Journal of Mathematics and Statistics 10 (3): 414-420, 2014
ISSN: 1549-3644

doi:10.3844/jmssp.2014.414.420 Published Online 10 (3) 2014 (http://www.thescipub.com/jmss.toc)

MODEL SELECTION VIA ROBUST VERSION OF R-SQUARED

Shokrya Saleh

Institute of Mathematical Sciences, University of Malaya, 50603 Kuala Lumpur, Malaysia

Received 2014-06-05; Revised 2014-07-08; Accepted 2014-09-17

ABSTRACT

R-squared (R^2) is a popular method for variable selection in linear regression models. R^2 based on Least Squares (LS) regression minimizes the sum of the squared residuals; *LS* is sensitive to outlier observation. Alternative criterion based on *M*-estimators, which is less sensitive to outlying observation has been proposed. In this study explicit expression for such criterion is obtained when the Least Trimmed Squares (LTS) estimator is used. The influence function of R^2 is also discussed. In our simulation study, the performance of proposed criterion is compared to the existing criteria based on *M*-estimators (R^2_M) and to the classical non-robust based on least squares estimators (R^2_{LS}). We observe that the proposed (R^2_{LTS}) selects more appropriate models in the case of bad leverage points (outliers in the *X*-direction) are present.

Keywords: Robust R^2-Coefficents, Least Trimmed Squares, Influence Function

1. INTRODUCTION

The use of squared multiple correlation coefficient, R^2, in choosing model is a common goal in econometrics analysis; it is a classical model selection criterion which has been widely used for centuries and it is still popular till today. Hahn (1973), Kvålseth (1985), Willett and Singer (1988) as well as Anderson-Sprecher (1994) have expounded on R^2. Consider a multiple linear regression model:

$$y_i = \alpha + X^T\beta + \varepsilon_i \quad (1)$$

where, $X = (x_{i1},...x_{ip})^T$ is a vector containing p explanatory variables, $i = 1,...,n$,y_i is the response variable, β is a vector of p parameters, a is the intercept parameter and ε_i is an independently and identically distributed (iid) random error with mean 0 and variance σ^2. The distribution of errors satisfying $F\sigma(x) = F_0(x/\sigma)$, where σ is the residual scale parameter and F_0 is symmetric with a strictly positive density function. With $SSE = \sum_{i=1}^{n}(r_i)^2$, where $r_i = y_i - \hat{\alpha}_{LS} + X^T\hat{\beta}_{LS}$, the residual from the Least Squares (LS) fit, the classical R^2 coefficient is given by:

$$R^2_{LS} = 1 - \frac{SSE}{SST} \quad (2)$$

where, $SST = \sum_{i=1}^{n}\left(y_i - \bar{y}_i\right)^2$ with $\bar{y}_i$ is the sample mean of the dependent variable. Selecting models on the basis of maximizing R^2_{LS} is equivalent to minimizing the residual mean square.

Note that the numerator in Equation 2 approximates the scale of the residuals in the full model, while the denominator is the scale of the residuals in the following reduced model:

$$y_i = \alpha_0 + \varepsilon_i \quad (3)$$

In fact the *LS* estimator of α_0 in Equation 3 equals $\bar{y}_i$. Then analogous to Equation 2 is:

$$R^2_{LS} = 1 - \frac{\text{var}(full)}{\text{var}(reduced)} \quad (4)$$

where, *var* is defined according to principles guide estimation:

We recognized that the multiple correlation coefficient R^2_{LS} statistics need special care in contaminated data. Since the classical procedure is *LS* fit-based, alternatives have been developed in the literature. For example McKean and Sievers (1987) have proposed a robust version of R^2, with respect to L_1regression estimator and the associated $R^2_{L_1}$ is given by:

$$1-\left(\sum_n^{i=1}\left|y_i - X^T\hat{\beta}_{L_1} - \hat{\alpha}_{L_1}\right| \Big/ \sum_n^{i=1}\left|y_i - median_i y_i\right|\right)$$

