WINSORIZED REGRESSION BASED ON STUDENTIZED RESIDUALS[1]

by

Ann Inez N. Gironella
Department of Statistics
University of the Philippines at Los Baños
College, Laguna, Philippines 3720

George A. Milliken
Department of Statistics
Kansas State University
Manhattan, Kansas 66506

Abstract

To reduce the effect of outlying data points in the estimation of the parameters in the simple linear regression model, Winsorization techniques are applied to Studentized residuals. Estimates obtained by this method are compared with those obtained by ordinary least squares and by the Yale-Forsythe methods using relative efficiency measurements obtained through Monte Carlo samples. It is found that estimators based on Winsorized Studentized residuals maintain high efficiencies over least squares estimators when the data are taken from scale contaminated normal, but when the data are not contaminated a loss in efficiency is observed. Estimates based on the proposed method are at least as efficient as those based on the Yale-Forsythe method depending on the sample size and the design matrix considered.

[1] Paper presented at the Second National Convention on Statistics, Philippine International Convention Center (PICC) Metro Manila, December 2-3, 1980.

I. Introduction

Consider the problem of estimating the parameters of the simple linear regression model

$$y_i = \alpha + Bx_i + \epsilon_i, \quad i = 1, 2, \ldots, n \tag{1.1}$$

which can be equivalently expressed in matrix notation as

$$\underset{\sim}{y} = \underset{\sim}{X}\,\underset{\sim}{B} + \underset{\sim}{\epsilon} \tag{1.2}$$

where

$$\underset{\sim}{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix}, \quad \underset{\sim}{X} = \begin{bmatrix} l & x_1 \\ l & x_2 \\ & \\ \cdot & \\ \cdot & \\ l & x_n \end{bmatrix}, \quad \underset{\sim}{B} = \begin{bmatrix} \alpha \\ B \end{bmatrix}, \text{ and } \underset{\sim}{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \cdot \\ \cdot \\ \cdot \\ \epsilon_n \end{bmatrix}$$

Classically, the least squares estimate of B is obtained by minimizing the sum of squares

$$(y - \underset{\sim}{X}\,\underset{\sim}{B})'\,(\underset{\sim}{y} - \underset{\sim}{X}\,\underset{\sim}{B})$$

which yields the solution

$$\underset{\sim}{\hat{B}} = (\underset{\sim}{X}'\underset{\sim}{X})^{-1}\,\underset{\sim}{X}'\underset{\sim}{Y}\,.$$

If the errors are from a $N(0, \sigma^2)$, this estimate has optimal properties. However, in many practical situations, the presence of outliers is suspected, in which case the population is said to be contaminated; or the distribution of the errors may have heavier tails than normal. In these cases, the ordinary least squares estimator may not provide good estimates of $\underset{\sim}{B}$ as these depend on the means of the y-values;

the mean being known to be highly sensitive to extreme values (Andrews, et al [2]).

Various approaches have been proposed to modify the least squares method in order to reduce the effect of outlying or "bad" data points, see [1-4, 6-8, 10-15]. A review of some of these works is given in Section 2. These modified procedures perform quite well under nonnormal conditions and yet maintain high efficiency, relative to least squares, under normality. Procedures of this kind are called robust procedures, and it is in this context of robustness that methods in this paper are developed.

One particular approach is Winsorization. Yale and Forsythe [14] have treated extreme residuals from contaminated normal distributions and by doing so have shown that the efficiency in estimating α and B was increased relative to least squares. The problem is: If the contaminated point happens to occur at the extremities the contribution of that point in estimating B is quite large, as can be seen from the formula

$$\hat{B} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) / \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

In small samples, this situation merits attention. To reduce the effect of outlying observations at the extremities, it is proposed to treat Studentized residuals instead and to determine the efficiency of this procedure relative to the Yale-Forsythe (*YF*) method.

2. Robust Estimation of Parameters in the Regression Model

Not until recently has robust estimation in the regression model been studied. Chen and Dixon [4] studied the effect of Winsorizing and trimming the residuals on the estimators of the parameters of a simple linear model with equal number of observations at each value of an equally spaced independent variable when contamination (location or scale) is present. They showed through Monte Carlo methods that Winsorization and trimming have increased efficiencies over the standard least squares procedure. However, when the samples come from the uncontaminated model, the two procedures cause some loss in the efficiency of estimating the parameters of the model. Yale and Forsythe [14] used two methods of computing residuals and three versions of Winsorization with the simple linear regression model and compared their efficiencies to least squares through computer simulation. They showed that Winsorizing was very efficient for contaminated samples and the loss in efficiency for the uncontaminated normal distribution was not more than 7%. Moussa-Hamouda and Leone [9] introduced the adjusted trimmed estimators. They have shown how to obtain the coefficients of these estimators so that their relative efficiency over the best linear unbiased estimators based on ordered observations (O-BLUE) is equal to one. They have also studied [10] the efficiency of the ordinary least squares estimator (OLSE) relative to the O-BLUE from trimmed and Winsorized and complete samples. For the standard normal and the scale contaminated normal distributions, the OLSE from trimmed samples have very high efficiencies relative to the O-BLUE. The same is true for the Winsorized OLSE, but as the tails of the distribution become heavier the loss in efficiency of the latter estimators becomes larger.

