MATHEMATICAL MODELING AND COMPUTING, Vol. 9, No. 1, pp. 26–35 (2022)

Robust bootstrap regression testing in the presence of outliers

Hassan S. U., Ali K. H.

Department of Statistics, University of Al-Qadisiyah, Aljamea'a Str., Diwanyiah, Iraq

(Received 7 July 2021; Accepted 16 November 2021)

Bootstrap is one of the random sampling methods with replacement, that was proposed to address the problem of small samples whose distributions are difficult to derive. The distribution of bootstrap samples is empirical or free and due to its random sampling with replacement, the probability of choosing a specific observation may be equal to one. Unfortunately, when the original sample data contains an outlier, there is a serious problem that leads to a breakdown OLS (Ordinary Least Squares) estimator, and robust regression methods should be recommended. It is well known that the best robust regression method has a high breakdown point is not more than 0.50, so the robust regression method would break down when the percentage of outliers in the bootstrap sample exceeds 0.50. It is well known that fixed-x bootstrap is resampled the residuals which probably are having outliers. Moreover, the leverage point(s) is an outlier that occurs in X-direction, so the effects of it on fixed-x bootstrap samples would be existence. However, the decision-making about the null hypothesis of bootstrap regression coefficients could not be reliable. In this paper, we propose using weighted fixed-x bootstrap with a probability approach to guarantee the percentage of outliers in the bootstrap samples will be very low. And then weighted M-estimate should be to tackle the problem of outliers and leverage points and taking a more reliable decision about bootstrap regression coefficients hypothesis test. The performance of the suggested method has been tested with others by using real data and simulation. The results show our proposed method is more efficient and reliable than the others.

Keywords: *bootstrap, robust regression, confidence intervals, point, WBP, weighted M, hypothesis test.*

2010 MSC: 62G10, 62G35, 62G09, 62J05 **DOI:** 10.23939/mmc2022.01.026

1. Introduction

The hypothesis test about regression coefficients is very important due to it helps to know which variables are more impact on the dependent variable and that can be used for prediction. It is well known that the estimates of the LS method are the best linear unbiased estimates when their assumptions are met. Unfortunately, in the real world of data is very hard to satisfy all of these assumptions (Uraibi et al., 2009). For instance, the presence of outliers violates the normality assumption of random errors and therefore robust regression methods are recommended. There are two types of outliers in regression data, one in the y-direction or in the regression residuals which are so-called outliers, and leverage points that are present in X-direction. So, the random error distribution F_ε is approximately normal and can be formulized as follows,

$$F_\varepsilon = (1-\varepsilon)N + \varepsilon H, \tag{1}$$

where N is normal with zero mean and constant variance, H may be another distribution, and $\varepsilon \in [0, 0.5]$. Sometimes, H is also normal distribution but with different parameters and in case Eq. (1) is considered mixture normal distribution. However, the parameters of H would determine the shape of the distribution, probably thin-tailed or heavy-tailed (thinner or heavier than exponential distribution), for more details about the effect of outliers, see (Uraibi et al.; 2015, Uraibi and Midi; 2019, Uraibi and Midi; 2017). Huber and Ronchetti (1981) introduced M-estimate that is an iterated and re-weighted LS

method to obtain robust regression coefficients. Rousseeuw (1984) Least Median Squares (LMS) which is ordering the squared of residuals from lower to upper values and then the estimation of regression coefficients is based on the half of data that analog the lower values of squared residuals. Rousseeuw and Leroy, (1987) considered dealing with 50% of data means losing a lot of information about the studied phenomena, therefore they suggested Least trimmed squared (LTS) which looks for the clean subset of data after trimming the proportion of outliers and then the regression coefficients should be estimated using LS. Least Absolute Deviation (Huber, 1987) were put forward to minimize the sum of absolute values of errors. The breakdown point of these methods is $1/n$ when the single leverage point is present in the dataset set (Croux et al., 2003). Yohai (1987) proposed MM-estimator which is a highly efficient and 50% breakdown point that is resistant to the presence of outliers and leverage points (Yohai and Maronna,1976).

Bootstrap is another approach that has not required any distributional assumptions for random errors (Efron, 1992). It is a random resampling procedure with replacement to construct free or empirical distribution for data. The fast and efficient non-parametric bootstrap method is fixed-x bootstrap or what is called residual bootstrap method which is fully dependent on resampling regression residuals. MacKinnon (2006) pointed out that when the model is linear with independent errors that are not correlated with independent variables then accurate inference can be done by residual bootstrap. On the contrary, Koenker (2005) reported that the residual bootstrap probably is independent but not identically distributed. Consequently, fixed-x bootstrap does not guarantee the homogenous of resampled residuals.

