Approximate MVE Estimator via the Random Subsampling Algorithm	278
B.5	Exact MCD Estimator via the Feasible Solution Algorithm	280
B.6	Stalactite Estimation	281
B.7	Hadi Forward Search	282
	Vita	285

Glossary of Acronyms

	Page
ANOVA	Analysis of Variance 11
BI	Bounded Influence 21
BLUE	Best Linear Unbiased Estimator 6
BP	Breakdown Point 18
CBI	Cluster-based Bounded Influence 73
cdf	Cumulative Distribution Function 173
DF	Degrees of Freedom 16
DFBETAS	Difference in Betas 13
DFFITS	Difference in Fits 13
FSA	Feasible Solution Algorithm 56
GM	Generalized M 22
GV	Generalized Variance 14
HBK	Hawkins, Bradu and Kass 194
IF	Influence Function 173
IQR	Interquartile Range 207
IRLS	Iterative Reweighted Least Squares 20
LMS	Least Median of Squares 60
LTS	Least Trimmed Squares 63
M1S	Mallows 1-Step 66
MAD	Median Absolute Deviation 20
MCD	Minimum Covariance Determinant 57
MLE	Maximum Likelihood Estimator 6
MLR	Multiple Linear Regression 16
MS	Mean Square 16
MSE	Mean Square Error 16
MVE	Minimum Volume Ellipsoid 55

		Page
OLS	Ordinary Least Squares	6
P-H	Pendleton-Hocking	183
RMSE	Root Mean Square Error	11
S1S	Schweppe 1-Step	67
SAS [®]	Statistical Analysis Software	44
SAS/IML [®]	SAS [®] / Interactive Matrix Language	45
SE	Standard Error	16
SLR	Simple Linear Regression	63
SS	Sums of Squares	16
SSCP	Sums of Squares and Cross Products	99
SSE	Sum of Squared Errors	11
WLS	Weighted Least Squares	25

Common Notation

Hat Notation

\hat{a} Represents an estimate of the quantity a ; a general

Transformed Data Notation (i.e. regression, scale or affine data transformations)

\tilde{a} The quantity a has been computed using or relates to transformed data; a general

Altered Data Notation (i.e. arbitrary movement; breakdown point discussion)

a^* The quantity a has been computed using or relates to altered data; a general

Constants

n	Number of observations
k	Number of regressor variables
p	Number of parameters in the linear model
h	Number of observations defining a halfset
c_H	Tuning constant (Huber)
c_B	Tuning constant (bisquare)
c	1. Tuning constant (CBI) 2. Scale transformation scalar
α	1. Type I error 2. Mallows weight constant
b	Mallows weight constant
N	Number of subsamples
β	1. Type II error 2. Tuning constant

m	1. The number of observations arbitrarily moved in breakdown point discussions 2. Subset size in the stalactite plot analysis
m_I	Number of observations in set I
g	1. Number of minor clusters 2. Index of observation being augmented to the anchor set
δ	1. Convergence constant 2. Largest $DFFITs_{+I}^2$ for minor cluster activation
K	1. Number of observations per subset in K-clustering procedure 2. Number of clusters in K-means procedure
ξ	Anchor set sizing constant

Parameters

μ	Mean
σ^2	Variance (random error)
μ_a	Mean for a ; a general
σ_a^2	Variance for a ; a general
θ	Parameter of interest; general
β	Regression parameter vector
β_j	Element of β relating to the j^{th} regressor

Data Form

y	The response variable of interest
y_i	The i^{th} response value
x_i	The i^{th} regressor variable
x_{ji}	The i^{th} observation for the j^{th} regressor
\mathbf{y}	The response vector
\mathbf{X}	The regressor matrix (including the intercept)
\mathbf{x}'_i	The i^{th} row of \mathbf{X}
\mathbf{Z}	The regressor matrix (no intercept)
\mathbf{z}'_i	The i^{th} row of \mathbf{Z}
\mathbf{X}_y	The regressor matrix (including the intercept) augmented with the response vector