Huber [8] proposed using as robust estimators of B in the linear regression model, $y = \underset{\sim}{X}\,\underset{\sim}{B} + \underset{\sim}{\epsilon}$, the M-estimates he introduced in [7]. He defined this to be the value which minimizes

$$\sum_{i=1}^{n} P(t_{ij} \quad \text{where } \underset{\sim}{t} = \underset{\sim}{y} - \underset{\sim}{X}\,\underset{\sim}{B} \,.$$

Huber suggested using of the form

$$p(t) \;=\; \begin{cases} \frac{1}{2}\; t^2 & |t| \leqslant c \\ c\,|\,t\,| - \frac{1}{2}\, c^2 & |t| > c \end{cases}$$

where the value of x may depend on the observations y_i in order to obtain scale invariance. Relles [12] applied Huber's [7] proposal to multiple regression problem using a modified least squares method. He presented an algorithm for computing the estimate of B and showed that this estimate is asymptotically normal. He studied small sample properties via Monte Carlo methods for various error distribution. Yohai [15] presented robust estimates of the regression parameters by proper choice of c so as to obtain greatest asymptotical efficiency. Andrews [1] used the sine estimator and proposed the use of medians to obtain better starting points in the iterative procedure he developed. Hinich and Talwar [6] obtained an initial estimator for the simple linear regression case by dividing the data into non-overlapping subsamples and computing the trimmed means of the ordinary least squares subsample regression coefficients. They showed that this method provides consistent and asymptotically normal initial estimates of the coefficients. Bickel [3] and Walsch [13] used one-step M-estimators for the multiple regression problem which are asymptotically efficient when the initial estimator is a $\sqrt{n}$ consistent estimator of $\underset{\sim}{B}$. These estimators are much easier to compute than the regular M-estimators.

3. The Proposal

Consider the simple linear regression model given in equation (1.1) or (1.2) where $\epsilon_i \sim N(O, \sigma^2)$ with probability $(l - p)$ $(O \leqslant p \leqslant 1)$ and $\epsilon_i \sim N(O, c^2\ \sigma^2)$ with probability p and $c \geqslant 1$. Note that the value of $c = 1$ implies no contamination.

A procedure is proposed for estimating α and B under this condition on the error distribution.

The first step in the proposed procedure is to estimate α and B by the ordinary least squares (OLS) method, thus obtaining estimates of the residuals, d_i, *i.e.*,

$$d_i = y_i - \hat{y}_i, \quad i = 1, 2, \ldots, n$$

where $\hat{y}_i = \hat{\alpha} + \hat{B}x_i$ and $\hat{\alpha}$ and $\hat{B}$ are the least squares estimates of α and B, respectively.

The Winsorization technique is applied to the Studentized residuals r_i where

$$r_i = d_i/s_i$$

$$s_i \;=\; \text{standard error of } d_i$$

$$s_i \;=\; \sqrt{\text{diag}\,[I_n \;-\; \underset{\sim}{X}(\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{X}']} \text{ and}$$

$\underset{\sim}{I}_n$ is the n x n identity matrix.

Let $r_{(1)} \leqslant r_{(2)} \leqslant \ldots \leqslant r_{(n)}$ be the ordered Studentized residuals, and let $s_{(i)}$ be the standard error associated with $d_{(i)}$ where $d_{(i)}$ is the ith ordered residual. The proposed Winsorized regression line is obtained by applying the ordinary least squares method on the treated sample of $n(x, y')$ points where y'_i is defined as $y_i' = \hat{y}_i + d_i'$,

$$d'_{(i)} = \begin{cases} r_{(g+1)}s_{(i)} & \text{if } r_{(g+1)}s_{(i)} \geqslant d_{(g+1)}, \quad i = 1, \ldots, g \\ d_{(g+1)} & \text{if } r_{(g+1)}s_{(i)} < d_{(g+1)}, \quad i = 1, \ldots, g \\ d_{(i)} & \text{if } = g + 1, \ldots, n-g \\ r_{(n-g)}\cdot s_{(i)} & \text{if } r_{(n-g)}\cdot s_{(i)} \leqslant d_{(n-g)}, \\ & \qquad i = n-g+1, \ldots, n \\ d_{(n-g)} & \text{if } r_{(n-g)}\cdot s_{(\mathrm{i})} > d_{(n-g)} \\ & \qquad i = n-g+1, \ldots, n. \end{cases} \tag{3.1}$$

Essentially, the proposed technique would replace the g smallest residuals by $d_{(g+1)}$ if

$$r_{(g+1)}\cdot s_{(i)} = d_{(g+1)}\cdot s_{(i)}/s_{(g\;+1)} < d_{(g+1)}$$

or by $r_{(g+1)}\cdot s_{(i)}$ if

$$r_{(g+1)}\cdot s_{(i)} = d_{(g+1)}\cdot s_{(i)}\cdot s_{(i)}/s_{(g+1)} \geqslant d_{(g+1)},$$

$$i = 1, 2, \ldots, g.$$

The former replacement is included in the modification to the *YF* method because the factor $s_{(i)}/s_{(g+1)}$ may be greater than one, thus decreasing the i^{th} residual of the set of the g smallest residuals instead of increasing it. On the other extreme, the proposed technique would replace the g largest residuals by $d_{(n-g)}$ if

$$r_{(n-g)\cdot s\,(i)} = (d_{(n-g)\cdot s\,(i)})/S_{(n-g)} > d_{(n-g)}$$

or by $r_{(n-g)\cdot s\,(i)}$ if

$$r_{(n-g)\cdot s\,(i)} = (d_{(n-g)\cdot s\,(i)})/s_{(n-g)} \leqslant d_{(n-g)},$$

$$i = .,2,\ldots,g.$$

The latter replacement is necessary because the factor $s_{(i)}/s_{(n-g)}$ may be greater than one, thus increasing the i^{th} residual of the set of the g largest residuals instead of decreasing it. This modification of the YF method was proposed so as to reduce the effect of a suspected outlier on the estimation of the parameters of the model, particularly when these outliers are farthest from $(\bar{x}, \bar{y})$ where their effect on the estimation is greatest.

Three methods, similar to those used by Yale and Forsythe, will be used for the estimation so that comparisons may be made. They are iteration method, the levels method, and the iteration at increasing levels method. These three methods are described below.

