Journal of Modern Applied Statistical Methods

Volume 13 | Issue 2 Article 7

11-2014

Robust Winsorized Shrinkage Estimators for Linear Regression Model

Nileshkumar H. Jadhav
D.R.K. College of Commerce, Kolhapur, India, n.nil08@gmail.com

D N. Kashid
Shivaji University, Kolhapur, India., dnk_stats@unishivaji.ac.in

Recommended Citation
Jadhav, Nileshkumar H. and Kashid, D N. (2014) "Robust Winsorized Shrinkage Estimators for Linear Regression Model," *Journal of Modern Applied Statistical Methods*: Vol. 13 : Iss. 2 , Article 7.
DOI: 10.22237/jmasm/1414814760

Journal of Modern Applied Statistical Methods
November 2014, Vol. 13, No. 2, 131-150.

ISSN 1538 − 9472

Robust Winsorized Shrinkage Estimators for Linear Regression Model

Nileshkumar H. Jadhav
D.R.K College of Commerce
Kolhapur, India

Dattatraya N. Kashid
Shivaji University
Kolhapur, India

In multiple linear regression, the ordinary least squares estimator is very sensitive to the presence of multicollinearity and outliers in the response variable. To handle these problems in the data, Winsorized shrinkage estimators are proposed and the performance of these estimators is evaluated through mean square error sense.

Keywords: Multicollinearity, outliers, contaminated normal error, Winsorization, mean square error, multiple linear regression

Introduction

In the multiple linear regression model

$$Y = X\beta + \varepsilon, \tag{1}$$

Y is an vector of n observations on the response variable, X is an $n \times p$ matrix of independent variables known as regressor variables, β is a $p \times 1$ vector of unknown regression parameters and ε is an $n \times 1$ vector of unobserved random errors. Classically, it is assumed that the ε_i, $i = 1, 2, ..., n$, are independent and identically normally distributed with zero mean and constant variance σ^2.

It is well known that when the normality assumption holds, the ordinary least squares (OLS) estimator becomes a maximum likelihood estimator and the best linear unbiased estimator of the unknown regression parameters and has the smallest variance in the class of all linear unbiased estimators. However, the real life data often may not satisfy these assumptions and the violation of assumptions dramatically affects the OLS estimation and consequently the prediction based on the OLS estimator. In the literature, the effect of violation of assumptions has been

Nileshkumar H. Jadhav is an Assistant Professor of Statistics at the D.R.K. College of Commerce. Email at n.nil08@gmail.com. Dr. Kashid is a Professor in the Department of Statistics. Email him at dnk_stats@unishivaji.ac.in.

discussed by many authors (see Birkes and Dodge, 1993; Draper and Smith, 1998; Montgomery, Peck and Vining, 2006).

The near linear dependency between the set of regressor variables produces the problem of multicollinearity in the data. Due to the presence of multicollinearity, the variance of the OLS estimator gets inflated. Consequently, the OLS estimates become unstable and may give misleading results. Various techniques are available in the literature to deal with the problem of multicollinearity. Hoerl and Kennard (1970a, b), Hoerl, Kennard and Baldwin (1975), Liu (1993), Liu (2003) are praiseworthy.

Another important problem that has received considerable attention is the presence of outliers in Y- space. Huber (1973) and Rousseeuw and Leroy (1987) pointed out that the presence of outliers significantly affect the performance of the OLS estimator. In most of the situations, outliers in Y- space are due to heavy tailed distribution of error variable. The least squares fit may be spoiled by small but reasonable deviation from normal error distribution (see Huber, 1973; Andrews, 1974). Many robust parameter estimation methods are available in the literature to handle the problem of outliers in the data.

A simultaneous occurrence of multicollinearity and outliers in Y-space due to non-normality of error variable is considered. To handle the problem of multicollinearity and outliers in the data, a class of Winsorized shrinkage estimators is proposed and the performance is evaluated through estimated mean square error (EMSE). An extensive simulation study was conducted to evaluate the performance of the proposed and existing estimators. Also, a real data example is used to illustrate the performance of the estimators.

Regression Model and Some Estimators

To reduce the notational complexity and lengthy expressions, various authors like Liu (1993), Liu (2003), Montgomery, Peck and Vining (2006), Gao and Liu (2011) used a canonical form of a multiple linear regression model. It is given as

$$Y = Z\alpha + \varepsilon, \tag{2}$$

where $Z = XQ$, $\alpha = Q'\beta$ and $Q = (q_1, q_2, \ldots, q_p)$ is an orthogonal matrix of eigenvectors $q_1, q_2, \ldots, q_p$ corresponding to eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_p \geq 0$ of $X'X$ matrix. Note that, the use of canonical form does not affect the mean square error (MSE) of the estimator (Liu, 2003).

Some existing estimators were examined to handle the problem of multicollinearity and problem of outliers individually present in the data.

Ordinary Least Squares (OLS) Estimator

It is well known that, when $\varepsilon \sim N(0, \sigma^2 I)$, then the optimal estimator of regression parameters is the OLS estimator. It is denoted by

$$\hat{\alpha}_{OLS} = \Lambda^{-1} Z'Y \tag{3}$$

where $\Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_p)$. It is widely used in regression analysis due to its computational ease. Because the OLS estimator is unbiased, the MSE of $\hat{\alpha}_{OLS}$ is given by

$$\begin{aligned} MSE\left(\hat{\alpha}_{OLS}\right) &= tr\left(Cov\left(\hat{\alpha}_{OLS}\right)\right) \\ &= \sigma^2 \sum\nolimits_{j=1}^{p} 1/\lambda_j \end{aligned} \tag{4}$$

where the error variance σ^2 is unknown and estimated by $\hat{\sigma}^2_{OLS} = \left(Y - Z\hat{\alpha}_{OLS}\right)'\left(Y - Z\hat{\alpha}_{OLS}\right)/\left(n-p\right)$.