Leroy and Rousseeuw (1987) have proposed a robust R^2 with respect to the overall measure of variation $med_i\left(y_i-\hat{y}_i\right)^2$ and this measure used in Equation 4 led directly to an analog of R^2, $1-\left[med_i\left(y_i-\hat{y}_i\right)^2 \Big/ med_i\left(y_i-C\right)^2\right]$ where C is a constant that minimizes $med_i\ (y_i\text{-}C)^2$. Croux and Dehon (2003) studied local robustness and confidence intervals of multiple coefficients which is based on *M*-estimators. In this study we explore whether improvements can occur if we apply a high breakdown scale estimate such as Leasttrimmed Squared (LTS) estimator proposed by Yohai (1987). $\hat{\beta}_{LTS}$ is computed by minimizing the H smallest squared residuals, defined as Equation 5:

$$\sum_{i=1}^{H} r_i^2(\beta) \tag{5}$$

Based on the ordered absolute residuals $\left|r_{(1)}\right| \le \ldots \le \left|r_{(n)}\right|$ *LTS* converages at the rate of $n^{(1/2)}$ with the same asymptotic efficiency under normality.

The paper is organized as follows: In section 2 the influence of outlier R^2 is illustrated through generated sample data set. In section 3 we have introduced an alternative robust version of R^2 based on *LTS* estimators. The influence function of R^2 is discussed in section 4.

2. EXPERIMENTAL

In this section, we set the idea of influence of outliers on R^2 through the presence of outliers in *y*-direction (called vertical outliers) and in the *X*-direction (called leverage points). We generated independent uniform variable x_i on [-1,1] according to $y_i = x_i+\epsilon_i$, i = 1,…,20, where the ϵ_i are iid normally distributed with expectation 0 and variance (0.1^2). A point with $(0,y_{10})$ is added. A similar approach is done in *X*-direction, by replacing added the value $(x_{10}, 0)$, **Fig. 1 and 2** shows both situations. **Figure 3** shows the effect of adding y_{10} and x_{10} on the values of the classical R^2 and, for comparison, those with the robust version of R^2 based on *M*-estimation developed by Croux and Dehon (2003) and R^2_{LTS} developed in section 3. The value of classical R^2 has high sensitivity in both contamination directions, whilerobust R^2 based on M-estimation decreases only as the size of contamination in *X*-direction increases. On the other hand, R^2_{LTS} are much more robust than others with both vertical and leverage points outliers.

