
[All ETDs from UAB](#)

[UAB Theses & Dissertations](#)

2008

A Statistical Approach Identifying and Limiting the Effect of Influential Observations in Linear Regression

Tamekia L. Jones
University of Alabama at Birmingham

Follow this and additional works at: <https://digitalcommons.library.uab.edu/etd-collection>

Recommended Citation

Jones, Tamekia L., "A Statistical Approach Identifying and Limiting the Effect of Influential Observations in Linear Regression" (2008). *All ETDs from UAB*. 6799.
<https://digitalcommons.library.uab.edu/etd-collection/6799>

This content has been accepted for inclusion by an authorized administrator of the UAB Digital Commons, and is provided as a free open access item. All inquiries regarding this item or the UAB Digital Commons should be directed to the [UAB Libraries Office of Scholarly Communication](#).

A STATISTICAL APPROACH IDENTIFYING AND LIMITING THE EFFECT OF
INFLUENTIAL OBSERVATIONS IN LINEAR REGRESSION

by

TAMEKIA L. JONES

DAVID T. REDDEN, COMMITTEE CHAIR
CHRISTOPHER COFFEY
LYNN GERALD
CHARLES KATHOLI
SHARINA PERSON

A DISSERTATION

Submitted to the graduate school of the University of Alabama at Birmingham,
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

BIRMINGHAM, ALABAMA

2008

A STATISTICAL APPROACH IDENTIFYING AND LIMITING THE EFFECT OF INFLUENTIAL OBSERVATIONS IN LINEAR REGRESSION

TAMEKIA L. JONES

BIOSTATISTICS

ABSTRACT

Outliers are observations with extreme standardized deviations between the observed dependent variable and the predicted value. Within linear regression, outliers are detected by using studentized residuals. Leverage is a measure of the standardized deviation of an observation's row vector of independent variables from the mean vector of the independent variables. Within linear regression, leverage may be assessed by using the diagonal of the projection matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. An observation that is both an outlier and a leverage point is an influential observation. Influential observations affect estimation of regression parameters and therefore lead to poor estimation of the regression line, to incorrect inference, and to inaccurate predictions. Detection of these observations is often difficult because of the masking and swamping effects. The masking effect occurs when some or all influential points are difficult to identify with the use of regression diagnostics because the extremeness of one observation obscures the extremeness of another. Swamping occurs when an observation is incorrectly identified as an outlier and/or leverage point. We present a robust regression (Robust Forward Detection) method that extends the concept of robust distances by using the minimum covariance determinant and the concept of least trimmed squares. By downweighting any observations considered atypical (outliers, leverage, and influential), we extend the use of the Robust Forward Detection method beyond the application of dichotomous weights and include all observations in the dataset. Results from utilizing our proposed method

are illustrated and compared with that of ordinary least squares via simulations and examples. We illustrate that our proposed approach is capable of overcoming masking and swamping, properly identifies influential observations (i.e., outliers and leverage), and is robust to their influence.

DEDICATION

This dissertation is dedicated to my parents, Allene M. H. Jones and the late Franklin D. Jones. My achievements are contributed to my parents for their continued support, love, and patience. I am extremely grateful to my mother, who has given me advice, encouraging words, hugs, and inspirational cards and who has supported me through all of the difficult and challenging times that I have encountered during this journey.

ACKNOWLEDGMENTS

I thank each of my committee members, Drs. David T. Redden, Christopher Coffey, Lynn Gerald, and Sharina Person, for their advice, support, and commitment throughout this entire process. I thank Dr. Redden for his hard work and dedication as my academic and dissertation advisor. Thanks to him for all of the advice that he has given me. I give special acknowledgments to Dr. Person, who wore many hats and held several capacities during this journey including those of professor, mentor, committee member, and friend. I thank her for listening and giving valuable guidance and support. Thanks go to Mr. Andy Westfall for his technical assistance. Also, I am extremely grateful to my parents, my family, and my friends, all of whom continuously encouraged, supported, and prayed for me.

TABLE OF CONTENTS

	<i>Page</i>
ABSTRACT.....	ii
DEDICATION.....	iv
ACKNOWLEDGMENTS	v
LIST OF TABLES.....	viii
LIST OF FIGURES	xi
INTRODUCTION	1
Ordinary Least Squares.....	1
Diagnostic Issues	4
Robust Regression	7
Influence and Breakdown	8
L_1 Regression.....	9
M-estimation.....	9
LMS Estimator.....	10
LTS Estimator.....	10
Hadi (1992, 1994).....	11
M1 Algorithm	13
Forward Search.....	14
A ROBUST FORWARD DETECTION METHOD FOR THE IDENTIFICATION OF INFLUENTIAL OBSERVATIONS IN LINEAR REGRESSION.....	17
LIMITING THE IMPACT OF INFLUENTIAL OBSERVATIONS IN LINEAR REGRESSION VIA WEIGHT FUNCTIONS.....	41
IMPLEMENTING THE ROBUST FORWARD DETECTION METHOD IN THE LUNG HEALTH STUDY	78
CONCLUSIONS.....	92
Concluding Remarks for Paper 1	92

TABLE OF CONTENTS (Continued)

	<i>Page</i>
Concluding Remarks for Paper 2.....	94
Concluding Remarks for Paper 3.....	95
Future Research	96
GENERAL LIST OF REFERENCES	98
APPENDIX	
A DERIVATION OF $\hat{\beta}_w$	100
B INSTITUTIONAL REVIEW BOARD FOR HUMAN USE APPROVAL FORM	107

LIST OF TABLES

<i>Table</i>		<i>Page</i>
A ROBUST FORWARD DETECTION METHOD FOR THE IDENTIFICATION OF INFLUENTIAL OBSERVATIONS IN LINEAR REGRESSION		
1	Simulated outlier and leverage results for the RFD method using equation (6) and for the LTS (ROBUSTREG) procedure	30
2	Simulation results for the RFD and LTS (ROBUSTREG) methods identifying influential observations	31
3	Simulation results for the proportion of true, masked, and swamped observations obtained by utilizing the RFD and LTS (ROBUSTREG) methods.....	31
4	Comparison of methods identifying atypical observations for the HBK data	33
5	Comparison of methods identifying atypical observations for the Star Cluster data.....	35
LIMITING THE IMPACT OF INFLUENTIAL OBSERVATIONS IN LINEAR REGRESSION VIA WEIGHT FUNCTIONS		
1	Average parameter estimates (std dev), MSE, and R^2 for 1000 simulations (100 observations per simulation) for outlier contamination for OLS, RFD1, and RFD2	52
2	Average squared deviations for parameter estimates for the true simulated parameter from OLS, RFD1, and RFD2 for 1000 simulations (100 observations per simulation) in the presence of outlier contamination	54
3	Average parameter estimates (std dev), MSE, and R^2 for 1000 simulations (100 observations per simulation) for leverage contamination for OLS, RFD1, and RFD2	55

LIST OF TABLES (Continued)

<i>Table</i>	<i>Page</i>
4 Average squared deviations for parameter estimates for the true simulated parameter from OLS, RFD1, and RFD2 for 1000 simulations (100 observations per simulation) in the presence of leverage contamination	56
5 Average parameter estimates (std dev), MSE, and R^2 for 1000 simulations (100 observations per simulation) for influence contamination for OLS, RFD1, and RFD2	57
6 Average squared deviations for parameter estimates for the true simulated parameter from OLS, RFD1, and RFD2 for 1000 simulations (100 observations per simulation) in the presence of influence contamination.....	59
7 Bootstrap results illustrating average estimated parameters (std dev) and corresponding 95% confidence intervals for the parameters when RFD1 and RFD2 are used in the presence of 10% outlier contamination (based on 10,000 iterations)	60
8 Bootstrap results illustrating average estimated parameters (std dev) and corresponding 95% confidence intervals for the parameters when RFD1 and RFD2 are used in the presence of 10% leverage contamination (based on 10,000 iterations)	60
9 Bootstrap results illustrating average estimated parameters (std dev) and corresponding 95% confidence intervals for the parameters when RFD1 and RFD2 are used in the presence of 10% influence contamination (based on 10,000 iterations)	61
10 Bootstrap results illustrating average estimated parameters (std dev) and corresponding 95% confidence intervals for the parameters when RFD1 and RFD2 are used in the presence of 20% outlier contamination (based on 10,000 iterations)	61
11 Bootstrap results illustrating average estimated parameters (std dev) and corresponding 95% confidence intervals for the parameters when RFD1 and RFD2 are used in the presence of 20% leverage contamination (based on 10,000 iterations)	62

LIST OF TABLES (Continued)

<i>Table</i>	<i>Page</i>
12	Bootstrap results illustrating average estimated parameters (std dev) and corresponding 95% confidence intervals for the parameters when RFD1 and RFD2 are used in the presence of 20% influence contamination (based on 10,000 iterations)62
13	Weights $g_k (= a_k * b_k)$ implemented by using weight functions via the RFD2 method for the HBK data64
14	Parameter estimates, MSE, and R^2 obtained when OLS, RFD1, and RFD2 are used to analyze the HBK data65
15	Bootstrap results illustrating average estimated parameters (std dev) and corresponding 95% confidence intervals for the parameters when RFD1 and RFD2 are used for the HBK data (based on 10,000 iterations)66
16	Weights $g_k (= a_k * b_k)$ implemented by using weight functions via the RFD2 method for the Star cluster67
17	Parameter estimates, MSE, and R^2 obtained when OLS, RFD1, and RFD2 are used to analyze the Star Cluster data68
18	Bootstrap results illustrating average estimated parameters (std dev) and corresponding 95% confidence intervals for the parameters when RFD1 and RFD2 are used for the Star Cluster data (based on 10,000 iterations)69
IMPLEMENTING THE ROBUST FORWARD DETECTION METHOD IN THE LUNG HEALTH STUDY	
1	Baseline characteristics for subjects 54 ⁺ years of age by lung cancer status during the 5 year follow-up period84
2	Weights $g_k (= a_k * b_k)$ implemented by using weight functions via the RFD2 method for the Lung Health Study data85
3	Parameter estimates (Std Dev), MSE, and R^2 obtained when OLS, RFD1, and RFD2 are used to analyze the for the Lung Health Study data.....86
4	Crosstab of leverage status (determined by using the RFD method) by lung cancer status for the Lung Health Study88

LIST OF FIGURES

<i>Figure</i>	<i>Page</i>
A ROBUST FORWARD DETECTION METHOD FOR THE IDENTIFICATION OF INFLUENTIAL OBSERVATIONS IN LINEAR REGRESSION	
1	Proportion of leverage identified on the basis of 1000 iterations per n and p when the $3p/n$ rule is used.....27
2	Plot of adjusted residuals vs. the rank of the robust distances for the HBK data34
3	Plot of adjusted residuals vs. the rank of the robust distances for the Star Cluster data.....35
LIMITING THE IMPACT OF INFLUENTIAL OBSERVATIONS IN LINEAR REGRESSION VIA WEIGHT FUNCTIONS	
1	Scatterplot of logarithmic light intensity of the star versus the logarithmic temperature at the surface of the star for the Star Cluster data68
2	Histograms for parameter estimates in the presence of 20% outlier contamination: $\hat{\beta}_0^*$ for OLS (a), RFD1 (b), and RFD2 (c); $\hat{\beta}_1^*$ for OLS (d), RFD1 (e), and RFD2 (f); $\hat{\beta}_2^*$ for OLS (g), RFD1 (h), and RFD2 (i); and $\hat{\beta}_3^*$ for OLS (j), RFD1 (k), and RFD2 (l).....71
3	Histograms for parameter estimates in the presence of 20% leverage contamination: $\hat{\beta}_0^*$ for OLS (a), RFD1 (b), and RFD2 (c); $\hat{\beta}_1^*$ for OLS (d), RFD1 (e), and RFD2 (f); $\hat{\beta}_2^*$ for OLS (g), RFD1 (h), and RFD2 (i); and $\hat{\beta}_3^*$ for OLS (j), RFD1 (k), and RFD2 (l)72
4	Histograms for parameter estimates in the presence of 20% influence contamination: $\hat{\beta}_0^*$ for OLS (a), RFD1 (b), and RFD2 (c); $\hat{\beta}_1^*$ for OLS (d), RFD1 (e), and RFD2 (f); $\hat{\beta}_2^*$ for OLS (g), RFD1 (h), and RFD2 (i); and $\hat{\beta}_3^*$ for OLS (j), RFD1 (k), and RFD2 (l)73

INTRODUCTION

Linear regression is used to describe relationships among variables, such as a response or dependent variable and one or more independent variables. The general linear model employed is $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where \mathbf{y} is an $n \times 1$ vector for which each row corresponds to an observation's response; \mathbf{X} is an $n \times p$ matrix of fixed, known constants for which each column corresponds to an independent variable or predictor including the intercept; $\boldsymbol{\beta}$ is a $p \times 1$ vector of the unknown parameters; and $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of the unobservable random errors. Note that n corresponds to the sample size and that p is the number of parameters specified in the model. Using matrix notation, it can be written such that

$$\mathbf{y}_{n \times 1} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \mathbf{X}_{n \times p} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}, \boldsymbol{\beta}_{p \times 1} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \text{ and } \boldsymbol{\varepsilon}_{n \times 1} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

Ordinary Least Squares

Within the statistical framework of linear regression, $\boldsymbol{\beta}$ is typically estimated by $\hat{\boldsymbol{\beta}}$ using ordinary least squares (OLS). The assumptions associated with OLS are homogeneity of the errors (i.e., $\text{Var}[\boldsymbol{\varepsilon}] = \sigma^2 \mathbf{I}$, where \mathbf{I} is an $n \times n$ identity matrix consisting of 1's on the diagonal and 0's off the diagonal), independence of error terms, linearity in terms of the parameters (i.e., $E[\mathbf{y}] = \mathbf{X}\boldsymbol{\beta}$), and finite variance (Muller and Fetterman, 2002). The random errors, $\boldsymbol{\varepsilon}$, are estimated by the residuals:

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}, \quad (1)$$

where $\hat{\mathbf{y}}$ is the $n \times 1$ vector of predicted values and where $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$. As illustrated in (1), the residuals measure the difference between the observed values and the predicted values.

Utilizing OLS and assuming that \mathbf{X} is full rank, the estimator for $\boldsymbol{\beta}$ can be computed by using the sums of squares of error (SSE). As indicated below, the sum of the squared deviations of the observed dependent values and the predicted values is minimized with respect to $\hat{\boldsymbol{\beta}}$:

$$\begin{aligned} \text{SSE} &= \mathbf{e}'\mathbf{e} \\ &= (\mathbf{y} - \hat{\mathbf{y}})'(\mathbf{y} - \hat{\mathbf{y}}) \\ &= (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\ &= \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - \mathbf{y}'\mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} \\ &= \mathbf{y}'\mathbf{y} - 2\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} \\ &= \Lambda \\ \frac{\partial \Lambda}{\partial \hat{\boldsymbol{\beta}}} &= -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}. \end{aligned} \quad (2)$$

Set (2) equal to zero and solve for $\hat{\boldsymbol{\beta}}$:

$$\begin{aligned} & -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = 0 \\ \Rightarrow & \quad 2\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = 2\mathbf{X}'\mathbf{y} \\ \Rightarrow & \quad \mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y} \\ \Rightarrow & \quad \hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}, \end{aligned} \quad (3)$$

where $\hat{\beta}$ is unbiased and has minimum variance (Neter et al., 1996). The properties of the estimator are derived as follows:

$$\begin{aligned}
 E(\hat{\beta}) &= E[(X'X)^{-1}X'y] \\
 &= E[(X'X)^{-1}X'(X\beta + \epsilon)] \\
 &= E[(X'X)^{-1}X'X\beta + (X'X)^{-1}X'\epsilon] \\
 &= E[\beta + (X'X)^{-1}X'\epsilon] \\
 &= E[\beta] + E[(X'X)^{-1}X'\epsilon] \\
 &= \beta + (X'X)^{-1}X' E[\epsilon] \text{ from } E[\epsilon]=0 \\
 &= \beta
 \end{aligned}$$

and

$$\begin{aligned}
 \text{Var}(\hat{\beta}) &= \text{Var}[(X'X)^{-1}X'y] \\
 &= (X'X)^{-1}X'\text{Var}[y](X'X)^{-1}X' \\
 &= (X'X)^{-1}X'(\sigma^2\mathbf{I})(X'X)^{-1}X' \text{ from } \text{Var}[\epsilon]=\sigma^2\mathbf{I} \\
 &= \sigma^2\mathbf{I}'X'X(X'X)^{-1} \\
 &= \sigma^2(X'X)^{-1}.
 \end{aligned}$$

Based on the previous derivations and assuming the unobservable random errors are normally distributed, with mean $\mathbf{0}$ and variance $\sigma^2\mathbf{I}$ such that $\epsilon \sim N(\mathbf{0}, \sigma^2\mathbf{I})$, it can be written such that $\hat{\beta} \sim N(\beta, \sigma^2(X'X)^{-1})$. It is noted that distributional assumptions are not necessary when estimating parameters using OLS. However, distributional assumptions are pertinent when hypotheses are being tested and when inferences are being made.

Diagnostic Issues

There are several diagnostic issues encountered when the OLS method is being applied: outliers, leverage points, influential observations, masking, and swamping.

Outliers are observations with extreme standardized deviations between the observed dependent variable and the predicted value. Within simple linear regression (regression on one predictor), these deviations in the y -direction can be visually identified via scatter plots such as plots of the dependent variable versus the independent variable and/or plots of the studentized residuals versus the predicted outcome. Two types of studentized residuals have been proposed. Gray and Woodall (1994) defined the internally studentized residuals, r_i , as

$$r_i = \frac{(y_i - \hat{y}_i)}{\sqrt{\text{MSE} * (1 - h_{ii})}}$$

and the externally studentized residuals, t_i , as

$$t_i = \frac{(y_i - \hat{y}_i)}{\sqrt{\text{MSE}_{(i)} * (1 - h_{ii})}},$$

where $\text{MSE} = \mathbf{e}'\mathbf{e}/(n-p)$; the mean squared error (MSE) is the sum of the squared residuals divided by $(n-p)$, where n is the sample size and where p is the number of parameters. $\text{MSE}_{(i)}$ is the mean squared error with the i^{th} case deleted from the calculation, and h_{ii} is the i^{th} diagonal element of the hat matrix.

The diagonal elements of \mathbf{H} , denoted h_{ii} , are utilized to assess leverage. Leverage is a measure of the standardized deviation of an observation's row vector of independent variables from the mean vector of the independent variables. From (3), we can write the predicted values such that

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$$

$$\begin{aligned}
&= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\
&= \mathbf{H}\mathbf{y},
\end{aligned}$$

where

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}', \quad (4)$$

$n \times n$

As indicated, the hat matrix, \mathbf{H} , involves matrix multiplication only of the data matrix \mathbf{X} .

Utilizing (4), it can be shown that $\mathbf{H}\mathbf{X} = \mathbf{X}$. This property identifies \mathbf{H} as a projection matrix (projects onto the column space of \mathbf{X}) which is both symmetric ($h_{ij} = h_{ji}$ for all i, j) and idempotent ($\mathbf{H}\mathbf{H} = \mathbf{H}$). Hence, the hat matrix \mathbf{H} has the following properties:

$$\text{i. } 0 \leq h_{ij} \leq 1$$

$$\text{ii. } \sum_{i=1}^n h_{ii} = p.$$

Muller and Fetterman (2002) suggested that $h_{ii} > 2\frac{p}{n}$ can be used as the “rule of thumb”

in order to indicate that the i^{th} observation has an extreme value in the predictor space and thus is a leverage point. In small datasets, high leverage points can have an unusually large effect on the estimated regression parameters. This effect can lead to distortion and poor approximations of the fitted regression line. As previously stated, these high leverage points are sometimes identified by evaluating the diagonal elements of the hat matrix.

An observation that is both an outlier and a leverage point is an influential observation. Influential observations affect estimation of regression parameters and therefore lead to poor estimation of the fitted regression line, to incorrect inference, and to inaccurate predictions, especially for small datasets. Influential observations may be detected by evaluating the changes in the estimated regression line and/or estimated

regression parameters that result from the inclusion/exclusion of those particular observations in the model. Influential observations may also be identified by using Cook's Distance, better known as Cook's D:

$$D = (\hat{\mathbf{y}} - \hat{\mathbf{y}}_{(i)})'(\hat{\mathbf{y}} - \hat{\mathbf{y}}_{(i)}) / (p * \text{MSE}),$$

where $\hat{\mathbf{y}}_{(i)}$ is the $n \times 1$ vector of predicted values when the i^{th} case is deleted from the calculation of the least squares estimate $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$. Cook's D measures the standardized deviation of the predicted values using all observations and the predicted values using all observations but with the i^{th} observation deleted.

Detection of extreme observations (outliers, leverage, and influential points) can be difficult because of masking and swamping. Masking occurs when some or all atypical (outliers, leverage, and influential points) observations are difficult to identify via regression diagnostics because the extremeness of one observation obscures the extremeness of another. Swamping is the type of effect that occurs when non-contaminated observations are incorrectly identified as atypical.

Because the primary objective of our research is to properly detect and manage the diagnostic issues that we encounter when using OLS, we utilize robust regression. Robust regression is a statistical approach originally designed to be robust toward outliers. Unlike ordinary least squares, it allows observations to be weighted unequally. Robust regression first fits a line to a subset of the data and then attempts to identify the outliers. Common robust regression approaches produce a fitted line and estimators that are not strongly affected by outliers. However, several forms of robust regression remain strongly affected by influential points. Within this dissertation, we propose robust

regression methods that are resistant to influential points, as well as to outliers and leverage points.

Robust Regression

Over the years, statisticians have defined robust regression differently. Huber (1981) stated,

The word ‘robust’ is loaded with many—sometimes inconsistent—connotations. We use it in a relatively narrow sense: for our purposes, *robustness signifies insensitivity to small deviations from the assumptions*. Primarily, we are concerned with *distributional robustness*: the shape of the true underlying distribution deviates slightly from the assumed model (usually the Gaussian law).

In addition to Huber, Hampel et al. (1986) also offered several definitions of ‘robustness’.

Hampel et al. in 1986 stated, “robust statistics is a body of knowledge...relating to deviations from idealized assumptions in statistics.” He and his co-authors made reference to a topic that is more of concern to us: “Robust statistics in the broad, vague, informal sense obviously encompasses rejection of outliers, although this field seems to lead an isolated life of its own, and only in recent years do some specialists for the rejection of outliers appreciate the close natural relationship (cf. Barnett and Lewis, 1978).”

Hampel et al. (1986) suggested that the two goals of robust statistics should consist of using the majority of the data and detecting outliers and highly influential points, which they denote as leverage points. They stated that identifying outliers can be problematic when the dataset is beyond two dimensions and when visual inspection is unreliable and sometimes no longer feasible. Furthermore, the greater the number of dimensions and the more complex the dataset, then the more one should rely on the safety

of robust methods instead of on simple methods like OLS (Hampel et al., 1986). Hampel et al. (1986) declared that identifying outliers can sometimes be difficult when typical regression diagnostics are being used and can be less difficult when robust regression methods are being used.

For the moment, it suffices to note that: (1) a single huge unnoticed gross error can spoil a statistical analysis completely (as in the case of least squares); (2) several percent gross errors are rather common; and (3) modern robust techniques can deal with outliers relatively easily, even better than classical methods for objective or subjective rejection of outliers (Hampel et al., 1986).

Note that “gross errors” are errors related to incorrect measurements and/or data entry inaccuracies.

Several statistical techniques have been developed that serve as contributions to the field of robust regression. However, only a few will be presented within this dissertation, including those authors whose work serves as a foundation for our proposed method and whose work is to an extent comparable to ours. We present methods by Hampel, Edgeworth, Huber, Rousseeuw, Hadi, Hadi and Simonoff, and Atkinson and Riani. Special attention should be given to Hadi (1992, 1994), Hadi and Simonoff (1993), and Atkinson and Riani (2000), whose contributions to robust regression are similar to but different from certain aspects of our proposed method.