Moreover, due to the fixed-x bootstrap is sampling with replacement, the outlier could appear in the residuals bootstrap sample probably one time, two times, or as a full dataset. In this case, the simulated bootstrap distribution perhaps influenced by these bootstrap samples because it is a higher proportion of outliers than in the original data set. Consequently, the classical fixed-x bootstrap is non-robustness (Shao, 1990) and it fails when the error distribution is the heavy tail (Athreya, 1987). Great efforts have been paid the literature for bootstrap robustness (see, e.g. Shao (1992), Stromberg (1997), Singh (1998), Willems and Van Aelst (2005), and Midi et al. (2009)). Amado and Pires (2004) suggested resampling bootstrap with probability to ascribe more importance to some sample than the other one. Midi et al. (2009) mentioned that the previous method is not for regression setting and therefore were proposed weighted bootstrap with probability (WBP). The WBP algorithm assigns very chance probability to outlier to be chosen in bootstrap samples.

The WBP algorithm assigns a very low chance of abnormal observation to be chosen in bootstrap samples. Therefore, using WBP with residual bootstrap would be resampling the normal residuals that possess a high probability to be selected in the bootstrap samples. Consequently, if there are no leverage points in the data, the classical LS method can be used to estimate the regression coefficients of each bootstrap sample. As we know that, the breakdown point of LS is $1/n$ in the presence of outliers (leverage point). So, when the data are having leverage points, the LS would not be a feasible choice. In this paper, we propose weighted the design matrix X to reduce the effect of leverage points. Employing a weighted design matrix with WBP should increase the resistance of LS to the effect of leverage points. Homogeneity of residuals is an essential issue therefore, the scaled residuals are used with WBP instead of residuals to get a constant variance.

Cherink and LaBudde (2011) stated that the relationship between hypotheses tests and confidence intervals make it possible to construct a bootstrap test to obtain bootstrap confidence interval. For instance, reject the null hypothesis $H_0 \colon \beta = 0$, where the significant level is α if and only if the value of zero lies outside the bootstrap confidence interval, even of using one or two tailed confidence intervals for one or two tailed tests. One of the prevalent incorrectness of the practitioners of statistics is that they considered the confidence intervals of resampling methods such as bootstrap do not have the interpretation, and believed that through one sample there is a $(1 - \alpha)$ chance the confidence interval around the estimated parameter contains the true one. Indeed, if the distribution of the estimated parameter is exactly or approximately derived, the exactly or approximately confidence intervals can be

formed, respectively. But when the distribution of the estimated parameter is unknown, the bootstrap method is one of the solutions to estimate the distribution that is used for forming approximate confidence intervals. Furthermore, the repeated samples may include $(1-\alpha)$ out of 100 confidence intervals that would be expected to contain the true parameter.

This paper suggests hypothesis testing of Weighted Fixed Bootstrap with Probability of WM-regression coefficients (WFBP.WM). The WFBP.WM is put forward improve the performance of WBP method in the presence of outliers and leverage points and obtaining accurate hypotheses test. This paper is organized to present the Weighted M-estimate in Section 2. Section 3 describes the algorithm of WFBP.WM hypotheses test. Section 4 and Section 5 illustrate numerical example and simulation study to assess the performance of the WFBP.WM algorithm. Section 6 presents the conclusion.

2. Weighted M-estimate

Consider the linear regression model

$$y_i = X_i\beta + e_i, \quad i = 1, 2, \ldots, n, \tag{2}$$

where X_i is the p dimensional of independent variables which may include intercept, β is a p-vector of unknown regression coefficients, e_i is the random errors with mean equals to zero and constant variance. By taking the expected value of Eq. (2) would result in,

$$\hat{y}_i = X_i^\intercal \hat{\beta}, \quad i = 1, 2, \ldots, n, \tag{3}$$

where $\hat{\beta}$ estimates are the best linear unbiased estimates minimizing the objective function of sum squared residuals.

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} e_i^2, \quad i = 1, 2, \ldots, n, \tag{4}$$

where $e_i = y_i - X_i^\intercal \hat{\beta}$. Suppose that the data set is having a leverage point in the x-direction and let the e_i terms follow the distribution of contaminated model Eq. (1). In this case, LS is not a practical choice and robust methods are recommended. One of the familiar robust methods is M-estimate which is resistant to outliers but it is sensitive to leverage point with zero breakdown. Our proposed algorithm takes into account weighted M-estimate (WM-estimate) to increase the breakdown point. WM-estimate can be described into steps:

Step 1. Weighted the design matrix X. In this step, the MCD (Rousseeuw and Van Driessen, 1999) location and scatter estimators have to computed and then calculation the vector of Robust Mahalanobis Distance RMD^2 as follows,

$$RMD^2 = (X - \hat{\mu}_{\mathrm{MCD}})^\intercal C_{\mathrm{MCD}}^{-1} (X - \hat{\mu}_{\mathrm{MCD}}). \tag{5}$$

It is obvious that, when the i^{th} observation is leverage point the i^{th} RMD^2 would be large value. So assigning low weight for leverage point requires inversely proportional of the i^{th} RMD^2 with the clean subset (Giloni et al.; 2006a, Uraibi; 2019), we adopt Giloni et al. wieghted function,

$$\omega_i = \min\left\{1, \frac{\chi^2_{(0.05,p)}}{RMD_i^2}\right\} \tag{6}$$

and the new weighted design matrix can be written as $x_\omega = \omega.x$, and the estimates of LS with x_ω can be formulize as follows,

$$\begin{aligned} \hat{\beta}_\omega &= \left(X_\omega^\intercal X_\omega\right)^{-1} X_\omega^\intercal y, \\ \hat{y}_\omega &= X_\omega^T \hat{\beta}_\omega, \\ \hat{e}_\omega &= y - X_\omega^\intercal \hat{\beta}_\omega, \end{aligned} \tag{7}$$

Step 2. Iteratively reweighted least squares (IRLS) for (x_ω, y). The $\hat{\beta}^M$ estimates are obtained by minimizing an objective function ρ that can be expressed as

$$\hat{\beta}^M = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} \rho(\hat{e}_{\omega i}), \quad i = 1, 2, \ldots, n, \tag{8}$$

where ρ is a symmetric function with a unique minimum at zero. Taking partial derivative with respect to β and setting them equal to zero, producing a system of normal equations that can solve this minimization problem. Thus, by letting $\psi = \rho'$, we would get

$$\sum \psi(\hat{e}_\omega i) X_{\omega i} = 0. \tag{9}$$

Several choices of ρ and ψ functions are available, in this paper we used bisquare functions (Tukey, 1964) as follows,

$$\rho(\hat{e}_{\omega i}) = \begin{cases} \left\{ 1 - \left[1 - \left(\dfrac{\hat{e}_{\omega i}}{k} \right)^2 \right]^3 \right\} & \text{if } |\hat{e}_{\omega i}| \leqslant k, \\ 1 & \text{if } |\hat{e}_{\omega i}| > k, \end{cases} \tag{10}$$

$$\psi(\hat{e}_{\omega i}) = \hat{e}_{\omega i} \left[1 - \left(\frac{\hat{e}_{\omega i}}{k} \right)^2 \right]^2 \quad \text{if } I(|\hat{e}_{\omega i}|) \leqslant k, \tag{11}$$

$$w(\hat{e}_{\omega i}) = \frac{\psi(\hat{e}_{\omega i})}{\hat{e}_{\omega i}} = \begin{cases} \left[1 - \left(\dfrac{\hat{e}_{\omega i}}{k} \right)^2 \right]^2 & \text{if } |\hat{e}_{\omega i}| \leqslant k, \\ 0 & \text{if } |\hat{e}_{\omega i}| > k, \end{cases} \tag{12}$$

where $I(\cdot)$ stands for indicator function, that is

$$I(\hat{e}_{\omega i}) = \begin{cases} 1 & \text{if } \hat{e}_{\omega i} > 0, \\ 0 & \text{if } \hat{e}_{\omega i} < 0. \end{cases}$$

Consequently, the estimation equation maybe written as,

$$\sum w(\hat{e}_{\omega i}) X_{\omega i} = 0. \tag{13}$$

Theses estimating equations require minimizing $\sum w_i^2 (\hat{e}_{\omega i})^2$ by using iteratively reweighted least squares (IRLS),

$$\hat{\beta}_\omega^{M(j)} = \left(X_\omega^\intercal w^{(j-1)} X_\omega \right)^{-1} X_\omega^\intercal w^{(j-1)} y. \tag{14}$$

In IRLS, the initial fit is calculated, and then a new set of weights is calculated based on the results of the initial fit. The iterations are continued until a convergence criterion is met.