Ordinary Ridge Regression (ORR) Estimator

To overcome the problem of multicollinearity, several methods are put forwarded in the literature, but the ordinary ridge regression estimator (ORR) proposed by Hoerl and Kennard (1970a, b) is one of the most popular biased estimators for regression parameters. It is defined as

$$\hat{\alpha}_{ORR} = \left(\Lambda + kI\right)^{-1} \Lambda \hat{\alpha}_{OLS} \tag{5}$$

where $k > 0$ is a ridge parameter and I is an identity matrix of an order $p \times p$. Because, the ORR estimator is biased, the MSE of ORR estimator is obtained as

$$\begin{aligned} MSE\left(\hat{\alpha}_{ORR}\right) &= tr\left(Cov\left(\hat{\alpha}_{ORR}\right)\right) + \left(Bias\left(\hat{\alpha}_{ORR}\right)\right)'\left(Bias\left(\hat{\alpha}_{ORR}\right)\right) \\ &= \sigma^2 \sum\nolimits_{j=1}^{p} \frac{\lambda_j}{\left(\lambda_j + k\right)^2} + k^2 \sum\nolimits_{j=1}^{p} \frac{\alpha_j^2}{\left(\lambda_j + k\right)^2} \end{aligned} \tag{6}$$

The ridge parameter *k* plays an important role in minimizing the MSE of the ORR estimator. Various choices for estimator of *k* are available in the literature, but the estimator proposed by Hoerl, Kennard and Baldwin (1975) is widely used. It is defined as

$$k = \frac{p\hat{\sigma}^2_{OLS}}{\hat{\alpha}'_{OLS}\hat{\alpha}_{OLS}} \tag{7}$$

where $\hat{\sigma}^2_{OLS}$ is the estimate of error variance based on the OLS estimator $\hat{\alpha}_{OLS}$. However, $\hat{\alpha}_{ORR}$ is nonlinear function of *k*. So, using some of the proposed methods to obtain the value of *k* becomes complicated.

Liu (LIU) Estimator

Liu (1993) proposed a new biased estimator of α called as LIU estimator and is given by

$$\hat{\alpha}_{LIU} = (\Lambda + I)^{-1}(\Lambda + dI)\hat{\alpha}_{OLS} \tag{8}$$

where 0 < *d* < 1, is a Liu parameter. The advantage of the LIU estimator is that $\hat{\alpha}_{LIU}$ is a linear function of *d*. Therefore, it is easier to choose *d* in $\hat{\alpha}_{LIU}$ than to choose *k* in $\hat{\alpha}_{ORR}$. Liu (1993) obtained the MSE of the LIU estimator as

$$\begin{aligned} MSE(\hat{\alpha}_{LIU}) &= tr(Cov(\hat{\alpha}_{LIU})) + (Bias(\hat{\alpha}_{LIU}))'(Bias(\hat{\alpha}_{LIU})) \\ &= \sigma^2 \sum_{j=1}^{p} \frac{(\lambda_j + d)^2}{\lambda_j(\lambda_j+1)^2} + (1-d)^2 \sum_{j=1}^{p} \frac{\alpha_j^2}{(\lambda_j+1)^2} \end{aligned} \tag{9}$$

where the optimal value of *d* is

$$d = 1 - \sigma^2 \left[\frac{\sum_{j=1}^{p} 1/\lambda_j(\lambda_j+1)}{\sum_{j=1}^{p} \alpha_j^2/(\lambda_j+1)^2} \right] \tag{10}$$

The unknown parameters *α* and σ^2 are replaced by their unbiased OLS estimates $\hat{\alpha}_{OLS}$ and $\hat{\sigma}^2_{OLS}$ respectively.

Linearized Ridge Regression (LRR) Estimator

Very recently, Liu and Gao (2011) proposed a linearized ridge regression (LRR) estimator to combat the problem of multicollinearity. It can be expressed as

$$\hat{\alpha}_{LRR} = \left(\Lambda + I\right)^{-1}\left(\Lambda + D\right)\hat{\alpha}_{OLS} \tag{11}$$

where $D = diag\left(d_1, d_2, \ldots, d_p\right), d_j \in \mathbb{R}, j = 1, 2, \ldots, p$. The optimal value of d_j proposed by Gao and Liu (2011) is given by

$$d_j = \frac{\lambda_j\left(\alpha_j^2 - \sigma^2\right)}{\sigma^2 + \lambda_j\alpha_j^2}, j = 1, 2, \ldots, p \tag{12}$$

and the unknown quantities α and σ^2 are replaced by their OLS estimates to obtain the estimate of d_j, $j = 1, 2, \ldots, p$. Gao and Liu (2011) showed that the LRR estimator attends the lower bound of the MSE of the generalized shrinkage estimators (GSE). The MSE of the LRR estimator is given by (Gao and Liu, 2011)

$$\begin{aligned} MSE\left(\hat{\alpha}_{LRR}\right) &= tr\left(Cov\left(\hat{\alpha}_{LRR}\right)\right) + \left(Bias\left(\hat{\alpha}_{LRR}\right)\right)'\left(Bias\left(\hat{\alpha}_{LRR}\right)\right) \\ &= \sigma^2 \sum_{j=1}^{p} \frac{\left(\lambda_j + d_j\right)^2}{\lambda_j\left(\lambda_j + 1\right)^2} + \sum_{j=1}^{p} \frac{\left(1 - d_j\right)^2 \alpha_j^2}{\left(\lambda_j + 1\right)^2} \end{aligned} \tag{13}$$

Here, σ^2 and α are replaced by their suitable estimates $\hat{\sigma}^2_{OLS}$ and $\hat{\alpha}_{OLS}$ respectively to obtain the estimate of the MSE of LRR estimator.

Winsorization Approach

Many robust parameter estimation methods have been developed in the literature to deal with the presence of outliers (see Huber, 1973; Birkes and Dodge, 1993). Winsorization is one of the robust techniques that aim to diminish the effect of outliers in the data. Dixon (1960), Bickel (1965), Dixon and Tukey (1968), Chen and Dixon (1972) discussed this approach. Mutan and Senoglu (2008) noted that the Winsorization does not worsen a good linear relationship on non-contaminated data. Winsorized regression is an effective alternative to the least squares estimation method which reduce the effect of contamination on the regression coefficient. To illustrate the advantage of Winsorization in estimation of regression coefficients, Yale and Forsythe (1976) introduced various methods of Winsorization and compared with each other and with the OLS estimates. Further study in Winsorization is done by Tan and Tabatabai (1988), Chen Welsh and Chan (2001). A general Winsorization procedure proposed by Yale and Forsythe (1976) is briefly introduced as follows.

Winsorization Methodology

Yale and Forsythe (1976) explained the Winsorization procedure for simple linear regression. It can be easily generalize to the multiple linear regression. In this article, following stepwise algorithm is used to obtain the least squares Winsorized (LSW) estimator. Step 1 to Step 5 are used to obtain least squares Winsorized (LSW) estimator for model given in (1) and Step 6 to Step 8 gives LSW estimator in canonical form of model defined in (2).

Stepwise Algorithm

Step 1. Using the model given in (1), obtain the OLS estimates and the predicted values ($\hat{Y}_i$) of Y_i, $i = 1, 2, ..., n$.

Step 2. Set number of points (g) to be Winsorized at each extreme.

Step 3. Obtain the residual values as $r_i = Y_i - \hat{Y}_i$ and order them. Let $r_1 \le r_2 \le \cdots \le r_n$ be ordered OLS residuals.