Influence and Breakdown

Hampel’s (1968) dissertation contributed to measures of robustness: the influence function and the breakdown point. The influence function (IF), formerly known as the influence curve, determines the amount of bias that an outlier causes. In other words, this function evaluates the asymptotic behavior of the estimator. The breakdown point

indicates the amount of contamination (i.e., outliers) that an estimator can withstand without “breaking down” or becoming unreliable; it reflects the sensitivity of a method to outliers. Based on the work of Hampel, many statisticians such as Rousseeuw (1984, 1985) began to focus on estimators with a high breakdown point.

L₁ Regression

Robust regression dates back to the 1800’s. Specifically, in 1887, Edgeworth proposed the least absolute values (L₁) regression method:

$$\text{Min}_{\hat{\beta}} \sum_{i=1}^n |y_i - \hat{y}_i|.$$

This method minimizes, with respect to the estimated regression parameter $\hat{\beta}$, the sum of the absolute value of the residuals whereas OLS minimizes the sum of the squared residuals. This estimator obtained from L₁ regression is robust to outliers in linear regression. It is known that, in comparison to the OLS approach, L₁ regression has a higher breakdown point. This difference implies that the L₁ estimator is more robust toward outliers than OLS is found to be. Unfortunately, L₁ regression is not protective against or robust to the impact of leverage points.

M-estimation

Huber is attributed for the development of several components such as the minimax approach, the gross-error model, and maximum likelihood type estimates (better known as M-estimators). However, only the M-estimators are reviewed in this dissertation. Huber (1964) defined an M-estimate as being any estimate T_n such

that $\sum \psi(x_i; T_n) = 0$, where $\psi(x; \theta) = (\partial / \partial \theta) \rho(x; \theta)$ and where $\rho(\cdot)$ is some arbitrary function. For the specific case of linear regression, Huber noted that the M-estimate could be derived from

$$\text{Min}_{\hat{\beta}} \sum_{i=1}^n \rho(y_i - \hat{y}_i).$$

This approach is similar to OLS but replaces the residuals with an arbitrary but symmetric function of the residuals, $\rho(\cdot)$. This estimator is insensitive to outliers but fails to be robust in the presence of leverage points.

LMS Estimator

Rousseeuw recognized the problems associated with OLS, such as outliers and the masking effect. After Huber's M-estimators and Hampel's influence function and breakdown point were developed, Rousseeuw introduced estimators that are also robust to outliers. In 1984, he developed the least median squares (LMS) estimator:

$$\text{Min}_{\hat{\beta}} \text{med}_i (y_i - \hat{y}_i)^2.$$

Instead of the sum, the median of the squared residuals is minimized with respect to $\hat{\beta}$.

This estimator is thought to be robust to outliers. However, it performs poorly in terms of statistical asymptotic efficiency because its convergence rate to its limiting distribution is slow.

LTS Estimator

To improve the poor performance of the LMS estimator, Rousseeuw (1985) introduced the use of the least trimmed squares (LTS) estimator:

$$\text{Min}_{\hat{\beta}} \sum_{i=1}^h ((y_i - \hat{y}_i)^2)_{i:n}.$$

This estimator indicates that the sum of the h smallest squared residuals is minimized with respect to $\hat{\beta}$. As suggested by Rousseeuw, the lower bound on h should be $[n/2] + 1$. This estimator is robust to outliers also.

Like Huber (1981) and Hampel et al. (1983), Rousseeuw and Leroy (1987) offered a definition of robust regression. In 1987, they stated, "...robust regression...tries to devise estimators that are not so strongly affected by outliers." In this dissertation, we are concerned about robustness toward influential observations, as well as toward outliers and leverage points. Hence, we model our research after Rousseeuw's (1985) concept of least trimmed squares and after Rousseeuw and van Zomeren's (1990) concept of robust distances using the minimum covariance determinant. However, before we define their concepts, we give a brief overview of the robust regression approaches by Hadi (1992, 1994), Hadi and Simonoff (1993), and Atkinson and Riani (2000).

Hadi (1992, 1994)

Hadi's (1992, 1994) proposed multivariate approach for the detection of outliers is as follows:

Step 0: Initial Ordering – Compute C_M , which is the vector consisting of the co-ordinatewise medians and $S_M = (n-1)^{-1} \sum_{i=1}^n (x_i - C_M)(x_i - C_M)'$. Arrange the data in ascending order based on the robust distances:

$$D_i(C_M, S_M) = [(x_i - C_M)' S_M^{-1} (x_i - C_M)]^{1/2}$$

for $i = 1, \dots, n$. Next, assign weights such that $w_i = 1$ if $i \leq \text{integer part of } (n+p+1)/2$ and such that $w_i = 0$ otherwise. Compute C_R and S_R , where

$$C_R = \left(\sum_{i=1}^n w_i \right)^{-1} \sum_{i=1}^n w_i x_i$$

and where

$$S_R = \left(\sum_{i=1}^n w_i - 1 \right)^{-1} \sum_{i=1}^n w_i (x_i - C_R)(x_i - C_R)'$$

Again, arrange the data in ascending order based on the robust distances but this time utilizing C_R and S_R such that

$$D_i(C_R, S_R) = [(x_i - C_R)' S_R^{-1} (x_i - C_R)]^{1/2} \text{ for } i = 1, \dots, n.$$

Step 1: Basic Subset – The ascending dataset based on C_R and S_R is divided into two subsets such that the first subset contains the first $p+1$ observations and such that the second contains the remaining $n-p-1$ observations. Hadi (1992) referred to the first set as the “basic” subset and the second subset as “non-basic”. Considering only full rank cases, compute another robust distance:

$$D_i(C_b, S_b) = [(x_i - C_b)' S_b^{-1} (x_i - C_b)]^{1/2} \text{ for } i = 1, \dots, n,$$

(5)

where C_b and S_b are the mean and covariance matrix, respectively, for the “basic” subset.

Step 2: Size of Basic Subset Increased – The observations are rearranged in ascending order based on (5). The number of observations in the “basic” subset is denoted as r . An observation is added to the “basic” subset so that the new “basic” subset contains the first $r+1$ observations. The other subset of the data contains the remaining $n-r-1$ observations.

Step 3 – Step 1 and Step 2 are repeated until the “basic” subset has h observations, where $h = [(n+p+1)/2]$; that is, h is the integer part of $(n+p+1)/2$.

Step 4 – The data are rearranged according to a new set of robust distances as seen in (5). Now, however, C_b and S_b are the mean and covariance matrix, respectively, for the new “basic” subset. Let $D_{(r+1)}^2$ be the $(r+1)^{\text{th}}$ order statistic with the modification such that S_b is multiplied by the correction factor

$$c = \left(1 + \frac{p+1}{n-p} + \frac{1}{n-h-p} \right)^2.$$

If the $D_{(r+1)}^2 \geq \chi_{p,\alpha/n}^2$, then this process should be stopped, and all observations with the modified robust distance (obtained by using the correction factor) greater than $\chi_{p,\alpha/n}^2$ are considered outliers. If not, then proceed to Step 5. Note that $\chi_{p,\alpha}^2$ is the $(1 - \alpha)$ percentile of the chi-square distribution with p degrees of freedom.

Step 5 – The observations are again divided into two subsets in which the first subset contains $r+1$ observations and the second contains the remaining $n-r-1$ observations. If $n = r+1$, stop, and proclaim that there are no outliers in the data. Otherwise, return to Step 4.

MI Algorithm

Hadi and Simonoff (1993) proposed an algorithm claimed to detect and test for multiple outliers in linear models. The procedure begins with applying OLS to the entire dataset. Thereafter, observations are ranked by some appropriately selected regression diagnostic tool. The data are divided into two subsets such that the initial subset consists of the first h observations, which is assumed to be “clean” or free of outliers. A new

regression model is fitted to the h observations in the initial subset. On the basis of this fit, the studentized residuals are computed and arranged in ascending order. The size of the initial subset is increased by one, and the $(h + 1)^{\text{th}}$ studentized residual is compared to $t(\alpha/2(h+1), h - p)$. If the studentized residual is found to be larger than the t statistic then testing is stopped, and all remaining observations are declared outliers. Otherwise, this process is continued until the first outlier is identified. Pena and Yohai (1999) noted, “According to a Monte Carlo study...the procedure works well for low-leverage outliers but may fail when the sample contains a set of high-leverage outliers.”

Forward Search

Atkinson and Riani (2000) utilized a “forward search” approach. They focused on parameter estimation and the removal of outliers. Atkinson and Riani (2000) stated,

...the emphasis is on using the forward search to find a single set of parameter estimates and of outliers. The emphasis in this book is very different: at each stage of the forward search we use information such as parameter estimates and residual plots to guide us to a suitable model.

The starting point of their approach uses the estimate obtained from Rousseeuw’s (1984) LMS estimator. The data are then divided into two sections such that the first is intended to be free of outliers and such that the second may contain outliers. The initial subset, which is thought to be free of outliers, contains m observations. The residuals are squared and then arranged in ascending order. The approach then moves forward by adding observations to the initial subset on the basis of the smallest squared residuals. (Note that they sometimes utilize the raw residuals.) This approach continues until all observations, which are not considered potential outliers, enter into the subset. Atkinson and Riani (2000) monitored the search by using the MSE and observing changes in the parameter

estimates. This method is known to be capable of detecting outliers. However, as Pena and Yohai (1999) mentioned in their article, it is not robust toward influential observations.

Many other robust estimators and techniques have been developed to accommodate the diagnostic issues (outliers, leverage, influence, masking, and swamping) that OLS does not overcome. The purpose of this dissertation is to propose a robust statistical technique that will be capable of detecting and limiting the impact of potential outliers, leverage points, and influential observations in linear regression. The objective discussed in Paper 1 was to develop the Robust Forward Detection (RFD) method that will properly detect atypical observations (outliers, leverage points, and influence), as well as overcome the masking and swamping effects. The RFD is an extension of and utilizes a combination of Rousseeuw's (1985) concept of least trimmed squares (LTS) and Rousseeuw and van Zomeren's (1990) concept of robust distances, using the minimum covariance determinant. In Paper 2, we downweight and limit the impact of the atypical observations identified by the detection tool proposed in Paper 1. We illustrate the sensitivities of OLS and the robustness of the RFD method by comparing the following methods: OLS, RFD with dichotomous weights of 0 and 1, and RFD with continuous weights ranging between 0 and 1, inclusive. An application of the RFD method using the Lung Health Study I data is demonstrated in Paper 3. The method is implemented to identify any potential atypical observations (outliers, leverage, and influential points). Thereafter, the RFD method with dichotomous and continuous weighting functions is applied. The results are compared to those obtained by using OLS. In addition, the RFD method is utilized to determine whether any study participants

identified as an outlier, leverage point, or influential observation in either the response or the predictors are more likely to have died of lung cancer during the five year follow-up period.

A ROBUST FORWARD DETECTION METHOD FOR THE IDENTIFICATION
OF INFLUENTIAL OBSERVATIONS IN LINEAR REGRESSION

by

TAMEKIA L. JONES AND DAVID T. REDDEN

In preparation for *Computational Statistics & Data Analysis*

Format adapted for disseration

Abstract

Diagnostic issues such as outliers, leverage points, influential observations, and masking and swamping effects are often encountered in linear regression when data are analyzed by using ordinary least squares (OLS). Robust regression methods are recommended for proper detection and management of these diagnostic issues. Although many robust regression methods are resistant to the effects of outliers or leverage points, the methods remain susceptible to influential points. We present a Robust Forward Detection (RFD) method within the framework of linear regression that is resistant to the effects of outliers, leverage points, and influential observations. Our RFD method is based on the concepts of robust distances, the minimum covariance determinant, least trimmed squares, and nearest neighbor multiple testing. The properties of the proposed method are evaluated via simulations. Examples are used to further illustrate both the robustness of the RFD method toward diagnostic issues and the method's ability to overcome the masking and swamping effects.

1. Introduction

The general linear model as applied in linear regression is $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where \mathbf{y} is an $n \times 1$ vector of the observed response values; \mathbf{X} is an $n \times p$ matrix of the predictors, including the intercept; $\boldsymbol{\beta}$ is a $p \times 1$ vector of the unknown regression parameters; $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of the unobservable random errors; and $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$. The random errors, $\boldsymbol{\varepsilon}$, are estimated by the residuals:

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}, \quad (1)$$

where $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\beta}}$ is the estimate for the unknown parameter $\boldsymbol{\beta}$, and $\hat{\mathbf{y}}$ is the vector containing the predicted values. As illustrated in equation (1), the residuals measure the difference between the observed values and the predicted values.

The most common approach used to estimate $\boldsymbol{\beta}$ in the general linear model is ordinary least squares (OLS). When the OLS method is applied in the linear regression framework, several diagnostic issues are often encountered: outliers, leverage points, influential observations, masking, and swamping. Outliers are observations with extreme standardized deviations between the observed dependent variable and the predicted value. Leverage is a measure of the standardized deviation of an observation's row vector of independent variables from the mean vector of the independent variables. An observation that is both an outlier and a leverage point is declared an influential observation. Detection of these extreme observations (outliers, leverage, and influential points) can be difficult because of masking and swamping. Masking is the effect of not being able to identify all true atypical observations because the extremeness of one observation obscures the extremeness of another. On the other hand, swamping occurs when one or more observations are incorrectly identified as atypical.

To address these diagnostic issues, we, like other statisticians, suggest utilizing a robust regression method. Unlike many of the current robust regression techniques, the robust statistical approach that we propose is robust not only to outliers but also robust to leverage and influential observations. In this paper, we develop a robust statistical approach modeled after Rousseeuw's (1985) concept of least trimmed squares (LTS) and after Rousseeuw and van Zomeren's (1990) concept of robust distances (RD), using the minimum covariance determinant (MCD). We first present Mahalanobis distance (MD), which is the foundation of RD. MD is a standardized measure of the distance of each independent row vector from the mean vector. Because there is a monotonic relationship between MD and h_{ii} (the diagonal elements of the hat matrix $\mathbf{H}=\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$), MD may be used to assess leverage:

$$h_{ii} = \frac{(\text{MD}_i)^2}{n-1} + \frac{1}{n}, \quad (2)$$

where $\text{MD}_i = [(\mathbf{x}_i - \mathbf{T}(\mathbf{X}))\mathbf{C}(\mathbf{X})^{-1}(\mathbf{x}_i - \mathbf{T}(\mathbf{X}))']^{1/2}$ and $\mathbf{x}_i = (x_{i1} \ x_{i2} \ \dots \ x_{i,p-1})$ for $i = 1, 2, \dots, n$.

Note that $\mathbf{T}(\mathbf{X})$ and $\mathbf{C}(\mathbf{X})$ are the mean vector and sample covariance matrix, respectively.

That is, $\mathbf{T}(\mathbf{X}) = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i$, and $\mathbf{C}(\mathbf{X}) = (n-1)^{-1} \sum_{i=1}^n (\mathbf{x}_i - \mathbf{T}(\mathbf{X}))(\mathbf{x}_i - \mathbf{T}(\mathbf{X}))'$. By using

equation (2) and by using Hoaglin and Welsch's (1978) recommendation that any observation with a diagonal element $h_{ii} > 2p/n$ should be considered a leverage point, we can establish the following:

$$\begin{aligned} h_{ii} &> 2p/n \\ \frac{(\text{MD}_i)^2}{n-1} + \frac{1}{n} &> \frac{2p}{n} \\ \frac{(\text{MD}_i)^2}{n-1} &> \frac{2p-1}{n} \end{aligned}$$

$$(\text{MD}_i)^2 > \frac{(2p-1)}{n} (n-1) . \quad (3)$$

The MD suffers from the masking effect because of the use of non-robust estimators, the mean and the covariance matrix. We, therefore, know that h_{ii} is not always an effective measure for assessing leverage. Because of this fact, Rousseeuw and van Zomeren (1990) introduced the concept of RD, which is the standardized distance of each independent row vector \mathbf{x}_i from the weighted mean vector:

$$\text{RD}_i = [(\mathbf{x}_i - \mathbf{T}_w(\mathbf{X}))\mathbf{C}_w(\mathbf{X})^{-1}(\mathbf{x}_i - \mathbf{T}_w(\mathbf{X}))']^{1/2}, \quad (4)$$

with weighted mean vector

$$\mathbf{T}_w(\mathbf{X}) = \left(\sum_{i=1}^n w_i \right)^{-1} \sum_{i=1}^n w_i \mathbf{x}_i$$

and with weighted covariance matrix

$$\mathbf{C}_w(\mathbf{X}) = \left(\sum_{i=1}^n w_i - 1 \right)^{-1} \sum_{i=1}^n w_i (\mathbf{x}_i - \mathbf{T}_w(\mathbf{X}))(\mathbf{x}_i - \mathbf{T}_w(\mathbf{X})).$$

The weighted mean $\mathbf{T}_w(\mathbf{X})$ and weighted covariance matrix $\mathbf{C}_w(\mathbf{X})$ from the RD are estimated by using Rousseeuw's (1984, 1985) originally proposed MCD, which is computed by randomly and repeatedly selecting at least half of the data and choosing the subset that has the smallest determinant of the covariance matrix. Rousseeuw and van Zomeren (1990) determined that an observation should be considered a leverage point if $\text{RD}_i > \chi_{p, 1-\alpha}^2$.

Our proposed RFD method utilizes familiar diagnostic tools, unlike Rousseeuw and van Zomeren's (1990) threshold for the robust distances, but transformed so they fit in the robust regression statistical framework. Rousseeuw and van Zomeren (1990) use a

$\chi^2_{p,1-\alpha}$ cutoff value to identify leverage points, which suggests that either \mathbf{X} or the RD have an underlying normal distribution. Of course, neither possibility is likely because the $n \times p$ matrix \mathbf{X} does not consist of random variables and therefore does not carry distributional assumptions. On the other hand, we compare the robust distances to the familiar but transformed regression leverage diagnostic rules $2p/n$ and $3p/n$.

We present our proposed method in section 2. In section 3, we present simulations to evaluate the performance of our method. In addition, we compare our method to LTS (ROBUSTREG), a procedure utilized in SAS[®] version 9.1.3, because LTS (ROBUSTREG) utilizes a combination of LTS (Rousseeuw, 1985) and RD's (Rousseeuw and van Zomeren, 1990) and because it is most similar to our proposed method. Last, in section 4, we illustrate our proposed method with applications based on two published datasets. Discussions are presented in section 5, and concluding remarks and future research are provided in section 6.

2. Method

On the basis of Rousseeuw and van Zomeren's (1990) RD and MCD concepts, we propose the Robust Forward Detection (RFD) approach to detect and properly manage diagnostic issues often encountered in linear regression when OLS is being used.

Step 1. Select a subset of m observations from the \mathbf{X} -space by using the minimum covariance determinant in order to obtain a leverage-free subset.

This first step of our method provides an initial subset free of leverage because we begin with an examination of our data with respect only to the \mathbf{X} -space and not to

y. Use the MCD to select m observations, such that $m \geq \left\lceil \frac{n}{2} \right\rceil + 1$ and $\lceil \cdot \rceil$ is the least integer, until the method finds the subset that has the smallest determinant of the covariance matrix. At this step, the weights w_i for $T_w(\mathbf{X})$ and $C_w(\mathbf{X})$ from equation (4) are determined:

$$w_i = \begin{cases} 0, & \text{if the } i^{\text{th}} \text{ point is excluded from the subset of } m \text{ observations} \\ 1, & \text{if the } i^{\text{th}} \text{ point is included within the subset of } m \text{ observations.} \end{cases}$$

Step 2. Order the data according to the robust distances in \mathbf{X} .

Note that, once step 1 is completed, the chosen subset is expected to be free of leverage points because the MCD yields the subset of the data consisting of m observations that are most compact in the \mathbf{X} -space.

2.1. For the remaining $n - m$ observations, compute RD's in equation (4) for each observation by setting $w_i = 1$ for the i^{th} observation for which the RD_i is being calculated. Then reset $w_i = 0$.

2.2. Arrange the robust distances RD_i in ascending order.

Step 3. Conduct nearest neighbor multiple testing for leverage.

The nearest neighbor is the j^{th} observation ($j = m+1, m+2, \dots, n$) that has the smallest RD_i of all of the observations excluded from the chosen subset (i.e., observations with $w_i = 0$). Conduct multiple testing by moving forward from the chosen subset (m^{th} observation) to the nearest neighbor ($(m+1)^{\text{th}}$ observation), and test whether the nearest neighbor is a leverage point.

3.1. Use a robust version of equation (3), to assess whether the nearest neighbor is a leverage point, by replacing MD_i with RD_i and by replacing n with $\sum_{i=1}^n w_i$. The nearest neighbor is considered a leverage point if

$$(RD_i)^2 > \frac{(2p-1)}{\sum_{i=1}^n w_i} \left(\sum_{i=1}^n w_i - 1 \right). \quad (5)$$

Instead of utilizing the $2p/n$ rule for leverage, transform the $3p/n$ rule (discussed later in section 3) such that

$$(RD_i)^2 > \frac{(3p-1)}{\sum_{i=1}^n w_i} \left(\sum_{i=1}^n w_i - 1 \right). \quad (6)$$

3.2. If the nearest neighbor is identified as a leverage point, then conclude that the j^{th} observation and all remaining observations are leverage points, and set $w_j = 0$ for the j^{th} observation and all remaining observations. Otherwise, set $w_j = 1$, repeat steps 2 and 3 until all observations have been evaluated.

Step 4. Use the LTS method to select a subset of h observations among the subset of non-leveraged points from the data in order to have a subset free of outliers, leverage, and influence.

At this point, we now begin examining the data with respect to the standardized deviation of $\mathbf{y} - E[\mathbf{y}|\mathbf{X}]$. Use the LTS (least trimmed squares) method to select h observations among the non-leveraged points (i.e., where $w_i = 1$), such that $h \geq$

$\left\lceil \frac{1}{2} \sum_{i=1}^n w_i \right\rceil + 1$, until the method finds the subset for which the fit minimizes the

least squares function. Initialize weights v_i such that

$$v_i = \begin{cases} 0, & \text{if the } i^{\text{th}} \text{ point is included in the subset of } h \text{ observations} \\ 1, & \text{if the } i^{\text{th}} \text{ point is excluded from the subset of } h \text{ observations.} \end{cases}$$

Define \mathbf{V} as an $n \times n$ diagonal matrix consisting of the weights v_i along the diagonal. At this step, the subset of h points is expected to be free of outliers, leverage points, and influential observations.

Step 5. Conduct nearest neighbor multiple testing for outliers.

5.1. Fit a linear regression model based on the subset of data for which $v_i = 1$.

5.2. Let $\mathbf{x}_k = (1 \ x_{k1} \ x_{k2} \ \dots \ x_{k,p-1})$, $\hat{y}_k = \mathbf{x}_k \tilde{\mathbf{b}}$, $\tilde{\mathbf{b}} = (\mathbf{X}' \mathbf{V} \mathbf{X})^{-1} \mathbf{X}' \mathbf{V} \mathbf{y}$, $v_{\text{sum}} = \left(\sum_{i=1}^n v_i \right)$,

$$\text{MSE}_{\mathbf{V}} = (\mathbf{y} - \mathbf{X} \tilde{\mathbf{b}})' \mathbf{V} (\mathbf{y} - \mathbf{X} \tilde{\mathbf{b}}) / (v_{\text{sum}} - p), \text{ and } d_k = \frac{y_k - \hat{y}_k}{\sqrt{\text{MSE}_{\mathbf{V}}}}. \text{ Compute}$$

standardized prediction residuals d_k for all observations excluded from the subset for which $v_i = 0$.