$\mathbf{x}'_{y,i}$	The i^{th} row of \mathbf{X}_y
\mathbf{Z}_y	The regressor matrix (no intercept) augmented with the response vector
$\mathbf{z}'_{y,i}$	The i^{th} row of \mathbf{Z}_y
\mathbf{X}_{-I}	The matrix formed by the removal of a set I of observations from \mathbf{X}
$\mathbf{X}_{y,-I}$	The matrix formed by the removal of a set I of observations from \mathbf{X}_y
\mathbf{X}^*	The matrix corresponding to \mathbf{X} when m observations are arbitrarily moved
\mathbf{x}'_j^*	the j^{th} row of \mathbf{X}^*
\mathbf{X}_y^*	The matrix corresponding to \mathbf{X}_y when m observations are arbitrarily moved
\mathbf{Z}_y^*	The matrix corresponding to \mathbf{Z}_y when m observations are arbitrarily moved
\mathbf{y}^*	The vector corresponding to \mathbf{y} when m observations are arbitrarily moved
y_j^*	the j^{th} element of \mathbf{y}^*

Hat Matrices

\mathbf{H}	The (ordinary) hat matrix
h_{ii}	The i^{th} diagonal of \mathbf{H}
h_{ij}	The i^{th} row, j^{th} column element of \mathbf{H}
\mathbf{H}_y	The altered hat matrix
$h_{y,ij}$	The i^{th} row, j^{th} column element of \mathbf{H}_y
\mathbf{H}_w	Hat matrix for WLS regression
$h_{ii,w}$	The i^{th} diagonal element of \mathbf{H}_w

Regression Setting

$f(\cdot)$	Model function
ε	Random error vector
ε_i	The i^{th} random error
$\hat{\beta}$	A vector of regression parameter estimates
\mathbf{b}	A possible solution for $\hat{\beta}$ when viewing an objective function
b_i	The element of \mathbf{b} relating to x_i
$\hat{\mathbf{y}}$	A vector of fitted values
\hat{y}_i	The fitted value at \mathbf{x}'_i
$\hat{y}_i(\cdot)$	Fitted value at \mathbf{x}'_i , fit based on the argument
\mathbf{r}	A vector of residuals
r_i	The residual at \mathbf{x}'_i

$r_i(\cdot)$ Residual for the i^{th} observation, based on the fit specified

$r_{[i]}$ The i^{th} ranked residual

Regression Diagnostics

s	Estimate of scale, typically RMSE
r'_i	Internally studentized residual for the i^{th} observation
$\hat{y}_{i,-i}$	PRESS fitted value at \mathbf{x}'_i (when i^{th} observation is removed)
$\hat{y}_{i,-I}$	Fitted value at \mathbf{x}'_i resulting from the removal of a set I of observations
$r_{i,-i}$	PRESS residual at \mathbf{x}'_i (when i^{th} observation is removed)
$Rstudent_i$	Rstudent diagnostic statistic for i^{th} observation
$DFFITs_i^2$	Difference in fits influence diagnostic statistic when i^{th} observation is removed
$DFFITs_I^2$	Difference in fits influence diagnostic statistic when a set I of observations is removed
$\hat{\beta}_{-i}$	PRESS regression estimator when i^{th} observation is removed

$\hat{\beta}_{-I}$	Regression estimator resulting from the removal of a set I of observations
$DFBETAS_{j,i}$	Difference in Betas influence diagnostic for the j^{th} coefficient when the i^{th} observation is removed
s_{-i}	RMSE for OLS regression when the i^{th} observation is removed
s_{-I}^2	MSE for OLS regression when a set I of observations is removed
$Cook's D_i$	Cook's D diagnostic statistic for i^{th} observation
$Cook's D_I$	Cook's D influence diagnostic statistic for a set I of observations
$CovRatio_i$	CovRatio diagnostic statistic for i^{th} observation
$CovRatio_I$	CovRatio influence diagnostic statistic for a set I of observations
R_I	Influence diagnostic statistic for a set I of observations
Q_I^2	Influence diagnostic statistic for a set I of observations

\mathbf{B}_{-I}	Intermediate matrix for use in computing Q_I^2
-------------------	--