Step 4. Obtain the least squares estimator using n observations on Y' and X, where

$$Y_i' = \hat{Y}_i + r_i'$$

and

$$r_i' = \begin{cases} r_{g+1} & i = 1, 2, \ldots, g \\ r_i & i = g+1, \ldots, n-g \\ r_{n-g} & i = n-g+1, \ldots, n \end{cases}$$

Step 5. Repeat the above Step 4 for fixed number of iteration (b). For each iteration, the baseline data on response variable (Y) has been modified as Y=Y' (after first iteration), Y=Y'' (after second iteration), Y=Y''' (after third iteration) and so on by generating new set of residuals (r', r'', r''' and so on).

Step 6. The modified dataset at the end of b^{th} iteration is denoted by (Y^*, X). Standardize the modified dataset in such a way that Y^*X denote the correlation between the modified response variable and the set of regressor variables.

Step 7. Convert the standardized modified dataset to canonical form using the matrix of eigenvectors (Q) of $X'X$ matrix.

Step 8. Using the canonical form of model, perform the OLS estimation to obtain the LSW estimates of unknown regression parameters.

In this article, 10% and 20% observations are considered for Winsorization ($g = 0.1n, 0.2n$). Nevitt and Tam (1998) conducted a pilot study to decide the number of iterations (b). They found that, after five iterations of data modification, the results shows very little change in parameter estimates. So, five iterations are considered to obtain the LSW estimator. Because, the Winsorization is done only in Y, the diagonal matrix of eigenvalues (Λ) and the corresponding matrix of eigenvectors (Q) of X remains unchanged. Using the canonical form of model given in (2), estimators of unknown regression parameters α are proposed to tackle the problem of multicollinearity and outliers simultaneously in the data.

Proposed Estimators

New estimators based on the LSW estimator are now proposed to handle the simultaneous occurrence of multicollinearity and outliers in the data. The proposed estimators are called as Winsorized shrinkage estimators because they reduce the impact of multicollinearity by shrinking the LSW estimator. The different forms of shrinkage quantity produce the different Winsorized shrinkage estimators. In the following subsections, some Winsorized shrinkage estimators are introduced and their modified MSE Expressions are obtained. The technique suggested by Kan, Alpu and Yazici (2013) is implemented to obtain the modified MSE of the proposed estimators.

Ordinary Ridge Regression Winsorized (ORRW) Estimator

The ordinary ridge regression Winsorized (ORRW) estimator of α, based on the ORR estimator (Hoerl and Kennard, 1970a, b), is defined as

$$\hat{\alpha}_{ORRW} = \left(\Lambda + k_{LSW} I\right)^{-1} \Lambda \hat{\alpha}_{LSW} \tag{14}$$

where k_{LSW} is the unknown ridge parameter. It is estimated by using the formula $\hat{k}_{LSW} = p\hat{\sigma}^2_{LSW} / \hat{\alpha}'_{LSW}\hat{\alpha}_{LSW}$, where p denote the number of regressor variables, the $\hat{\alpha}_{LSW}$ denote the LSW estimator of α and $\hat{\sigma}^2_{LSW} = \left(Y - Z\hat{\alpha}_{LSW}\right)' \left(Y - Z\hat{\alpha}_{LSW}\right) / \left(n - p\right)$ is the estimator of σ^2 based on the LSW estimator. The modified MSE of the ORRW estimator is given by

$$MSE\left(\hat{\alpha}_{ORRW}\right) = \sigma^2 \sum\nolimits_{j=1}^{p} \frac{\lambda_j}{\left(\lambda_j + k_{LSW}\right)^2} + k^2_{LSW} \sum\nolimits_{j=1}^{p} \frac{\alpha_j^2}{\left(\lambda_j + k_{LSW}\right)^2} \tag{15}$$

The unknown parameters σ^2, α and k_{LSW} are replaced by $\hat{\sigma}^2_{LSW}$, $\hat{\alpha}_{LSW}$ and $\hat{k}_{LSW}$ respectively.

Liu Winsorized (LIUW) Estimator

The Liu Winsorized estimator (LIUW), based on the Liu estimator (Liu, 1993), is defined as

$$\hat{\alpha}_{LIUW} = \left(\Lambda + I\right)^{-1} \left(\Lambda + d_{LSW} I\right) \hat{\alpha}_{LSW} \tag{16}$$

where d_{LSW} is a Liu parameter and it is obtained by using the following formula

$$d_{LSW} = 1 - \sigma^2 \left[\frac{\sum\nolimits_{j=1}^{p} 1/\lambda_j \left(\lambda_j + 1\right)}{\sum\nolimits_{j=1}^{p} \alpha_j^2 / \left(\lambda_j + 1\right)^2} \right] \tag{17}$$

The estimate of d_{LSW} denoted by $\hat{d}_{LSW}$ is obtained by replacing the unknown parameters σ^2 and α in (17) by their estimates based on the LSW estimator. The modified MSE of LIUW estimator is obtained by

$$MSE\left(\hat{\alpha}_{LIUW}\right)=\sigma^2\sum_{j=1}^{p}\frac{\left(\lambda_j+d_{LSW}\right)^2}{\lambda_j\left(\lambda_j+1\right)^2}+\left(1-d_{LSW}\right)^2\sum_{j=1}^{p}\frac{\alpha_j^2}{\left(\lambda_j+1\right)^2} \tag{18}$$

and the unknown parameters are replaced by their corresponding estimates based on the LSW estimator.

Linearized Ridge Regression Winsorized (LRRW) Estimator

The LRRW estimator based on the LSW estimator, (Liu and Gao, 2011) is defined as

$$\hat{\alpha}_{LRRW}=\left(\Lambda+I\right)^{-1}\left(\Lambda+D_{LSW}\right)\hat{\alpha}_{LSW} \tag{19}$$

where $\left(\Lambda+I\right)^{-1}\left(\Lambda+D_{LSW}\right)$ is a shrinkage matrix and a diagonal matrix D_{LSW} is an order of $p\times p$ with diagonal elements $d_{LSW_j}, j=1,2,...,p$ such that $d_{LSW_j}\in\mathbb{R}$ is obtained by using the formula

$$d_{LSW_j}=\frac{\lambda_j\left(\alpha_j^2-\sigma^2\right)}{\sigma^2+\lambda_j\alpha_j^2},\quad j=1,2,...,p \tag{20}$$

where α and σ^2 are estimated using the LSW estimators $\hat{\alpha}_{LSW}$ and $\hat{\sigma}^2_{LSW}$. The modified MSE of the LRRW estimator is given by

$$MSE\left(\hat{\alpha}_{LRRW}\right)=\sigma^2\sum_{j=1}^{p}\frac{\left(\lambda_j+d_{LSW_j}\right)^2}{\lambda_j\left(\lambda_j+1\right)^2}+\sum_{j=1}^{p}\frac{\left(1-d_{LSW_j}\right)^2\alpha_j^2}{\left(\lambda_j+1\right)^2} \tag{21}$$

Here, σ^2 and α are replaced by their suitable estimates based on the LSW estimator.