5.3. The nearest neighbor is the k^{th} observation ($k = h+1, h+2, \dots, n$) that has the smallest standardized prediction residual d_k of all the observations excluded from the chosen subset (i.e., observations with $v_i = 0$). Compute the $100(1-\alpha)\%$ adjusted prediction interval for y_k , the k^{th} observation using

$$\mathbf{h}_k = \mathbf{x}_k (\mathbf{X}' \mathbf{V} \mathbf{X})^{-1} \mathbf{x}_k' :$$

$$\hat{y}_k \pm t \left(1 - \frac{\alpha}{2(v_{\text{sum}} + 1)}, v_{\text{sum}} - p \right) (\text{MSE}_{\mathbf{V}} * (1 + \mathbf{h}_k))^{1/2}.$$

$$\hat{y}_k - t \left(1 - \frac{\alpha}{2(v_{\text{sum}} + 1)}, v_{\text{sum}} - p \right) (\text{MSE}_{\mathbf{V}} (1 + \mathbf{h}_k))^{1/2} < y_k < \hat{y}_k + t \left(1 - \frac{\alpha}{2(v_{\text{sum}} + 1)}, v_{\text{sum}} - p \right) (\text{MSE}_{\mathbf{V}} (1 + \mathbf{h}_k))^{1/2},$$

$$-t \left(1 - \frac{\alpha}{2(v_{\text{sum}} + 1)}, v_{\text{sum}} - p \right) (\text{MSE}_{\mathbf{V}} (1 + \mathbf{h}_k))^{1/2} < y_k - \hat{y}_k < t \left(1 - \frac{\alpha}{2(v_{\text{sum}} + 1)}, v_{\text{sum}} - p \right) (\text{MSE}_{\mathbf{V}} (1 + \mathbf{h}_k))^{1/2},$$

$$-t\left(1 - \frac{\alpha}{2(v_{\text{sum}} + 1)}, v_{\text{sum}} - p\right)(1 + h_k)^{1/2} < \frac{y_k - \hat{y}_k}{\text{MSE}_V} < t\left(1 - \frac{\alpha}{2(v_{\text{sum}} + 1)}, v_{\text{sum}} - p\right)(1 + h_k)^{1/2},$$

$$-t\left(1 - \frac{\alpha}{2(v_{\text{sum}} + 1)}, v_{\text{sum}} - p\right)(1 + h_k)^{1/2} < d_k < t\left(1 - \frac{\alpha}{2(v_{\text{sum}} + 1)}, v_{\text{sum}} - p\right)(1 + h_k)^{1/2}.$$

5.5. Conduct multiple testing by moving forward from the chosen subset (h^{th} observation) to the nearest neighbor ($(h+1)^{\text{th}}$ observation), and test whether the nearest neighbor is an outlier. If the standardized prediction residual d_k for the nearest neighbor, which is the k^{th} observation, is beyond the lower and upper bounds seen above, then v_k remains equal to zero and the nearest neighbor is an outlier. Otherwise, update the diagonal weight matrix \mathbf{V} such that $v_k = 1$. Repeat step 5 until all points have been evaluated.

Step 6. Note that, if the observation is identified as both a leverage point ($w_k = 0$) and an outlier ($v_k = 0$), it is declared an influential observation on the basis of the aforementioned steps.

3. Simulation results

3.1. Leverage rules: $2p/n$ versus $3p/n$

We now address the issue of identifying leverage by utilizing equations (5) and (6), which are associated with the $2p/n$ and the $3p/n$ rules, respectively. Belsley, Kuh, and Welsch (1980) stated, “For small p , $2p/n$ tends to call a few too many points to our attention.” Velleman and Welsch (1981) suggest using the $3p/n$ rule to identify leverage points when $p > 6$ and $(n - p) > 12$. In general, the decision of whether to use the $2p/n$ versus the $3p/n$ rule should be decided on the basis of the ratio of the number of parameters to the sample size. Figure 1 illustrates this point.

Figure 1 depicts the proportion of leverage points identified within a dataset on the basis of a simulation of 1000 iterations per n and p . The data were simulated from a multivariate normal distribution, and the correlation among each variable was set to 0.25. An observation was identified as a leverage point if the diagonal element of the hat matrix $h_{ii} > 3p/n$. When $n = 30$ and $p < 4$, Figure 1 suggests that 8% of the data is the maximum proportion of leverage points expected when the $3p/n$ rule is being used. On the other hand, when the same n and p criteria are considered, 20% – 26% of the data are detected as leverage points under the $2p/n$ rule. The simulation results suggest that, on average, at least six observations are classified as leverage points if utilizing the $2p/n$ rule. The results from the $2p/n$ rule (not shown) identify more observations as leverage points when p is small.

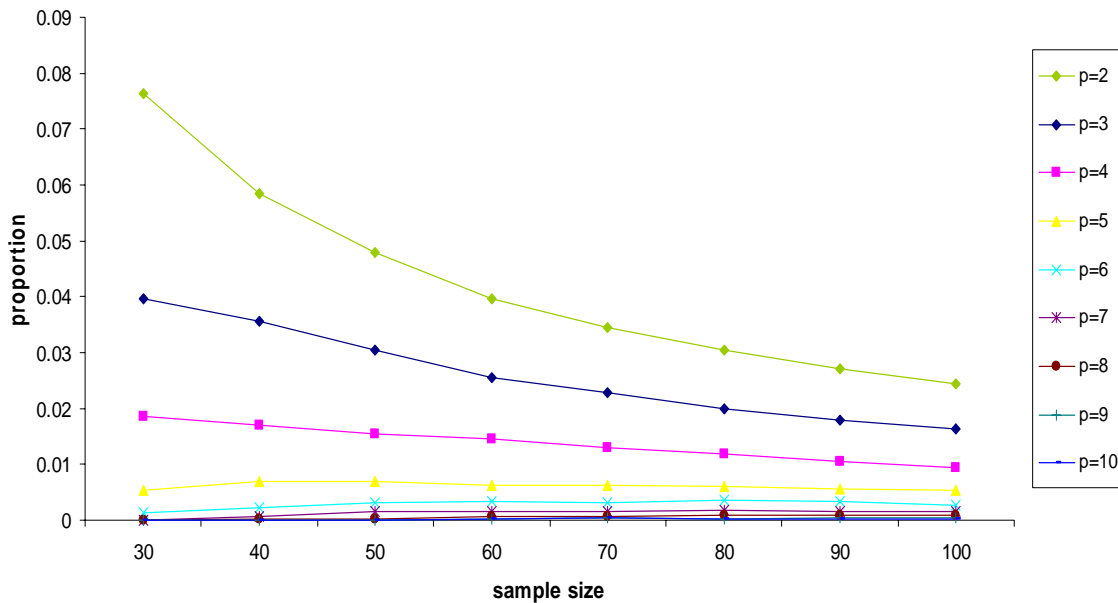


Figure 1. Proportion of leverage identified on the basis of 1000 iterations per n and p when the $3p/n$ rule is used.

3.2. Evaluation of the RFD method

Simulations were performed to evaluate the proposed RFD method described in section 2 and to compare this method to the LTS (ROBUSTREG) procedure as presented in SAS[®] version 9.1.3. The MCD was computed in SAS[®] by using Rousseeuw and van Driessen's (1999) FAST-MCD algorithm to aid in the computation of the RD's for the RFD method. The LTS method that we utilize for the RFD method is based on Rousseeuw and van Driessen's (2006) FAST-LTS algorithm. The LTS (ROBUSTREG) procedure in SAS[®] also utilizes a combination of Rousseeuw and van Driessen's (1999) FAST-LTS and Rousseeuw and van Driessen's (2006) FAST-MCD algorithms. For the LTS (ROBUSTREG) procedure, the default cutoff value to identify outliers was ± 3 standard deviations below/above the value of the residuals. The default cutoff value to identify leverage points for LTS (ROBUSTREG) was $\chi^2_{p, 1-\alpha}$, with $\alpha = 0.025$. To make valid comparisons between the RFD and LTS (ROBUSTREG) approaches, we programmed both methods so that $[n/2]+1$ observations were selected for the FAST-MCD algorithm.

Both methods were examined under four conditions ([1] no contamination, [2] outliers, [3] leverage, [4] influence) and four different contamination levels (10%, 20%, 30%, 40%). We generated 100 observations for each of the 1000 iterations from the linear regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$; $\boldsymbol{\beta} = (25 \ 2 \ 2 \ 2)'$; $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3 \sim N(0,1)$; and $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{I})$. For scenario [2], we generated 10% outliers by increasing the errors by 6 units. To ensure that we were generating outliers, we specified $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ to be distributed as $U(-1,1)$ for the same subset of the data. For scenario [3], each predictor was increased by 5 units to simulate leverage for 10% of the data. To ensure that influence was being generated for

scenario [4] with a contamination level of 10%, we shifted 10% of the observations to the right in \mathbf{X} (creating leverage), and we shifted those same observations downward in \mathbf{y} (creating outliers); this shifting guaranteed placement of the contaminated observations below the fitted regression line which is based on the majority of the data. Note that, when generating influence, we specified $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ to be distributed as $U(0,1)$. We continued in a similar manner for the remaining contamination levels for scenarios [2] to [4].

The simulation results are presented in Tables 1, 2, and 3. As shown in Table 1, when a dataset contains no simulated outliers or leverage points, the RFD method, on average, identifies $0.05\% \pm 0.22\%$ (mean \pm standard deviation) of the observations as outliers and $1.24\% \pm 1.31\%$ of the observations as leverage points. On the other hand, the LTS (ROBUSTREG) procedure identifies 0.99% (standard deviation = $\pm 1.28\%$) of the observations as outliers and 7.74% (standard deviation = $\pm 4.04\%$) as leverage points. The remainder of Table 1 and Table 2 can be interpreted in a similar manner. Table 3 demonstrates the simulation results for the probability of the number of true/correct, masked, and swamped atypical observations identified by the RFD and LTS (ROBUSTREG) methods. Note that the leverage results presented in Tables 1, 2, and 3 for the RFD method are based on equation (6) because equation (5) tends to overestimate the mean proportion of leverage points identified for each level of contamination.

Table 1. Simulated outlier and leverage results for the RFD method using equation (6) and for the LTS (ROBUSTREG) procedure

Simulation Specification	RFD method, equation (6)		LTS (ROBUSTREG)	
	Outlier Mean % (Std Dev %)	Leverage Mean % (Std Dev %)	Outlier Mean % (Std Dev %)	Leverage Mean % (Std Dev %)
No contamination	0.05 (0.22)	1.24 (1.31)	0.99 (1.28)	7.74 (4.04)
10% outliers	9.92 (0.55)	1.69 (1.61)	10.55 (0.96)	8.86 (4.02)
20% outliers	19.10 (3.83)	2.33 (1.93)	20.09 (0.93)	9.94 (4.17)
30% outliers	26.77 (9.04)	3.16 (2.27)	28.88 (3.40)	11.33 (4.06)
40% outliers	32.99 (15.05)	3.89 (2.51)	21.09 (15.62)	12.16 (3.84)
10% leverage	0.04 (0.22)	11.09 (1.23)	1.21 (1.84)	15.22 (2.98)
20% leverage	0.04 (0.20)	20.97 (1.16)	1.00 (1.29)	23.30 (2.09)
30% leverage	0.04 (0.21)	30.82 (1.08)	0.95 (1.22)	31.93 (1.46)
40% leverage	0.04 (0.19)	40.66 (1.00)	0.95 (1.22)	41.15 (0.99)

Note: Std Dev = Standard Deviation.

As indicated in Tables 1 to 3, our RFD method is comparable to and even outperforms the LTS (ROBUSTREG) procedure in certain cases. When no true leverage exists, the LTS (ROBUSTREG) procedure tends to overestimate the proportion of leverage points more than our proposed RFD method (see Table 1). Table 2 shows that, in comparison with the LTS (ROBUSTREG) procedure, the RFD method is better at detecting influence at a level of 20% or more. This difference results from the fact that the LTS (ROBUSTREG) procedure masks many of the outliers among the influential observations, as indicated in Table 2. In addition, the LTS (ROBUSTREG) procedure masks almost half of the simulated outliers when the data are contaminated with 40% outliers and masks more than 70% of the simulated influential observations when the data are contaminated with 30% or more of influence (see Table 3).

Table 2. Simulation results for the RFD and LTS (ROBUSTREG) methods identifying influential observations

Simulation specification	RFD method using equation (6)			LTS (ROBUSTREG)		
	Influence Mean % (Std Dev %)	Outlier Mean % (Std Dev %)	Leverage Mean % (Std Dev %)	Influence Mean % (Std Dev %)	Outlier Mean % (Std Dev %)	Leverage Mean % (Std Dev %)
10% influence	9.97 (0.55)	10.03 (0.61)	10.01 (0.13)	9.74 (1.63)	10.34 (1.78)	12.56 (3.38)
20% influence	19.98 (0.63)	20.03 (0.68)	20.01 (0.13)	15.26 (8.44)	15.68 (8.31)	20.82 (1.66)
30% influence	29.97 (0.95)	30.02 (0.98)	30.01 (0.13)	7.77 (12.93)	8.16 (12.89)	30.17 (0.54)
40% influence	39.92 (1.79)	39.97 (1.81)	40.01 (0.15)	0.64 (3.80)	0.97 (3.86)	40.05 (0.24)

Note: Std Dev = Standard Deviation.

Table 3. Simulation results for the proportion of true, masked, and swamped observations obtained by utilizing the RFD and LTS (ROBUSTREG) methods

Simulation specification	RFD method using equation (6)			LTS (ROBUSTREG)		
	True Proportion (%)	Masked Proportion (%)	Swamped Proportion (%)	True Proportion (%)	Masked Proportion (%)	Swamped Proportion (%)
10% outliers	98.79	1.21	0.04	99.49	0.51	0.66
20% outliers	95.34	4.66	0.06	98.95	1.05	0.37
30% outliers	89.11	10.89	0.05	95.84	4.16	0.18
40% outliers	82.37	17.63	0.07	52.70	47.30	0.02
10% leverage	100.00	0.00	1.21	100.00	0.00	5.80
20% leverage	100.00	0.00	1.21	100.00	0.00	4.12
30% leverage	100.00	0.00	1.16	100.00	0.00	2.76
40% leverage	100.00	0.00	1.65	100.00	0.00	1.91
10% influence	99.70	0.30	0.00	97.21	2.79	0.03
20% influence	99.90	0.10	0.00	76.23	23.77	0.02
30% influence	99.90	0.10	0.00	25.79	74.21	0.003
40% influence	99.80	0.20	0.00	1.60	98.40	0.00

4. Examples

Two datasets, which have been previously used to illustrate the non-robustness of OLS, are utilized to compare our RFD method to LTS (ROBUSTREG). We begin with the Hawkins-Bradru-Kass (HBK) data and follow up with the Hertzsprung-Russell Star Cluster data. We specify the RFD method to utilize adjusted confidence and prediction intervals to identify outliers and to utilize equation (6) to identify leverage points for both examples. When the results based on the RFD method are presented, “interval” refers to

adjusted confidence intervals if $h \leq \left[\frac{1}{2} \sum_{i=1}^n w_i \right] + 1$ and adjusted prediction intervals

otherwise, as explained in section 2. Also, as previously defined in section 2, an adjusted residual is a studentized residual if the observation is in the initially chosen subset based on the FAST-LTS algorithm and a standardized prediction residual otherwise. Results from the LTS (ROBUSTREG) method are based on default cutoff values for leverage and outliers.

4.1. HBK data

Hawkins, Bradu, and Kass (1984) generated the HBK dataset. This dataset consists of 75 observations with three independent variables and one dependent variable. Hawkins, Bradu, and Kass (1984) intentionally constructed the data so that observations 1 to 10 are influential and observations 11 to 14 are leverage points. Furthermore, the data provide an example of the masking effect. Table 4 indicates which observations in the HBK data are identified as outliers, leverage points, and influential observations according to OLS, RFD, and LTS (ROBUSTREG). OLS suggests that observation 11 is an outlier, that observation 14 is a leverage point, and that observations 12 and 13 are

influential (i.e., both an outlier and a leverage point). Unlike OLS, our proposed RFD method and LTS (ROBUSTREG) properly identify observations 1 to 10 as influential and observations 11 to 14 as leverage points; these findings duplicate the simulations developed by Hawkins, Bradu, and Kass (1984).

Table 4. Comparison of methods identifying atypical observations for the HBK data

Method	Outliers	Leverage	Influential
OLS	11 – 13	12 – 14	12 – 13
RFD	1 – 10	1 – 14	1 – 10
LTS	1 – 10	1 – 14	1 – 10

Figure 2 illustrates the results of using the RFD approach with the adjusted residuals plotted against the rank of the robust distances. The Upper Bound and Lower Bound represent the upper and lower adjusted confidence/prediction bounds, respectively, for the intervals previously defined in section 2. As stated in step 5 in section 2, any observation with an adjusted residual beyond the bounds of its interval is identified as an outlier. Figure 2 shows observations 1 to 10 beyond the Upper Bound band.

On the basis of the sorted robust distances, observation 1 is the first to be identified as a leverage point by the RFD method. Any observation with a robust distance that exceeds that of observation 1 is also identified as a leverage point. Therefore, the RFD approach identifies observations 1 through 14 as leverage points on the basis of the robust distances and equation (6).

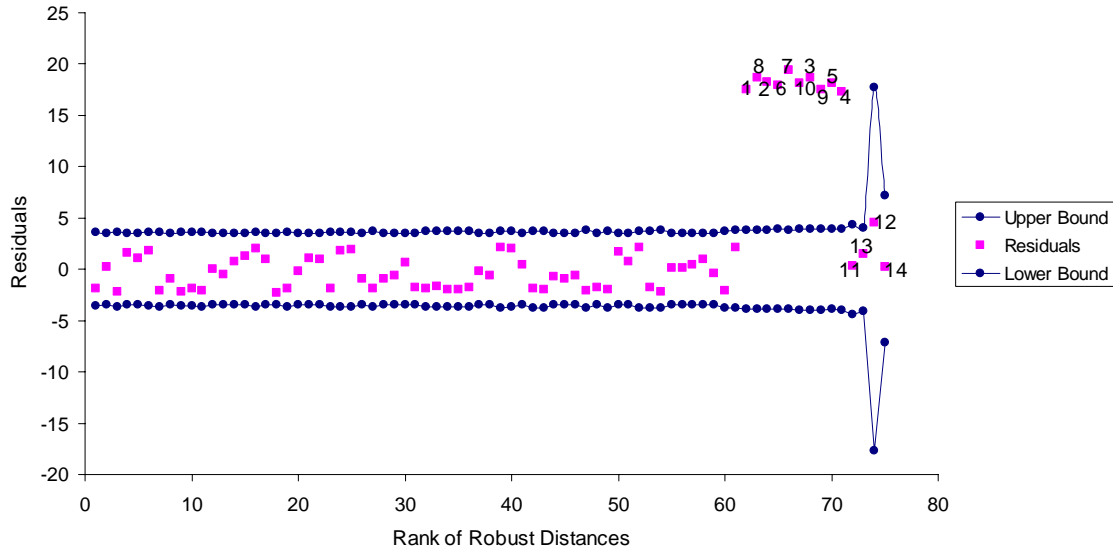


Figure 2. Plot of adjusted residuals vs. the rank of the robust distances for the HBK data.

4.2. Star Cluster CYG OB1 data

The Hertzsprung-Russell diagram of the Star Cluster CYG OB1 data (Rousseeuw and Leroy, 1987), containing 47 observations, describes the logarithm of the light intensity of the star (y) by using the logarithm of the effective temperature at the surface of the star (x). OLS identifies observations 11, 20, 30, and 34 as leverage points only (see Table 5). Unlike the RFD and LTS (ROBUSTREG) methods, OLS does not detect any outliers or influential observations within the dataset. Our RFD approach declares the same set of leverage points and influential observations as declared by the LTS (ROBUSTREG) procedure, with one exception; observation 7 is detected as an influential point by the LTS (ROBUSTREG) procedure but only as a leverage point by our RFD method.

5. Discussion

We propose a robust regression method for the proper detection and management of diagnostic issues often encountered in linear regression when OLS is being utilized. Many robust regression techniques are resistant to the effects of outliers or leverage points but remain susceptible to influential observations (Hadi and Simonoff, 1993; Pena and Yohai, 1999; Wisnowski et al., 2001). Our RFD method is based on the concepts of robust distances, the minimum covariance determinant, least trimmed squares, and nearest neighbor multiple testing. Note that we utilize adjusted confidence and prediction intervals. Prediction intervals are adjusted according to the size of the non-outlier subset (v_{sum}) plus one, which is due to the nearest neighbor (or the observation with the smallest standardized prediction residual outside the non-outlier subset) being tested as an outlier.

For the simulations in subsection 3.2 and the examples in section 4, we set $m =$

$[n/2] + 1$ in step 1 and $h = \left\lceil \frac{1}{2} \sum_{i=1}^n w_i \right\rceil + 1$ in step 4 of the RFD method. We wanted the

MCD to contain at least the majority of the data but no more than $[n/2] + 1$. By including more than our threshold, we would decrease the chance of the initially chosen subset being free of leverage points; however, recall that we want to begin our proposed RFD method with a subset free of leverage points in step 1 and later free of outliers in step 4. The default quantile or number of observations used in the minimization process for the MCD and LTS algorithms in LTS (ROBUSTREG) in SAS[®] is $((3n+p)/4)$. However, we specified $[n/2] + 1$ observations for minimization for LTS (ROBUSTREG) in order to make valid comparisons between LTS (ROBUSTREG) and our proposed RFD method. Again, LTS (ROBUSTREG) was the method selected for comparison purposes because it

is the method found to be most similar to our proposed RFD because of its use of a combination of robust distances and the least trimmed squares method.

As previously seen, with few exceptions, our proposed RFD method is comparable to that of LTS (ROBUSTREG) when atypical observations (outliers, leverage, and influential points) are being identified and when the masking and swamping effects are being revealed. The results in subsection 4.1 for the HBK data illustrate that the RFD method and LTS (ROBUSTREG) identify the same set of leverage points and influence. In subsection 4.2, in which the Star Cluster data are presented, the RFD and LTS (ROBUSTREG) yield the same results, with the exception of observation 7 being detected as an outlier when the LTS (ROBUSTREG) procedure is used.

However, there are instances when our RFD method outperforms the LTS (ROBUSTREG) procedure. The simulation results (see section 3), indicate that, in the presence of no contamination and in the presence of outlier contamination only, the LTS (ROBUSTREG) procedure tends to overestimate the proportion of leverage points more than what our RFD method does. In the presence of 40% outlier contamination, the LTS (ROBUSTREG) procedure breaks down quicker and masks more of the outliers than the RFD method. As we hoped, our proposed RFD method, in the presence of influential observations, performs better than the LTS (ROBUSTREG) procedure. Furthermore, the RFD method captures 99% of the influential observations for each specified level of contamination.

6. Concluding remarks and future research

In this paper, we were interested in developing a detection method capable of identifying atypical observations (outliers, leverage, and influential points) and overcoming the masking and swamping effects. Our RFD approach, as previously seen via simulations and examples, properly detects outliers and leverage points. In addition, the method we proposed, the RFD, is resistant to influential observations and overcomes both the masking and swamping effects in linear regression.

Further investigation of additional weights and weighting functions, such as those presented by Draper and Smith (1998), will be explored. We will implement weighting functions in order to downweight any observations declared atypical by the RFD method. Properties of the corresponding estimators will be examined.

References

- Belsley, D., Kuh, E., Welsch, R., 1980. Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. Wiley, New Jersey.
- Draper, N., Smith, H., 1998. Applied Regression Analysis. Wiley, New York.
- Gray, B., Woodall, W., 1994. The maximum size of standardized and internally studentized residuals in regression analysis. *The American Statistician*, 48, 111–113.
- Hadi, A., Simonoff, J., 1993. Procedures for the identification of multiple outliers in linear models. *Journal of the American Statistical Association*, 88, 1264–1272.
- Hawkins, D., Bradu, D., Kass, G., 1984. Location of several outliers in multiple-regression data using elemental sets. *Technometrics*, 26, 197–208.
- Hoaglin, D., Welsch, R., 1978. The hat matrix in regression and ANOVA. *The American Statistician*, 32, 17–22.
- Pena, D., Yohai, V., 1999. A fast procedure for outlier diagnostics in large regression problems. *The Journal of the American Statistical Association*, 94, 434–445.
- Rousseeuw, P., 1984. Least median of squares regression. *Journal of the American Statistical Association*, 79, 871–880.
- Rousseeuw, P., 1985. Multivariate estimation with high breakdown point. In: Grossmann, W., Pflug, G., Vincze, I., Wertz, W., Eds., *Mathematical Statistics and Applications*, Vol. B. Reidel, Dordrecht, 283–297.
- Rousseeuw, P., Leroy, A., 1987. Robust Regression and Outlier Detection. Wiley, New York.