Regression Methods

$\hat{\beta}_0$	Generic representation of an initial regression estimator
$\hat{\beta}_{OLS}$	OLS regression estimator
$\hat{\beta}_M$	M regression estimator
$\hat{\beta}_{BI}$	BI regression estimator
$\hat{\beta}_{LMS}$	LMS regression estimator
$\hat{\beta}_{LTS}$	LTS regression estimator
$\hat{\beta}_{MIS}$	M1S regression estimator
$\hat{\beta}_{SIS}$	S1S regression estimator
$\hat{\beta}_{CBI}$	CBI regression estimator
π_i	Leverage weight for the i^{th} observation
\mathbf{W}	Weight matrix
w_i	Weight for the i^{th} observation
$\hat{\sigma}$	General estimate of scale
$\hat{\sigma}_0$	LMS scale estimate (see also CBI initial scale estimate)
$\hat{\sigma}_{LTS}$	LTS scale estimate (equals $\hat{\sigma}_0$)
\mathbf{g}_0	Intermediate vector used in 1-step (M1S, S1S) derivation

\mathbf{H}_0	Intermediate vector used in 1-step (M1S, S1S) derivation
\mathbf{W}	Weight matrix used in 1-step (M1S, S1S) improvement
\mathbf{B}	Leverage weight matrix used in 1-step (M1S, S1S) improvement
\mathbf{M}_0	Intermediate matrix used in 1-step (M1S, S1S) standard error derivation
\mathbf{V}	Intermediate matrix used in 1-step (M1S, S1S) estimated covariance matrix derivation
$Cov(\hat{\boldsymbol{\beta}})$	Asymptotic covariance matrix for $\hat{\boldsymbol{\beta}}$

CBI Regression

C_0	Main cluster (initial or final)
C_1, C_2, \dots, C_g	Minor clusters
$\boldsymbol{\Omega}$	Anchor set
\mathbf{B}	Point-addition OLS regression coefficient matrix
\mathbf{b}'_i	The i^{th} row of \mathbf{B}
\mathbf{S}	Similarity matrix
s_{ij}	The similarity between the i^{th} and j^{th} observations
$\hat{\sigma}_0$	Initial CBI scale estimate

$\hat{\sigma}_1$	Intermediate scale estimate in CBI
$\hat{\sigma}_{CBI}$	Final CBI scale estimate (not used for inference)
v^2	Scale estimate (BI and CBI)
n_w	Effective sample size (i.e. sum of observation weights)
v_w^2	Scale estimate based on effective sample size (CBI)
H	Set of observations to be used to base second clustering
$\boldsymbol{\omega}$	Weight vector based on membership in H
ω_i	Weight for i^{th} observation
$\mathbf{m}_H(\cdot)$	Mean vector of argument using only elements of H
$\mathbf{C}_H(\cdot)$	covariance matrix of argument using only elements of H
I	A minor cluster to be evaluated for potential activation
$\hat{\boldsymbol{\beta}}_{+I}$	Regression estimator when adding set I to C_0
$\hat{y}_{i,+I}(\cdot)$	Fitted value at \mathbf{x}'_i when using $\hat{\boldsymbol{\beta}}_{+I}$