(i) *The iteration method*

From a sample of $n(x, y)$ points, estimates of α and B are obtained by the OLS method, thus obtaining $n(x, \hat{y})$ from which the residuals, d_i, and the standard errors of the residuals, s_i, are computed. One point is then Winsorized at each extreme according to equation (3.1) to obtain new or adjusted y-values, i.e., the y-primes. Using the set of $n(x, y')$ points, new estimates of α and B are calculated by OLS. The process can be repeated by continually updating the data to obtain more refinement of the residuals. Each repetition of the process is an iteration or a degree of Winsorization. At each iteration, adjusted residuals (from preceding iteration) are used. By Winsorizing g points at each extreme at each iteration, this procedure could be generalized. However, only the case $g = 1$ will be considered in this paper.

(ii) *The levels method*

From a sample of $n(x, y)$ points, estimates of α and B are obtained by the OLS method from which the residuals, d_i. and their standard errors, s_i, are computed. The "levels" of Winsorization is defined to be the number of residuals Winsorized at each end. Here, a level is defined as a degree of Winsorization. For example: four degrees of Winsorization means that four residuals at each extreme are Winsorized according to equation (3.1) or four levels of Winsorization. The process is not repeated for this method.

(iii) *The iteration of increasing levels method*

This is a combination of the iteration and levels methods. From a sample of $n(x, y)$ points, estimates of α and B are first computed by OLS from which the $d's$ and the $s's$ are obtained. One point at each extreme is then Winsorized according to equation (3.1) to obtain new estimates of α and B. The process may be repeated by continually updating the y-values to obtain more refinement of the residuals. Unlike the iteration method, the number of points, g, Winsorized at each extreme increases with each iteration. An iteration at a level is defined as a degree of Winsorization. For example: two degrees of Winsorization using this method means two iterations with one point Wisorized at each extreme on the first iteration and two points Winsorized at each extreme on the second iteration.

4. Simulation Procedure

The efficiency of the proposed technique in estimating α and B is compared with the OLS and the YF techniques for the three methods (iteration, levels, and iterations at increasing levels) using the relative mean square error (RMSE) measure, defined as

$$\text{RMSE}\ (\hat{\alpha}_{pk}) = \sum_{j=1}^{S} (\hat{\alpha}_{kj} - \alpha)^2 \ / \ \Sigma_{j=1}^{S} (\hat{\alpha}_{pj} - \alpha)^2$$

$$\text{RMSE}\ (\hat{B}_{pk}) = \Sigma_{j=1}^{S} (\hat{B}_{kj} - B)^2 / \ \Sigma_{j=1}^{S} (\hat{B}_{pj} - B)^2 \qquad (4.1)$$

where

$$k = \begin{cases} O \text{ for ordinary least squares technique} \\ YF \text{ for Yale-Forsythe technique} \end{cases}$$

$\hat{\alpha}_{pj}$ and $\hat{B}_{pj}$ are the estimates of α and B, respectively, using the proposed technique in the j^{th} simulation $\hat{\alpha}_{kj}$ and $\hat{B}_{kj}$ are the estimates of α and B, respectively, using the OLS technique (k=0) or the YF technique (k=YF) in the j^{th} simulation, and S is the number of simulations per set.

A FORTRAN program was written for the computer simulation study. Data were generated using the model

$$y_i = \alpha + Bx_i + \epsilon_i, \quad i = 1, \ldots, n$$

where the $x_i's$ were taken from the $N(O, l)$ and the $\epsilon_i's$ were taken from the $N(O, 1)$ with probability $(1 - p)$ and from the $N(O, c^2)$ with probability p using the normal random number generator Super Duper [9]. The values of α and B were set at O and 1, respectively. The size of the simulation study was taken to be 400.

In order to get an estimate of the standard error of the RMSE, the 400 simulations were considered in sets of 20. For each set of 20 simulations the RMSE's were calculated. These were then used to compute an estimate of the standard error of the RMSE's. The results are shown in Tables 2a, 2b, 2c, 3a, 3b, and 3c.

5. Size of the Study

The same combinations of the parameters used by Yale and Forsythe [14] were tested to enable comparisons to be made. These are given in Table 1.

Table 1.

TEST CASES STUDIED.

Test	*n*	*p*	c^2
1	10	.00	1.0
2	10	.10	16.0
3	10	.20	16.0
4	20	.00	1.0
5	20	.20	16.0
6	50	.00	1.0
7	50	.08	16.0
8	50	.20	16.0

The cases p = .00 and c = 1.0 for n = 10, 20, 50 are the non-contaminated cases (c^2 is the variance of the contaminating population).

6. Results and Discussion

Tables 2a, 2b, and 2c show the RMSE of the proposed procedure in estimating α over the OLS method and the YF technique for iteration, levels, and iteration at increasing levels using samples of size n = 10, 20, and 50, respectively. The numbers in parentheses are the standard errors of the calculated relative efficiencies. Tables 3a, 3b, and 3c show the relative efficiency of the proposed method in estimating B over the OLS and YF procedures using samples of size n = 10, 20, and 50, respectively. The numbers in parentheses are the standard errors of the RMSE's.

In Tables 2a and 3a the entries for levels and iterations at increasing levels with 5 degrees of Winsorization are blank because with sample size n = 10, a maximum of only four degrees of Winsorization is possible.

To judge how good a method is, three factors are considered:

(i) Robustness – the greater the RMSE in the contaminated case, the more robust the method;

(ii) Efficiency – the smaller the loss in efficiency compared to least squares in the uncontaminated case, the more efficient the proposed method;

(iii) Stability of the RMSE as measured by its standard error.

From Tables 2a, 2b, and 2c, it can be seen that RMSE ($\hat{\alpha}_{po}$) is most stable through five degrees of Winsorization when the iteration method is used. The RMSE ($\hat{\alpha}_{pYF}$) is most stable through five degrees of Winsorization when the levels methods is used for n = 10, but for n = 20 or 50 the three methods yield standard errors of RMSE ($\hat{\alpha}_{pYF}$) less than 1%.