Rousseeuw, P., van Driessen, K., 1999. A fast algorithm for the minimum covariance determinant. *Technometrics*, 41, 212–223.

Rousseeuw, P., van Driessen, K., 2006. Computing LTS regression for large datasets. *Data mining and Knowledge Discovery*, 12, 29–45.

Rousseeuw, P., van Zomeren, B., 1990. Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, 85, 633–639.

SAS Institute Inc., 2004. SAS Online[®] 9.1.3. SAS Institute Inc., North Carolina.

Velleman, P., Welsch, R., 1981. Efficient computing of regression diagnostics. *The American Statistician*, 35, 234–242.

Wisnowski, J., Montgomery, D., Simpson, J., 2001. A comparative analysis of multiple outlier detection procedures in the linear regression model. *Computational Statistics and Data Analysis*, 36, 351–382.

LIMITING THE IMPACT OF INFLUENTIAL OBSERVATIONS IN LINEAR
REGRESSION VIA WEIGHT FUNCTIONS

by

TAMEKIA L. JONES AND DAVID T. REDDEN

In preparation for *Computational Statistics & Data Analysis*

Format adapted for dissertation

Abstract

When utilizing ordinary least squares (OLS) within linear regression, statisticians often encounter diagnostic issues such as outliers, leverage points, and influential observations. These diagnostic issues can affect estimation of regression parameters and therefore can lead to poor estimation of the regression line, to incorrect inference, and to inaccurate predictions in linear regression when OLS is being used. In this paper, we utilize weight functions that allow unequal weighting of atypical observations identified by the Robust Forward Detection method. Simulations are conducted to evaluate performance of weighting functions, and bootstrapping procedures are utilized to make inferences about parameter estimates. As an application of the proposed weighting approach, two datasets well known to the robust statistical literature are presented. We illustrate that, without completely discarding any of the points, our proposed weighting approach allows robust estimation of parameters by downweighting any observations identified as atypical (outliers, leverage, or influential) by the Robust Forward Detection method.

1. Introduction

In linear regression, ordinary least squares (OLS) is typically utilized to estimate β in the general linear model $y = X\beta + \epsilon$, where y is an $n \times 1$ vector of the observed response values; X is an $n \times p$ matrix of the predictors, including the intercept; β is a $p \times 1$ vector of the unknown regression parameters; and ϵ is an $n \times 1$ vector of the unobservable random errors. However, it is well known that OLS is not robust toward outliers, leverage points, and influential observations and that OLS suffers from both masking and swamping effects. Outliers are observations with extreme standardized deviations of $y - E[y|X]$, where $E[y|X]$ is the expected value of y conditional upon X ; leverage points are those observations considered to be far away from the center of the data in the X -space; influential observations are those observations that are extreme in both $(y - E[y|X])$ and X (i.e., both an outlier and a leverage point). Masking occurs when some or all of the outliers, leverage points, and influential observations are “masked” or hard to detect via regression diagnostics; swamping occurs when observations are incorrectly identified as influential. Because they cause poor parameter estimation, these diagnostic issues can lead to inaccurate predictions and inaccurate inference.

Jones and Redden (2007) proposed a detection tool, the Robust Forward Detection (RFD) method, capable of identifying outliers, leverage points, and influential observations while overcoming the masking and swamping effects associated with OLS. However, it has remained unclear to investigators whether observations declared atypical should be deleted from the analysis and/or included in the analysis but with an appropriate warning detailing their potential impact. Some investigators use a form of weighted ordinary least squares (WOLS) that has the same assumptions as OLS, with the

exception that the variances of the error terms are no longer necessarily homogeneous; that is, $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{W}^{-1})$. Weights are either known or unknown. In the simple case in which weights w_i are known, it can be shown that

$$\hat{\boldsymbol{\beta}}_{\text{WOLS}} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1} \mathbf{X}'\mathbf{W}\mathbf{y} ,$$

where $\mathbf{W}_{n \times n} = \begin{bmatrix} w_1 & 0 & \dots & 0 \\ 0 & w_2 & \dots & 0 \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ 0 & 0 & \dots & w_n \end{bmatrix}$ is a diagonal matrix (Neter et al., 1996) and the distribution

of $\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1})$. If the weights are unknown, they can be estimated (Neter et al., 1996; Carroll and Cline, 1988; Draper and Smith, 1998; Davidian and Carroll, 1987; Carroll, 1982), mostly via iterative algorithms. Several weighting functions have already been proposed (see Hampel et al., 1986; Rousseeuw and Leroy, 1987; Draper and Smith, 1998; McKean, 2004). However, estimating weights on the basis of the data complicates the analysis regarding the expectation and variance of the weighted estimator. Seber (1977) suggested that the notion of the weights being random should be ignored and that the variance of the OLS estimator should be utilized for the variance of the weighted estimator.

We are interested in downweighting atypical observations via weight functions still in the form of WOLS. However, we use the fact that our weights are random because they are functions of random variables, and we acknowledge that the properties of our weighted estimator are no longer exactly the same as the properties of OLS-based estimators. Furthermore, unlike weighting functions in WOLS, our weighting functions are not based on iterative algorithms.

The purpose of this paper is twofold: (1) to illustrate a weighting method that will allow atypical observations (outliers, leverage, and influential points) identified by the RFD method to be included in the data analysis and (2) to evaluate the properties of the weighted estimators via bootstrapping procedures. We also compare OLS, RFD with dichotomous weights, and RFD with continuous weights. In section 2, we implement weight functions and continue with a brief overview of the properties of the parameter estimates obtained by using the RFD method with weight functions consisting of random variables. Section 3 presents results based on simulations and bootstrapping procedures for the RFD method in comparison to the results from OLS. Examples are used to illustrate the application of our method in section 4. A discussion and conclusions are given in sections 5 and 6, respectively.

2. Method

We implement unequal weights by using weight functions based on the extremeness of observations detected as outliers and/or leverage points by the RFD method. Unlike the dichotomous weights of 0 and 1, the continuous weights are unequal and range between 0 and 1, inclusive. With this particular weighting scheme, no observation is excluded from the statistical analyses (i.e., no weight of 0 is allowed). Each observation is allowed to have a different impact in the analysis, and this impact depends upon the observation's distance from the center of the data within its respective space.

We first review necessary concepts from Jones and Redden's (2007) RFD method in order to identify atypical observations: weighted mean squared error (MSE_v),

standardized prediction residuals, adjusted prediction intervals, Rousseeuw and van Zomeren's (1990) concept of robust distances (RD), the minimum covariance determinant (MCD), and the RD threshold. Let $\mathbf{x}_k = (x_{k1} \ x_{k2} \ \dots \ x_{k,p-1})$, $\hat{y}_k = \mathbf{x}_k \tilde{\mathbf{b}}$, $\tilde{\mathbf{b}} = (\mathbf{X}'\mathbf{V}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}\mathbf{y}$, $v_{\text{sum}} = \left(\sum_{i=1}^n v_i \right)$, $\text{MSE}_{\mathbf{V}} = (\mathbf{y} - \mathbf{X}\tilde{\mathbf{b}})' \mathbf{V} (\mathbf{y} - \mathbf{X}\tilde{\mathbf{b}}) / (v_{\text{sum}} - p)$, and $d_k = \frac{y_k - \hat{y}_k}{\sqrt{\text{MSE}_{\mathbf{V}}}}$, where d_k are standardized prediction residuals. Note that \mathbf{V} is a diagonal

matrix with diagonal elements

$$v_i = \begin{cases} 0, & \text{if the } i^{\text{th}} \text{ observation is detected as an outlier by the RFD method} \\ 1, & \text{otherwise} \end{cases}$$

for $i = 1, \dots, n$. Thus, v_{sum} is the number of observations in the dataset not detected as outliers by the RFD approach. The adjusted $100(1-\alpha)\%$ prediction interval for y_k , the k^{th} observation using $h_k = \mathbf{x}_k (\mathbf{X}'\mathbf{V}\mathbf{X})^{-1} \mathbf{x}_k'$, was shown by Jones and Redden (2007) to be

$$\left(-t \left(1 - \frac{\alpha}{2(v_{\text{sum}} + 1)} \right), v_{\text{sum}} - p \right) (1 + h_k)^{1/2}, \quad t \left(1 - \frac{\alpha}{2(v_{\text{sum}} + 1)} \right), v_{\text{sum}} - p \right) (1 + h_k)^{1/2}.$$

The upper bound of this adjusted prediction interval, denoted as $UpperPI_{\text{adj}}$, is utilized to define outliers and to design weights v_i , previously defined.

Rousseeuw and van Zomeren (1990) defined RD such that

$$RD_i = [(\mathbf{x}_i - \mathbf{T}_w(\mathbf{X})) \mathbf{C}_w(\mathbf{X})^{-1} (\mathbf{x}_i - \mathbf{T}_w(\mathbf{X}))']^{1/2},$$

with weighted mean vector

$$\mathbf{T}_w(\mathbf{X}) = \left(\sum_{i=1}^n w_i \right)^{-1} \sum_{i=1}^n w_i \mathbf{x}_i$$

and with weighted covariance matrix

$$C_w(\mathbf{X}) = \left(\sum_{i=1}^n w_i - 1 \right)^{-1} \sum_{i=1}^n w_i (\mathbf{x}_i - T_w(\mathbf{X}))(\mathbf{x}_i - T_w(\mathbf{X})).$$

The w_i for the initial weighted mean $T_w(\mathbf{X})$ and for the initial weighted covariance matrix $C_w(\mathbf{X})$ are dependent upon the MCD, which randomly selects subsets consisting of at least half of the data and outputs the subset with the smallest determinant of the covariance matrix. However, for the final update and computation of the RD_i 's after all observations have been examined for leverage as given in Jones and Redden (2007), we can define

$$w_i = \begin{cases} 0, & \text{if the } i^{\text{th}} \text{ observation is detected as a leverage point by the RFD method} \\ 1, & \text{otherwise} \end{cases}$$

for $i = 1, \dots, n$. Jones and Redden (2007) determined that an appropriate upper threshold for RD 's in order to properly identify leverage points should be set such that

$$Rule = \frac{(3p-1)}{\sum_{i=1}^n w_i} \left(\sum_{i=1}^n w_i - 1 \right).$$

Unlike other methods that compute weights iteratively on the basis of β , our method does not compute weights iteratively by using mathematical algorithms; iterative weights are commonly used in weighted ordinary least squares. Dichotomous weights in this paper are based on weights v_i and w_i , previously defined. Continuous weights are generated on the basis of weight functions by using the previously defined standardized prediction residuals, adjusted prediction intervals, and RD 's instead of by using an iterative approach.

We create diagonal matrices $\mathbf{A}_{n \times n}$ and $\mathbf{B}_{n \times n}$ with a_k and b_k , respectively, along the diagonals. We define diagonal elements a_k such that

$$a_k = \begin{cases} (UpperPI_{adj}/d_k)^2 & \text{if the } k^{\text{th}} \text{ observation is an outlier} \\ 1, & \text{otherwise} \end{cases}$$

and diagonal elements b_k such that

$$b_k = \begin{cases} (Rule/RD_k)^2 & \text{if the } k^{\text{th}} \text{ observation is a leverage point} \\ 1, & \text{otherwise.} \end{cases}$$

Now let $\mathbf{G}_{n \times n} = \mathbf{A}_{n \times n} * \mathbf{B}_{n \times n}$ be a diagonal matrix consisting of weights g_k on the diagonal and 0's off the diagonal. Then,

$$\mathbf{G}_{n \times n}^{1/2} = \text{diag}(g_1^{1/2} \dots g_n^{1/2})$$

$$= \begin{bmatrix} g_1^{1/2} & 0 & \dots & 0 \\ 0 & g_2^{1/2} & \dots & 0 \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & g_n^{1/2} \end{bmatrix}.$$

Note that $\mathbf{G}^{1/2} = (\mathbf{G}^{1/2})'$. We use g_k to downweight and limit the effect of outliers, leverage points, and influential observations identified by the RFD. Note that $g_k = 1$ for those observations that are neither outliers or leverage points.

Because we are now using a form of WOLS, we begin with the general linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ and assume that \mathbf{X} is full rank. We further assume that the unobservable random errors are distributed like before such that $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{G}^{-1})$. We make this assumption to account for the portion of data that is contaminated with outliers, leverage, and/or influential observations. Now, we can downweight each observation that is declared by the RFD method to be atypical (outlier, leverage point, and influential). We rewrite the general linear model by implementing the continuous weights from matrix $\mathbf{G}_{n \times n}^{-1}$ such that our new linear model is denoted as

$$\mathbf{G}^{1/2} \mathbf{y} = \mathbf{G}^{1/2} \mathbf{X}\boldsymbol{\beta} + \mathbf{G}^{1/2} \boldsymbol{\varepsilon}. \quad (1)$$

Note that

$$\mathbf{y}_1 = \mathbf{X}_1\boldsymbol{\beta} + \boldsymbol{\varepsilon}_1 \text{ where } \boldsymbol{\varepsilon}_1 \sim N(\mathbf{0}, \sigma^2\mathbf{G}_1^{-1})$$

represents the portion of the data consisting of atypical observations and that

$$\mathbf{y}_2 = \mathbf{X}_2\boldsymbol{\beta} + \boldsymbol{\varepsilon}_2 \text{ where } \boldsymbol{\varepsilon}_2 \sim N(\mathbf{0}, \sigma^2\mathbf{I}_2)$$

represents the portion of the data not contaminated. Thus, we can rewrite the general

linear model in the form of partitioned matrices such that $\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix}$, $\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix}$, $\boldsymbol{\varepsilon} = \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \end{bmatrix}$,

and $\mathbf{G}_{n \times n}^{-1} = \begin{bmatrix} \mathbf{G}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_2 \end{bmatrix}^{-1}$, where \mathbf{y}_1 is a $k \times 1$ vector, where \mathbf{y}_2 is a $(n-k) \times 1$ vector, where \mathbf{G}_1

is a $k \times k$ matrix, where $\mathbf{0}$ is an $(n-k) \times k$ matrix of 0's, and where \mathbf{I}_2 is an $(n-k) \times (n-k)$ identity matrix.

If we let $\mathbf{y}_w = \mathbf{G}^{1/2} \mathbf{y}$, $\mathbf{X}_w = \mathbf{G}^{1/2} \mathbf{X}$, and $\boldsymbol{\varepsilon}_w = \mathbf{G}^{1/2} \boldsymbol{\varepsilon}$, then we can rewrite the linear model in (1), which incorporates the weights, as

$$\mathbf{y}_w = \mathbf{X}_w\boldsymbol{\beta}_w + \boldsymbol{\varepsilon}_w.$$

Following the same pattern as that of OLS, we solve for $\boldsymbol{\varepsilon}_w = \mathbf{y}_w - \mathbf{X}_w\boldsymbol{\beta}_w$. Weighting each element of the general linear model adjusts *SSE*, with residuals \mathbf{e}_w , in the following manner:

$$\begin{aligned} \text{SSE}_w &= \mathbf{e}_w' \mathbf{e}_w \\ &= (\mathbf{y}_w - \hat{\mathbf{y}}_w)' (\mathbf{y}_w - \hat{\mathbf{y}}_w) \\ &= (\mathbf{y}_w - \mathbf{X}_w \hat{\boldsymbol{\beta}}_w)' (\mathbf{y}_w - \mathbf{X}_w \hat{\boldsymbol{\beta}}_w) \\ &= \mathbf{y}_w' \mathbf{y}_w - \hat{\boldsymbol{\beta}}_w' \mathbf{X}_w' \mathbf{y}_w - \mathbf{y}_w' \mathbf{X}_w \hat{\boldsymbol{\beta}}_w + \hat{\boldsymbol{\beta}}_w' \mathbf{X}_w' \mathbf{X}_w \hat{\boldsymbol{\beta}}_w \\ &= \mathbf{y}_w' \mathbf{y}_w - 2\hat{\boldsymbol{\beta}}_w' \mathbf{X}_w' \mathbf{y}_w + \hat{\boldsymbol{\beta}}_w' \mathbf{X}_w' \mathbf{X}_w \hat{\boldsymbol{\beta}}_w \\ &= \Lambda_w. \end{aligned}$$

Continuing from above, we derive the new estimator β_w by minimizing the sum of the weighted squared deviations of the observed and predicted values with respect to $\hat{\beta}_w$:

$$\frac{\partial \Lambda_w}{\partial \hat{\beta}_w} = -2X_w' y_w + 2X_w' X_w \hat{\beta}_w. \quad (2)$$

We then set (2) equal to zero and solve for $\hat{\beta}_w$:

$$\begin{aligned} & -2X_w' y_w + 2X_w' X_w \hat{\beta}_w = 0 \\ \Rightarrow & 2X_w' X_w \hat{\beta}_w = 2X_w' y_w \\ \Rightarrow & X_w' X_w \hat{\beta}_w = X_w' y_w \\ \Rightarrow & \hat{\beta}_w = (X_w' X_w)^{-1} X_w' y_w \\ \Rightarrow & \hat{\beta}_w = (X'GX)^{-1} X'Gy. \end{aligned} \quad (3)$$

As can be seen in (3), $\hat{\beta}_w$ consists of functions of random variables G and y .

Using known facts and the Sherman-Morrison Woodbury Theorem (Hager 1989), we can rewrite the estimator $\hat{\beta}_w$ (see Appendix A). On the basis of the derivation of $\hat{\beta}_w$, we can conclude that our new weighted estimated parameter $\hat{\beta}_w$ is biased because of the additional terms. Thus, using the properties of WOLS would seem to be inappropriate and unreliable, particularly in the case of small sample sizes. Because of the complex nature of the estimator $\hat{\beta}_w$ and because it consists of functions of several random variables, we will not directly derive the exact properties of this estimator. However, simulation results are presented in section 3 showing the average of the weighted estimator $\hat{\beta}_w$, along with corresponding 95% confidence intervals.

3. Simulations and results

3.1 Simulations

We conducted simulations ranging from 0% to 40% of outlier, leverage, and influence contamination. However, only results illustrating 10% and 20% contamination are presented. Simulations were performed to assess whether implementing different weighting schemes resulted in accuracy and parameter estimation that were superior to those produced by OLS. Data were generated so that 100 observations per 1000 simulations, as described in Jones and Redden (2007), had a true parameter $\beta = (25 \ 2 \ 2 \ 2)'$. The results are presented in Tables 1 to 12.

3.2 Notation

RFD1 denotes the application of the RFD method using dichotomous weights of 0 and 1 only, and RFD2 represents the RFD method utilizing continuous weights on the interval of (0,1] via weight functions that were explained in section 2. Further notation is as follows: $MSE_{OLS}(\hat{\beta}^*)$, $MSE_{RFD1}(\hat{\beta}^*)$, and $MSE_{RFD2}(\hat{\beta}^*)$ indicate the average squared deviations between the true parameter of the simulated data and the estimated parameters from OLS, RFD1, and RFD2, respectively.

3.3 Results

Average parameter estimates, MSE, and R^2 for 1000 simulations (100 observations per simulation) for OLS, RFD1, and RFD2 are presented in Table 1. When there are no simulated outliers or leverage points, these estimates do not deviate far from each other or from the true parameter with the use of any of the three methods—OLS,

RFD1, and RFD2—as expected. Table 1 shows that, when the same method of simulating data is used but with outlier contamination, the parameter estimates computed by using OLS, RFD1, and RFD2 are similar to the true parameters for both 10% and 20% outlier contamination. However, the impact of outlier contamination is seen in the inaccurate estimation of the intercept by using OLS compared to the true parameter for the intercept.

Table 1. Average parameter estimates (std dev), MSE, and R^2 for 1000 simulations (100 observations per simulation) for outlier contamination for OLS, RFD1, and RFD2

	OLS	RFD1	RFD2
10% outliers			
$\hat{\beta}_0^*$	25.6130764	25.0045134	25.2509982
(Std Dev)	(0.10465473)	(0.11314764)	(0.12090112)
$\hat{\beta}_1^*$	2.00434705	1.99883883	1.99870325
(Std Dev)	(0.16773101)	(0.11414173)	(0.12640626)
$\hat{\beta}_2^*$	1.99728055	1.9979373	1.99840157
(Std Dev)	(0.16673811)	(0.11435645)	(0.12647563)
$\hat{\beta}_3^*$	2.00033737	2.0004015	2.00187875
(Std Dev)	(0.16506273)	(0.11648385)	(0.12382746)
MSE	4.3213	1.0078	2.1952
R^2	0.7242	0.9186	0.8295
20% outliers			
$\hat{\beta}_0^*$	26.2229517	25.0539535	25.5577183
(Std Dev)	(0.10479193)	(0.26137389)	(0.19759345)
$\hat{\beta}_1^*$	1.99845648	1.99763873	1.99777125
(Std Dev)	(0.23460129)	(0.13346761)	(0.1632396)
$\hat{\beta}_2^*$	1.99738343	1.99749041	1.99737992
(Std Dev)	(0.22213519)	(0.13356535)	(0.15677151)
$\hat{\beta}_3^*$	2.00763734	2.00291025	2.00568105
(Std Dev)	(0.22139617)	(0.13314137)	(0.1566355)
MSE	6.8969	1.2237	3.3658
R^2	0.6052	0.9044	0.7406

Note: Std Dev = Standard Deviation.

This difference exemplifies the finding that the parameter estimates obtained by using OLS seem to be biased. Note that the parameter estimate associated with the intercept

also exhibits bias under RFD2. In addition, although the weights for RFD1 are dichotomous, they are still a function of the random variables. However, when there exists both 10% and 20% outlier contamination, there is much less bias produced by using the RFD1 method than results from using OLS or RFD2 (see Table 1). Similar findings were evident in the simulations for which 30% and 40% outlier contamination were generated.

Again, note that the average squared deviations of the true parameter from the simulations and of the parameter estimates obtained by using OLS is denoted as $MSE_{OLS}(\hat{\beta}^*)$. Similar notation is used also for the deviations determined by comparing the true simulated parameters to the parameter estimates obtained by using RFD1 and RFD2. These average squared deviations (see Table 2) indicate that using either dichotomous or continuous weights with the RFD method produces parameter estimates more accurate than OLS does when no adjustments are made to offset outlier contamination of observations.

Simulation results based on 10% and 20% leverage contamination are presented in Tables 3 and 4. Parameter estimates obtained by utilizing either dichotomous weights via RFD1 or continuous weights via RFD2 are comparable to those obtained by the use of OLS, as indicated in Table 3. Because the RFD method is able to properly detect leverage points, more consistency is found among the parameter estimates, especially for the intercept, than is found when there is outlier contamination for RFD1 and RFD2. The MSE for RFD2 gradually decreases because the level of contamination increases and because RFD2 is able to capture and downweight all of the simulated leverage points. Because the dichotomous weights of 0 and 1 cause RFD1 to exclude several observations

Table 2. Average squared deviations for parameter estimates for the true simulated parameter from OLS, RFD1, and RFD2 for 1000 simulations (100 observations per simulation) in the presence of outlier contamination

		$MSE_{OLS}(\hat{\beta}^*)$	$MSE_{RFD1}(\hat{\beta}^*)$	$MSE_{RFD2}(\hat{\beta}^*)$
10% outliers				
	$\hat{\beta}_0^*$	0.3868043	0.0128100	0.0776025
	$\hat{\beta}_1^*$	0.0281245	0.0130167	0.0159642
	$\hat{\beta}_2^*$	0.0277812	0.0130686	0.0159826
	$\hat{\beta}_3^*$	0.0272186	0.0135551	0.0153214
20% outliers				
	$\hat{\beta}_0^*$	1.5065811	0.0711590	0.3500539
	$\hat{\beta}_1^*$	0.0549851	0.0178014	0.0266255
	$\hat{\beta}_2^*$	0.0493015	0.0178282	0.0245596
	$\hat{\beta}_3^*$	0.0490256	0.0177174	0.0245424

Note: $MSE_{OLS}(\hat{\beta}^*)$, $MSE_{RFD1}(\hat{\beta}^*)$, and $MSE_{RFD2}(\hat{\beta}^*)$ denote the average squared deviations between the true parameter of the simulated data and the parameter estimates obtained by using OLS, RFD1, and RFD2, respectively.

from the analysis, the degrees of freedom for RFD1 differ from those for OLS and RFD2; therefore, caution is warranted when MSE is being compared. The results for the simulations with 30% and 40% of leverage points follow the same pattern.