Simulation Study

A simulation study was carried out to evaluate the performance of proposed estimators. First, the estimated MSE's (EMSE) of the different estimators are obtained and based on the average EMSE (AEMSE), the existing and proposed estimators are compared. Secondly, the relative average EMSE's (RAEMSE) of estimators with respect to the OLS estimator are obtained and the average reduction in the estimated MSE's of the estimators with respect to the OLS estimator for the different Winsorization proportions is noted.

Comparison of Estimators through Estimated MSE

The regressor variables are generated using a simulation design proposed by McDonald and Galarneau (1975) as

$$x_{ij} = \left(1-\rho^2\right)^{\frac{1}{2}} \zeta_{ij} + \rho\zeta_{i(p+1)}, \qquad i = 1,2,\ldots,n,\ j = 1,2,\ldots,p \tag{22}$$

where ζ_{ij} are independent pseudo random numbers generated from standard normal distribution and ρ^2 is the correlation between any two regressor variables. The following regression model is used to generate n observations on the response variables

$$Y = 10 + 4X_1 + 6X_2 + 2X_3 + 8X_4 + \varepsilon$$

where the error variable ε is generated using the contaminated normal distribution. The δ% contamination is done using the following mixture of normal distributions

$$\varepsilon_i \sim f_{\varepsilon_i}(\cdot) = (1-\delta) \times N(0,1) + \delta \times N\left(0,10^2\right).$$

For δ = 0%, 10%, 20% and 30%, and n = 20, 30, and 50, the different degrees of multicollinearity have been achieved by generating regressor variables using the model given in (22) for ρ = 0.9, 0.99, 0.999 and 0.9999. The 0.1n and 0.2n points are Winsorized at each extreme to reduce the effect of outlier observations. Hence, the 10% and 20% Winsorized estimators of OLS, ORR, LIU and LRR are denoted by LSW10, ORRW10, LIUW10, LRRW10 and LSW20, ORRW20, LIUW20, LRRW20 respectively.

The EMSE of OLS, ORR, LIU, LRR, OLSW10, ORRW10, LIUW10, LRRW10, OLSW20, ORRW20, LIUW20 and LRRW20 estimators are obtained by replacing the values of unknown parameters with their suitable estimates in their respective MSE expressions. Note that, the EMSE of the LIU, LIUW10 and LIUW20 is considered corresponding to those iterations where the estimate of Liu parameter

(d) lies between 0 and 1. For each combination of sample size (n), degree of multicollinearity (ρ) and contamination proportion (δ), the above simulation experiment is repeated 10,000 times and the AEMSE of these estimators are obtained and reported in Table 1. Also, for sample size n = 30, the AEMSE of each estimator was plotted for all combinations of ρ and δ. They are depicted graphically in Figure 1.

Table 1. AEMSE of Estimators

***n* = 20**	**δ = 0%**				**δ = 10%**			
	0.9	**0.99**	**0.999**	**0.9999**	**0.9**	**0.99**	**0.999**	**0.9999**
OLS	0.0031	0.0254	0.2512	2.4896	0.0315	0.2629	2.6386	26.5099
ORR	0.0030	0.0203	0.1254	1.1043	0.0226	0.1301	1.1880	11.8995
LIU	0.0029	0.0199	0.1804	1.7630	0.0209	0.1525	1.7741	18.4332
LRR	0.0024	0.0133	0.0934	0.8641	0.0153	0.0979	0.9179	9.2182
LSW10	0.0020	0.0169	0.1671	1.6480	0.0081	0.0676	0.6723	6.7148
ORRW10	0.0020	0.0144	0.0979	0.8729	0.0072	0.0476	0.4125	4.0837
LIUW10	0.0020	0.0138	0.1219	1.1870	0.0069	0.0500	0.5392	5.5197
LRRW10	0.0017	0.0098	0.0727	0.6783	0.0053	0.0349	0.3179	3.1419
LSW20	0.0013	0.0105	0.1045	1.0304	0.0043	0.0354	0.3482	3.4815
ORRW20	0.0013	0.0095	0.0704	0.6329	0.0040	0.0291	0.2608	2.5935
LIUW20	0.0013	0.0091	0.0789	0.7568	0.0039	0.0292	0.2931	2.9764
LRRW20	0.0011	0.0068	0.0527	0.4941	0.0032	0.0221	0.2048	2.0340
	δ = 20%				**δ = 30%**			
	0.9	**0.99**	**0.999**	**0.9999**	**0.9**	**0.99**	**0.999**	**0.9999**
OLS	0.0584	0.4960	4.8625	48.0399	0.0849	0.7186	7.0523	69.7408
ORR	0.0383	0.2290	2.1818	21.7413	0.0526	0.3279	3.1753	31.3082
LIU	0.0371	0.3185	3.3957	34.1454	0.0550	0.4875	4.9287	48.9810
LRR	0.0261	0.1766	1.6910	16.7676	0.0363	0.2545	2.4522	24.1812
LSW10	0.0193	0.1598	1.5703	15.5570	0.0345	0.2865	2.8498	28.0075
ORRW10	0.0156	0.0998	0.9268	9.1832	0.0263	0.1691	1.6315	16.0833
LIUW10	0.0148	0.1131	1.2099	12.2463	0.0255	0.2045	2.1224	20.8768
LRRW10	0.0112	0.0754	0.7156	7.0553	0.0185	0.1298	1.2524	12.3462
LSW20	0.0092	0.0755	0.7412	7.2817	0.0162	0.1347	1.3280	12.9899
ORRW20	0.0082	0.0582	0.5489	5.4090	0.0140	0.0989	0.9536	9.4040
LIUW20	0.0080	0.0606	0.6169	6.0481	0.0136	0.1064	1.0715	10.4862
LRRW20	0.0063	0.0448	0.4308	4.2137	0.0104	0.0768	0.7429	7.3156

Table 1, continued.