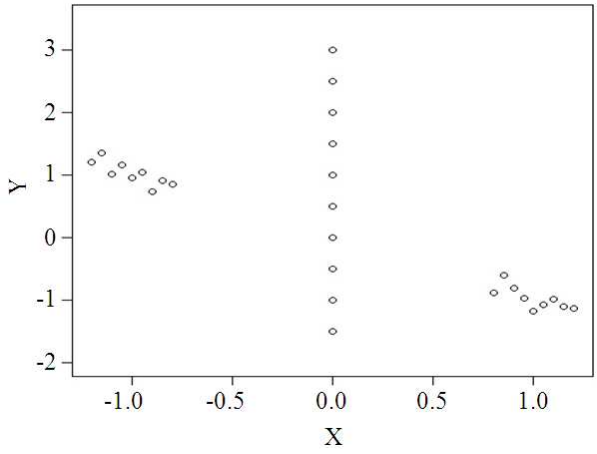

Fig. 1. Data and positions for y_{10} points

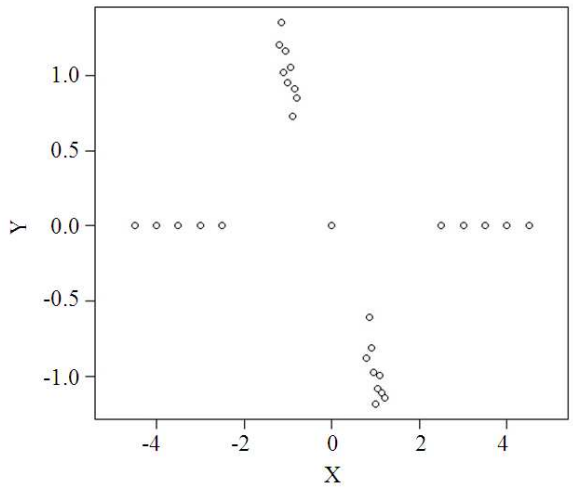

Fig. 2. Data and positions for x_{10} points

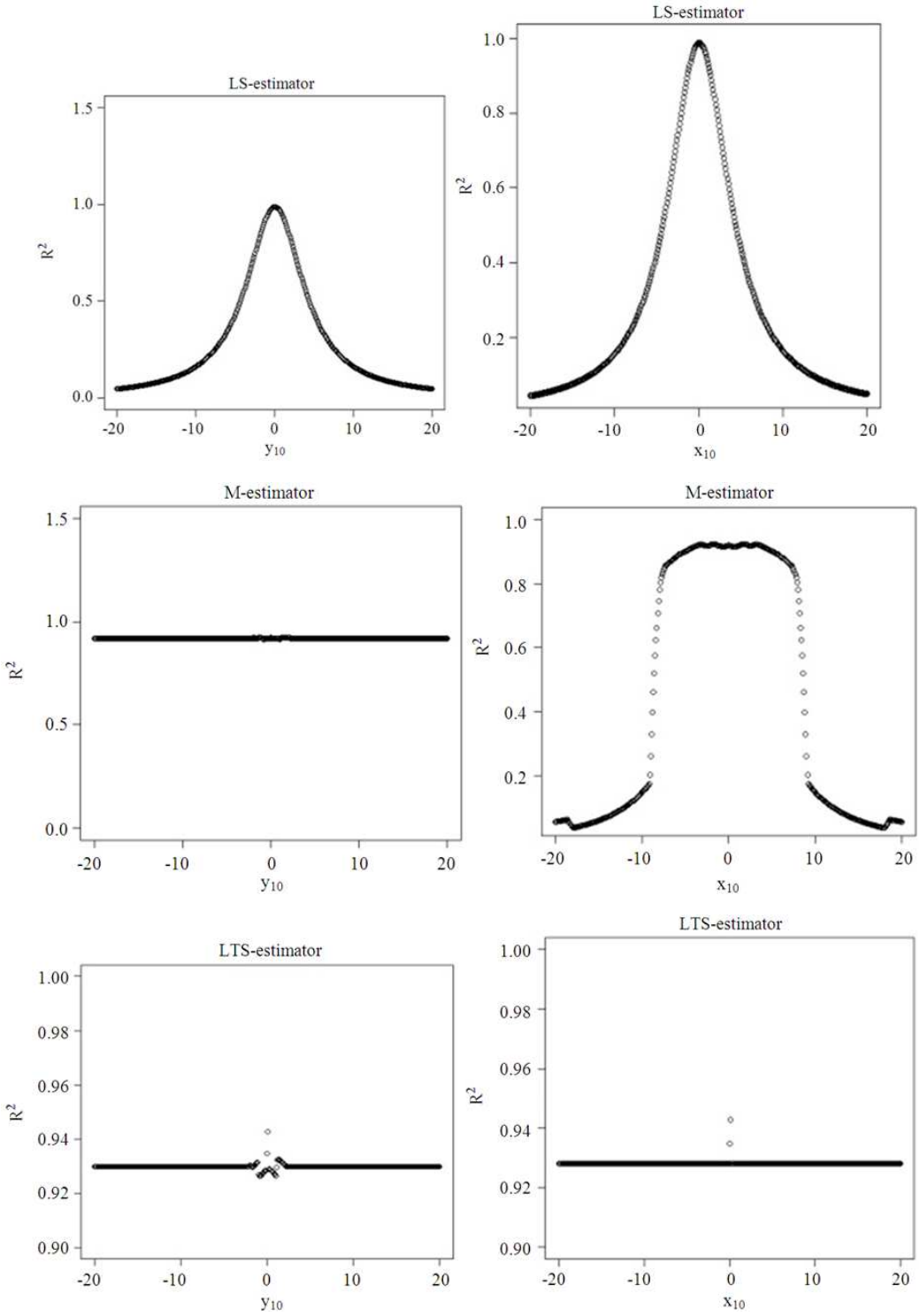

Fig. 3. The different values of R^2 after adding one observation $(0,y_{10})$ (left figures) and effect of adding one observation $(x_{10},0)$ (right figures)

3. ROBUST VERSION OF R^2 BASED ON LTS SCALE ESTIMATE

3.1. LTS -Estimators

An algorithm of obtaining the *LTS* estimator and its scale is as follows:

- Draw a random H subset where $|h|=n+p+1/2$
- Compute $\hat{\beta}_0 = (X_{old}^T X_{old})^{-1} X_{old}^T y_{old}$
- Compute the residual $e_0 = y_i - x_i\hat{\beta}_0$
- Order absolute of the residual, $|r_0|_{1:n} \le \ldots \le |r_0|_{n:n}$
- Choose the new *h* observation
- Compute the $\hat{\beta}_1 = (X_{new}^T X_{new})^{-1} X_{new}^T y_{new}$
- Compute the residual $e_1 = y_i - x_i\hat{\beta}_1$
- Because new *h* correspond to the *h* smallest absolute residual out of *n*, we have:

$$\sum_{i=newh}(e_1(i))^2 \le \sum_{i=oldh}(e_0(i))^2 = Q_0$$

And because the *LS* estimator $\hat{\beta}_1$ of these *h* observations is such that it minimizes Q_1 so we find that $Q_1 \le Q_0$. Repeatthe 7-steps. If $Q_1 = Q_0$, else we stop.