Unlike the weighting functions used when there is outlier contamination, the weighting functions used in the presence of leverage contamination are not random. Therefore, the estimators produced under the utilization of both RFD1 and RFD2 are expected to produce unbiased estimates. This notion is supported in Table 3, which contains a comparison of the parameter estimates obtained using RFD1 and RFD2 to the true simulated parameters. Table 4 presents the average squared deviations for each method from the truth, which was obtained by using the simulations previously described, for each parameter estimate in the presence of 10% and 20% leverage contamination. These average squared deviations indicate how close each parameter

Table 3. Average parameter estimates (std dev), MSE, and R^2 for 1000 simulations (100 observations per simulation) for leverage contamination for OLS, RFD1, and RFD2

	OLS	RFD1	RFD2
10% leverage			
$\hat{\beta}_0^*$	25.0001710	25.0001894	25.0002722
(Std Dev)	(0.1060576)	(0.10902695)	(0.10743667)
$\hat{\beta}_1^*$	2.00109725	1.99979039	2.00062167
(Std Dev)	(0.08684433)	(0.11227716)	(0.1039884)
$\hat{\beta}_2^*$	1.9988195	1.99778778	1.99848404
(Std Dev)	(0.08608599)	(0.11150812)	(0.10388387)
$\hat{\beta}_3^*$	1.99971058	2.00114872	2.00138489
(Std Dev)	(0.08580192)	(0.11385609)	(0.10607359)
MSE	0.9975	0.9912	0.8904
R^2	0.9897	0.9211	0.9319
20% leverage			
$\hat{\beta}_0^*$	25.0005430	25.0002478	25.0003427
(Std Dev)	(0.10948341)	(0.11244732)	(0.11070164)
$\hat{\beta}_1^*$	2.0009665	1.99941147	2.0004938
(Std Dev)	(0.08554408)	(0.11964304)	(0.1063941)
$\hat{\beta}_2^*$	1.9992210	1.99705907	1.99771486
(Std Dev)	(0.08448195)	(0.11749118)	(0.10562747)
$\hat{\beta}_3^*$	1.99925902	2.00130211	2.0011754
(Std Dev)	(0.08556984)	(0.12056428)	(0.10910281)
MSE	0.9975	0.99095	0.7907
R^2	0.9939	0.9211	0.9403

Note: Std Dev = Standard Deviation.

estimate produced by utilizing OLS, RFD1, and RFD2 is to the true parameters from the simulated data. We note that the average squared deviations for each parameter produced by using RFD1 and RFD2 are close in value, especially in the presence of 10% leverage contamination. Again, the weighting functions utilized for both the RFD1 and RFD2 are no longer based on random variables. Instead, the weights produced are functions of the design matrix \mathbf{X} , which consists of fixed, known constants. Utilizing these non-random weights, we noticed that the average squared deviations for the true parameter and the

Table 4. Average squared deviations for parameter estimates for the true simulated parameter from OLS, RFD1, and RFD2 for 1000 simulations (100 observations per simulation) in the presence of leverage contamination

		$MSE_{OLS}(\hat{\beta}^*)$	$MSE_{RFD1}(\hat{\beta}^*)$	$MSE_{RFD2}(\hat{\beta}^*)$
10% leverage	$\hat{\beta}_0^*$	0.0112370	0.0118750	0.0115312
	$\hat{\beta}_1^*$	0.0075356	0.0125936	0.0108032
	$\hat{\beta}_2^*$	0.0074048	0.0124265	0.0107834
	$\hat{\beta}_3^*$	0.0073547	0.0129516	0.0112423
20% leverage	$\hat{\beta}_0^*$	0.0119749	0.0126318	0.0122427
	$\hat{\beta}_1^*$	0.0073114	0.0143005	0.0113086
	$\hat{\beta}_2^*$	0.0071307	0.0137990	0.0111512
	$\hat{\beta}_3^*$	0.0073154	0.0145229	0.0118929

Note: $MSE_{OLS}(\hat{\beta}^*)$, $MSE_{RFD1}(\hat{\beta}^*)$, and $MSE_{RFD2}(\hat{\beta}^*)$ denote the average squared deviations between the true parameter of the simulated data and the parameter estimates obtained by using OLS, RFD1, and RFD2, respectively.

parameter estimates produced under OLS are slightly smaller than those resulting from using the RFD1 and RFD2 methods.

It is apparent in Tables 5 and 6 that OLS is not robust to influential observations; this lack of robustness can cause distortion in parameter estimates and therefore can lead to incorrect inference. The parameter estimate $\hat{\beta}_0^*$ associated with the intercept obtained by using OLS is not as close to the true parameter as that obtained by utilizing RFD1 and RFD2. In addition, all other parameter estimates obtained by using OLS do not resemble the true parameters. The results suggest that the estimated parameters are no longer positively associated with the outcome. On the other hand, RFD1 and RFD2 are able to accurately estimate all four parameters (see Table 5). In addition, the results from the two methods are comparable to each other. Again, because the degrees of freedom differ, caution is warranted when the MSE and R^2 of RFD1 are being compared to those

Table 5. Average parameter estimates (std dev), MSE, and R^2 for 1000 simulations (100 observations per simulation) for influence contamination for OLS, RFD1, and RFD2

	OLS	RFD1	RFD2
10% influence			
$\hat{\beta}_0^*$	28.2163596	24.9916008	24.9969275
(Std Dev)	(0.18350817)	(0.33420571)	(0.33312225)
$\hat{\beta}_1^*$	-0.1588182	2.01371168	2.00955023
(Std Dev)	(0.38659137)	(0.37421373)	(0.37322495)
$\hat{\beta}_2^*$	-0.1826541	1.99587374	1.99271331
(Std Dev)	(0.38759689)	(0.36545691)	(0.36552697)
$\hat{\beta}_3^*$	-0.1747713	2.00316059	1.99968924
(Std Dev)	(0.39041338)	(0.38824705)	(0.38642813)
MSE	2.1895	0.9887	0.8944
R^2	0.2371	0.5136	0.5107
20% influence			
$\hat{\beta}_0^*$	28.2277201	24.9935508	25.0081207
(Std Dev)	(0.19698372)	(0.3554853)	(0.35647496)
$\hat{\beta}_1^*$	-0.1739830	2.01154553	2.00088364
(Std Dev)	(0.38381836)	(0.39381616)	(0.39255028)
$\hat{\beta}_2^*$	-0.1900160	1.99514232	1.98520851
(Std Dev)	(0.3907799)	(0.3875676)	(0.38901811)
$\hat{\beta}_3^*$	-0.1876814	2.00049627	1.99168781
(Std Dev)	(0.39086025)	(0.41861528)	(0.41744265)
MSE	2.1959	0.9869	0.7913
R^2	0.3758	0.5150	0.5105

Note: Std Dev = Standard Deviation.

obtained using OLS and RFD2. The results obtained from RFD2 indicate that it consistently produces a smaller MSE and a larger R^2 than what OLS yields; this finding is expected because, unlike OLS, RFD2 is able to properly detect and able to downweight influential observations. In addition, the MSE of RFD2 decreases as contamination level increases. This condition is expected because observations are downweighted causing a smaller numerator in the MSE while the degrees of freedom do not change and because the RFD method captures 99% of the contaminated observations. We note that, across all

three methods, the results obtained under both 30% and 40% influential contamination are similar to those produced under 10% and 20% influential contamination.

Unlike those for leverage, the weighting functions implemented, when influence exists among the data, are functions of random variables as described in section 2. On the basis of the derivation of our weighted estimator $\hat{\beta}_w$, we expect our estimator to be biased when RFD1 and RFD2 are being implemented. However, the results in Table 5 suggest that there is no severe bias in parameter estimates when either RFD1 or RFD2 is implemented. On the other hand, the results from using OLS indicate that there is bias, especially in the intercept, and that the estimated parameters are inaccurate which can lead to incorrect inference.

Table 6 further supports the results shown in Table 5. The average squared deviations indicate that, in the presence of 10% and 20% influence, the estimates from both RFD1 and RFD2 are closer to the true parameter of the simulated data than that from using OLS. The average squared deviations indicate that both RFD1 and RFD2 yield parameter estimates that are close to the true simulated parameters underlying the uncontaminated data. The average squared deviations seen in Table 6 under both 10% and 20% influence contamination indicate that OLS produces parameter estimates that are not close to the true parameters underlying the uncontaminated data. Both Tables 5 and 6 indicate that OLS is not the most appropriate method to utilize when there is influential contamination in the data. These results indicate that OLS will distort parameter estimates, and this distortion can lead to incorrect inference. The results are similar for the simulated datasets containing 30% and 40% influence contamination.

Table 6. Average squared deviations for parameter estimates for the true simulated parameter from OLS, RFD1, and RFD2 for 1000 simulations (100 observations per simulation) in the presence of influence contamination

		$MSE_{OLS}(\hat{\beta}^*)$	$MSE_{RFD1}(\hat{\beta}^*)$	$MSE_{RFD2}(\hat{\beta}^*)$
10% influence				
	$\hat{\beta}_0^*$	10.3786108	0.1116523	0.1108689
	$\hat{\beta}_1^*$	4.8097993	0.1400839	0.1392488
	$\hat{\beta}_2^*$	4.9140602	0.1334422	0.1335295
	$\hat{\beta}_3^*$	4.8819004	0.1505950	0.1491775
20% influence				
	$\hat{\beta}_0^*$	10.4569411	0.1262850	0.1270133
	$\hat{\beta}_1^*$	4.8733714	0.1550694	0.1539424
	$\hat{\beta}_2^*$	4.9487262	0.1500820	0.1514025
	$\hat{\beta}_3^*$	4.9385690	0.1750638	0.1741532

Note: $MSE_{OLS}(\hat{\beta}^*)$, $MSE_{RFD1}(\hat{\beta}^*)$, and $MSE_{RFD2}(\hat{\beta}^*)$ denote the average squared deviations between the true parameter of the simulated data and the parameter estimates obtained by using OLS, RFD1, and RFD2, respectively.

Because our weights are functions of random variables, we did not derive direct calculations of the expectation and variance of our parameter estimator. Salibian-Barrera and Zamar (2002) proposed using bootstrapping approaches to make inferences for robust regression methods. Therefore, we utilized bootstrapping procedures, which randomly selects observations from the data with replacement. This procedure is a resampling technique that can be used to provide robust estimates for the mean of our weighted parameter because the estimator is involved and because exact derivation is not computed.

We conducted 10,000 bootstrap recalculations based on one sample ($n = 100$) of 1000 iterations from the simulations previously mentioned. Tables 7 – 12 provide estimates for the mean (standard deviation) and corresponding 95% confidence intervals for the parameter estimates for RFD1 and RFD2. To maintain consistency, we included

only the results for 10% and 20% contamination. Table 7 illustrates the bootstrap results for the condition when there is 10% outlier contamination. Because the 95% confidence intervals for the average estimated parameters for both RFD1 and RFD2 are narrow, the parameter estimates obtained via bootstrapping procedures are considered to be stable and therefore reliable. This parallelism is also exhibited in Tables 8 – 12.

Table 7. Bootstrap results illustrating average estimated parameters (std dev) and corresponding 95% confidence intervals for the parameters when RFD1 and RFD2 are used in the presence of 10% outlier contamination (based on 10,000 iterations)

10% outliers	RFD1	RFD1 95% CI	RFD2	RFD2 95% CI
$\hat{\beta}_0^*$	24.7386248	(24.5187, 24.9481)	24.943871	(24.6974, 25.2094)
(Std Dev)	(0.11499146)	——	(0.13225057)	——
$\hat{\beta}_1^*$	1.76465105	(1.5145, 1.9909)	1.77914618	(1.5431, 2.0014)
(Std Dev)	(0.13552229)	——	(0.12314166)	——
$\hat{\beta}_2^*$	1.82296745	(1.5957, 2.0664)	1.77230737	(1.5410, 1.9932)
(Std Dev)	(0.12757483)	——	(0.11819354)	——
$\hat{\beta}_3^*$	2.0672577	(1.8259, 2.3575)	2.11721222	(1.9024, 2.3543)
(Std Dev)	(0.17861244)	——	(0.13809592)	——

Note: Std Dev = Standard Deviation.

Table 8. Bootstrap results illustrating average estimated parameters (std dev) and corresponding 95% confidence intervals for the parameters when RFD1 and RFD2 are used in the presence of 10% leverage contamination (based on 10,000 iterations)

10% leverage	RFD1	RFD1 95% CI	RFD2	RFD2 95% CI
$\hat{\beta}_0^*$	24.7358469	(24.5125, 24.9467)	24.7314601	(24.5233, 24.9358)
(Std Dev)	(0.13195725)	——	(0.11316298)	——
$\hat{\beta}_1^*$	1.77141887	(1.5128, 1.9959)	1.81007861	(1.5961, 2.0084)
(Std Dev)	(0.15821)	——	(0.11679315)	——
$\hat{\beta}_2^*$	1.82406522	(1.5936, 2.0633)	1.83811203	(1.6457, 2.0291)
(Std Dev)	(0.14810129)	——	(0.10713411)	——
$\hat{\beta}_3^*$	2.06506364	(1.8236, 2.3427)	2.08897823	(1.9018, 2.2863)
(Std Dev)	(0.2273945)	——	(0.1384549)	——

Note: Std Dev = Standard Deviation.

Table 9. Bootstrap results illustrating average estimated parameters (std dev) and corresponding 95% confidence intervals for the parameters when RFD1 and RFD2 are used in the presence of 10% influence contamination (based on 10,000 iterations)

10% influence	RFD1	RFD1 95% CI	RFD2	RFD2 95% CI
$\hat{\beta}_0^*$	25.5202477	(24.7715, 26.2537)	25.5315321	(24.7875, 26.2636)
(Std Dev)	(0.44231619)	—	(0.39614002)	—
$\hat{\beta}_1^*$	1.76328731	(1.0328, 2.5186)	1.75228651	(1.0280, 2.5017)
(Std Dev)	(0.54149699)	—	(0.4280524)	—
$\hat{\beta}_2^*$	1.35093058	(0.6138, 2.1122)	1.34404026	(0.6148, 2.0983)
(Std Dev)	(0.4251826)	—	(0.39118199)	—
$\hat{\beta}_3^*$	1.62933624	(0.9257, 2.3573)	1.62238345	(0.9225, 2.3508)
(Std Dev)	(0.38802613)	—	(0.36875902)	—

Note: Std Dev = Standard Deviation.

Table 10. Bootstrap results illustrating average estimated parameters (std dev) and corresponding 95% confidence intervals for the parameters when RFD1 and RFD2 are used in the presence of 20% outlier contamination (based on 10,000 iterations)

20% outliers	RFD1	RFD1 95% CI	RFD2	RFD2 95% CI
$\hat{\beta}_0^*$	24.8025359	(24.5785, 25.0184)	25.2764087	(24.9591, 25.6508)
(Std Dev)	(0.13049551)	—	(0.18198965)	—
$\hat{\beta}_1^*$	1.80312619	(1.5492, 2.0456)	1.70782935	(1.4245, 1.9580)
(Std Dev)	(0.13829204)	—	(0.13799236)	—
$\hat{\beta}_2^*$	1.89184562	(1.6303, 2.1535)	1.85793274	(1.5696, 2.1252)
(Std Dev)	(0.16225068)	—	(0.14289782)	—
$\hat{\beta}_3^*$	2.03494299	(1.7512, 2.3305)	2.13596181	(1.8497, 2.4342)
(Std Dev)	(0.15872077)	—	(0.15077055)	—

Note: Std Dev = Standard Deviation.

Table 11. Bootstrap results illustrating average estimated parameters (std dev) and corresponding 95% confidence intervals for the parameters when RFD1 and RFD2 are used in the presence of 20% leverage contamination (based on 10,000 iterations)

20% leverage	RFD1	RFD1 95% CI	RFD2	RFD2 95% CI
$\hat{\beta}_0^*$	24.7939242	(24.5710, 25.0118)	24.79184	(24.5812, 24.9993)
(Std Dev)	(0.13418223)	—	(0.11586054)	—
$\hat{\beta}_1^*$	1.80944449	(1.5560, 2.0362)	1.84484736	(1.6302, 2.0373)
(Std Dev)	(0.14271638)	—	(0.11202608)	—
$\hat{\beta}_2^*$	1.8879437	(1.6471, 2.1368)	1.89770276	(1.7080, 2.0878)
(Std Dev)	(0.14097053)	—	(0.10474466)	—
$\hat{\beta}_3^*$	2.04289263	(1.7806, 2.3100)	2.07514349	(1.8793, 2.2658)
(Std Dev)	(0.20611349)	—	(0.13022677)	—

Note: Std Dev = Standard Deviation.

Table 12. Bootstrap results illustrating average estimated parameters (std dev) and corresponding 95% confidence intervals for the parameters when RFD1 and RFD2 are used in the presence of 20% influence contamination (based on 10,000 iterations)

20% influence	RFD1	RFD1 95% CI	RFD2	RFD2 95% CI
$\hat{\beta}_0^*$	25.7045929	(24.8651, 26.4934)	25.7324604	(24.9131, 26.5202)
(Std Dev)	(0.44292862)	—	(0.41555077)	—
$\hat{\beta}_1^*$	1.62270814	(0.8282, 2.4701)	1.60079286	(0.8111, 2.4351)
(Std Dev)	(0.4488016)	—	(0.41836524)	—
$\hat{\beta}_2^*$	1.25156098	(0.4264, 2.1127)	1.2333293	(0.4299, 2.0688)
(Std Dev)	(0.49530289)	—	(0.43758561)	—
$\hat{\beta}_3^*$	1.51005453	(0.7856, 2.2791)	1.4915569	(0.7672, 2.2469)
(Std Dev)	(0.4319931)	—	(0.38560809)	—

Note: Std Dev = Standard Deviation.

4. Examples

We demonstrate the effect of downweighting atypical observations with two well-known datasets commonly used to demonstrate robust statistical methods: Hawkins-Bradru-Kass (HBK) and Hertzsprung-Russell Star Cluster data. These datasets are used to

illustrate and compare results obtained by using OLS to results obtained by applying the weighting functions previously mentioned for RFD1 and RFD2.

4.1. HBK data

Hawkins, Bradu, and Kass (1984) simulated the well-known HBK dataset, which contains 75 observations, three predictors, and one response. Observations 1 – 10 were intentionally generated to be influential, and observations 11 – 14 were intentionally generated to be leverage points. Using the RFD1 method previously explained, observations 1 to 14 are downweighted with weights of zero and are therefore excluded from the analysis. When RFD2 is employed, outliers are downweighted by using weights a_k , and leverage points are downweighted by using weights b_k . Weights g_k , which is the product of a_k and b_k , are utilized to substantially downweight influential observations such as observations 1 – 10. Table 13 demonstrates the limited impact that the weights a_k , b_k , and g_k caused each leverage point and influential observation to have in the analysis.

Observations identified by the RFD method as influential have smaller weights than leverage points. As a result, observations 1 to 10 have a more restricted or limited impact in the analysis. In addition, the magnitude of each weight is an indication of an observation's extremeness in the standardized deviation of $y - E[y|X]$ or an observation's extremeness in X relative to the remaining observations. For instance, on the basis of its weight a_k (see Table 13), observation 7 has the most extreme standardized deviation in $y - E[y|X]$. This finding is consistent with the graph supplied in Jones and Redden (2007). Observation 14 is considered to be the most extreme observation in the predictor or X -

space because it has the smallest b_k . Furthermore, on the basis of the weights g_k , observation 5 is considered to be the most influential observation in the HBK data.

Table 13. Weights $g_k (= a_k * b_k)$ implemented by using weight functions via the RFD2 method for the HBK data

Id	g_k	a_k	b_k	Contamination type
1	.000007550	0.04871	.000155006	Influential
2	.000006139	0.04422	.000138828	Influential
3	.000005085	0.04507	.000112841	Influential
4	.000005110	0.05162	.000098984	Influential
5	.000004948	0.04669	.000105969	Influential
6	.000006276	0.04702	.000133477	Influential
7	.000005239	0.03970	.000131970	Influential
8	.000006342	0.04315	.000146985	Influential
9	.000005610	0.05037	.000111376	Influential
10	.000006051	0.04738	.000127699	Influential
11	.000064966	1.00000	.000064966	Leverage
12	.000055797	1.00000	.000055797	Leverage
13	.000062076	1.00000	.000062076	Leverage
14	.000041061	1.00000	.000041061	Leverage

Table 14 displays the parameter estimates, MSE, and R^2 for the HBK data after weights g_k are implemented in order to downweight those observations identified as atypical by the RFD method. The magnitude and the direction of some of the parameter estimates are inconsistent among the three methods. Because the RFD method is capable of detecting outliers, leverage points, and influential observations, we assume that the parameter estimates based on the RFD1 method can serve as the baseline for comparison of and with OLS and RFD2. It is noted that the results obtained by using RFD2 is more similar to the results obtained by using the baseline RFD1 than to that of OLS. This finding is expected since it is well known that OLS is not robust to outliers, leverage points, or influential observations. However, caution must be taken when the MSE and R^2 for the HBK data are being compared across all three methods. Because the RFD1

method excludes observations 1 to 14 from the analysis by implementing dichotomous weights of 0 and 1, the degrees of freedom for RFD1 differ from those for OLS and RFD2. Nevertheless, when parameters were estimated by implementing the RFD2 method versus implementing OLS, the MSE decreased by 94.95% ($= (5.063 - 0.2556)/5.063$).

Table 14. Parameter estimates, MSE, and R^2 obtained when OLS, RFD1, and RFD2 are used to analyze the HBK data

	OLS	RFD1	RFD2
$\hat{\beta}_0^*$	-0.3876	-0.0105	-0.0122
$\hat{\beta}_1^*$	0.2392	0.0624	0.0625
$\hat{\beta}_2^*$	-0.3346	0.0119	0.0123
$\hat{\beta}_3^*$	0.3833	-0.1070	-0.1064
MSE	5.063	0.3183	0.2556
R^2	0.6018	0.0472	0.0469

We conducted 10,000 bootstrap recalculations for the HBK data. Table 15 provides the average parameter estimates (standard deviation) and corresponding 95% confidence intervals for RFD1 and RFD2. Overall, the 95% confidence intervals for the average parameter estimates indicate that the results utilizing RFD1 and RFD2 are stable. However, based on bootstrap recalculations, the 95% confidence intervals computed for the average parameter estimates under RFD2 are slightly narrower than those provided by RFD1.

Table 15. Bootstrap results illustrating average estimated parameters (std dev) and corresponding 95% confidence intervals for the parameters when RFD1 and RFD2 are used for the HBK data (based on 10,000 iterations)

	RFD1	RFD1 95% CI	RFD2	RFD2 95% CI
$\hat{\beta}_0^*$	0.0263	(-0.4636, 1.1467)	0.0088	(-0.4439, 0.7216)
(Std Dev)	(0.3789)	——	(0.2845)	——
$\hat{\beta}_1^*$	0.06210	(-0.1846, 0.2358)	0.0643	(-0.1044, 0.2246)
(Std Dev)	(0.1196)	——	(0.0890)	——
$\hat{\beta}_2^*$	0.0061	(-0.1946, 0.1680)	0.0076	(-0.1693, 0.1588)
(Std Dev)	(0.1254)	——	(0.0928)	——
$\hat{\beta}_3^*$	-0.1157	(-0.3847, 0.0665)	-0.1117	(-0.3043, 0.0566)
(Std Dev)	(0.1283)	——	(0.0985)	——

Note: Std Dev = Standard Deviation.