$DFFITS_{+I}^2$	Difference in fits statistic when adding set I (a single minor cluster) to main cluster	$\tilde{\mathbf{A}}$	Special matrix constructed to apply generalized affine transformation results
J	Union of all activated minor clusters	\mathbf{v}	Vector defining regression transformation
$\hat{\boldsymbol{\beta}}_{+J}$	Regression estimator when adding set J to C_0	\mathbf{v}_{-0}	Vector defining a regression transformation, but without the intercept term
$DFFITS_{+J}^2$	Difference in fits statistic when adding set J (the collection of all activated minor clusters) to main cluster	v_i	The element of \mathbf{v} corresponding to x_i
$\boldsymbol{\pi}_{C_0}$	Vector of leverage weights for the main cluster	$\mathbf{A}_{-1,-1}$	The \mathbf{A} matrix without the first row or first column
$\boldsymbol{\pi}_{(C_0, C_I)}$	Vector of leverage weights for the union of C_0 and C_I	$\hat{\boldsymbol{\beta}}^*$	A regression estimator based on altered data
$\hat{\boldsymbol{\beta}}_1$	Intermediate CBI regression estimator	$MAD(\cdot, \cdot, \cdot)$	Median absolute deviation scale estimate, specifying arguments
$\hat{\boldsymbol{\beta}}_2$	Intermediate CBI regression estimator	$\boldsymbol{\Omega}_x$	The columns of $\boldsymbol{\Omega}$ relating to the regressors
CBI Proofs		$\boldsymbol{\Omega}_y$	The column of $\boldsymbol{\Omega}$ relating to the response
\mathbf{A}	Matrix used to define an affine transformation	$\boldsymbol{\Omega}_{(g)x}$	$\boldsymbol{\Omega}_x$ augmented with \mathbf{z}'_g
\mathbf{D}	Generic data matrix	$\boldsymbol{\Omega}_{(g)y}$	$\boldsymbol{\Omega}_y$ augmented with y_g
$\tilde{\mathbf{D}}$	Generic data matrix under a transformation	\mathbf{m}_x	Mean vector based on regressors only
		\mathbf{C}_x	Covariance matrix based on regressors only

$(\mathbf{\Omega}'_{(g)x}\mathbf{\Omega}_{(g)y})_i$	The i^{th} element of the vector computed by $\mathbf{\Omega}'_{(g)x}\mathbf{\Omega}_{(g)y}$
C_h	The set of h original observations that remains unchanged when data are altered (during a breakdown point discussion)
Δ_{\max}	The maximum similarity measure amongst non-altered observations
h_g	Number of good observations left unchanged when data are altered (during a breakdown point discussion)
Θ	The set of all possible OLS anchor set regression estimators that would result in an observation being included in the main cluster
$BP(\cdot, \cdot)$	The breakdown point, specifying the estimator and dataset
x_0	A point of which is placed extra emphasis
$F(x)$	cdf for a random variable, x
$\tilde{F}(x)$	Altered cdf for a random variable, x

$\delta_{x_0}(x)$	Binary function based on x_0
$IF(x_0; \theta)$	Influence function of θ as a function of x_0
$E_F[\cdot]$	Expected value of the argument under F
F_n	A finite sample (of size n) cdf
G	Generic cdf
$V_F[\cdot]$	Variance of the argument under F
F_x	Marginal distribution of \mathbf{x}'_i
$F_{y x}$	Conditional distribution of y_i given \mathbf{x}'_i
\mathbf{R}^a	The a -dimensional real domain; a general
F_ε	cdf for random error ε_i
$\gamma(F)$	Representation of the quantity $E_F[\mathbf{x} \mathbf{y}]$
$\Sigma(F)$	Representation of the quantity $E_F[\mathbf{x} \mathbf{x}']$
$IF_{\mathbf{T}, F}(\mathbf{x}', y)$	Influence function of the functional parameter \mathbf{T} , under the true cdf F , as a function of \mathbf{x}' and y
$T_w(F)$	The explicit WLS functional

$\gamma_w(F)$	A component of the WLS functional, equaling $E_F[w\mathbf{x}y]$	\hat{v}_w	Monte Carlo estimate of expected standard error (CBI)
$\Sigma_w(F)$	A component of the WLS functional, equaling $E_F[w\mathbf{x}\mathbf{x}']$	Functions	
$\hat{\mathbf{W}}$	A data-driven weight matrix	$\rho(\cdot)$	Function used in M regression objective function
$\hat{\boldsymbol{\beta}}_{\hat{\mathbf{W}},n}$	The WLS estimator, written as being based on data-driven weights and a finite sample of size n	$\psi(\cdot)$	Function used in altered normal equations for M regression
\mathbf{C}	A matrix used in the computation of the influence function for the explicit WLS estimator	$\psi^{(1)}$	First derivative of ψ
$\mathbf{S}_w(F)$	The implicit WLS estimator	$\eta(\cdot, \cdot)$	Function used in WLS altered normal equations
$IF_{\mathbf{S}_w, F}(\mathbf{x}', y)$	Influence function of the implicit WLS estimator, under the true cdf F , as a function of \mathbf{x}' and y	$u(\cdot)$	Function of regressors used in BI regression altered normal equations
\mathbf{M}	A matrix used in the computation of the influence function for the implicit WLS estimator	$v(\cdot)$	Function of regressors used in BI regression altered normal equations
\mathbf{Q}	A matrix used in the computation of the asymptotic covariance matrix for the implicit WLS estimator	$\varphi(\cdot)$	Function used in GM regression altered normal equations
		$\chi(\cdot)$	Function used in GM regression for scale estimation
		$K(\cdot)$	Weight function, transformed 1-step weighted dispersion matrix
		$w(\cdot)$	Weight function (location and scale estimation)