Table 2a.

RMSE AND ITS STANDARD ERROR[1] IN ESTIMATING α.
SAMPLE SIZE N = 10.

		Iteration		*Levels*		*Iter/Lv*	
Test	*Degree of Winsorization*	*Over L.S.*	*Over Y-F*	*Over L.S.*	*Over Y-F*	*Over L.S.*	*Over Y-F*
	1	0.98 (.02)	1.00 (**)[2]	0.98 (.02)	1.00 (**)	0.98 (.02)	1.00 (**)
	2	0.97 (.02)	0.99 (.01)	0.94 (.04)	1.00 (**)	0.93 (.04)	1.00 (.01)
p = .00	3	0.96 (.03)	0.99 (.01)	0.88 (.04)	1.00 (**)	0.86 (.04)	1.01 (.01)
c^2 = 1.0	4	0.96 (.03)	0.99 (.01)	0.79 (.04)	1.00 (**)	0.80 (.05)	1.01 (.01)
	5	0.95 (.03)	0.98 (.01)	–	–	–	–
	1	1.32 (.07)	1.02 (.01)	1.32 (.07)	1.02 (.01)	1.32 (.07)	1.02 (.01)
	2	1.36 (.08)	1.02 (.02)	1.46 (.10)	1.03 (.01)	1.52 (.10)	1.03 (.02)
p = .10	3	1.38 (.08)	1.02 (.03)	1.39 (.11)	1.00 (.01)	1.55 (.12)	1.02 (.03)

Table 2a.

RMSE AND ITS STANDARD ERROR IN ESTIMATING α.
SAMPLE SIZE N = 20.

		Iteration		*Levels*		*Iter/Lv*	
Test	*Degree of Winsorization*	*Over L.S.*	*Over Y-F*	*Over L.S.*	*Over Y-F*	*Over L.S.*	*Over Y-F*
c^2 = 16.0	4	1.38 (.09)	1.01 (.03)	1.26 (.12)	1.00 (.01)	1.41 (.14)	1.03 (.02)
	5	1.40 (.09)	1.01 (.03)	–	–	–	–
	1	1.23 (.06)	1.00 (.01)	1.27 (.06)	1.00 (.01)	1.27 (.06)	1.00 (.01)
	2	1.31 (.06)	1.00 (.01)	1.55 (.08)	1.00 (.01)	1.59 (.09)	1.01 (.01)
p = .20	3	1.32 (.06)	1.00 (.01)	1.57 (.10)	1.00 (**)	1.67 (.13)	1.01 (.01)
c^2 = 16.0	4	1.33 (.06)	1.00 (.01)	1.54 (.10)	1.00 (**)	1.73 (.14)	1.00 (.01)
	5	1.34 (.06)	1.01 (.01)	–	–	–	–

[1] Standard error of RMSE is based on the variation of the 20 observed RMSE's, where each observation is computed using equation (4.1) with S = 20.

[2] (**) means the standard error of the average RMSE is less than 1%.

Table 2b.

RMSE AND ITS STANDARD ERROR IN ESTIMATING α.
SAMPLE SIZE N = 20.

		Iteration		*Levels*		*Iter/Lv*	
Test	*Degree of Winsorization*	*Over L.S.*	*Over Y-F*	*Over L.S.*	*Over Y-F*	*Over L.S.*	*Over Y-F*
	1	0.93 (.01)	1.00 (**)	0.97 (.01)	1.00 (**)	0.97 (.01)	1.00 (**)
	2	0.97 (.01)	1.00 (**)	0.95 (.02)	1.00 (**)	0.95 (.02)	1.00 (**)
p= .00	3	0.97 (.01)	1.00 (**)	0.92 (.02)	1.00 (**)	0.92 (.02)	1.00 (**)
c^2 = 1.00	4	0.97 (.01)	1.00 (**)	0.88 (.03)	1.00 (**)	0.87 (.02)	1.00 (**
	5	0.97 (.01)	1.00 (**)	0.85 (.03)	1.00 (**)	0.83 (.02)	1.01 (**
	1	1.19 (.04)	1.00 (**)	1.16 (.04)	1.00 (**)	1.16 (.04)	1.00 (**)
	2	1.20 (.04)	1.01 (**)	1.33 (.07)	1.00 (**)	1.34 (.07)	1.00 (**)
p = .20	3	1.20 (.04)	1.01 (**)	1.66 (.08)	1.00 (**)	1.73 (.08)	1.00 (**)
c^2 = 16.0	4	1.20 (.04)	1.01 (**)	1.85 (.09)	1.00 (**)	1.92 (.09)	1.00 (**)
	5	1.21 (.04)	1.01 (**)	1.93 (.10)	1.00 (**)	2.10 (.11)	1.00 (**)

Table 2c.

RMSE AND ITS STANDARD ERROR IN ESTIMATING α.
SAMPLE SIZE N = 50.