3. The WFBP.WM method

$\hat{\beta}_\omega^{M*}$, Eq. (14) is robust against outliers and leverage points, that is not the residuals of WM are free from outliers, where "*" refers to j^{th} iteration in which the convergence criterion is met. That is because the residuals are the differences between y and $\hat{y}_\omega^* = X_\omega \hat{\beta}_\omega^{M*}$, $\hat{e}_\omega = y - \hat{y}_\omega^*$, as a result $\hat{e}_\omega$ is having outliers due to y is already having outliers based our assumption in above. To protect the whole bootstrapping procedure, the WBP method suggested to weighted bootstrap with probability. Thus the i^{th} residual will get the selection probability of $P_i = \frac{w_i^{(*)}}{\sum_{i=1}^n w_i^{(*)}}$. In this case, these probabilities would control the mechanism of sampling with replacement whereby the residual outliers are ascribed less

importance to than the clean ones (Ramli, 2009). In the other word, only $\hat{e}_{\omega}^{(-D)}$ subject to bootstrap procedure, where D is the number of residuals that poses zero weights according to Eq. (12) and zero probability. This procedure will allow us to use WLS with bootstrap samples to get robust regression coefficients and computing the standard errors. The confidence intervals and the hypotheses test of WBP regression coefficients algorithm can be summarized as follows,

1. Let the observed t values of robust regression coefficients are computed using MM-estimator (Yohai, 1987) such that,
$$\hat{\tau} = \frac{\hat{\beta}^{\mathrm{MM}} - \beta}{\mathrm{SD}(\hat{\beta}^{\mathrm{MM}})}, \tag{15}$$
where β is the regression population parameter, $\hat{\beta}^{\mathrm{MM}}$ is robust estimated regression coefficients and $\mathrm{SD}(\hat{\beta}^{MM})$ the standard deviation of it.
2. Calculate the residuals of WM method $\hat{e}_{\omega}$.
3. Sampling n residual of bootstrap sample $e_{\omega}^{(b)}$ with probability from $\hat{e}_{\omega}$ and then attach it to $\hat{y}_{\omega}^{*}$, $\hat{y}_{\omega}^{(b)} = \hat{y}_{\omega}^{*} + e_{\omega}^{(b)}$.
4. Regress the bootstrapped values of $\hat{y}_{\omega}^{(b)}$ on the fixed X_{ω} to get $\hat{\beta}_{\omega}^{(b)} = \left(X_{\omega}^{\intercal} X_{\omega}\right)^{-1} X_{\omega}^{\intercal} \hat{y}_{\omega}^{(b)}$.
5. Due to the t statistic is pivotal quantity under normality assumption a centered and standardized bootstrap version of $\hat{\beta}_{\omega}^{(b)}$ which is denoted as $t^{(b)}$ is asymptotical pivotal under the same assumption, where
$$t^{(b)} = \frac{\hat{\beta}_{\omega}^{(b)} - \beta}{\mathrm{SD}(\hat{\beta}_{\omega}^{(b)})}. \tag{16}$$
6. Repeat the steps (3–5), B times.
7. The bootstrap percentile intervals can be used the empirical quantity $t^{(b)}$ to form the confidence intervals for β; such that $t_{(1)}^{(b)} < t_{(2)}^{(b)} < \ldots < t_{(B)}^{(b)}$ are the order of WBP replicqations of the $t^{(b)}$ statistic. The lower and upper confidence interval for β can be computed as follows,
$$\hat{\beta}^{\mathrm{MM}} - t^{(b)}_{\left[(B+1)\frac{\alpha}{2}\right]} \mathrm{SD}\big(\hat{\beta}_{\omega}^{(b)}\big) \leqslant \beta \leqslant \hat{\beta}^{\mathrm{MM}} + t^{(b)}_{\left[(B+1)\frac{\alpha}{2}\right]} \mathrm{SD}\big(\hat{\beta}_{\omega}^{(b)}\big), \tag{17}$$
where the square brackets indicate rounding to the nearest integer.
The equal-tail bootstrap p-value is another test that can perform two tests on the same time, against values in the lower and upper tail of the distribution, respectively, as follows,
$$p(\hat{\tau}_j) = 2 \min\left(\frac{1}{B}\sum_{j}^{B} I\big(t_j^{(b)} \leqslant \hat{\tau}_j\big), \frac{1}{B}\sum_{j}^{B} I\big(t_j^{(b)} > \hat{\tau}_j\big)\right). \tag{18}$$

4. The modified market value of Iraq's trade banks

The data are collected from the official website of the Iraqi Stock Market for nine local trade banks which are the most traded than others for the period $(2011 - 2015)$. The researchers are considered six (Trading Rate (X1), Earning per share (EPS) (X2), share turnover ratio (X3), Annual Average price $(X4)$, the Assets (X5), and Undistributed earnings (X6). We modified this data by replacing the 5^{th} observation of $X1$ and 15^{th} of $X2$ with random observations that have been generated from $\chi^2_{(0.05,50)}$ distribution to contaminate both variables by leverage points. The y_{30} and y_{45} observations are replaced with two values from $\chi^2_{(0.05,50)}$ distribution too, to contaminate the response variable, where y is the banks market value that we expect it is affected with these variables according to the multiple linear regression model that can be described as follows:
$$y = X_{(45)}\beta_{(7\times 1)} + e_{(45\times 1)}. \tag{19}$$

Table 1 summarizes the results of the WFBP.M method, as we note that the percentile confidence intervals (L.C.I, U.C.I) and P.value have agreed, that the two variables of assets and the annual price

rate are the most significant than others in determining the banks market value in the Iraqi Stock Exchange. While the results of the WFBP.WM method that is presented in Table 2 find five out of six variables that are most significant in determining the banks market value were identified since the trading Rate variable (X1) is excluded from that significance.