Distributions

$N[a, b]$	Normal distribution having mean a and variance b ; a, b general.
$U[a, b]$	Uniform distribution over the interval $[a, b]$; a, b general.
t_a	t distribution with a degrees of freedom; a general
$t_{b,a}$	b quantile for the t distribution with a degrees of freedom; a, b general
$\chi^2_{b,a}$	b quantile for a chi-square distribution with a degrees of freedom; a, b general
χ^2_a	Chi-square distribution with a degrees of freedom; a general

Multivariate Space

m	Generic representation of a location vector estimator
C	Generic representation of a covariance matrix estimator
$\bar{\mathbf{z}}_J$	Regressor average vector for observations in set J
d_i	Mahalanobis or other statistical distance

RD_i Robust Mahalanobis distance
(d_i also used for this purpose)

u_i Robust distance for i^{th} observation using Stahel-Donoho projection method

$D_i(w^{(k)})$ Robust distance for i^{th} observation based on k^{th} iteration weights using FSA for MVE

MVE₁ Generic representation of the minimum volume ellipsoid location vector estimator

MVE₂ Generic representation of the minimum volume ellipsoid dispersion matrix estimator

MVE₁(·) Minimum volume ellipsoid location vector estimator for the argument specified

MVE₂(·) Minimum volume ellipsoid scale matrix estimator for the argument specified

MCD₁ Generic representation of the minimum covariance determinant location vector estimator

MCD₂ Generic representation of the minimum covariance

	determinant scale matrix estimator	E	Matrix containing the eigenvectors (as columns)
m_j^2	Median squared robust diestance	\mathbf{e}_i	the i^{th} eigenvector
$\hat{\boldsymbol{\mu}}$	Robust location matrix	e_{ij}	The i^{th} row, j^{th} column element of E
V	Intermediate dispersion matrix (Transformed 1-step weighted dispersion matrix)	$\boldsymbol{\lambda}$	Vector of eigenvalues
U	Robust dispersion matrix	λ_i	Eigenvalue for i^{th} eigenvector
Analysis of Variance		R^2	Coefficient of Determination
SS_{MODEL}	Model sum of squares	$\mathbf{I}_{p \times p}$	Identity matrix with p rows and p columns
SS_{ERROR}	Error sum of squares (SSE)	$\mathbf{1}_{n \times 1}$	Vector of ones (with n rows and 1 column)
SS_{TOTAL}	Total sum of squares	$\mathbf{0}_{k \times 1}$	Vector of zeroes (with k rows and 1 column)
df_{MODEL}	Model degrees of freedom	$\mathbf{0}_{1 \times k}$	Row vector of zeroes (with 1 row and k columns)
df_{ERROR}	Error degrees of freedom	$\hat{E}[\cdot]$	Monte Carlo estimated expected value of argument
MS_{MODEL}	Model mean square	u_{1i}, u_{2i}, u_{3i}	Random variables used for the generation of outliers during a simulation
MS_{ERROR}	Mean square error (MSE)		
Miscellaneous			
c_{jj}	The j^{th} diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$		
T	An estimator		
C_a^b	Number of combinations of a elements out of b elements; a, b general		