4.3. Star Cluster CYG OB1 data

The Hertzsprung-Russell diagram of the Star Cluster CYG OB1 data (Rousseeuw and Leroy 1987), containing 47 observations, describes the logarithm of the light intensity of the star (y) by using the logarithm of the effective temperature at the surface of the star (x). The RFD method in Jones and Redden (2007) declared observations 7 and 14 as leverage points and observations 11, 20, 30, and 34 as influential. Therefore, these observations are excluded from the RFD1 analysis because of the dichotomous weights of 0 and 1 that are implemented. Table 16 illustrates weights a_k , b_k , and g_k implemented toward each observation identified as atypical. The weights were implemented to limit the impact of these observations within the linear regression analysis. The leverage impact of observations 7 and 14 is limited in the analysis via weights b_k . Because observations 11, 20, 30, and 34 were identified by the RFD as outlying in both ($y - E[y|X]$) and the X -space, these observations are downweighted by using weights g_k . Furthermore, consistent with Jones and Redden (2007), we found that observation 34 was

the most extreme observation in $(\mathbf{y} - E[\mathbf{y}|\mathbf{X}])$ and that observation 30 was the most extreme observation in the \mathbf{X} -space or predictor space. On the basis of weights g_k given in Table 16, we can conclude that observation 34 is the most influential observation in the Star Cluster dataset.

Table 16. Weights $g_k (= a_k * b_k)$ implemented by using weight functions via the RFD2 method for the Star Cluster data

Id	g_k	a_k	b_k	Contamination type
7	0.03116	1.00000	0.03116	Leverage
11	0.00251	0.54404	0.00461	Influential
14	0.12836	1.00000	0.12836	Leverage
20	0.00224	0.48371	0.00464	Influential
30	0.00190	0.43136	0.00439	Influential
34	0.00171	0.37047	0.00462	Influential

Parameter estimates, MSE, and R^2 obtained for the Star Cluster data by using OLS, RFD1, and RFD2 are provided in Table 17. Again, RFD1 is utilized as the baseline measurement for comparison of parameter estimates. The direction of the fitted regression line is indicated by the slope $\hat{\beta}_1^*$. The results from OLS indicate that the fitted regression line slopes downward, whereas the RFD1 and RFD2 results indicate that the fitted regression line should slope upward. When utilizing OLS, the logarithmic light intensity of the star increases as the logarithmic temperature at the surface of the star decreases. On the other hand, the results from RFD1 and RFD2 indicate that as the logarithmic temperature at the surface of the star increases then the logarithmic light intensity of the star increases. Table 17 further indicates that the MSE decreased by 59.72% $(= (0.3188 - 0.1284)/0.3188)$ when estimating parameters by using the RFD2 method compared to using OLS.

Table 17. Parameter estimates, MSE, and R^2 obtained when OLS, RFD1, and RFD2 are used to analyze the Star Cluster data

	OLS	RFD1	RFD2
$\hat{\beta}_0^*$	6.79	-8.21	-7.61
$\hat{\beta}_1^*$	-0.41	2.98	2.85
MSE	0.3188	0.1435	0.1284
R^2	0.0443	0.4287	0.4171

In addition, a basic scatterplot of the Star Cluster data can be utilized to illustrate the relationship between the response and the predictor. Figure 1 reveals that, with the exception of the atypical observations identified by the RFD method, the data form a positive linear relationship between the response and the predictor. This finding supports the results obtained by using the RFD1 and RFD2 methods provided in Table 17, and this finding further suggests that the fitted regression line from OLS is affected by the masked outliers (11, 20, 30, and 34) which are identified by the RFD.

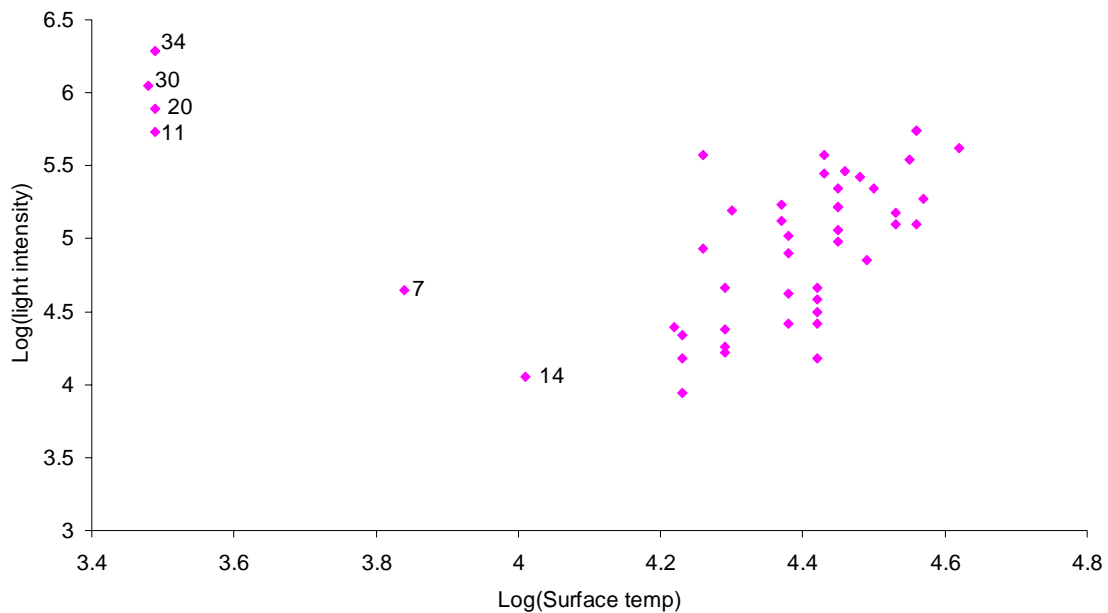


Figure 1. Scatterplot of logarithmic light intensity of the star versus the logarithmic temperature at the surface of the star for the Star Cluster data.

Table 18 provides results for the 10,000 recalculations of the bootstrapping procedure for the average parameter estimates, as well as the corresponding 95% confidence intervals, for the Star Cluster data obtained by utilizing both the RFD1 and RFD2 methods. The confidence intervals provided for the intercept are fairly large when both the RFD1 and RFD2 methods are used. On the other hand, the 95% confidence intervals are much narrower for the logarithm of the effective temperature at the surface of the star. This finding suggests that the parameter estimates provided for the one predictor are more stable. In addition, based on the bootstrapping results in Table 18, we can conclude that $\hat{\beta}_1^*$ is statistically significantly different from zero.

Table 18. Bootstrap results illustrating average estimated parameters (std dev) and corresponding 95% confidence intervals for the parameters when RFD1 and RFD2 are used for the Star Cluster data (based on 10,000 iterations)

	RFD1	RFD1 95% CI	RFD2	RFD2 95% CI
$\hat{\beta}_0^*$	-10.35	(-21.19, -1.60)	-8.32	(-16.17, -2.40)
(Std Dev)	(18.96)	—	(6.55)	—
$\hat{\beta}_1^*$	3.46	(1.48, 5.92)	3.01	(1.68, 4.78)
(Std Dev)	(4.28)	—	(1.48)	—

Note: Std Dev = Standard Deviation.

5. Discussion

Any of the three methods—OLS, RFD1, RFD2—can be utilized when there is no contamination in the data. However, the previous results indicate that, when any type of contamination exists, RFD1 or RFD2 is a more appropriate method to utilize than OLS. The parameter estimates for RFD1 and RFD2 were comparable across all contamination levels for the given simulations, as well as for the HBK and Star Cluster examples.

According to the results from the simulations, parameter estimates based on RFD2 are closer to the true parameters than parameter estimates based on RFD1 when the data were contaminated with leverage points only, whereas RFD1 provided better parameter estimates when only outlier contamination existed. When it is known that only outlier contamination exists, the RFD1 method yields a smaller MSE than OLS and RFD2. However, we must take caution when comparing the MSE and R^2 for the RFD1 method with those for OLS and RFD2. This caution is necessary because the MSE's are not based on the same degrees of freedom. The RFD1 method excludes each observation considered to be an outlier, leverage point, or an influential observation. This exclusion causes a decrease in the sample size and a change in the degrees of freedom for the MSE. No observations are excluded from the analysis when OLS or RFD2 is being applied. Instead, as seen in this paper, the impact of atypical observations is downweighted via weight functions when using RFD2.

The weight functions for outliers and influence consist of random variables. Although the histograms for RFD2 in Figures 2 and 4 suggest normality of the estimated parameters, caution is warranted when inference is being made by utilizing RFD2. Weight functions specified for leverage points for RFD2 are based on fixed values. Thus, we can continue with the assumption that, as the histograms in Figure 3 suggest, our estimated parameter follows a normal distribution. In addition, Figure 2 indicates that the parameter estimate associated with the intercept for RFD1 is bimodal because the RFD masks several of the contaminated observations when there is 20% outlier contamination. This bimodality becomes even more apparent as outlier contamination increases to 30% and 40%.

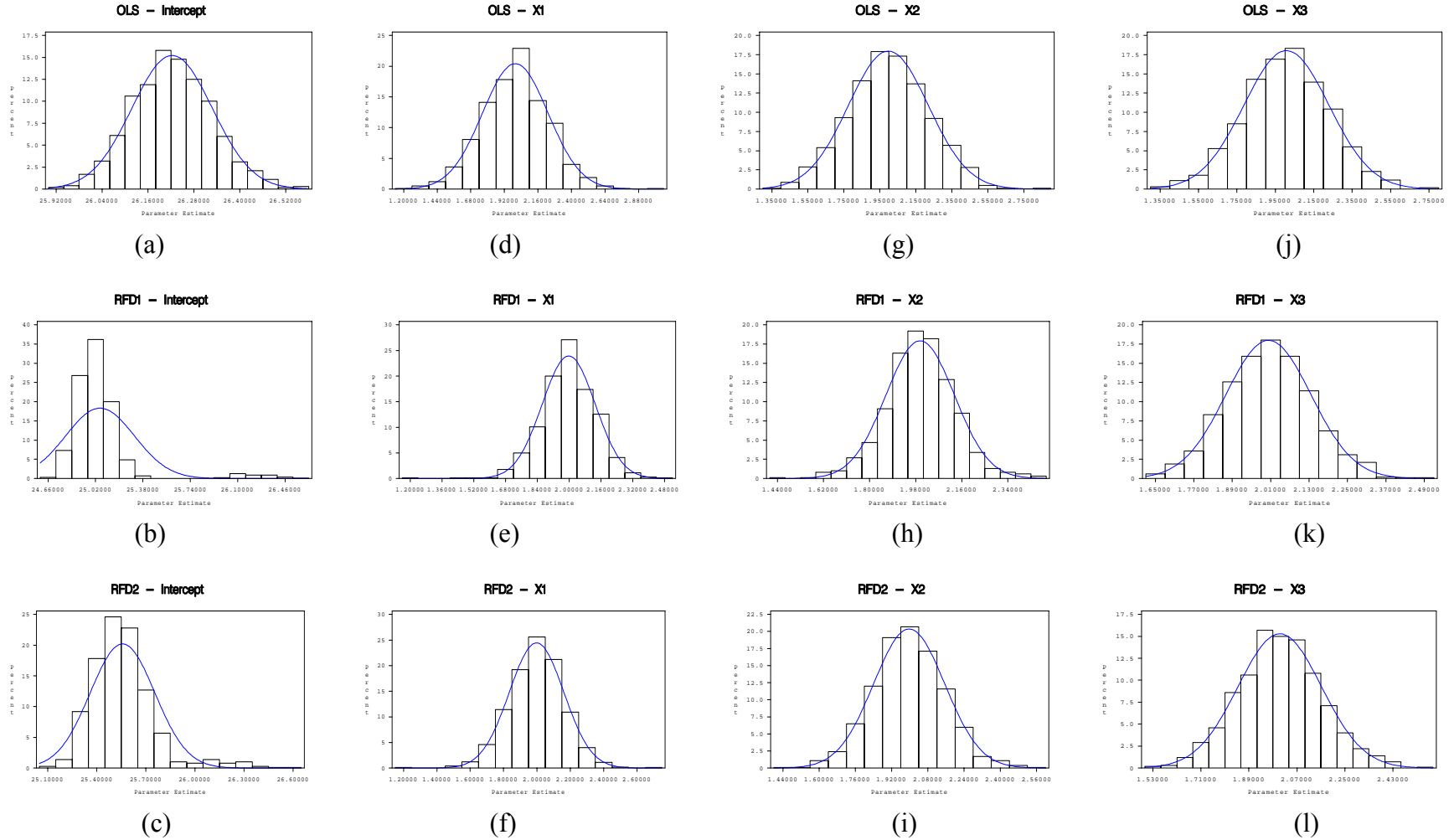


Figure 2. Histograms for parameter estimates in the presence of 20% outlier contamination: $\hat{\beta}_0^*$ for OLS (a), RFD1 (b), and RFD2 (c); $\hat{\beta}_1^*$ for OLS (d), RFD1 (e), and RFD2 (f); $\hat{\beta}_2^*$ for OLS (g), RFD1 (h), and RFD2 (i); and $\hat{\beta}_3^*$ for OLS (j), RFD1 (k), and RFD2 (l).

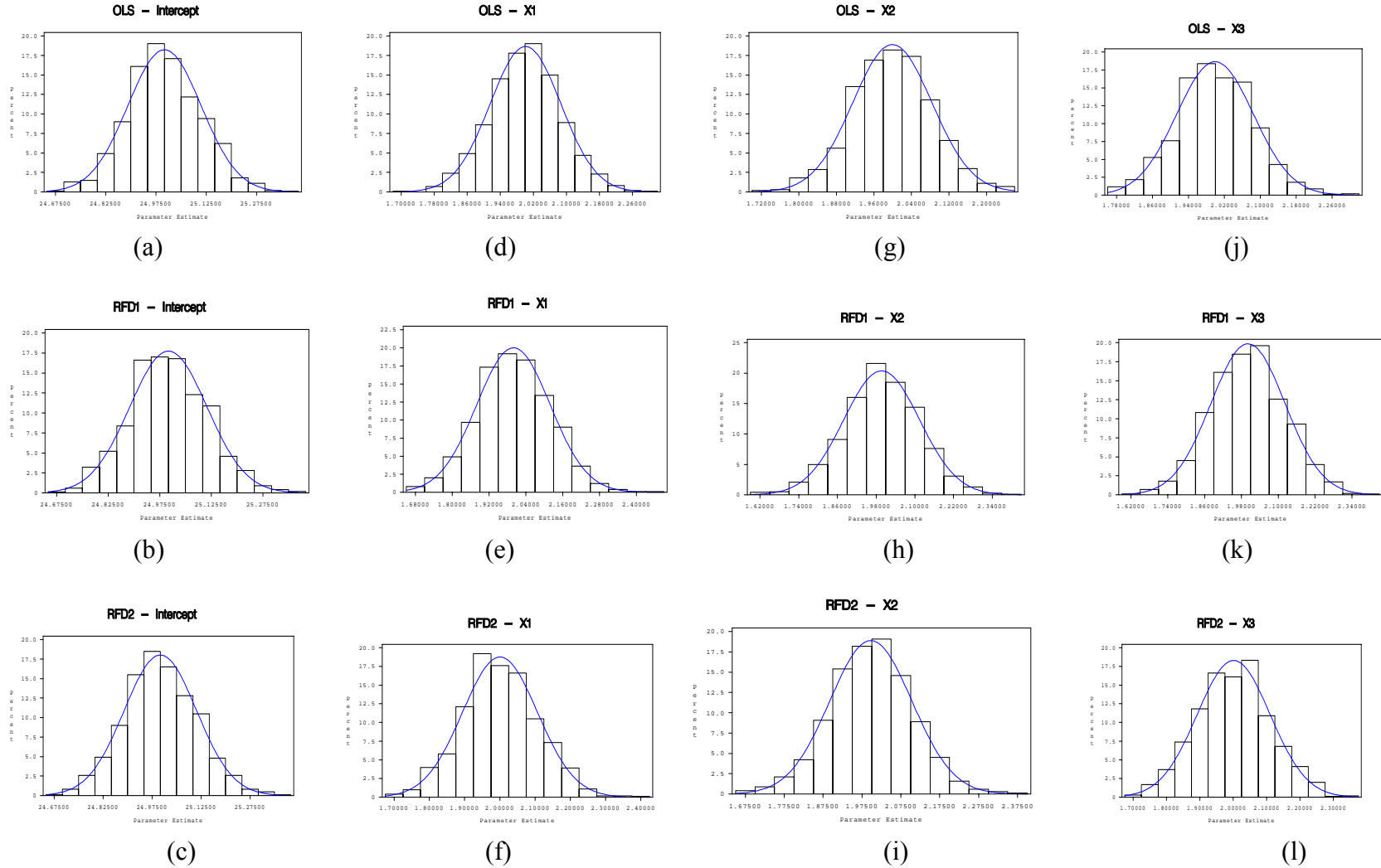


Figure 3. Histograms for parameter estimates in the presence of 20% leverage contamination: $\hat{\beta}_0^*$ for OLS (a), RFD1 (b), and RFD2 (c); $\hat{\beta}_1^*$ for OLS (d), RFD1 (e), and RFD2 (f); $\hat{\beta}_2^*$ for OLS (g), RFD1 (h), and RFD2 (i); and $\hat{\beta}_3^*$ for OLS (j), RFD1 (k), and RFD2 (l).

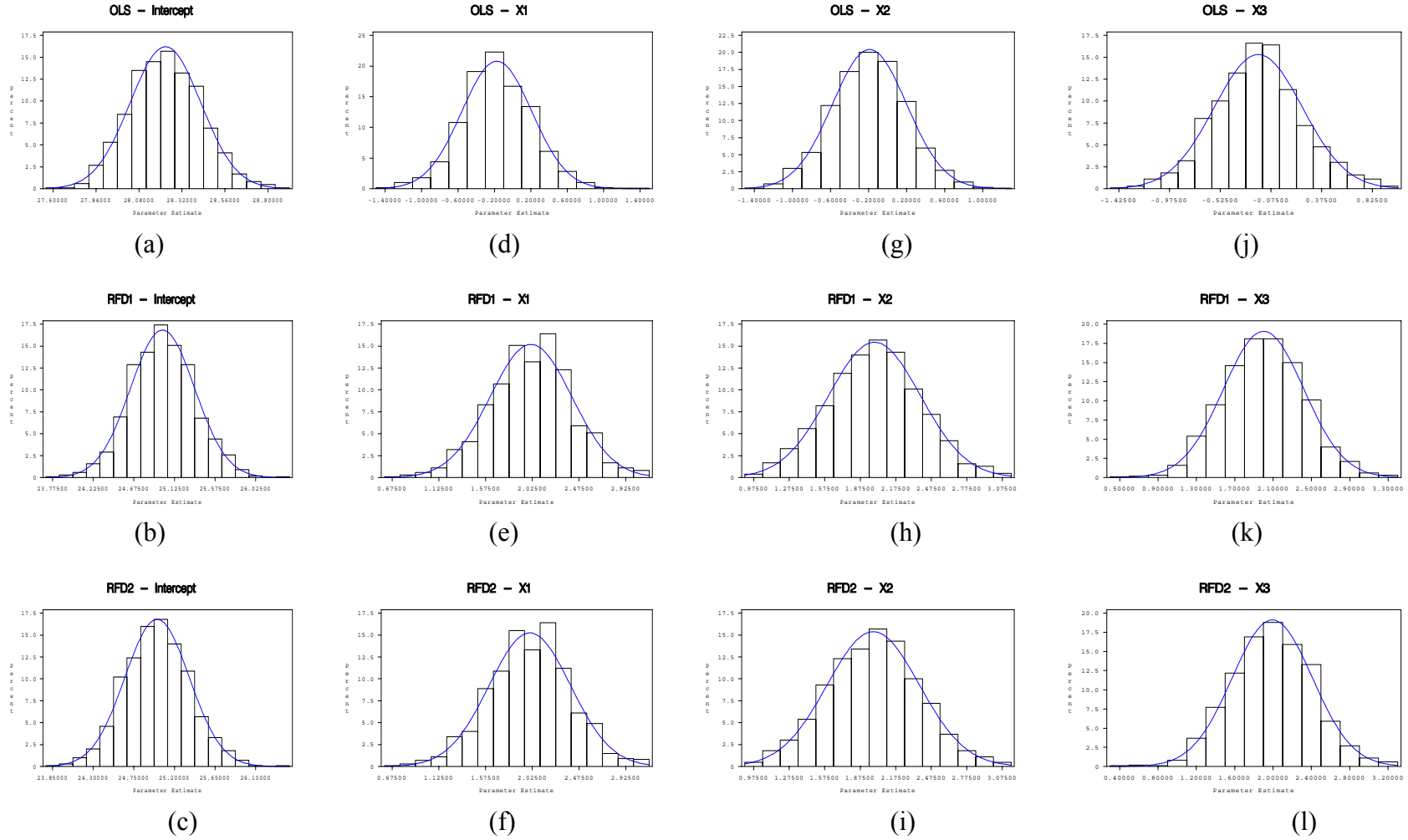


Figure 4. Histograms for parameter estimates in the presence of 20% influence contamination: $\hat{\beta}_0^*$ for OLS (a), RFD1 (b), and RFD2 (c); $\hat{\beta}_1^*$ for OLS (d), RFD1 (e), and RFD2 (f); $\hat{\beta}_2^*$ for OLS (g), RFD1 (h), and RFD2 (i); and $\hat{\beta}_3^*$ for OLS (j), RFD1 (k), and RFD2 (l).

6. Conclusion

It is well known that OLS is not robust to atypical observations. This lack of robustness causes distortion of parameter estimates and therefore leads to inaccurate predictions and inference, both of which were illustrated in the HBK and the Star Cluster data examples. Because of this problem, we implemented a method that downweights the impact of influential observations within linear regression. Unlike an ordinary least squares analysis, this robust regression method weights observations unequally. Observations were downweighted based on their extremeness in $(\mathbf{y} - E[\mathbf{y}|\mathbf{X}])$ and/or the \mathbf{X} -space.

Weights were implemented to limit the impact of each outlying or atypical observation on the overall fit and on the estimation of the parameters. Two versions of the Robust Forward Detection method were utilized: RFD1 and RFD2. The RFD1 method was used to implement dichotomous weights of 0 and 1 and therefore exclude each observation identified as an outlier, a leverage point, or an influential observation. The RFD2 method utilized continuous weights. As a result, the sample size is maintained because all observations are included in the analysis but with a different weight contribution.

When a robust procedure is implemented, it should produce results similar to, if not the same, as those produced by using the ordinary least squares approach when no atypical observations (outliers, leverage, and influential points) are present within the dataset. This fact was evident in the results from utilizing OLS, RFD1, and RFD2 in the simulations with no contamination (not shown). The weighting approach presented in this paper is a robust application that does not allow atypical observations (outliers, leverage,

and influential observations) to have a large impact on the estimation of parameters. The RFD2 method was shown to limit the impact of extreme observations in the \mathbf{X} -space and in the standardized deviations of $\mathbf{y} - E[\mathbf{y}|\mathbf{X}]$. The weight functions for the RFD2 method are based on the standardized prediction residuals, adjusted prediction intervals, and the robust distances. In contrast, the weights for some forms of weighted least squares are computed iteratively on the basis of the estimation of β .

Because of the random weights, we were unable to derive an exact calculation of the mean and variance of the weighted estimator. Instead, we utilized bootstrapping procedures, which are typically used to provide robust estimates of the mean and variance of statistics for which there are doubts regarding assumptions or for which the direct derivation is not straightforward. The 95% confidence intervals for the average parameter estimates from the simulations were presented. The confidence intervals indicated whether the estimates were stable and whether the estimates were statistically significant from zero.