		Iteration		*Levels*		*Iter/Lv*	
Test	*Degree of Winsorization*	*Over L.S.*	*Over Y-F*	*Over L.S.*	*Over Y-F*	*Over L.S.*	*Over Y-F*
	1	0.99	1.00	0.99	1.00	0.99	1.00
		(.01)	(**)	(.01)	(**)	(.01)	(**)
	2	0.99	1.00	0.99	1.00	0.99	1.00
		(.01)	(**)	(.01)	(**)	(.01)	(**)
p = .00	3	0.99	1.00	0.99	1.00	0.99	1.00
		(.01)	(**)	(.02)	(**)	(.02)	(**)
$c^2 = 1.00$	4	0.99	1.00	0.98	1.00	0.98	1.00
		(.01)	(**)	(.02)	(**)	(.02)	(**)
	5	0.99	1.00	0.99	1.00	0.98	1.00
		(.01)	(**)	(.02)	(**)	(.02)	(**)
	1	1.12	1.00	1.09	1.00	1.09	1.00
		(.02)	(**)	(.03)	(**)	(.03)	(**)
	2	1.12	1.00	1.29	1.00	1.29	1.00
		(.02)	(**)	(.05)	(**)	(.05)	(**)
p = .08	3	1.12	1.00	1.49	1.00	1.49	1.00
		(.02)	(**)	(.09)	(**)	(.09)	(**)
$c^2 = 16.0$	4	1.12	1.00	1.51	1.00	1.50	1.00
		(.02)	(**)	(.09)	(**)	(.10)	(**)
	5	1.13	1.01	1.56	1.00	1.57	1.00
		(.02)	(**)	(.11)	(**)	(.11)	(**)
	1	1.02	1.00	1.00	1.00	1.00	1.00
		(.02)	(**)	(.02)	(**)	(.02)	(**)
	2	1.02	1.00	1.08	1.00	1.08	1.00
		(.02)	(**)	(.02)	(**)	(.02)	(**)
p = .20	3	1.02	1.00	1.30	1.00	1.31	1.00
		(.02)	(**)	(.04)	(**)	(.04)	(**)
$c^2 = 16.0$	4	1.02	1.00	1.52	1.00	1.51	1.00
		(.02)	(**)	(.06)	(**)	(.06)	(**)
	5	1.02	1.00	1.71	1.00	1.74	1.00
		(.02)	(**)	(.09)	(**)	(.09)	(**)

Table 3a.

RMSE AND ITS STANDARD ERROR IN ESTIMATING β.
SAMPLE SIZE N = 10.

		Iteration		*Levels*		*Iter/Lv*	
Test	*Degree of Winsorization*	*Over L.S.*	*Over Y-F*	*Over L.S.*	*Over Y-F*	*Over L.S.*	*Over Y-F*
	1	0.98 (.02)	0.99 (**)	0.98 (.02)	0.99 (**)	0.98 (.02)	0.99 (**)
	2	0.96 (.02)	0.99 (.01)	0.95 (.02)	1.00 (.01)	0.92 (.03)	0.99 (.01)
p = .00	3	0.94 (.02)	0.98 (.01)	0.97 (.01)	1.00 (.01)	0.88 (.03)	0.99 (.01)
c^2 = 1.00	4	0.92 (.02)	0.97 (.01)	0.99 (.01)	0.99 (**)	0.87 (.03)	0.99 (.01)
	5	0.92	0.97	–	–	–	–
	1	1.34 (.06)	1.04 (.01)	1.34 (.06)	1.04 (.01)	1.34 (.06)	1.04 (.01)
	2	1.51 (.11)	1.07 (.02)	1.33 (.04)	1.02 (.01)	1.54 (.10)	1.05 (.02)
p = .10	3	1.60 (.15)	1.08 (.03)	1.20 (.03)	1.00 (.01)	1.64 (.14)	1.06 (.02
c^2 = 16.0	4	1.66 (.20)	1.09 (.04)	1.06 (.01)	1.00 (**)	1.68 (.14)	1.06 (.03
	5	1.68 (.24)	1.08 (.04)	–	–	–	–
	1	1.24 (.04)	1.01 (.01)	1.24 (.04)	1.01 (.01)	1.24 (.04)	1.01 (.01)
	2	1.33 (.06)	1.02 (.01)	1.27 (.03)	1.00 (.01)	1.37 (.07)	1.01 (.01)
p = .20	3	1.37 (.07)	1.02 (.02)	1.19 (.03)	1.00 (.01)	1.44 (.10)	1.02 (.01)
c^2 = 16.0	4	1.39 (.08)	1.03 (.02)	1.07 (.02)	1.00 (**)	1.45 (.12)	1.02 (.01)
	5	1.40 (.09)	1.02 (.02)	–	–	–	–

Table 3b.

RMSE AND ITS STANDARD ERROR IN ESTIMATING β.
SAMPLE SIZE N = 20.

		Iteration		*Levels*		*Iter/Lv*	
Test	*Degree Winsorization*	*Over L.S.*	*Over Y-F*	*Over L.S.*	*Over Y-F*	*Over L.S.*	*Over Y-F*
	1	0.96 (.02)	0.99 (**)	0.96 (.02)	0.99 (**)	0.96 (.02)	0.99 (**)
	2	0.95 (.02)	0.99 (.01)	0.95 (.02)	1.00 (**)	0.93 (.02)	0.99 (.01)
p=.00	3	0.95 (.02)	0.99 (.01)	0.94 (.02)	1.00 (**)	0.89 (.02)	0.99 (.01)
c^2=1.00	4	0.95 (.02)	0.99 (.01)	0.94 (.02)	1.00 (.01)	0.86 (.03)	0.99 (.01)
	5	0.94 (.02)	0.99 (.01)	0.94 (.02)	1.00 (**)	0.82 (.03)	0.99 (.01)
	1	1.30 (.05)	1.01 (.01)	1.37 (.06)	1.01 (**)	1.37 (.06)	1.01 (**)
	2	1.34 (.07)	1.02 (.01)	1.47 (.08)	1.00 (**)	1.58 (.11)	1.01 (.01)
p=.20	3	1.35 (.07)	1.02 (.01)	1.50 (.08)	0.99 (**)	1.70 (.13)	1.00 (.01)
c^2=16.0	4	1.35 (.07)	1.02 (.01)	1.46 (.07)	0.99 (**)	1.72 (.13)	0.99 (.01)
	5	1.35 (.07)	1.02 (.01)	1.38 (.06)	0.97 (**)	1.73 (.13)	0.99 (.01)

Table 3c.

RMSE AND ITS STANDARD ERROR IN ESTIMATING β.
SAMPLE SIZE N = 50.