Leroy and Rousseeuw (1987) suggest that might be selected as $h = [n(1-\alpha)]+1$, where α is the level of trimming. As for example, if we trimmed 10% observation then $\alpha = 0.1$ and hence $h = [n(0.9)]+1$. So we can control the level of trimming when we suspect that the data contain nearly 10% outliers and we can increase the level of trimming if we suspect there are more outliers in the data.

3.2. Robust R^2 Based on LTS Estimators

An analogue to the classical formula in Equation 2 is Equation 6:

$$R_S^2 = 1 - \frac{S_1^2(r_i)}{S_0^2(y_i - \alpha_0)} \tag{6}$$

Consider scale estimate of error defined by: $S_1 = \sum_{i=1}^{n}\left(y_i - \hat{\alpha}_{LTS} + X^T\hat{\beta}_{LTS}\right)$. $S_0 = \sum_{i=1}^{n}(y_i - \hat{\alpha}_{LTS})$ we get the following robust R^2 Equation 7:

$$R_{LTS}^2 = 1 - \frac{S_1^2(y_i - \hat{\alpha}_{LTS} + X^T\hat{\beta}_{LTS})}{S_0^2(y_i - \hat{\alpha}_{LTS})} \tag{7}$$

Since Equation 3 is a sub model in Equation 1, we readily see that $0 \le R_S^2 \le 1$. The model with the value of R_{LTS}^2 closed to 1 indicated the preferred model.

4. INFLUENCE FUNCTION

In this section we discuss the influence function of R^2 using any scale functional verifying a certain smoothness condition as available in (Croux and Dehon 2003) with some extension of effect of *LTS* estimator.