References

- Carroll, R., Cline, D., 1988. An asymptotic theory for weighed least squares with weights estimated by replication. *Biometrika*, 75, 35–43.
- Carroll, R., 1982. Adapting for heteroscedasticity in linear models. *Annals of Statistics*, 10, 1224–1233.
- Davidian, M., Carroll, R., 1987. Variance function estimation. *Journal of the American Statistical Association*, 82, 1079–1091.
- Draper, N., Smith, H., 1998. *Applied Regression Analysis*. Wiley, New York.
- Hager, W., 1989. Updating the inverse of a matrix. *Society for Industrial and Applied Mathematics*, 31, 221–239.
- Hawkins, Bradu, Kass, 1984. Location of several outliers in multiple-regression data using elemental sets. *Techometrics*, 26, 197–208.
- Jones, T., Redden, D. A robust forward detection method for the identification of influential observations in linear regression. Ph.D. Dissertation. (Dept. of Biostatistics, University of Alabama at Birmingham, Birmingham, AL, 2007).
- Neter, J., Kutner, M., Nachtsheim, C., and Wasserman, W., 1996. *Applied Linear Statistical Models*. WCB/McGraw-Hill, New York.
- Rice, J., 1995. *Mathematical Statistics and Data Analysis*. Duxbury Press, California.
- Rousseeuw, P., Leroy, A., 1987. *Robust Regression and Outlier Detection*. Wiley, New York.
- Rousseeuw, P., van Zomeren, B., 1990. Unmasking multivariate outliers and leverage points,” *Journal of the American Statistical Association*, 85, 633–639.

Salibian-Barrera, M., Zamar, R., 2002. Bootstrapping robust estimates of regression.

The Annals of Statistics, 30, 556-582.

Seber, G., 1977. Linear Regression Analysis. Wiley, New York.

IMPLEMENTING THE ROBUST FORWARD DETECTION METHOD IN THE
LUNG HEALTH STUDY

by

TAMEKIA L. JONES AND DAVID T. REDDEN

In preparation for *CHEST*

Format adapted for dissertation

Abstract

The Lung Health Study data collected from October 1986 to April 1994 are utilized to illustrate the potential clinical applications of the Robust Forward Detection (RFD) method proposed by Jones and Redden.¹¹ The goal in this paper is to identify atypical observations (outliers, leverage, and influential points) by using the RFD method. We believe that this method may assist researchers in identifying atypical clinical observations. The RFD method is used to determine the relationship of atypical observations to lung cancer status by using forced expiratory volume in one second (FEV_1) as the response and by using age, smoke duration, and average number of cigarettes smoked per day as the predictors. The RFD method identified 111 leverage points. However, the RFD method did not declare any observations to be outliers. Therefore, no observations were identified as influential. The RFD identified 6.56% subjects to be extreme in the predictors (age, smoke duration, and average number of cigarettes smoked per day). In addition, an association between lung cancer status and leverage status ($p\text{-value}=0.0324 < 0.05=\alpha$) was found. The odds ratio indicated that the odds of possessing high values of the risk factors age, smoke duration, or average number of cigarettes smoked per day is 3.19 times higher for participants who died of lung cancer during the five year follow-up period than for those who did not do so during that time. On the basis of our results, we cannot conclude that extreme values of FEV_1 are significantly related to lung cancer. However, we did reconfirm that there is a relationship between dying of lung cancer and having extreme values of risk factors (age, smoke duration, and/or average number of cigarettes smoked per day), which are declared leverage points.

1. Introduction

As an important assessment of lung function, the forced expiratory volume in one second (FEV_1) measures the amount of air exhaled in a forcible manner in the first second. Typically, FEV_1 is measured in liters via a spirometer. Wise¹ indicated that FEV_1 is an important predictor of chronic obstructive pulmonary disease. Other researchers have examined the relationship of FEV_1 to lung cancer.^{2,3,4,5} From previous studies and their meta-analysis, Wasswa-Kintu et al⁴ concluded that there is “a strong inverse relationship between FEV_1 and lung cancer which applies to all levels of FEV_1 . The risk increases even with a relatively modest reduction in FEV_1 , especially among women.” Because it is one of the leading causes of cancer-related deaths in the United States,⁶ lung cancer is an important topic and a type of cancer that needs much attention. Furthermore, cigarette smoking is, at present, known to be one of the main risk factors associated with this seemingly preventable condition.⁷

The objective in this paper is to utilize the Lung Health Study (LHS) data to examine the relationship of FEV_1 and lung cancer via robust statistical techniques. The Robust Forward Detection method (RFD), the robust statistical technique proposed by Jones and Redden,⁸ is a statistical tool that identifies potential outliers, leverage points, or influential observations while overcoming the masking and swamping effects that are difficulties encountered with using traditional methods. Masking refers to incorrectly declaring an observation to be non-atypical, and swamping is falsely identifying an observation as atypical (outlier, leverage, or influential). For the purpose of this paper, the RFD identifies as outliers those subjects whose extreme standardized deviations in the response FEV_1 differ greatly from the bound of the observation’s corresponding adjusted

prediction interval. If a subject is extreme or has high values in any of the risk factors in comparison with the majority of the data, that subject is declared a leverage point according to the RFD method. Any subject extreme in both the response FEV_1 and the predictors is identified as an influential observation.

The goal in this paper is to detect atypical observations by using the RFD method in the LHS. We hypothesize that the atypical observations identified by the RFD method will identify those subjects who died of lung cancer within a five year follow-up period.

2. Methods

2.1 Study design

The Lung Health Study (LHS) is a randomized multi-center clinical trial conducted from October 1986 to April 1994. The study was designed to determine the effectiveness of intervention via smoking cessation and the use of a bronchodilator among 5887 cigarette smokers aged 35 to 60 at the time of enrollment. Any persons with FEV_1 values ranging between 55% and 90% were allowed to participate in the study. There were 1962 participants randomized to the no intervention group, 1962 randomized to the smoking intervention program with a placebo inhaler, and 1963 randomized to the smoking intervention program with an inhaler containing bronchodilator ipratropium bromide.⁹ Questionnaires and spirometry tests were administered yearly.

Of the entire sample of 5887 subjects in the LHS, more than 94% reported to the annual visit at the fifth year.⁷ For the purpose of this paper, we restricted our analyses to data obtained within the five year follow-up period and to subjects who were at least 54 years of age (i.e., the 75th percentile of age). This restriction was implemented because it

is believed that lung cancer is more prevalent in the older population. Therefore, our sample included 1691 of the 5887 smokers initially enrolled in the LHS. The first annual FEV₁ was utilized as the response variable predicted by age, smoke duration, and the average number of cigarettes smoked per day. Age, at baseline, was defined as the age of the subject on the day of randomization into one of the three groups. Smoke duration was considered the length of time between the person's age when he/she first began smoking cigarettes and his/her age at enrollment. The average number of cigarettes smoked per day was based on the average across the entire time that the subject smoked. Because of the limitation of using categorical variables in the RFD method, we were unable to include dichotomous variables such as gender in our model.

2.2 Statistical methods

The Robust Forward Detection method,⁸ is a robust regression method appropriate for the linear regression framework, is a screening tool utilized to detect atypical observations (outliers, leverage, and influential points) that Ordinary Least Squares (OLS) is not always capable of detecting. In addition, the RFD is capable of overcoming masking and swamping. To ensure that masking or swamping does not occur, the RFD method uses robust distances, the minimum covariance determinant, and Rousseeuw's¹⁰ concept of least trimmed squares (LTS) to begin with a subset free of atypical observations. Note that the method begins with a leverage-free subset and moves forward and continuously updates the robust distances in order to identify leverage points. After this step, the procedure continues by using LTS to select a subset of the observations among the non-leveraged points that is free of outliers. The method

identifies outliers via standardized prediction residuals. An observation is declared an outlier if its standardized prediction residual is located beyond the bounds of the adjusted prediction intervals. This step continues until all observations have been evaluated and tested for outliers. Furthermore, any observation that is detected as an outlier and a leverage point is declared an influential observation.

Our statistical analyses exemplify results obtained by utilizing the RFD in two capacities: RFD1 and RFD2. RFD1 is the RFD method with the application of dichotomous weights. That is, any subject identified by the RFD method as an outlier and/or a leverage point, is given a weight of 0 and therefore excluded from the analyses. Otherwise, the subject receives a weight of 1 and is included in the statistical analyses. The RFD2 method is the RFD approach that utilizes continuous weights. These weights range from 0 to 1, inclusive, and are defined on the basis of the adjusted prediction intervals, standardized prediction residuals, and robust distances, all of which are defined in Jones and Redden.¹¹ All statistical analyses were performed using SAS[®] version 9.1.3.¹²

3. Results

Table 1 provides baseline characteristics by lung cancer status for our sample of 1691 subjects. Many subjects lived beyond the five year follow-up period. There were approximately 4.85% ($= 82/1691 \times 100$) deaths of which 34.15% ($= 28/82 \times 100$) were attributed to lung cancer. Our sample consisted of 1098 males (64.93%) and 593 females (35.07%). Table 1 indicates that 71% of the participants in the sample who died of lung cancer within the five year follow-up period were males and that 29% were females. The

sample consisted of 95.03 % (n = 1607) Caucasians and 4.73% (n = 80) African Americans. Because this distribution was skewed, we did not utilize race as a variable in our model when predicting FEV₁. As can be seen in Table 1, no statistically significant differences were found between those participants, using baseline characteristics, who died of lung cancer and those who did not die of lung cancer during the five year follow-up period.

Table 1. Baseline characteristics for subjects 54⁺ years of age by lung cancer status during the 5 year follow-up period

Baseline characteristics	Died of lung cancer (n=28)		Did not die of lung cancer (n=1663)		p-value
	Mean	Std Dev	Mean	Std Dev	
Age (years)	56.57	1.85	56.57	1.78	0.9956
Gender (%)					
Males	71.43	—	64.82	—	0.5525
Race (%)					
Caucasians	89.29	—	95.13	—	0.1591
African Americans	10.71	—	4.63	—	0.1433
Average cigarettes/day	28.89	12.22	25.37	9.86	0.1399
FEV ₁	2.49	0.5046	2.45	0.5669	0.6942
Smoke duration	39.46	7.53	38.49	4.56	0.5020
Age started smoking (years)	17.11	7.15	18.08	4.29	0.4807
Treatment group (%)					
No intervention	28.57	—	32.95	—	0.6902
Intervention+placebo	42.86	—	33.67	—	0.3184
Intervention+active	28.57	—	33.37	—	0.6889

Note: Std Dev = Standard Deviation.

For this paper, the RFD method was implemented to identify atypical (outliers, leverage, and influential points) observations among the 1691 subjects in the LHS. The RFD method identified 111 leverage points. However, the RFD method did not declare any observations to be outliers; therefore, no observations were identified as influential. That is, no observations were identified by the RFD method to have extreme values of

FEV₁. On the other hand, the RFD identified 6.56% subjects as having extreme values in the predictors (age, smoke duration, and average number of cigarettes smoked per day).

Downweighting is important because it prevents a single observation from unduly influencing the analysis. When RFD2 is being utilized, outliers are intended to be downweighted by using weights a_k , and leverage points are intended to be downweighted using weights b_k (see Table 2 in which a partial list of the observations identified by the RFD as leverage points is provided). Weights g_k ranged from 0.00664 to 0.98258, as indicated in Table 2. The magnitude of each weight is an indication of a subject's degree of extremeness in the predictors (age, smoke duration, average number of cigarettes smoked per day) relative to the remaining subjects. On the basis of the information provided in Table 2, we are able to determine that the subject with id 5733 was identified as the subject most extreme in the predictors. We can make this conclusion because this subject has the smallest weight. Among the atypical observations, the subject with id 125 was the least extreme in the predictors.

Table 2. Weights $g_k (= a_k * b_k)$ implemented by using weight functions via the RFD2 method for the Lung Health Study data

id	g_k	a_k	b_k	Contamination type
5733	0.00664	1	0.00664	Leverage
485	0.01462	1	0.01462	Leverage
5863	0.01741	1	0.01741	Leverage
...
5707	0.97668	1	0.97668	Leverage
5877	0.98203	1	0.98203	Leverage
125	0.98258	1	0.98258	Leverage

Table 3 displays the parameter estimates (standard deviation), MSE, and R^2 obtained when OLS, RFD1, and RFD2 are used on the Lung Health Study data. Note that, under OLS, all variables were statistically significant (p-value < 0.0001) in

predicting FEV_1 in the presence of each other. Parameter estimates for the RFD2 are obtained by implementing weights g_k to limit the impact of those observations identified as leverage points. We assume that, because the RFD method is capable of detecting leverage points and overcoming the masking effect, the parameter estimates based on the RFD1 method can serve as the baseline for comparison of and with OLS and RFD2. It is noted that all three methods produce similar parameter estimates. Because the LHS was found to be contaminated with leverage points only, the weights implemented by using RFD1 and RFD2 are not functions of random variables. Therefore, we know that the standard deviations produced by SAS[®] version 9.1.3¹² are valid. Furthermore, while controlling for the length of time that a subject smoked and for the average number of cigarettes a subject smoked per day, the results from each method suggest that, as a person ages, there is a reduction in his/her FEV_1 .

Table 3. Parameter estimates (Std Dev), MSE, and R^2 obtained when OLS, RFD1, and RFD2 are used to analyze the Lung Health Study data

	OLS	RFD1	RFD2
Intercept (Std Dev)	3.56264 (0.43579)	3.59196 (0.46929)	3.54137 (0.44605)
Age (Std Dev)	-0.03263 (0.00812)	-0.03887 (0.00918)	-0.03638 (0.00855)
Smk_dur (Std Dev)	0.01555 (0.00315)	0.02297 (0.00439)	0.02107 (0.00387)
F31ctavg (Std Dev)	0.00542 (0.00138)	0.00695 (0.00169)	0.00620 (0.00151)
MSE	0.31097	0.31324	0.30126
R^2	0.0304	0.0338	0.0335

Note: Std Dev = Standard Deviation.

Caution must be taken when the MSE and R^2 for the LHS are compared across all three methods, especially because the RFD1 method implements dichotomous weights of

0 and 1. This implementation causes the degrees of freedom for RFD1 to differ from those computed using OLS and RFD2. The MSE under OLS is lower than the MSE under RFD1. This difference occurs because the sample size is large and because $SSE_{OLS} = 524.61$ ($df_{OLS} = 1687$) and $SSE_{RFD1} = 493.67$ ($df_{RFD1} = 1576$).

We were interested in knowing whether the RFD method is capable of identifying lung cancer patients based on those subjects with extreme FEV_1 values and/or extreme values in the predictors (age, smoke duration, and average number of cigarettes smoked per day). We were able to link the atypical observations back to participants who died of lung cancer within the five year follow-up period. As Table 1 indicates, 28 subjects died of lung cancer before the end of the five year follow-up period. Table 4 indicates that less than 0.5% of our sample who had extreme values in the predictors died of lung cancer. In addition, only 5 of the 111 subjects (4.5%) identified as leverage points by the RFD method died of lung cancer. The corresponding Fisher's Exact test (utilized because of the small expected cell counts) for Table 4 suggests there is an association between lung cancer status and leverage status ($p\text{-value}=0.0324 < 0.05=\alpha$). The odds ratio indicates that the odds of possessing extreme values of age, smoke duration, or average number of cigarettes smoked per day is 3.19 times higher for participants who died of lung cancer during the five year follow-up period than for those who did not do so during that time period. The 95% confidence interval for the odds ratio (1.19, 8.57) further supports the conclusion based on Fisher's Exact test and allows us to reject the null hypothesis. Therefore, we can conclude that there exists an association between whether a subject died of lung cancer by the end of the fifth year follow up and whether a subject had extreme values on any of the predictors in comparison with the majority of the data.

Table 4. Crosstab of leverage status (determined by using the RFD method) by lung cancer status for the Lung Health Study

	Died of lung cancer	Did not die of lung cancer	Total
Leverage	5	106	111
No leverage	23	1557	1580
Total	28	1663	1691

4. Discussion

The LHS data from the National Heart, Lung, and Blood Institute was examined for influential observations by utilizing the RFD method. OLS, RFD1, and RFD2 produced similar results for the parameter estimates, the MSE, and the R^2 . However, because the degrees of freedom obtained by using the RFD1 method differ from those obtained using OLS and RFD2, caution is warranted when the MSE and R^2 for RFD1 are being directly compared with those found for OLS and for RFD2. Jones and Redden's⁸ RFD method identified 111 observations to have extreme values in at least one of the predictors (age, smoke duration, and average number of cigarettes smoked per day). Of the 111 subjects identified as leverage points, only 5 of 28 subjects were identified to have died of lung cancer by the end of the five year follow-up period.

Table 4 can be viewed in terms of specificity and sensitivity where the truth is lung cancer status and where the test outcome is whether a subject was identified by the RFD as being extreme in any of the predictors (i.e., leverage). Specificity, which measures the proportion of people who did not die of lung cancer and were not captured as leverage, is 93.6% ($= [1557 \cdot 100 / 1663]$). Sensitivity, which measures the proportion of

subjects who died of lung cancer and were identified by the RFD to be extreme in the predictors, is approximately 18%. This information indicates that FEV_1 is not the best screening tool to use to determine which patients are at highest risk of dying of lung cancer.

5. Concluding remarks

In this paper, we utilized the Robust Forward Detection method to identify atypical observations in the Lung Health Study, with continuous variables (age, smoke duration, and average number of cigarettes smoked per day) predicting FEV_1 . Weight functions were used that allowed unequal weighting of atypical observations identified by the RFD method. We illustrated and compared parameter estimates obtained by utilizing OLS with those obtained by using RFD1 and RFD2. There was a statistically significant association between leverage status and lung cancer related death status. Based on our results, we cannot conclude that extreme values of FEV_1 are significantly related to lung cancer related deaths. However, we did reconfirm that subjects who died of lung cancer is related to their having extreme levels of risk factors (age, smoke duration, and/or average number of cigarettes smoked per day), which are declared leverage points.

6. Acknowledgments

We thank Drs. Bailey, Gerald, and Soong for guidance in examining the most appropriate variables. We acknowledge Dr. Connett for his help with details for the data within the database of the Lung Health Study.

References

- ¹ Wise R. The value of forced expiratory volume in 1 second decline in the assessment of chronic obstructive pulmonary disease progression. *The American Journal of Medicine* 2006; 119:S4 – S11.
- ² Eberly L, Ockene J, Sherwin R, et al. Pulmonary function as a predictor of lung cancer mortality in continuing cigarette smokers and in quitters for the Multiple Risk Factor Intervention Trial Research Group. *International Journal of Epidemiology* 2003; 32:592–599
- ³ Van Den Eeden SK, Friedman GD. Forced expiratory volume (1 second) and lung cancer incidence and mortality. *Epidemiology* 1992; 3:253–257.
- ⁴ Wasswa-Kintu S, Gan W, Man S, Pare P, Sin D. Relationship between reduced forced expiratory volume in one second and the risk of lung cancer: a systematic review and meta-analysis. *Thorax* 2005; 60:570 – 575.
- ⁵ Young RP, Hopkins R, Eaton TE. Forced expiratory volume in one second: not just a lung function test but a marker of premature death from all causes. *Eur Respir J* 2007; 30: 616–622.
- ⁶ Guessous I, Cornuz J, Paccaud F. Lung cancer screening: current situation and perspective. *Swiss Medical Weekly* 2007; 137:304 – 311.
- ⁷ Gerald G, Miller D, Anthonisen N, et al. Risk factors for lung cancer in the Lung Health Study. Unpublished, 1999.
- ⁸ Jones T, Redden D. A robust forward detection method for the identification of influential observations in linear regression. Unpublished, 2007.

- ⁹ Connett J, Kusek J, Bailey W, O'Hara P, Wu M. Design of the lung health study: a randomized clinical trial of early intervention for chronic obstructive pulmonary disease. *Controlled Clinical Trials* 1993; 14:3S – 19S.
- ¹⁰ Rousseeuw P. Multivariate estimation with high breakdown point. In: Grossmann, W., Pflug, G., Vincze, I., Wertz, W., Eds. *Mathematical Statistics and Applications*, Vol. B. Reidel, Dordrecht, 283 – 297.
- ¹¹ Jones T, Redden D. Limiting the impact of influential observations in linear regression via weight functions. Unpublished, 2007.
- ¹² SAS Institute Inc. SAS Online[®] 9.1.3. North Carolina: SAS Institute Inc., 2004.

CONCLUSIONS

Diagnostic issues such as atypical observations (outliers, leverage, and influential points), masking, and swamping, which are often encountered in linear regression via ordinary least squares (OLS), were the motivation for this dissertation. Outliers are observations with extreme standardized deviations of $(y - E[y|X])$; leverage points are observations with extreme deviations in the predictor or X -space, and influential observations are those observations considered to be both outliers and leverage points. The masking effect occurs when a contaminated observation is not identified as atypical, whereas swamping occurs when an observation is incorrectly identified as atypical. It is well known that OLS is not robust or resistant to these diagnostic issues and that their presence can have a large impact on estimation, prediction, and inference.

Concluding Remarks for Paper 1

We developed the Robust Forward Detection method (RFD) under the robust regression statistical framework. This method uses a combination of Rousseeuw and van Zomeren's (1990) concept of robust distances (RD), which uses the minimum covariance determinant (MCD), and Rousseeuw's (1985) concept of least trimmed squares (LTS). At first glance, our RFD method may seem to resemble the Forward Search (FS) method by Atkinson and Riani (2000), but the two approaches have their differences. Both methods begin with a robust procedure to obtain an initial subset. However, our RFD method utilizes Rousseeuw and van Driessen's (1999) FAST-MCD to obtain an initial subset that

is free of leverage points, whereas Atkinson and Riani's (2000) FS method utilizes either Rousseeuw's (1984, 1985) LMS or LTS method to attain an initial subset free of outliers. In addition, both methods monitor the forward progression of adding observations to the initial subset. Unlike Atkinson and Riani's (2000) FS method, which adds observations to the initial subset based on the smallest squared unadjusted residuals, our RFD subset allows an observation to enter the subset on the basis of its robust distance and its standardized prediction residuals.

Simulations were conducted using various levels of contamination ranging from 0% to 40% of outliers, leverage, and influential observations. It was demonstrated via simulations that our proposed method is comparable to LTS (ROBUSTREG) as presented in SAS[®] version 9.1.3; the latter method is known to be robust to outliers and leverage points. The two methods are similar but yet differ in how the combination of RD and LTS is utilized. Unlike LTS (ROBUSTREG), the RFD method begins with both leverage-free and outlier-free subsets. Leverage points are identified on the basis of the threshold given in Jones and Redden (2007) for the robust distances. The initial subset for the LTS method in the RFD approach is based on the non-leveraged observations. That is, an outlier-free subset is obtained after all of the leverage and non-leveraged points are identified within the data. By proceeding in this manner, the RFD approach is capable of detecting 99% influential observations, even when there is 40% contamination. The LTS (ROBUSTREG) method does not possess this capability. We do note that LTS (ROBUSTREG) has better detection capability when there is 30% or less outlier contamination. However, in the presence of 40% outlier contamination, LTS (ROBUSTREG) captures approximately 53% of the outliers and thus masks 47%. On the

other hand, when outlier contamination is present, our proposed RFD is capable of detecting 82% of the outliers and masks approximately 18%. We conclude that the RFD approach can overcome swamping and masking to a certain extent while properly detecting influential observations.

Concluding Remarks for Paper 2

The RFD approach utilizes dichotomous weights of 0 and 1 when computing final parameter estimates; this approach is also the RFD1 method. We were interested in the effects of implementing continuous weights on atypical observations so that all observations could be included in the analyses; this approach is denoted RFD2. Based on simulation results, we noticed that there was little bias in the parameter estimates under RFD1 and RFD2 in the presence of outlier and leverage contamination. Therefore, we can conclude that parameter estimates based on using both the RFD1 and RFD2 methods in the presence of outlier and leverage contamination were comparable to those of the true simulated parameter underlying the uncontaminated data. When estimating parameters, both the RFD1 and RFD2 methods provide parameter estimates closer to the true simulated parameter than OLS does. In addition, the simulation results indicated that the parameter estimates obtained by using the RFD1 and RFD2 methods when there was influence were more reliable than those obtained by using OLS. Note that, because there is a change in the degrees of freedom under RFD1, caution is advised when the MSE is being compared across all three methods.

In the presence of leverage contamination only, weights were based on non-random variables, and the usual distributional assumptions of the parameter β held.