Table 1. The estimate of WFBP.M method for the modified market value of Iraq's trade Banks data.

Variable	Obs.t	$\hat{\beta}^{MM}$	L.C.I.	U.C.I.	Sig	*P*.value	Sig
Intercept	−11.383	−0.232	−0.452	0.075		0.000	**
X1	2.786	0.033	−0.158	0.517		0.483	
X2	−5.913	0.105	−0.591	0.366		0.292	
X3	−0.623	−0.018	−0.202	0.324		0.268	
X4	10.607	0.190	0.071	0.720	**	0.090	**
X5	7.991	0.402	0.004	0.790	**	0.000	**
X6	7.430	0.061	−0.028	0.928		0.12	

On the other hand, Table 2 depicts that perfect similarity between the tests of C.I. and p-values, since only one case can be observed that the value zero value lies outside of C.I. of X1 which its *P*.value of regression coefficient is greater than 0.05, consequently, both tests reject the null hypothesis about the $\hat{\beta}^{\mathrm{MM}}$.

Table 2. The estimate of WFBP.WM method for the modified market value of Iraq's trade Banks data.

Variable	Obs.t	$\hat{\beta}^{\mathrm{MM}}$	L.C.I.	U.C.I.	Sig	*P*.value	Sig
Intercept	−11.383	−0.092	−0.374	−0.069	**	0.011	**
X1	2.786	0.024	−0.058	0.333		0.310	
X2	−5.913	−0.331	0.450	0.105	**	0.000	**
X3	−0.623	−0.262	0.125	0.181	**	0.000	**
X4	10.607	0.471	0.167	0.543	**	0.000	**
X5	7.991	0.802	0.113	0.574	**	0.000	**
X6	7.430	0.209	0.113	0.667	**	0.010	**

It is noticed in Table 2 that three of the regression coefficients are negative, which are the intercept, the share turnover ratio, and EPS. It indicates that these variables have inverse relationships with the market value. In other words, the higher the market value, for instance, results in the lower of both variables (the share turnover ratio and EPS), and vice versa. As for the rest of the variables, they maintained a positive relationship with the market value of the banks.

5. Simulation

The simulation studies have been done to know the performance of WFBP.WM algorithm compared with WFBP.M. The WBP.M algorithm is similar to WFBP.WM except WBP combined with M-estimate. Another comparison will be done inside WFBP.WM algorithm to know with which test this algorithm will be stable. The design matrix $X_{(n\times 6)}$ of the six independent variables is generated randomly from multivariate normal distribution with zero means and $\rho^{|i-j|}$ variance and covariance matrix, $\rho = 0.20$. The maximum value of X_1 is replaced by value is generated from $\chi^2_{(0.05,50)}$ to create the high leverage point, and $m = \alpha \times n$ observations of X_4 are contaminated by using the previous contamination mechanism to create another leverage points, where α the percentage of outlying observation. The first m of random errors $e_{(m\times 1)}$ vector are generated from chi-square distribution with 50 degree of freedom and the remaining $e_{[(n-m)\times 1]}$ are generated from random normal distribution $N(0, 2)$. Suppose that population regression coefficients which is denoted asis known, and $\beta_{(7\times 1)} = (1, 1, 1, 1, 0, 0, 0)$, the response variable $y_{(n\times 1)}$ can be computed as follows,

$$y = X_{(n\times 7)}\beta_{(7\times 1)} + e_{(n\times 1)}. \tag{20}$$

Table 3. The simulation result of WFBP.M method, where $n = 45$, $\alpha = 0.05$.

Variable	Obs.t	$\hat{\beta}^{MM}$	Ave.L.C.I.	Ave.U.C.I.	Sig	$\bar{P}$ value	Sig
Intercept	2.136	51.81	−10.334	25.729		0.393	
X1	1.813	−2049.26	−11.126	27.345		0.320	
X2	1.820	108.79	−10.485	26.209		0.560	
X3	1.236	-75.75	−9.573	23.715		0.461	
X4	0.009	8.29	−10.513	22.920		0.489	
X5	0.156	187.12	−11.243	24.649		0.489	
X6	0.669	77.32	−9.424	20.590		0.488	

Table 4. The simulation result of WFBP.WM method, where $n = 45$, $\alpha = 0.05$.