***n* = 30**	***δ* = 0%**				***δ* = 10%**			
	0.9	**0.99**	**0.999**	**0.9999**	**0.9**	**0.99**	**0.999**	**0.9999**
OLS	0.0018	0.0152	0.1504	1.5042	0.0188	0.1618	1.5641	15.4151
ORR	0.0018	0.0133	0.0829	0.6727	0.0154	0.0851	0.6966	6.9166
LIU	0.0018	0.0126	0.1122	1.1217	0.0145	0.1003	1.0987	11.5955
LRR	0.0015	0.0091	0.0586	0.5276	0.0104	0.0619	0.5456	5.3711
LSW10	0.0012	0.0102	0.1008	1.0052	0.0037	0.0309	0.2943	2.9777
ORRW10	0.0012	0.0093	0.0646	0.5373	0.0035	0.0244	0.1856	1.8383
LIUW10	0.0012	0.0088	0.0763	0.7571	0.0034	0.0241	0.2343	2.4833
LRRW10	0.0011	0.0066	0.0457	0.4190	0.0028	0.0173	0.1428	1.4246
LSW20	0.0008	0.0063	0.0621	0.6184	0.0019	0.0159	0.1519	1.5370
ORRW20	0.0008	0.0059	0.0455	0.3881	0.0018	0.0141	0.1177	1.1710
LIUW20	0.0008	0.0056	0.0486	0.4768	0.0018	0.0137	0.1290	1.3243
LRRW20	0.0007	0.0044	0.0327	0.3036	0.0015	0.0106	0.0924	0.9255
	***δ* = 20%**				***δ* = 30%**			
	0.9	**0.99**	**0.999**	**0.9999**	**0.9**	**0.99**	**0.999**	**0.9999**
OLS	0.0354	0.3013	2.9141	29.3682	0.0511	0.4325	4.1930	42.7257
ORR	0.0268	0.1462	1.3014	12.9487	0.0363	0.2013	1.8649	19.0030
LIU	0.0256	0.2055	2.1456	21.7532	0.0365	0.3080	3.1180	31.5423
LRR	0.0177	0.1105	1.0168	10.0854	0.0239	0.1550	1.4581	14.8004
LSW10	0.0090	0.0761	0.7419	7.4299	0.0176	0.1485	1.4250	14.6328
ORRW10	0.0081	0.0523	0.4451	4.4012	0.0149	0.0924	0.8218	8.4375
LIUW10	0.0078	0.0559	0.5759	5.8704	0.0142	0.1084	1.1069	11.2143
LRRW10	0.0060	0.0383	0.3453	3.4063	0.0105	0.0690	0.6388	6.5194
LSW20	0.0039	0.0325	0.3183	3.1710	0.0071	0.0592	0.5764	5.8364
ORRW20	0.0038	0.0274	0.2455	2.4276	0.0066	0.0470	0.4311	4.3970
LIUW20	0.0037	0.0274	0.2684	2.6864	0.0064	0.0484	0.4791	4.8429
LRRW20	0.0030	0.0209	0.1938	1.9226	0.0051	0.0359	0.3402	3.4536

***n* = 50**	***δ* = 0%**				***δ* = 10%**			
	0.9	**0.99**	**0.999**	**0.9999**	**0.9**	**0.99**	**0.999**	**0.9999**
OLS	0.0010	0.0084	0.0832	0.8325	0.0107	0.0893	0.8805	8.6261
ORR	0.0010	0.0078	0.0536	0.3722	0.0097	0.0540	0.3981	3.8131
LIU	0.0010	0.0074	0.0646	0.6398	0.0092	0.0606	0.6416	6.5740
LRR	0.0009	0.0057	0.0351	0.2932	0.0069	0.0366	0.3126	2.9891
LSW10	0.0007	0.0057	0.0560	0.5618	0.0015	0.0124	0.1236	1.2125
ORRW10	0.0007	0.0054	0.0408	0.3007	0.0015	0.0111	0.0847	0.7510
LIUW10	0.0007	0.0052	0.0442	0.4342	0.0015	0.0105	0.0996	0.9978
LRRW10	0.0006	0.0041	0.0273	0.2347	0.0013	0.0079	0.0628	0.5858
LSW20	0.0004	0.0035	0.0344	0.3439	0.0008	0.0066	0.0655	0.6442
ORRW20	0.0004	0.0034	0.0279	0.2184	0.0008	0.0062	0.0539	0.4973
LIUW20	0.0004	0.0033	0.0279	0.2713	0.0008	0.0060	0.0560	0.5517
LRRW20	0.0004	0.0027	0.0193	0.1703	0.0007	0.0048	0.0415	0.3962

Table 1, continued.

***n* = 50**	***δ* = 20%**				***δ* = 30%**			
	0.9	**0.99**	**0.999**	**0.9999**	**0.9**	**0.99**	**0.999**	**0.9999**
OLS	0.0200	0.1674	1.6402	16.3004	0.0288	0.2411	2.3471	23.6339
ORR	0.0170	0.0894	0.7280	7.1945	0.0234	0.1218	1.0353	10.4008
LIU	0.0161	0.1205	1.2570	12.4734	0.0224	0.1807	1.8062	18.3140
LRR	0.0114	0.0638	0.5749	5.6323	0.0153	0.0907	0.8127	8.1273
LSW10	0.0038	0.0314	0.3116	3.0516	0.0082	0.0682	0.6680	6.7232
ORRW10	0.0037	0.0249	0.1946	1.8493	0.0076	0.0480	0.3886	3.8843
LIUW10	0.0036	0.0245	0.2442	2.4453	0.0073	0.0513	0.5133	5.2782
LRRW10	0.0029	0.0173	0.1497	1.4394	0.0056	0.0342	0.3015	3.0003
LSW20	0.0015	0.0126	0.1242	1.2294	0.0028	0.0230	0.2242	2.2560
ORRW20	0.0015	0.0115	0.1006	0.9742	0.0027	0.0201	0.1762	1.7555
LIUW20	0.0015	0.0111	0.1067	1.0578	0.0027	0.0197	0.1902	1.9220
LRRW20	0.0013	0.0087	0.0795	0.7768	0.0022	0.0152	0.1390	1.3828