List of Tables

Table	Title	Page
1.1	<i>Analysis of variance table and parameter estimates summary for the OLS analysis of the stackloss data.</i>	16
1.2	<i>Diagnostics for the OLS analysis of the stackloss data.</i>	17
2.1	<i>Robust analysis of variance table and parameter estimate summary for the M-Huber regression of the stackloss data, using a tuning parameter of 1.345.</i>	26
2.2	<i>Robust analysis of variance table and parameter estimate summary for the M-bisquare regression of the stackloss data, using M-Huber as the initial estimate and a tuning parameter of 4.685.</i>	27
2.3	<i>Robust analysis of variance table and parameter estimate summary for the M-bisquare regression of the stackloss data, using a tuning parameter of 3.5 for the bisquare function and initial M-Huber estimates with a 1.01 tuning parameter.</i>	27
2.4	<i>Comparison of final observation weights for the three M regression analyses of the stackloss data.</i>	28
2.5	<i>Raw data and hat diagonals for Example 2.1.</i>	31
2.6	<i>Regression coefficients for Example 2.1.</i>	31
2.7	<i>Final weights given to the nine observations of Example 1 for each of the six regression methods.</i>	32
2.8	<i>Two local minima for M-bisquare analysis of Example 2.1.</i>	33
2.9	<i>Raw data and hat diagonals for Example 2.2.</i>	35
2.10	<i>Regression coefficients for Example 2.2.</i>	36
2.11	<i>Final observation weights for the six competing regression procedures in the analysis of Example 2.2.</i>	36
2.12	<i>Raw data and hat diagonals for Example 2.3.</i>	37
2.13	<i>Regression coefficients for Example 2.3.</i>	37
2.14	<i>Final observation weights for the six competing regression procedures in the analysis of Example 2.3.</i>	39
2.15	<i>Influence diagnostics for observation 10 of Example 2.3, Section 2.5.3.</i>	45
2.16	<i>Joint influence diagnostics for observations 9 and 10 of Example 2.3.</i>	45
2.17	<i>Joint influence diagnostics for observations 4 and 21 of stackloss data.</i>	46
4.1	<i>High breakdown regression for stackloss data.</i>	70
5.1	<i>Simple linear regression example.</i>	84
5.2	<i>The 12×2 \mathbf{B} matrix; the OLS regression estimators from anchor set regressions.</i>	87
5.3	<i>Similarity matrix using $s_{ij} = (\mathbf{b}_i - \mathbf{b}_j)' (\mathbf{MVE}_2(\mathbf{B}))^{-1} (\mathbf{b}_i - \mathbf{b}_j)$.</i>	87
5.4	<i>Cluster history for example data.</i>	88

List of Tables (continued)