Variable	Obs.t	$\hat{\beta}^{MM}$	Ave.L.C.I.	Ave.U.C.I.	Sig	$\bar{P}$ value	Sig
Intercept	2.136	1.020	−0.349	4.484		0.024	**
X1	1.813	1.068	−0.414	4.660		0.032	**
X2	1.820	1.080	−0.369	4.548		0.035	**
X3	1.236	1.147	−0.248	4.211		0.033	**
X4	0.009	0.001	−1.256	3.222		0.261	
X5	0.156	−0.003	−1.345	3.487		0.279	
X6	0.669	0.039	−1.127	2.898		0.276	

This simulation scenario has been considered when α equals 0.05 for $n = 45, 65, 85, 100$, where n is the samples. This simulation study is designed to be having three non-zero coefficients and three zero coefficients. The best method is the one that diagnostic the correct significant and non-significant coefficients and is more stable than others. For this purpose, both methods were used to get the results of 1000 datasets that each one replicated 200 times. The average of WFBPP coefficients $\hat{\beta}^{\mathrm{MM}}$, an average of WFBP Lower and Upper bounds of Confidence Intervals Ave.L.C.I. and Ave.U.C.I., respectively, and the average of $p(\hat{\tau}_j)$ is denoted as $\bar{P}$.value are computed for both methods' overall datasets. The decision-making about rejects or accepts the null hypothesis that assumes the regression coefficients are equal to zero would be for two statistics, percentile confidence intervals, and p-values. If zero lies outside the interval between Ave.L.C.I. and Ave.U.C.I., the null hypothesis has to reject, and when the ($\bar{P}$.value< 0.05), the null hypothesis should reject too. When the alternative hypothesis of specific regression coefficient is accepted that means it is different from the zero. The method will recognize the significant coefficient is the result by two stars (**). The best method is the one that diagnostic the correct significant and non-significant coefficients. In another hand, the best test is that one is more stable than the other.

Table 5. The simulation result of WFBP.M method, where $n = 65$, $\alpha = 0.05$.

Variable	Obs.t	$\hat{\beta}^{MM}$	Ave.L.C.I.	Ave.U.C.I.	Sig	$\bar{P}$ value	Sig
Intercept	2.457	7.07	−6.113	17.633		0.393	
X1	2.798	35.82	−6.639	18.541		0.297	
X2	1.597	−4.74	−6.285	17.701		0.576	
X3	2.226	6.76	−5.617	16.440		0.410	
X4	0.427	−10.96	−6.637	15.454		0.480	
X5	−0.296	−11.38	−7.097	16.670		0.457	
X6	−0.174	4.95	−5.951	13.786		0.501	

The results in tables (3 and 4) show the simulation result when $n = 45$ was contaminated by 0.05 outliers and leverage points. It is clear that percentile confidence intervals statistic of WFBR.M and WFBR.WM could not diagnose the significant variables, so the null hypotheses of both methods are accepted. It is notable that zero lies in the interval (Ave. L.C.I. and Ave. U.C.I.) of both methods. The

Table 6. The simulation result of WFBP.WM method, where $n = 65$, $\alpha = 0.05$.

Variable	Obs.t	$\hat{\beta}^{MM}$	Ave.L.C.I.	Ave.U.C.I.	Sig	$\bar{P}$ value	Sig
Intercept	2.457	1.048	0.131	3.710	**	0.005	**
X1	2.798	1.084	0.072	3.827	**	0.013	**
X2	1.597	1.076	0.104	3.720	**	0.014	**
X3	2.226	1.095	0.175	3.494	**	0.010	**
X4	0.427	0.011	−0.846	2.479		0.267	
X5	−0.296	0.001	−0.883	2.694		0.264	
X6	−0.174	0.037	−0.736	2.236		0.298	

$\bar{P}$.value statistic of WFBR.WM method appears in table (2) the $X1$, $X2$, $X3$ variables are significant, but all variables are non-significant in Table 3 which is the result of the WFBR.M method.

Table 7. The simulation result of WFBP.M method, where $n = 85$, $\alpha = 0.05$.