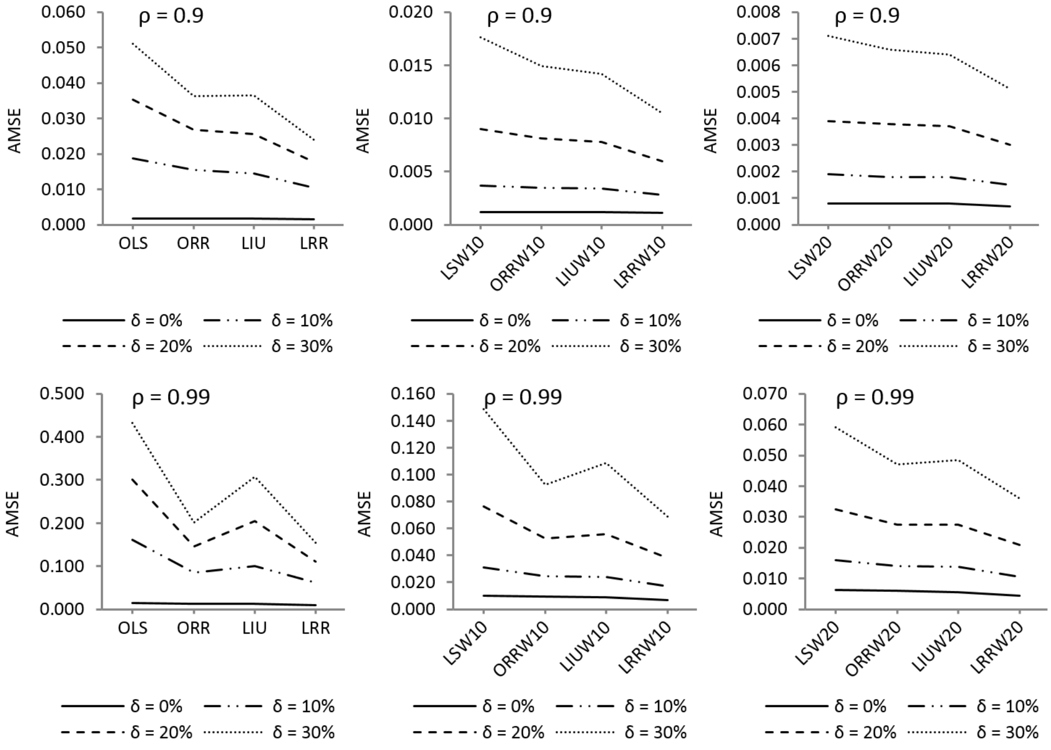

Figure 1. AEMSE plot of various estimators for different combinations of *ρ* and *δ*

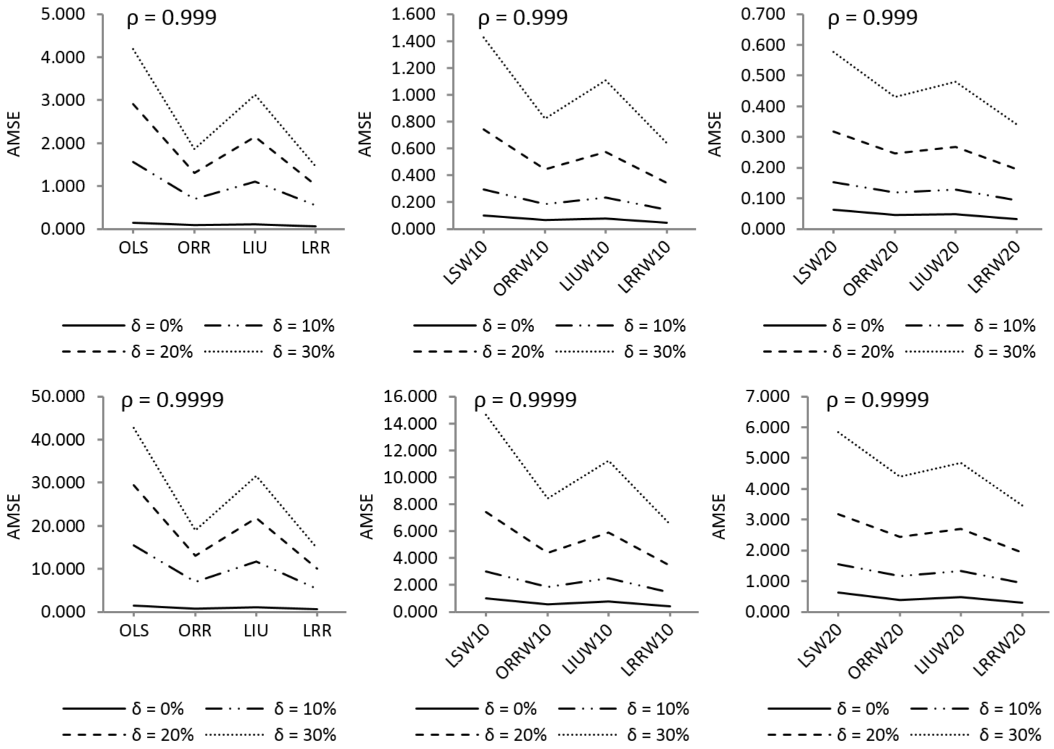

Figure 1, continued.

Table 1 and Figure 1 clearly indicate that

- For each combination of *n*, *ρ* and *δ*, the LRRW20 estimator has consistently smaller AEMSE value than that of the other estimators. It clearly indicates that the estimator LRRW20 shows better performance as compare to the other estimators in the EMSE sense.
- The AEMSE of each estimator decreases with increase in sample size (*n*), but it increases with increase in the proportion of contamination (*δ*) in the error variable.
- When degree of multicollinearity increases, the AEMSE of each estimator is also increases.
- For any combination of *n*, *ρ* and *δ*, as degree of Winsorization (*δ*) increases, the AEMSE of each estimator goes on decreases.

Relative AEMSE (RAEMSE) comparison

The RAEMSE is one of the suitable measure to evaluate the performance of estimators. The RAEMSE of estimator 'T' with respect to the OLS estimator is obtained by using the formula

$$\text{RAEMSE}_{\text{T}} = (\text{AEMSE}_{\text{OLS}} - \text{AEMSE}_{\text{T}}) / \text{AEMSE}_{\text{OLS}}$$

where $\text{AEMSE}_{\text{OLS}}$ and AEMSE_{T} denote the AEMSE of the OLS estimator and considered estimator 'T'. The maximum value of RAEMSE_{T} is one. RAEMSE_{T} greater than zero indicates the corresponding estimator 'T' performs better than the OLS estimator in AEMSE sense. The RAEMSE_{T} close to one indicates the corresponding estimator 'T' outperforms as compare to the OLS estimator.

Using the AEMSE's of the OLS, ORR, LIU, LRR, OLSW10, ORRW10, LIUW10, LRRW10, OLSW20, ORRW20, LIUW20 and LRRW20 estimators obtained in Table 1, the RAEMSE of each estimator was computed with respect to the OLS estimator. For all combinations of ρ and δ with n = 30, the RAEMSE of each estimator is presented in Table 2. Also, RAEMSE of all considered estimators with respect to the OLS estimator is plotted in Figure 2.