		Iteration		*Levels*		*Iter/Lv*	
Test	*Degree of Winsorization*	*Over L.S.*	*Over Y-F*	*Over L.S.*	*Over Y-F*	*Over L.S.*	*Over Y-F*
	1	0.98	1.00	0.98	1.00	0.98	1.00
		(.01)	(**)	(.01)	(**)	(.01)	(**)
	2	0.98	1.00	0.98	1.00	0.98	1.00
		(.01)	(**)	(.01)	(**)	(.01)	(**)
p = .00	3	0.98	1.00	0.98	1.00	0.98	1.00
		(.01)	(**)	(.02)	(**)	(.02)	(**)
c^2=1.00	4	0.98	1.00	0.98	1.00	0.97	1.00
		(.01)	(**)	(.02)	(**)	(.02)	(**)
	5	0.98	1.00	0.98	1.00	0.96	1.00
		(.01)	(**)	(.02)	(**)	(.02)	(**)
	1	1.30	1.01	1.29	1.01	1.29	1.01
		(.06)	(**)	(.04)	(**)	(.04)	(**)
	2	1.32	1.01	1.54	1.01	1.57	1.01
		(.06)	(**)	(.05)	(.01)	(.05)	(**)
p=.08	3	1.32	1.01	1.61	1.00	1.70	1.00
		(.06)	(**)	(.06)	(**)	(.07)	(**)
c^2=16.0	4	1.32	1.01	1.60	1.00	1.72	1.00
		(.06)	(**)	(.06)	(**)	(.09)	(**)
	5	1.33	1.01	1.58	1.00	1.71	1.00
		(.06)	(**)	(.06)	(**)	(.09)	(**)
	1	1.20	1.01	1.20	1.01	1.20	1.01
		(.02)	(**)	(.03)	(**)	(.03)	(**)
	2	1.22	1.01	1.41	1.01	1.43	1.01
		(.03)	(**)	(.06)	(**)	(.07)	(**)
p=.20	3	1.22	1.01	1.60	1.01	1.68	1.01
		(.03)	(**)	(.07)	(**)	(.11)	(**)
c^2=16.0	4	1.22	1.02	1.71	1.01	1.86	1.01
		(.03)	(**)	(.08)	(**)	(.12)	(**)
	5	1.23	1.02	1.77	1.00	2.00	1.01
		(.03)	(**)	(.10)	(**)	(.15)	(**)

Tables 3a, 3b, and 3c show that for $n = 10$ the standard error of RMSE $(\hat{B}_{po})$ is least when the levels method is used but for $n = 20$ or 50 the iteration method is most stable, with the levels method following very closely. The standard error of RMSE $(\hat{B}_{pYF})$ is least when the levels method is used for $n = 10$, but for $n = 20$ or 50 the three methods are practically equally stable with the levels method showing slightly less standard error for $n = 20$.

Tables 2a, 2b, 2c and also 3a, 3b, 3c show that when there is no contamination the least squares procedure is a better method of estimating the parameters, but the loss in efficiency is not too great (at most 8% when $n = 10$, at most 6% when $n = 20$, at most 2% when $n = 50$ using the iteration method). Compared to the YF technique, the proposed technique shows a loss in efficiency of at most 3% when there is no contamination for samples of size 10, but for $n = 20$ and 50 these two methods have practically the same efficiency. When contamination is present, the proposed technique is found to be a more efficient method compared to least squares with the iterations at increasing levels showing the greatest efficiency. Although this method (iter/lv) is most efficient in the presence of contamination, it shows the greatest loss in efficiency in the noncontaminated case. When $n = 10$, the proposed technique is more efficient than the YF technique, particularly when the contamination factor is 10%, but when $n = 20$ or 50 the two techniques have practically the same efficiency.

Overall, the proposed procedure showed only a slight gain in efficiency over the YF technique. It appears that when the sample size is small the distribution is not too heavily contaminated, and when the contamination factor is quite large $(c^2 = 16.0)$ the proposed technique is better –– the gain in efficiency using the iteration method when estimating B is at most 9%.

7. Further Considerations

The discussion in Section 6 indicated that when the x-values are drawn from the $N(0, 1)$, the proposed method is as efficient as the YF method for all practical purposes in estimating the parameters of the simple linear regression model. This is because the residuals are