However, the weights obtained from the weight functions in the presence of outlier or influential contamination were random because of the use of standardized prediction residuals and because of the use of the upper bound of the adjusted prediction interval. This randomness caused changes in the distributional assumptions for our parameter; as a result, we exercise caution when making inference. Due to indicated derivations, we expected our new parameter $\hat{\beta}_w$ to be biased. We also expected that the variance of our parameter estimates under the RFD1 and RFD2 methods would differ from what would be obtained by implementing the theory of WOLS as if the weights are known. Therefore, we recommend against using WOLS to approximate standard errors of robust regression estimators when weights are based on functions of random variables and if inference is required. We agree with Salibian-Barrera and Zamar (2002) that 95% confidence intervals for parameter estimates should be estimated from bootstrapping approaches. Our simulation tables indicated that the confidence intervals were narrow. This finding suggests that the parameter estimates obtained via bootstrapping are stable and reliable. Note that, if researchers are concerned about the normality assumption, inference can be made by using the appropriate bootstrapped confidence intervals. If a p-value is desired, permutation tests can be utilized.

Concluding Remarks for Paper 3

A comparison of the RFD method with OLS was performed by using the Lung Health Study data conducted from October 1986 to April 1994. There were 5887 smokers who were 35 to 60 years old. For the purpose of this dissertation, we restricted our sample so that it consisted only of subjects at least 54 years of age; this criterion yielded a

sample size of 1691 subjects. Our statistical model predicted FEV_1 based on the continuous predictors age, smoke duration, and average number of cigarettes smoked per day. The RFD method was implemented, and it declared no outliers but identified 111 observations as leverage. Unequal weighting of observations was allowed via the RFD2 method. OLS, RFD1, and RFD2 yielded similar results in terms of the parameter estimates and standard deviations. Although caution is warranted when RFD1 is being compared with OLS and RFD2, we noticed that the MSE and R^2 were similar across all three methods. Based on the results from the RFD method and lung cancer related death status, we were able to determine that approximately 7% of the 1691 subjects were identified as leverage points, less than 2% of the sample died of lung cancer, and that less than 0.5% of the sample died of lung cancer and was extreme in at least one of the predictors. Based on the results in Paper 3, we are not able to conclude that there is a statistically significant relationship between extreme values of FEV_1 and deaths due to lung cancer.

Future Research

Future research needs to be conducted to obtain a better understanding of the limitation of using dichotomous variables under the RFD method. In addition, it is appropriate to extend the development of the RFD method by utilizing both dichotomous and continuous weights that are functions of random variables in order to conduct hypothesis testing and make inference. Exact or asymptotic derivations and computations of the distribution of the weighted parameter can lead to the development of test statistics and corresponding p-values when inference needs to be made. In addition, further

investigation of the relationship between FEV1 and lung cancer via robust statistical methods beyond linear regression can contribute to the current body of knowledge and possibly lead to improved statistical models.

GENERAL LIST OF REFERENCES

- Atkinson, Riani, 2000. Robust Diagnostic Regression Analysis. Springer-Verlag, New York.
- Gray, B., Woodall, W., 1994. The maximum size of standardized and internally studentized residuals in regression analysis. *The American Statistician*, 48, 111–113.
- Hadi, A., 1992. Identifying multiple outliers in multivariate data. *Journal of the Royal Statistical Society, Ser. B*, 54, 761–771.
- Hadi, A., 1994. A modification of a method for the detection of outliers in multivariate samples. *Journal of the Royal Statistical Society, Ser. B*, 56, No. 2, 393–396.
- Hadi, A., Simonoff, J., 1993. Procedures for the identification of multiple outliers in linear models. *Journal of the American Statistical Association*, 88, 1264–1272.
- Hampel, F., 1968. Contributions to the Theory of Robust Estimation, Ph.D. thesis, University of California, Berkeley.
- Hampel, F, Ronchetti, E., Rousseeuw, P, Stahel, W., 1986. Robust Statistics: the Approach Based on Influence Functions. Wiley, New York.
- Hoaglin, D., Welsch, R., 1978. The hat matrix in regression and ANOVA. *The American Statistician*, 32, 17–22.
- Huber, P., 1964. Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 35, 73–101.

- Huber, P., 1981. Robust Statistics. Wiley, New York.
- Jones, T., Redden, D., 2007. A robust forward detection method for the identification of influential observations in linear regression. Unpublished.
- Muller, Fetterman, 2002. Regression and ANOVA: An Integrated Approach Using SAS Software. Wiley, New York.
- Neter, J., Kutner, M., Nachtsheim, C., Wasserman, W., 1996. Applied Linear Statistical Models. WCB/McGraw-Hill, New York.
- Pena, D., Yohai, V., 1999. A fast procedure for outlier diagnostics in large regression Problems. The Journal of the American Statistical Association, 94, 434–445.
- Rousseeuw, P., 1984. Least median of squares regression. Journal of the American Statistical Association, 79, 871–880.
- Rousseeuw, P., 1985. Multivariate estimation with high breakdown point. In: Grossmann, W., Pflug, G., Vincze, I., Wertz, W. (Eds.), Mathematical Statistics and Applications, Vol. B. Reidel, Dordrecht, pp. 283–297.
- Rousseeuw, P., Leroy, A., 1987. Robust Regression and Outlier Detection. Wiley, New York.
- Rousseeuw, P., van Zomeren, B., 1990. Unmasking multivariate outliers and leverage points. Journal of the American Statistical Association, 85, 633–639.
- Rousseeuw, P., van Driessen, K., 1999. A fast algorithm for the minimum covariance determinant. Technometrics, 41, 212–223.
- Salibian-Barrera, M., Zamar, R., 2002. Bootstrapping robust estimates of regression. The Annals of Statistics, 30, 556–582.

APPENDIX A

DERIVATION OF $\hat{\boldsymbol{\beta}}_w$

We use previous information from Paper 2 and the Sherman-Morrison Woodbury Theorem (Hager 1989) to provide the derivation of $\hat{\beta}_w = (\mathbf{X}'\mathbf{G}\mathbf{X})^{-1}\mathbf{X}'\mathbf{G}\mathbf{y}$ that yields more insight into computing the expectation and variance when the weights are functions of random variables. From Jones and Redden (2007) and Hadi and Simonoff (1993), we know that

$$d_k \sim t\left(1 - \frac{\alpha}{v_{\text{sum}} + 1}, v_{\text{sum}} - p\right).$$

Basic statistical textbooks such as Rice (1995) teach us that

$$t^2 \sim F(1, v_{\text{sum}} - p)$$

and that

$$(1/F) \sim F(v_{\text{sum}} - p, 1).$$

Therefore, we can write

$$(1/d_k)^2 \sim F(v_{\text{sum}} - p, 1).$$

Note that the adjusted upper prediction interval is a constant dependent on α , v_{sum} , and

$\mathbf{h}_k = \mathbf{x}_k (\mathbf{X}'\mathbf{V}\mathbf{X})^{-1} \mathbf{x}_k'$. Hence,

$$\begin{aligned} a_k &= \left(\text{UpperPI}_{\text{adj}} / d_k \right)^2 \\ &= c_* F(v_{\text{sum}} - p, 1), \end{aligned}$$

where c_* is some constant.

We are interested in the form of the weighted estimator in the presence of atypical observations. If we look at the case in which there is only one outlier, then we can generalize our results to see how the random weights affect the estimator. Let

$$\mathbf{G} = \mathbf{I} - \begin{bmatrix} 1-g_1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ \cdot & & \cdot & & \\ \cdot & & & \cdot & \\ 0 & \dots & 0 & 0 & \dots & 0 \end{bmatrix}$$

$$= \mathbf{I} - \mathbf{K}.$$

Then, we can write

$$\begin{aligned} (\mathbf{X}'\mathbf{G}\mathbf{X}) &= \mathbf{X}'(\mathbf{I} - \mathbf{K})\mathbf{X} \\ &= \mathbf{X}'\mathbf{X} - \mathbf{X}'\mathbf{K}\mathbf{X} \\ &= \mathbf{X}'\mathbf{X} - \mathbf{X}' \begin{bmatrix} 1-g_1 & \mathbf{0}' \\ \mathbf{0} & 0 \end{bmatrix} \mathbf{X} \\ &= \mathbf{X}'\mathbf{X} - \begin{bmatrix} x_{11}(1-g_1) & \mathbf{0}' \\ (1-g_1)\mathbf{0} & 0 \end{bmatrix} \begin{bmatrix} x_{11} & \mathbf{c}' \\ \mathbf{f} & \mathbf{X}_2 \end{bmatrix} \\ &= \mathbf{X}'\mathbf{X} - \begin{bmatrix} x_{11}^2(1-g_1) & x_{11}(1-g_1)\mathbf{c}' \\ x_{11}(1-g_1)\mathbf{c} & (1-g_1)\mathbf{c}\mathbf{c}' \end{bmatrix} \\ &= \mathbf{X}'\mathbf{X} - (1-g_1) \begin{bmatrix} x_{11}^2 & x_{11}\mathbf{c}' \\ x_{11}\mathbf{c} & \mathbf{c}\mathbf{c}' \end{bmatrix} \\ &= \mathbf{X}'\mathbf{X} - (1-g_1) \begin{bmatrix} x_{11} \\ \mathbf{c} \end{bmatrix} \begin{bmatrix} x_{11} \\ \mathbf{c} \end{bmatrix}'. \end{aligned} \quad (1)$$

Continuing with (1), we use the Sherman-Morrison Woodbury Theorem (Hager 1989) to find

$$\begin{aligned} (\mathbf{X}'\mathbf{G}\mathbf{X})^{-1} &= \left(\mathbf{X}'\mathbf{X} - (1-g_1) \begin{bmatrix} x_{11} \\ \mathbf{c} \end{bmatrix} \begin{bmatrix} x_{11} \\ \mathbf{c} \end{bmatrix}' \right)^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1} + \frac{(\mathbf{X}'\mathbf{X})^{-1}(1-g_1) \begin{bmatrix} x_{11} \\ \mathbf{c} \end{bmatrix} \begin{bmatrix} x_{11} \\ \mathbf{c} \end{bmatrix}' (\mathbf{X}'\mathbf{X})^{-1}}{1 - \begin{bmatrix} x_{11} \\ \mathbf{c} \end{bmatrix}' (\mathbf{X}'\mathbf{X})^{-1} \begin{bmatrix} x_{11} \\ \mathbf{c} \end{bmatrix} (1-g_1)} \end{aligned}$$

$$\begin{aligned}
&= (\mathbf{X}'\mathbf{X})^{-1} + \frac{(\mathbf{X}'\mathbf{X})^{-1} \begin{bmatrix} x_{11} \\ \mathbf{c} \end{bmatrix} \begin{bmatrix} x_{11} \\ \mathbf{c} \end{bmatrix}' (\mathbf{X}'\mathbf{X})^{-1}}{\frac{1}{1-g_1} - \begin{bmatrix} x_{11} \\ \mathbf{c} \end{bmatrix}' (\mathbf{X}'\mathbf{X})^{-1} \begin{bmatrix} x_{11} \\ \mathbf{c} \end{bmatrix}} \\
&= (\mathbf{X}'\mathbf{X})^{-1} + \frac{\mathbf{M}_1 \mathbf{M}_1'}{\frac{1}{1-g_1} - \begin{bmatrix} x_{11} \\ \mathbf{c} \end{bmatrix}' \mathbf{M}_1},
\end{aligned}$$

where $\mathbf{M}_1 = (\mathbf{X}'\mathbf{X})^{-1} \begin{bmatrix} x_{11} \\ \mathbf{c} \end{bmatrix}$. Now, rewrite $(\mathbf{X}'\mathbf{G}\mathbf{X})^{-1} = (\mathbf{X}'\mathbf{X})^{-1} + \mathbf{M}^*$, where

$$\begin{aligned}
\mathbf{M}^* &= \mathbf{M}_1 \mathbf{M}_1' \left[\frac{1-g_1}{1-(1-g_1) \begin{bmatrix} x_{11} \\ \mathbf{c} \end{bmatrix}' \mathbf{M}_1} \right] \\
&= \mathbf{M}_1 \mathbf{M}_1' \left[\frac{1-\mathbf{c}_*F}{1+(1-\mathbf{c}_*F)\mathbf{u}} \right] \\
&= \mathbf{M}_1 \mathbf{M}_1' \mathbf{K},
\end{aligned}$$

where $g_1 = \mathbf{c}_*F(v_{\text{sum}} - p, 1)$, $\mathbf{u} = -\begin{bmatrix} x_{11} \\ \mathbf{c} \end{bmatrix}' \mathbf{M}_1$, and $\mathbf{K} = \left[\frac{1-\mathbf{c}_*F}{1+(1-\mathbf{c}_*F)\mathbf{u}} \right]$. Now, we can rewrite \mathbf{K} and solve for α and β in the following manner:

$$\Rightarrow \alpha + \frac{\beta}{1+\mathbf{u}-\mathbf{u}\mathbf{c}_*F} = \frac{1-\mathbf{c}_*F}{1-\mathbf{u}\mathbf{c}_*F+\mathbf{u}} \quad (2)$$

$$\Rightarrow \alpha(1+\mathbf{u}-\mathbf{u}\mathbf{c}_*F) + \beta = 1-\mathbf{c}_*F$$

$$\Rightarrow \alpha(1+\mathbf{u}) - \alpha\mathbf{u}\mathbf{c}_*F + \beta = 1-\mathbf{c}_*F$$

$$\Rightarrow \text{(i) } \alpha(1+\mathbf{u}) + \beta = 1 \quad \text{and} \quad \text{(ii) } -\alpha\mathbf{u}\mathbf{c}_*F = -\mathbf{c}_*F.$$

We can solve (ii) such that $\alpha = \frac{1}{\mathbf{u}}$. By substitution, solve (i) by using the solution from (ii) such that

$$\alpha(1+\mathbf{u}) + \beta = 1$$

$$\Rightarrow \frac{1}{u}(1+u) + \beta = 1$$

$$\Rightarrow \frac{1}{u} + 1 + \beta = 1$$

$$\Rightarrow \beta = -\frac{1}{u}.$$

So, we can now rewrite (2) such that

$$\begin{aligned} \alpha + \frac{\beta}{1+u - uc_*F} &= \frac{1 - c_*F}{1 - uc_*F + u} \\ \Rightarrow \frac{1}{u} - \frac{1/u}{1+u - uc_*F} &= \frac{1 - c_*F}{1 - uc_*F + u}. \end{aligned} \quad (3)$$

Now let

$$z = \frac{1/u}{1+u - uc_*F}. \quad (4)$$

We can rewrite (3) in the following manner:

$$\Rightarrow z + zu - zuc_*F = \frac{1}{u}$$

$$\Rightarrow z + zu - \frac{1}{u} = zuc_*F$$

$$\Rightarrow \frac{1}{zuc_*} \left(z + zu - \frac{1}{u} \right) = F$$

$$\Rightarrow \frac{1+u}{uc_*} - \frac{1}{zu^2c_*} = g^{-1}(z).$$

From basic statistical textbooks such as Rice (1995), we know that the probability distribution function of a random variable following an F -distribution with η and 1 degrees of freedom is as follows:

$$f_X(x|\eta,1) = \frac{\Gamma\left(\frac{\eta+1}{2}\right)}{\Gamma\left(\frac{\eta}{2}\right)\Gamma\left(\frac{1}{2}\right)} \left(\frac{\eta}{1}\right)^{\eta/2} \frac{x^{(\eta-2)/2}}{\left(1+\left(\frac{\eta}{1}\right)x\right)^{(\eta+1)/2}}.$$

Using this information and the fact z is a function of F as seen in (4), we can find its probability distribution function:

$$f_Z(z) = \frac{\Gamma\left(\frac{\eta+1}{2}\right)}{\Gamma\left(\frac{\eta}{2}\right)\Gamma\left(\frac{1}{2}\right)} \left(\frac{\eta}{1}\right)^{\eta/2} \frac{\left(\frac{1+u}{uc_*} - \frac{1}{zu^2c_*}\right)^{(\eta-2)/2}}{\left(1+\left(\frac{\eta}{1}\right)\left(\frac{1+u}{uc_*} - \frac{1}{zu^2c_*}\right)\right)^{(\eta+1)/2}} \left|\frac{d}{dz}g^{-1}(z)\right|$$

$$f_Z(z) = \frac{\Gamma\left(\frac{\eta+1}{2}\right)}{\Gamma\left(\frac{\eta}{2}\right)\Gamma\left(\frac{1}{2}\right)} \left(\frac{\eta}{1}\right)^{\eta/2} \frac{\left(\frac{1+u}{uc_*} - \frac{1}{zu^2c_*}\right)^{(\eta-2)/2}}{\left(1+\left(\frac{\eta}{1}\right)\left(\frac{1+u}{uc_*} - \frac{1}{zu^2c_*}\right)\right)^{(\eta+1)/2}} \left|\frac{1}{z^2u^2c_*}\right|.$$

Combining the above results, we can write the following equation.

$$\begin{aligned} (\mathbf{X}'\mathbf{GX})^{-1} &= (\mathbf{X}'\mathbf{X})^{-1} + \mathbf{M}_1\mathbf{M}_1'\left(\frac{1-c_*F}{1+u(1-c_*F)}\right) \\ &= (\mathbf{X}'\mathbf{X})^{-1} + \mathbf{M}_1\mathbf{M}_1'\left(\frac{1}{u} - z\right) \end{aligned}$$

Then, we can write the new weighted estimator in the following manner:

$$\begin{aligned} \hat{\boldsymbol{\beta}}_w &= (\mathbf{X}'\mathbf{GX})^{-1}\mathbf{X}'\mathbf{G}\mathbf{y} \\ &= (\mathbf{X}'\mathbf{GX})^{-1}\mathbf{X}'\mathbf{G}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) \\ &= (\mathbf{X}'\mathbf{GX})^{-1}\mathbf{X}'\mathbf{GX}\boldsymbol{\beta} + (\mathbf{X}'\mathbf{GX})^{-1}\mathbf{X}'\mathbf{G}\boldsymbol{\varepsilon} \\ &= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{GX})^{-1}\mathbf{X}'\mathbf{G}\boldsymbol{\varepsilon} \end{aligned}$$

$$\begin{aligned}
&= \boldsymbol{\beta} + \left((\mathbf{X}'\mathbf{X})^{-1} + \mathbf{M}_1\mathbf{M}_1' \left(\frac{1}{u} - z \right) \right) \mathbf{X}'\mathbf{G}\boldsymbol{\varepsilon} \\
&= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{G}\boldsymbol{\varepsilon} + \mathbf{M}_1\mathbf{M}_1' \left(\frac{1}{u} - z \right) \mathbf{X}'\mathbf{G}\boldsymbol{\varepsilon}.
\end{aligned}$$

In the presence of only one outlier, we have

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_w &= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \begin{bmatrix} \mathbf{g}_1 & \mathbf{0}' \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \boldsymbol{\varepsilon} + \mathbf{M}_1\mathbf{M}_1' \left(\frac{1}{u} - z \right) \mathbf{X}' \begin{bmatrix} \mathbf{g}_1 & \mathbf{0}' \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \boldsymbol{\varepsilon} \\
&= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \begin{bmatrix} g_1 \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} + \mathbf{M}_1\mathbf{M}_1' \left(\frac{1}{u} - z \right) \mathbf{X}' \begin{bmatrix} g_1 \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \\
&= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \begin{bmatrix} g_1 \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} + \mathbf{M}_1\mathbf{M}_1' \left(\frac{1}{u} \right) \mathbf{X}' \begin{bmatrix} g_1 \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} - \mathbf{M}_1\mathbf{M}_1' z \mathbf{X}' \begin{bmatrix} g_1 \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix},
\end{aligned}$$

where \mathbf{M}_1 , u , and z are previously defined. From this derivation, we noticed that, in the presence of one outlier, functions of random variables are found in several components of our new estimator $\hat{\boldsymbol{\beta}}_w$. Therefore, we can conclude that, in the presence of multiple outliers, $\hat{\boldsymbol{\beta}}_w$ will contain functions of random variables throughout its derivation.

APPENDIX B

INSTITUTIONAL REVIEW BOARD FOR HUMAN USE APPROVAL FORM

5/11/'07

OMB No. 0990-0263
Approved for use through 11/30/2008

Protection of Human Subjects Assurance Identification/IRB Certification/Declaration of Exemption (Common Rule)

Policy: Research activities involving human subjects may not be conducted or supported by the Departments and Agencies adopting the Common Rule (56FR28003, June 18, 1991) unless the activities are exempt from or approved in accordance with the Common Rule. See section 101(b) of the Common Rule for exemptions. Institutions submitting applications or proposals for support must submit certification of appropriate Institutional Review Board (IRB) review and approval to the Department or Agency in accordance with the Common Rule.

Institutions must have an assurance of compliance that applies to the research to be conducted and should submit certification of IRB review and approval with each application or proposal unless otherwise advised by the Department or Agency.

1. Request Type <input type="checkbox"/> ORIGINAL <input checked="" type="checkbox"/> CONTINUATION <input type="checkbox"/> EXEMPTION	2. Type of Mechanism <input type="checkbox"/> GRANT <input type="checkbox"/> CONTRACT <input type="checkbox"/> FELLOWSHIP <input type="checkbox"/> COOPERATIVE AGREEMENT <input type="checkbox"/> OTHER: _____	3. Name of Federal Department or Agency and, if known, Application or Proposal Identification No.
4. Title of Application or Activity Doctoral Dissertation: Novel Statistical Approaches for Identifying and Limiting the Effect of Influential Observations in Linear Regression		5. Name of Principal Investigator, Program Director, Fellow, or Other JONES, TAMEKIA

6. Assurance Status of this Project (Respond to one of the following)

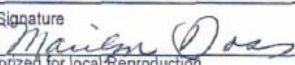
- ☒ This Assurance, on file with Department of Health and Human Services, covers this activity:
 Assurance Identification No. FWA00005960, the expiration date 2/14/09 IRB Registration No. IRB00000196
- ☐ This Assurance, on file with (agency/dept) _____, covers this activity.
 Assurance No. _____, the expiration date _____ IRB Registration/Identification No. _____ (if applicable)
- ☐ No assurance has been filed for this institution. This institution declares that it will provide an Assurance and Certification of IRB review and approval upon request.
- ☐ Exemption Status: Human subjects are involved, but this activity qualifies for exemption under Section 101(b), paragraph _____.

7. Certification of IRB Review (Respond to one of the following IF you have an Assurance on file)

- ☒ This activity has been reviewed and approved by the IRB in accordance with the Common Rule and any other governing regulations.
 by: ☐ Full IRB Review on (date of IRB meeting) _____ or ☒ Expedited Review on (date) 05/11/07
☐ If less than one year approval, provide expiration date _____
- ☐ This activity contains multiple projects, some of which have not been reviewed. The IRB has granted approval on condition that all projects covered by the Common Rule will be reviewed and approved before they are initiated and that appropriate further certification will be submitted.

8. Comments Protocol subject to Annual continuing review.	Title <u>X060504009</u> Doctoral Dissertation: Novel Statistical Approaches for Identifying and Limiting the Effect of Influential Observations in Linear Regression
--	---

 IRB Approval Issued: 05/11/07

9. The official signing below certifies that the information provided above is correct and that, as required, future reviews will be performed until study closure and certification will be provided.		10. Name and Address of Institution University of Alabama at Birmingham 701 20th Street South Birmingham, AL 35294	
11. Phone No. (with area code) (205) 934-3789	12. Fax No. (with area code) (205) 934-1301	13. Email: <u>smoore@uab.edu</u>	
14. Name of Official Marilyn Doss, M.A.		15. Title Vice Chair, IRB	
16. Signature 		17. Date <u>5-11-07</u> Sponsored by HHS	

Public reporting burden for this collection of information is estimated to average less than an hour per response. An agency may not conduct or sponsor, and a person is not required to respond to, a collection of information unless it displays a currently valid OMB control number. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to: OS Reports Clearance Officer, Room 503 200 Independence Avenue, SW., Washington, DC 20201. Do not return the completed form to this address.