Table	Title	Page
5.5	<i>The updated 12×2 \mathbf{B} matrix; the OLS regression estimators from updated anchor set regressions.</i>	91
5.6	<i>The updated 12×12 similarity matrix using $s_{ij} = (\mathbf{b}_i - \mathbf{b}_j)' (\mathbf{C}_H(\mathbf{B}))^{-1} (\mathbf{b}_i - \mathbf{b}_j)$.</i>	92
5.7	<i>The updated cluster history for example data.</i>	92
5.8	<i>Final observation weights for the CBI regression estimator.</i>	94
5.9	<i>The initial 21×4 \mathbf{B} matrix for the stackloss case study.</i>	112
5.10	<i>Similarity matrix using $s_{ij} = (\mathbf{b}_i - \mathbf{b}_j)' (\mathbf{MVE}_2(\mathbf{B}))^{-1} (\mathbf{b}_i - \mathbf{b}_j)$.</i>	114
5.11	<i>Cluster history for example data.</i>	115
5.12	<i>The updated 21×4 \mathbf{B} matrix for the stackloss case study.</i>	118
5.13	<i>Updated similarity matrix using $s_{ij} = (\mathbf{b}_i - \mathbf{b}_j)' (\mathbf{C}_H(\mathbf{B}))^{-1} (\mathbf{b}_i - \mathbf{b}_j)$.</i>	119
5.14	<i>Updated cluster history for example data.</i>	120
5.15	<i>Final observation weights for the CBI regression estimator.</i>	122
5.16	<i>Robust analysis of variance table and parameter estimates summary for the CBI analysis of the Section 5.3 example dataset.</i>	126
5.17	<i>Robust analysis of variance table and parameter estimates summary for the CBI analysis of the stackloss dataset.</i>	128
7.1	<i>Analysis of variance table and parameter estimate summary for the OLS analysis of the Pendleton-Hocking data.</i>	184
7.2	<i>Diagnostics for the OLS analysis of the Pendleton-Hocking data.</i>	185
7.3	<i>Robust analysis of variance table and parameter estimates summary for the BI-bisquare regression of the Pendleton-Hocking data.</i>	186
7.4	<i>Diagnostics for the BI-bisquare regression of the Pendleton-Hocking data.</i>	187
7.5	<i>High breakdown regression estimation of the Pendleton-Hocking data.</i>	189
7.6	<i>Results for five high breakdown analyses of the Pendleton-Hocking data.</i>	190
7.7	<i>Summary of the CBI regression analysis of the Pendleton-Hocking dataset.</i>	192
7.8	<i>Analysis of variance table and parameter estimates summary for the OLS analysis of the HBK data.</i>	195
7.9	<i>Diagnostics for the OLS analysis of the HBK data. Only the first 20 observations are shown due to size of data.</i>	196
7.10	<i>OLS estimation on reduced HBK data. Column 1 used the full dataset; column 2 removed the four good leverage points; column 3 removed the ten bad leverage points; column 4 removed the fourteen high leverage points.</i>	196

List of Tables (continued)

Table	Title	Page
7.11	<i>Robust analysis of variance table and parameter estimates for the BI-bisquare analysis of the HBK data.</i>	197
7.12	<i>Diagnostics for the BI-bisquare regression of the HBK data. First twenty observations only.</i>	198
7.13	<i>High breakdown regression estimators for the HBK data.</i>	199
7.14	<i>Results of five simulations of high breakdown analysis.</i>	199
7.15	<i>Summary of the CBI regression analysis of the HBK dataset.</i>	201
8.1	<i>Format of a Monte Carlo simulation summary table. An “x” represents a numerical entry. Shaded cells are not applicable. The following is representative of a $p = 3$ regressor situation.</i>	206
8.2	<i>True parameter values for the Monte Carlo simulation studies.</i>	209
8.3	<i>Simulation results for Monte Carlo study #1A, the fixed regressor case.</i>	211
8.4	<i>Simulation results for Monte Carlo study #1B, the random regressor case.</i>	213
8.5	<i>Simulation results for Monte Carlo study #2A, the fixed regressor case.</i>	216
8.6	<i>Simulation results for Monte Carlo study #2B, the random regressor case.</i>	218
8.7	<i>Simulation results for Monte Carlo study #3.</i>	221
8.8	<i>Simulation results for Monte Carlo study #4.</i>	224
8.9	<i>Simulation results for Monte Carlo study #5A, fixed regressor case.</i>	228
8.10	<i>Simulation results for Monte Carlo study #5B, random regressor case.</i>	230
8.11	<i>Simulation results for Monte Carlo study #6A, fixed regressor case.</i>	233
8.12	<i>Simulation results for Monte Carlo study #6B, random regressor case.</i>	235
8.13	<i>Simulation results for Monte Carlo study #7A, fixed regressor case.</i>	239
8.14	<i>Simulation results for Monte Carlo study #7B, random regressor case.</i>	241
8.15	<i>Simulation results for Monte Carlo study #8A, fixed regressor case.</i>	246
8.16	<i>Simulation results for Monte Carlo study #8B, random regressor case.</i>	247