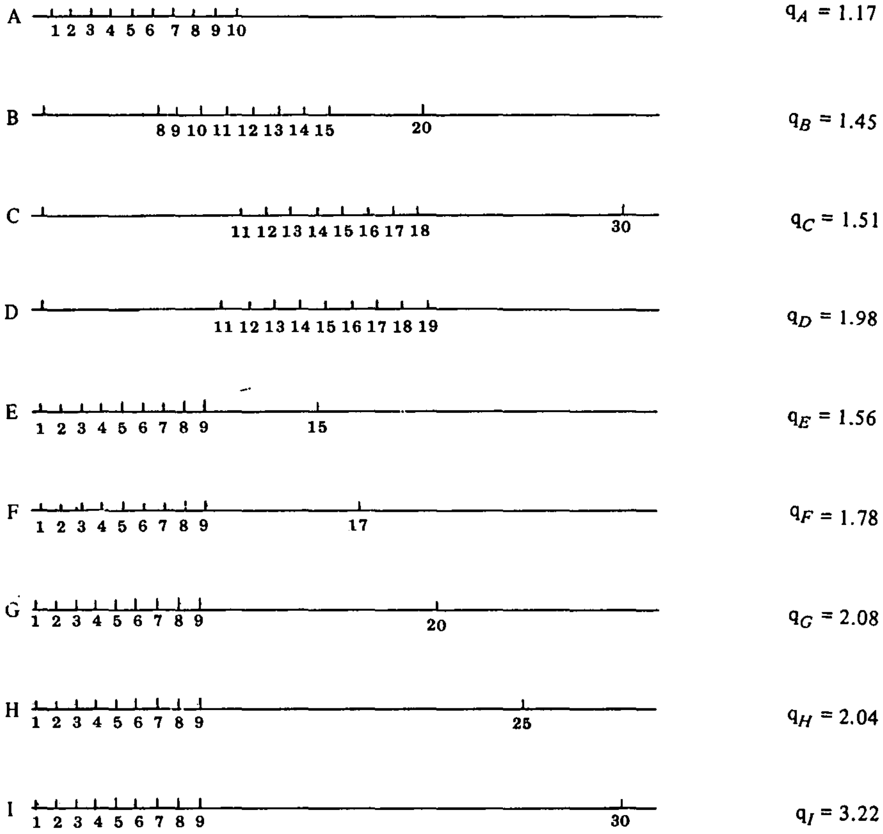
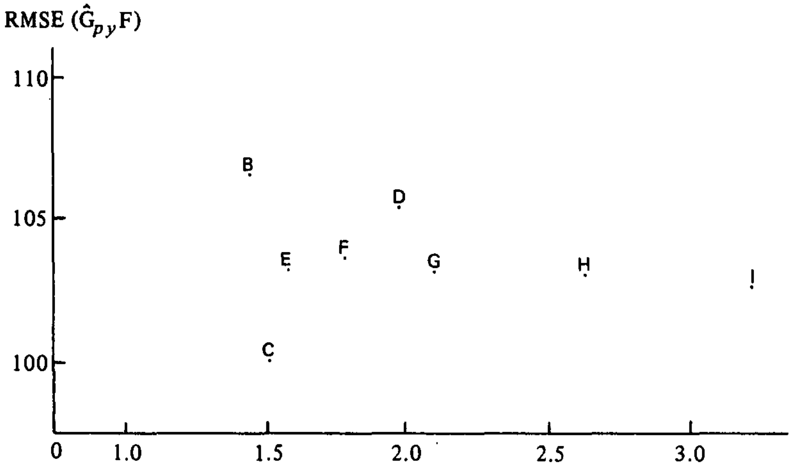


Figure I. Nine cases studied. Ticked marks are the values of the x-vester in the design-metrix.

$$q_j = \frac{\max s_j}{\min s_j} \quad j = A, B, \ldots\ldots I, \quad i = 1, 2, \ldots\ldots 10$$