Variable	Obs.t	$\hat{\beta}^{\mathrm{MM}}$	Ave.L.C.I.	Ave.U.C.I.	Sig	$\bar{P}$ value	Sig
Intercept	3.961	5.73	−3.520	12.982		0.300	
X1	3.794	36.30	−3.696	13.478		0.221	
X2	2.254	−7.07	−3.621	13.064		0.451	
X3	2.992	2.16	−3.174	11.892		0.318	
X4	0.854	3.61	−4.132	10.939		0.422	
X5	0.210	1.26	−4.502	11.820		0.431	
X6	−0.278	−4.85	−3.742	9.957		0.411	

Table 8. The simulation result of WFBP.WM method, where $n = 85$, $\alpha = 0.05$.

Variable	Obs.t	$\hat{\beta}^{\mathrm{MM}}$	Ave.L.C.I.	Ave.U.C.I.	Sig	$\bar{P}$ value	Sig
Intercept	3.961	1.031	0.329	3.316	**	0.002	**
X1	3.794	1.093	0.309	3.395	**	0.003	**
X2	2.254	1.039	0.282	3.296	**	0.003	**
X3	2.992	1.067	0.356	3.081	**	0.002	**
X4	0.854	0.013	−0.631	2.105		0.273	
X5	0.210	−0.018	−0.693	2.267		0.258	
X6	−0.278	0.028	−0.563	1.917		0.303	

Increasing the sample size to (65) in Tables 5 and 6 has not changed the performance of the WFBR.M method as the result of Table 5 is shown but in Table 6 the Ave.L.C.I. of WFBR.WM method of regression coefficients of $X1$, $X2$, $X3$ variables is greater than zero and less than zero for $X4, X5, X6$ variables. The WFBR.WM method has kept its high performance and showed more stability than the WFBP.M method when the $n = 85, 100$ according to the results that are displayed in Tables 7–10.

Table 9. The simulation result of WFBP.M method, where $n = 100$, $\alpha = 0.05$.

Variable	Obs.t	$\hat{\beta}^{\mathrm{MM}}$	Ave.L.C.I.	Ave.U.C.I.	Sig	$\bar{P}$ value	Sig
Intercept	3.340	1.38	−1.716	11.141		0.194	
X1	2.741	64.70	−1.911	11.559		0.127	
X2	2.869	−17.16	−1.793	11.312		0.330	
X3	5.028	4.70	−1.520	10.317		0.209	
X4	−0.359	−3.65	−2.504	9.291		0.361	
X5	−1.134	−6.26	−2.685	10.131		0.375	
X6	−0.963	0.06	−2.301	8.501		0.395	

Table 10. The simulation result of WFBP.WM method, where $n = 100$, $\alpha = 0.05$.

Variable	Obs.t	$\hat{\beta}^{\mathrm{MM}}$	Ave.L.C.I.	Ave.U.C.I.	Sig	$\bar{P}$ value	Sig
Intercept	3.340	1.010	0.392	3.127	**	0.001	**
X1	2.741	1.053	0.356	3.191	**	0.001	**
X2	2.869	1.087	0.387	3.170	**	0.001	**
X3	5.028	1.087	0.451	2.968	**	0.000	**
X4	−0.359	0.003	−0.564	1.952		0.275	
X5	−1.134	−0.004	−0.615	2.123		0.252	
X6	−0.963	−0.010	−0.524	1.779		0.289	

6. Conclusion

This paper suggests the weighted M-Huber method to tackle the problem of leverage points present in the design matrix X. Due to the M-Huber being resistant to outliers in regression residuals, the WM-Huber is resistant to outliers and leverage points. Furthermore, based on the results of real data and simulation, the evidence points are almost exclusive to the high performance and robustness of the WM-Huber. At the same time, the M-Huber method has been unreliable when the high leverage points are present in the dataset. These findings encourage us to recommend using the WM-Huber in scientific applications.

[1] Amado C., Pires A. M. Robust bootstrap with non random weights based on the influence function. Communications in Statistics – Simulation and Computation. **33** (2), 377–396 (2004).

[2] Athreya K., Hinkley D. V. Bootstrap of the mean in the infinite variance case. Annals of Statistics. **15** (2), 724–731 (1987).

[3] Croux C., Filzmoser P., Pison G., Rousseeuw P. J. Fitting multiplicative models by robust alternating regressions. Statistics and Computing volume. **13**, 23–36 (2003).

[4] Efron B. Bootstrap methods: another look at the jackknife. In Breakthroughs in Statistics. 569–593 (1992).

[5] Giloni A., Simonoff J. S., Sengupta B. Robust weighted LAD regression. Computational Statistics & Data Analysis. **50** (11), 3124–3140 (2006).

[6] Huber P. J. The place of the L1-norm in robust estimation. Computational Statistics & Data Analysis. **5** (4), 255–262 (1987).

[7] Huber P. J., Ronchetti E. M. Robust statistics. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, Inc. (1981).