Appendix A Tables

A.1	<i>Stackloss data.</i>	265
A.2	<i>Pendleton-Hocking data.</i>	266
A.3	<i>Hawkins, Bradu and Kass data.</i>	267
A.4	<i>Fixed regressor values for Monte Carlo study #1A.</i>	268
A.5	<i>Fixed regressor values for Monte Carlo study #2A.</i>	268
A.6	<i>Fixed regressor values for Monte Carlo study #5A.</i>	269
A.7	<i>Fixed regressor values for Monte Carlo study #6A.</i>	270
A.8	<i>Fixed regressor values for Monte Carlo study #7A.</i>	271
A.9	<i>Fixed regressor values for Monte Carlo study #8A.</i>	272

List of Figures

Figure	Title	Page
1.1	<i>Scatterplot illustrating leverage and influence. Point A is a low leverage outlier, point B is a good leverage point, and point C is a bad leverage point.</i>	7
1.2	<i>The “pulling” effect of a high influence point in OLS regression.</i>	8
2.1	<i>Huber(solid line) and bisquare (dashed line) weight functions.</i>	24
2.2	<i>Various regression fits for Example 2.1. OLS is the solid line, M-bisquare is the dashed line, and BI-bisquare is the dotted line.</i>	30
2.3	<i>Various fits to the Example 2.2 data. OLS is the solid line, M-bisquare is the dashed line, and BI-bisquare is the dotted line.</i>	35
2.4	<i>Various regression fits for Example 2.3. OLS is the solid line, M-bisquare is the dashed line, and BI-bisquare is the dotted line.</i>	38
2.5	<i>Illustration of joint influence in SLR. (A) and (B) show a jointly influential pair; (C) and (D) show a non-influential pair.</i>	41
2.6	<i>Best stalactite plot for stackloss data.</i>	49
4.1	<i>Possible configuration of the four blocks used in the second LMS subsampling algorithm, given eight observations on a scatterplot.</i>	62
5.1	<i>Scatterplot of Example 1 data, which contains a high influence cluster of two observations (10 and 11) and also a good leverage point (12).</i>	84
5.2	<i>Illustration of the anchor point locations (first cluster phase of the CBI algorithm) for the Section 5.3 example dataset.</i>	86
5.3	<i>Dendrogram for initial clustering of Section 5.3 example dataset.</i>	88
5.4	<i>Updated anchor points in relation to the original dataset.</i>	90
5.5	<i>Dendrogram for final clustering of Section 5.3 dataset.</i>	93
5.6	<i>CBI regression fit (solid line) for the Section 5.3 example dataset. Also shown are the LTS fit (dashed line) and the SIS fit (dotted line). The MIS fit is identical to the SIS fit. The OLS fit (omitting observations 10 and 11) is virtually identical (graphically indistinguishable) to the CBI fit.</i>	95
5.7	<i>IRLS weights via Mallows weighting philosophy. Illustration uses a bisquare ψ-function, $\sigma = 1$, and the robust distance (rd) critical value is 10.</i>	109
5.8	<i>IRLS weights via Schweppe weighting philosophy. Illustration uses a bisquare ψ-function, $\sigma = 1$, and the robust distance (rd) critical value is 10.</i>	109

List of Figures (continued)

Figure	Title	Page
5.9	<i>Illustration of the anchor point layout for the stackloss data, as well as the visual representation of the high influence points. Center of the anchor points a square, other anchor points are asterisks, observations 1, 3, 4 and 21 are circles, observation 2 a solid triangle, remaining data are solid circles.</i>	111
5.10	<i>Dendrogram for initial clustering of the stackloss data.</i>	113
5.11	<i>Illustration of the revised anchor point layout for the stackloss data, as well as the visual representation of the high influence points. Center of the anchor points a square, other anchor points are asterisks, observations 1, 3, 4 and 21 are circles, observation 2 a solid triangle, remaining data are solid circles.</i>	116
5.12	<i>Dendrogram for final clustering of the stackloss data.</i>	120
5.13	<i>Final clustering dendrogram and final CBI observation weights for the Section 5.3 example dataset.</i>	127
5.14	<i>Final clustering dendrogram and final CBI observation weights for the Section 5.3 example dataset.</i>	129
7.1	<i>Final clustering of the Pendleton-Hocking data and the final CBI regression observation weights.</i>	193
7.2	<i>Final clustering of the HBK data and the final CBI regression observation weights.</i>	202