Table 2. RAEMSE of estimators with the OLS estimator for *n* = 30

	ρ = 0.9				ρ = 0.99			
	δ = 0%	δ = 10%	δ = 20%	δ = 30%	δ = 0%	δ = 10%	δ = 20%	δ = 30%
ORR	0.0000	0.1809	0.2429	0.2896	0.1250	0.4740	0.5148	0.5346
LIU	0.0000	0.2287	0.2768	0.2857	0.1711	0.3801	0.3180	0.2879
LRR	0.1667	0.4468	0.5000	0.5323	0.4013	0.6174	0.6333	0.6416
LSW10	0.3333	0.8032	0.7458	0.6556	0.3289	0.8090	0.7474	0.6566
ORRW10	0.3333	0.8138	0.7712	0.7084	0.3882	0.8492	0.8264	0.7864
LIUW10	0.3333	0.8191	0.7797	0.7221	0.4211	0.8511	0.8145	0.7494
LRRW10	0.3889	0.8511	0.8305	0.7945	0.5658	0.8931	0.8729	0.8405
LSW20	0.5556	0.8989	0.8898	0.8611	0.5855	0.9017	0.8921	0.8631
ORRW20	0.5556	0.9043	0.8927	0.8708	0.6118	0.9129	0.9091	0.8913
LIUW20	0.5556	0.9043	0.8955	0.8748	0.6316	0.9153	0.9091	0.8881
LRRW20	0.6111	0.9202	0.9153	0.9002	0.7105	0.9345	0.9306	0.9170

Table 2, continued.

	ρ = 0.999				*ρ* = 0.9999			
	δ = 0%	*δ* = 10%	*δ* = 20%	*δ* = 30%	*δ* = 0%	*δ* = 10%	*δ* = 20%	*δ* = 30%
ORR	0.4488	0.5546	0.5534	0.5552	0.5528	0.5513	0.5591	0.5552
LIU	0.2540	0.2976	0.2637	0.2564	0.2543	0.2478	0.2593	0.2617
LRR	0.6104	0.6512	0.6511	0.6523	0.6492	0.6516	0.6566	0.6536
LSW10	0.3298	0.8118	0.7454	0.6601	0.3317	0.8068	0.7470	0.6575
ORRW10	0.5705	0.8813	0.8473	0.8040	0.6428	0.8807	0.8501	0.8025
LIUW10	0.4927	0.8502	0.8024	0.7360	0.4967	0.8389	0.8001	0.7375
LRRW10	0.6961	0.9087	0.8815	0.8477	0.7214	0.9076	0.8840	0.8474
LSW20	0.5871	0.9029	0.8908	0.8625	0.5889	0.9003	0.8920	0.8634
ORRW20	0.6975	0.9247	0.9158	0.8972	0.7420	0.9240	0.9173	0.8971
LIUW20	0.6769	0.9175	0.9079	0.8857	0.6830	0.9141	0.9085	0.8867
LRRW20	0.7826	0.9409	0.9335	0.9189	0.7982	0.9400	0.9345	0.9192

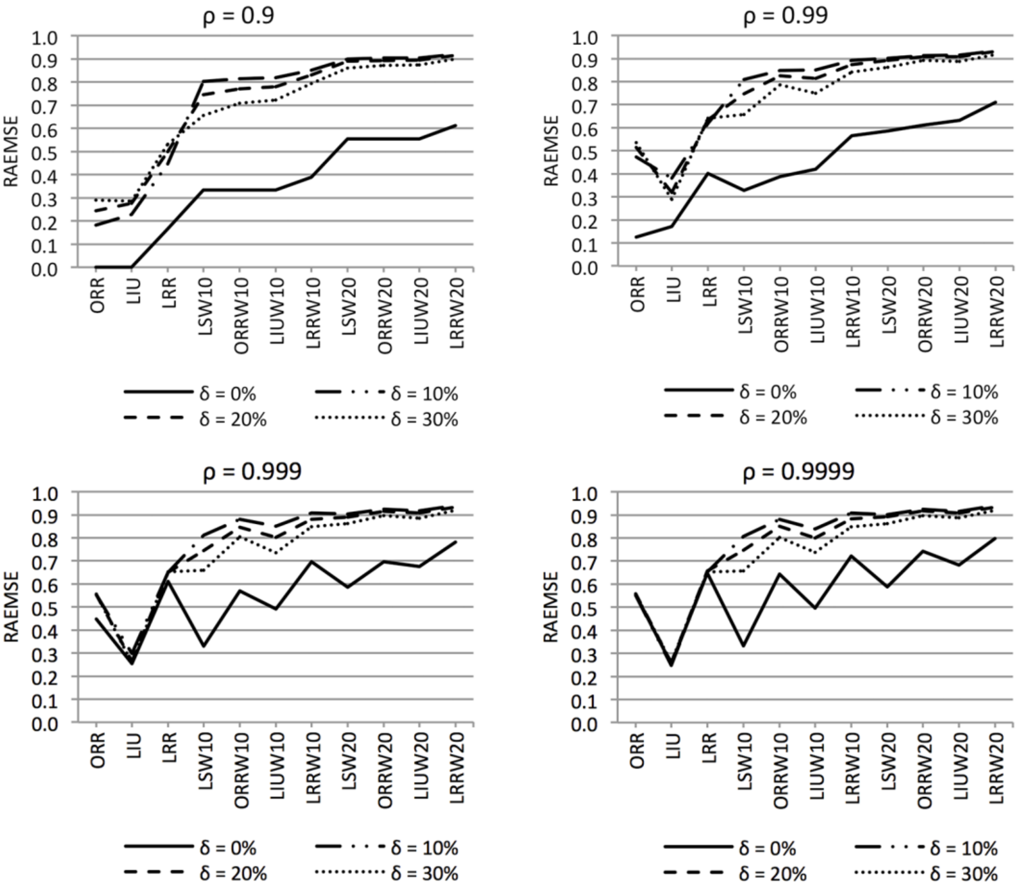

Figure 2. Line plot plot of RAEMSE of estimators with respect to the OLS estimator

Table 2 and Figure 2 show that

- For 0% contamination, as degree of multicollinearity (ρ) increases, the RAEMSE of each estimator with respect to the OLS estimator is also increases.
- With 0% contamination and for $\rho = 0.9$, on an average, 10% Winsorized shrinkage estimators (LSW10, ORRW10, LIUW10 and LRRW10) shows 34.72% reduction in AEMSE with respect to the OLS estimator. Similarly, for $\rho = 0.99$, 0.999 and 0.9999, it shows 42.60%, 52.23% and 54.82% reduction respectively. Also for ρ = 0.9, 0.99, 0.999 and 0.9999, the 20% Winsorized shrinkage estimators (LSW20, ORRW20, LIUW20 and LRRW20), on an average shows 56.94%, 63.49%, 68.60% and 70.30% reduction in AEMSE respectively.
- On the similar line, for $\delta = 30\%$, the 10% Winsorization shows on an average 72.02%, 75.82%, 76.20% and 76.12% reduction in AEMSE for ρ = 0.9, 0.99, 0.999 and 0.9999 and for 20% Winsorization, it is 87.67%, 88.99%, 89.11% and 89.16% respectively.