[8] Yohai J. V., Maronna A. R. Location estimators based on linear combinations of modified order statistics. Communications in Statistics – Theory and Methods. **5** (5), 481–486 (1976).

[9] Koenker R. Quantile regression. Cambridge University Press (2005).

[10] Midi H., Uraibi H. S., Talib B. A. Dynamic robust bootstrap method based on LTS estimators. European Journal of Scientific Research. **32** (3), 277–287 (2009).

[11] Habshah M., Norazan M. R., Rahmatullah Imon A. H. M. The performance of diagnostic-robust generaliozed potential approach for the identification of multiple high leverage points in linear regression. Journal of Applied Statistics. **36** (5), 507–520 (2009).

[12] Rousseeuw P. J. Least median of squares regression. Journal of the American Statistical Association. **79** (388), 871–880 (1984).

[13] Rousseeuw P. J., Leroy A. M. Robust Regression and Outlier Detection. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, Inc. (1987).

[14] Shao J. Bootstrap estimation of the asymptotic variances of statistical functionals. Annals of the Institute of Statistical Mathematics. **42**, 737–752 (1990).

[15] Shao J. Bootstrap variance estimators with truncation. Statistics & Probability Letters. **15** (2), 95–101 (1992).

[16] Singh K. Breakdown theory for bootstrap quantiles. Annals of Statistics. **26** (5), 1719–1732 (1998).

[17] Stromberg A. J. Robust covariance estimates based on resampling. Journal of Statistical Planning and Inference. **57** (2), 321–334 (1997).

[18] Uraibi H. S. Weighted Lasso Subsampling for High Dimensional Regression. Electronic Journal of Applied Statistical Analysis. **12** (1), 69–84 (2019).

[19] Uraibi H. S., Midi H. On Robust Bivariate and Multivariate Correlation Coefficient. Economic computation and economic cybernetics studies and research. **53** (2/2019), 221–239 (2019).

[20] Uraibi H. S., Midi H., Rana S. Robust stability best subset selection for autocorrelated data based on robust location and dispersion estimator. Journal of Probability and Statistics. **2015**, Article ID 432986, 8 pages (2015).

[21] Uraibi H. S., Midi H., Rana S. Robust multivariate least angle regression. ScienceAsia. **43** (1), 56–60 (2017).

[22] Uraibi H. S., Midi H., Talib B. A., Yousif J. H. Linear regression model selection based on robust bootstrapping technique. American Journal of Applied Sciences. **6** (6), 1191–1198 (2009).

[23] Willems G. S., Aelst S. Fast and Robust Bootstrap for LTS. Computational Statistics & Data Analysis. **48** (4), 703–715 (2005).

Робастне бутстрап регресійне тестування за наявності викидів

Хассан С. У., Алі К. Х.

Кафедра статистики, Університет Аль-Кадисія, вул. Альямея, Діванья, Ірак

Бутстрап — це один із методів вибору випадкової вибірки із заміною, який був запропонований для вирішення проблеми малих вибірок, розподіли яких важко отримати. Розподіл бутстрап-вибірок є емпіричним або вільним, і завдяки його випадковому відбору із заміною ймовірність вибору конкретного спостереження може дорівнювати одиниці. На жаль, коли вихідні дані вибірки містять викиди, виникає серйозна проблема, яка призводить до некоректності оцінки за допомогою звичайних найменших квадратів, тому слід рекомендувати робастні методи регресії. Добре відомо, що найкраща робастна регресійна модель має високу точку пробою не більше ніж 0.50, тому робастний регресійний метод не буде працювати, якщо відсоток викидів у вибірці перевищує 0.50. Добре відомо, що бутстрап-процес з фіксованим x робить перевибірку залишків, які, ймовірно, мають викиди. Більше того, точка(и) важеля є викидом, який виникає в X-напрямку, тому буде існувати його вплив на бутстрап-вибірки з фіксованим x. Тому прийняття рішення щодо нульової гіпотези коефіцієнтів бутстрап-регресії не може бути надійним. У цій статті пропонується використовувати зважений бутстрап із фіксованим x із ймовірнісним підходом, щоб гарантувати, що відсоток викидів у бутстрап-вибірках буде дуже низьким. А потім зважена M-оцінка повинна бути спрямована на розв'язання проблеми викидів і важливих точок та прийняття більш надійного рішення щодо перевірки гіпотези про коефіцієнти бутстрап-регресії. Ефективність запропонованого методу була порівняна з іншими методами на реальних та змодельованих даних. Результати показують, що запропонований нами метод є ефективнішим та надійнішим за інші.

Ключові слова: *бутстрап, надійна регресія, довірчі інтервали, точка, зважений бутстрап з ймовірністю, зважене M, перевірка гіпотези.*