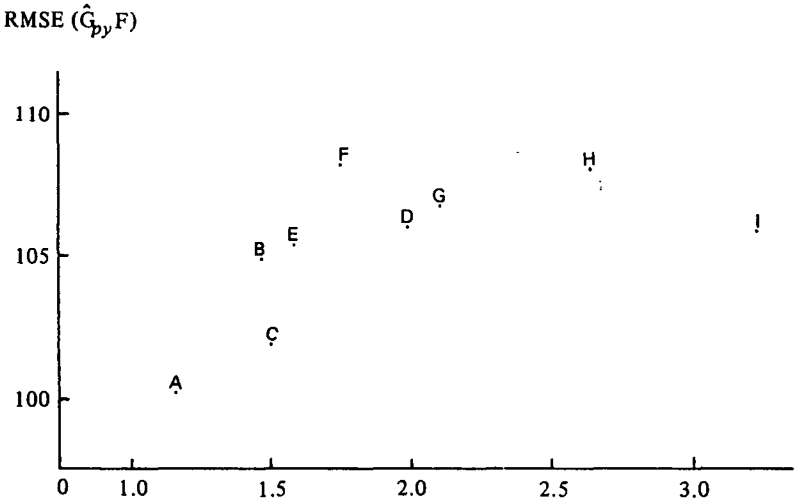


Figure 2. Mean (over 5 degrees of Winaorization) RMSE vs q

n = 10, p = 10%, c = 40

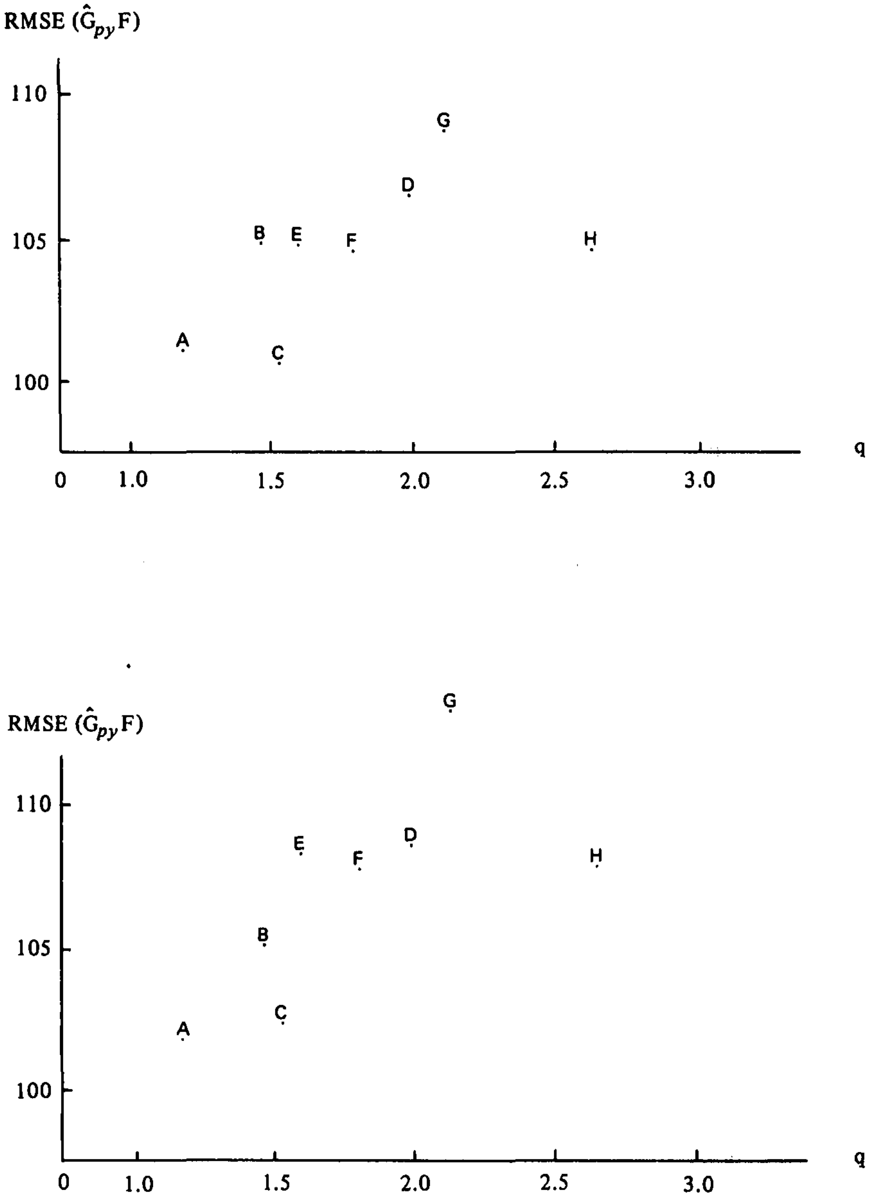


Figure 3. Mean (over 5 degrees of Winaorization) RMSE vs q

given almost the same weight, particularly when the sample size is large.

It is then of interest to determine conditions on the design matrix for which the proposed method is more efficient.

Nine configurations of the design matrix (Figure 1) were studied for a sample of size $n = 10$ with contaminating fraction $p = 10\%$ and $p = 20\%$ and scale contamination factor $c = 4.0$ using the iteration method.

Only these tests were considered because of the results obtained from the earlier discussion (Section 6): that is, the iteration method showed a fairly stable value of the relative efficiency through five degrees of Winsorization; the proposed method is more efficient when n is small, the distribution not too heavily contaminated, and the contamination factor quite large.

8. Discussion

The simulation results show that when the x-values are equally spaced the proposed method is as efficient as the YF method but is considerably more efficient than the least squares method. Also, no gain in efficiency over the YF method is observed with Case C and Case I (with 20% contamination). For all other forms of the design matrix studied, the proposed method shows an increase in efficiency over the YF method, with the efficiency increasing as the degree of Winsorization is increased.

The average over the five degrees of Winzorization of the RMSE (α_{pYF}) and the RMSE ($\hat{B}_{pYF}$) were plotted against the ratio

$$q = \frac{\max\ s_i}{\min\ s_i}$$

where s_i = standard error of the i^{th} residual, d_i, for each case to show how the RMSE behaves with varying forms of the design matrix as measured by q (Figures 2a, 2b, 3a, and 3b).

The case studied can essentially be classified into two groups:

Group 1: Cases A, B, and C. These are the cases where the

x-values are symmetrically spaced; case A being the special case when the x-values are equally spaced.

Group 2: Cases D, E, F, G, H, I. These are the cases where the x-values are assymmetrically spaced.

The quadratic nature of the RMSE is evident in Group 2. As a point x_i is moved farther away from the "cluster" of x-values, the value of q is increased, and more weight is given the residual at the extremities. However, as x_i is pulled farther out, large residuals at the extremities are seldom experienced because the estimated regression line tends to be defined by the "cluster" and the extreme point. This, too, explains the decreasing value of RMSE $(\hat{B}_{po})$ after a high point is reached. Note that in case I the RMSE $(\hat{B}_{po})$ is close to RMSE $(\hat{B}_{pYF})$, implying that Winsorization is no longer effective in estimating B in the presence of outliers. A glance at its RMSE (α_{po}) and RMSE(α_{pYF}) indicates that Winsorization has the effect of shifting the line up or down with very little change in its slope. But even in this case the proposed method is still more efficient than the other two methods.

Group I follows the same pattern of behavior in RMSE. Case C appears interesting. Even while its value of q is greater than that of case B (yet only slightly: $q = 1.45$, $q_C = 1.51$), a drastic drop in the RMSE is observed. One reason is, as in Case H, that the extreme x-values tend to define the estimated line. The other reason is that the x-values are more symmetrically placed, thus giving residuals to the left of the $\bar{x}$ – line the same weight as corresponding residuals to the right of $\bar{x}$ – line.

The discussion above holds for both when the contamination factor $p = 10\%$ and when $p = 20\%$. But as shown in Figure 3b, with heavier contamination the RMSE $(\hat{B}_{pYF})$ increases sharply as q is increased until a high point is reached ($q = 2.09$, RMSE $(\hat{B}_{pYF}) =$ 112.8 and then decreases sharply in the same fashion.

In conclusion, when the x-values are equally spaced, the proposed method and the YF method are equally efficient in estimating α and B and both methods are much more efficient than the OLS method. For all other forms of the design matrix studied, a gain in efficiency is attained with the proposed technique. When the ν-values are assymmetrically positioned with the extreme x-values far from the cluster of x-values or when a point x_i is markedly distant from the cluster of x-values, very little gain in efficiency is observed (only about 2%); otherwise, the gain in efficiency is not less than 4%. With 10% contamination, the maximum value RMSE $(\hat{B}_{pYF})$ 108% when q = 1.78. With heavier contamination (p = 20%) the maximum gain in efficiency is 12.8% when q = 2.09.

Acknowledgements

Contribution no. 79-162-j, Department of Statistics, and the Statistical Laboratory of the Kansas Agricultural Experiment Station, Manhattan, Kansas, 66506, U.S.A. This paper is a portion of Ann Inez N. Gironella's dissertation for the Ph. D. in Statistics at Kansas State University.