Real Data Example

A real data set on tobacco blends given by Myers (1990) is used to evaluate the performance of various estimators. The response variable Y measures the heat evolved from tobacco during the smoking process. This data set contains 30 observations and four regressor variables namely X_1, X_2, X_3, and X_4. Arslan and Billor (2000) noted that the tobacco blends data suffers from the problem of multicollinearity and outliers simultaneously. The variance inflation factor (VIF) values for each term are 324.1412, 45.1728, 173.2577 and 138.1753. It indicates the severe problem of multicollinearity.

For this real data, the estimate of the bias (EBIAS), variance (EVAR) and MSE (EMSE) of the OLS, ORR, LIU, LRR, OLSW10, ORRW10, LIUW10, LRRW10, OLSW20, ORRW20, LIUW20 and LRRW20 estimators were obtained and are reported in Table 3. Also, the relative EMSE (REMSE) of each estimator with respected to the OLS estimator is computed and presented in Table 3. Positive value of REMSE implies the performance of the corresponding estimator is better than the OLS estimator.

Table 3. EBIAS, EVAR, EMSE and REMSE of Estimators

Estimators	EBIAS	EVAR	EMSE	REMSE	REMSE (in %)
OLS	0.000000	1.120600	1.120600	-	-
ORR	0.352000	0.482000	0.937300	0.163573	16.357300
LIU	-0.058100	0.544200	0.883900	0.211226	21.122600
LRR	0.078300	0.647300	0.850400	0.241121	24.112100
LSW10	0.000000	0.607500	0.607500	0.457880	45.788000
ORRW10	0.125700	0.325700	0.480600	0.571123	57.112300
LIUW10	-0.157600	0.382400	0.507900	0.546761	54.676100
LRRW10	0.120200	0.367800	0.469400	0.581117	58.111700
LSW20	0.000000	0.088100	0.088100	0.921381	92.138100
ORRW20	0.045000	0.080400	0.086700	0.922631	92.263100
LIUW20	-0.012700	0.083900	0.086000	0.923255	92.325500
LRRW20	0.018600	0.083200	0.085600	0.923612	92.361200

From Table 3, it seems that the increase in Winsorization proportion reduces the EVAR and EMSE of each estimator. 10% and 20% Winsorization on an average shows 53.92% and 92.27% reduction in the EMSE with respect to the OLS estimator respectively. Also, LRRW20 shows smaller EMSE as compare to other estimators.

Conclusion

A Winsorized form of the OLS estimator, ORR estimator, LIU estimator and LRR estimators are introduced. A simulation study and a real data example show that the Winsorization procedure reduces the EMSE of estimators and improve the performance of the estimators. Also, it reveals that the LRR estimator with 20% Winsorization shows consistently minimum EMSE among the all other considered estimators.

References

Andrews, D. F. (1974). A Robust Method for Multiple Linear Regression. *Technometrics, 16*(4), 523-531.

Arslan, O., & Billor, N. (2000). Robust Liu estimator for regression based on an M-estimator. *Journal of Applied Statistics, 27*(1), 39-47.

Bickel, P. J. (1965). On Some Robust Estimates of Location. *Annals of Mathematical Statistics, 36*, 847-858.

Birkes, D., & Dodge, Y. (1993). *Alternative methods of regression*. New York: John Wiley & Sons.

Chen, E. H., & Dixon, W. J. (1972). Estimates of Parameters of a Censored Regression Sample. *Journal of the American Statistical Association, 67*, 664-671.

Chen, L. A., Welsh, A. H., & Chan, W. (2001). Estimators for the Linear Regression Model based on Winsorized Observations. *Statistica Sinica, 11*, 147-172.

Dixon, W. J. (1960). Simplified Estimation from Censored Normal Samples. *Annals of Mathematical Statistics, 3*, 385-391.

Dixon, W. J., & Tukey, J. W. (1968). Approximate Behaviour of the Distribution of Winsorized t (Trimming / Winsorization 2). *Technometrics, 10*, 83-98.

Draper, N. R., & Smith, H. (1998). *Applied Regression Analysis* (3rd Ed.). New York: John Wiley & Sons.

Gao, F., & Liu, X. Q. (2011). Linearized Ridge Regression Estimator under the Mean Square Error Criterion in a Linear Regression Model. *Communications in Statistics – Simulation and Computation, 40*, 1434-1443.

Hoerl, A. E., & Kennard, R. W. (1970a). Ridge regression: Biased Estimation for Nonorthogonal Problems. *Technometrics, 12*, 55-67.

Hoerl, A. E., & Kennard, R. W. (1970b). Ridge regression: Applications to Nonorthogonal Problems. *Technometrics, 12*, 69-82.

Hoerl, A. E., Kennard, R. W., & Baldwin, K. F. (1975). Ridge regression: Some Simulations. *Communications in Statistics – Theory and Methods, 4*(2), 105-123.

Huber, P. J. (1973). Robust Regression: Asymptotics, Conjectures, and Monte Carlo. *Annals of Statistics, 1*, 799–821.

Kan, B., Alpu, O., & Yazici, B. (2013). Robust ridge and robust Liu estimator for regression based on the LTS estimator. *Journal of Applied Statistics, 40*(3), 644-655.

Liu, K. (1993). A New Class of Biased Estimate in Linear Regression. *Communications in Statistics – Theory and Methods*, *22*, 393-402.

Liu, K. (2003). Using Liu-Type Estimator to Combat Collinearity. *Communications in Statistics – Theory and Methods*, *32*(5), 1009-1020.

Liu, X. Q., & Gao, F., (2011). Linearized Ridge Regression Estimator in Linear Regression. *Communications in Statistics – Theory and Methods*, *40*(12), 2182-2192.

McDonald, G. C., & Galarneau, D. I. (1975). A Monte Carlo Evaluation of Some Ridge-type Estimators. *Journal of the American Statistical Association, 70*(350), 407-416.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2006). *Introduction to Linear Regression Analysis* (3rd ed.). New York: John Wiley & Sons.

Mutan, O. C., & Senoglu, B. (2008). A Monte Carlo Comparison of Regression Estimators When the Error Distribution is Long-Tailed Symmetric. *Journal of Modern Applied Statistical Methods*, *8*(1), 161-172. http://digitalcommons.wayne.edu/jmasm/vol8/iss1/14/

Myers, R. H. (1990). *Classical and Modern Regression with Applications* (2nd ed.). Boston: Duxbury.

Nevitt, J., & Tam, H. P. (1998). A Comparison of Robust and Nonparametric Estimators under the Simple Linear Regression Model. *Multiple Regression Viewpoints*, *25*, 54-69.

Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust Regression and Outlier Detection*. New York: John Wiley & Sons.

Tan, W. Y., & Tabatabai, M. A. (1988). A Modified Winsorized Regression Procedure for Linear Models. *Journal of Statistical Computation and Simulation, 30*, 299-313.

Yale, C., & Forsythe, A. B. (1976). Winsorized Regression. *Technometrics, 18*, 291-300.