Let *X* and *y* be independent stochastic variables with distribution *D*. The functional *T* is Fisher-consistent for the parameters (α,β) at the model distribution *D*, that is Equation 8:

$$T(D) = \begin{pmatrix} a(D) \\ b(D) \end{pmatrix} = \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \tag{8}$$

And, consider allocation model $F_\sigma = F\left(\frac{X}{\sigma}\right)$ here F is called the model distribution. We want our estimator to be Fisher-consistent, which mean $T(F_\sigma) = \sigma$ for all $\sigma>0$. The influence function of the criterion functional for the distribution *F* is given by Equation 9:

$$IF\left((X,y),T,D\right) = \lim_{\varepsilon\to 0}\frac{T\left((1-\varepsilon)D + \varepsilon\Delta_{(X,y)}\right) - T(D)}{\varepsilon} = \frac{\partial}{\partial\varepsilon}\left(T\left(\Delta_{(X,y)}\right)\right) \tag{9}$$

where, *T*(*D*) be the functional defined as the solution of the objective model, $\Delta_{(X,y)}$ is the distribution which (*X*,*y*) contain outliers. The influence function measures the effect of a possible outlier (*X*,*y*) on the R^2 statistic. It represents the magnitude of change in R^2 arising from minute degrees of contamination. For more details about influence function, Hampel *et al*. (2011).

The following theorem is given in Croux and Dehon (2003).

Theorem 1

Let the model distribution *D* verify (*D*). Take (*X*,*y*) ~*D* and denote ε the error term of the model. Assume that *S* has the property that $(a,b)\to S(\varepsilon+b^t X+a)$ is differentiable with partial derivatives equal to zero at the origin (0,0). The Equation 10:

$$IF\left((X,y),R_S^2,D\right)=$$
$$\frac{2\sigma}{\sigma_y^3}\left(IF\left(\frac{y-\mu_y}{\sigma_y},S,F_0\right)-IF\left(r_i/\sigma,S,F_0\right)\right) \quad (10)$$

The proof of Theorem 1.

$$IF\left((x,y),R_S^2,H\right)=\frac{\partial}{\partial\varepsilon}(1-\frac{S_1^2(H_\varepsilon)}{S_0^2(H_\varepsilon)})|_{\varepsilon=0}$$

Where:

$$\begin{aligned}
&R_S^2(H)=1-\frac{S_1^2(H)}{S_0^2(H)}=1-\frac{\sigma}{\sigma_y}\\
&\frac{\partial}{\partial\varepsilon}\left(1-\frac{S_1^2(H_\varepsilon)}{S_0^2(H_\varepsilon)}\right)|_{\varepsilon=0}\\
&\quad -2S_1(H).S_0^2(H)\frac{\partial}{\partial\varepsilon}S_1(H_\varepsilon)|_{\varepsilon=0}\\
&=\frac{+2S_1^2(H).S_0(H)\frac{\partial}{\partial\varepsilon}S_0(H_\varepsilon)|_{\varepsilon=0}}{\left[S_0^2(H)\right]^2}\\
&=\frac{-2\sigma\sigma_y^2\frac{\partial}{\partial\varepsilon}S_1(H_\varepsilon)|_{\varepsilon=0}+2\sigma^2\sigma_y\frac{\partial}{\partial\varepsilon}S_0(H_\varepsilon)|_{\varepsilon=0}}{\sigma_y^4}\\
&=\frac{2\sigma}{\sigma_y^3}\left[\sigma\frac{\partial}{\partial\varepsilon}S_0(H_\varepsilon)|_{\varepsilon=0}-\sigma_y\frac{\partial}{\partial\varepsilon}S_1(H_\varepsilon)|_{\varepsilon=0}\right]
\end{aligned} \quad (11)$$

Where:

$$\begin{aligned}
&\frac{\partial}{\partial\varepsilon}S_1(H_\varepsilon)|_{\varepsilon=0}=IF(r_i,S,F_\sigma)\\
&=\frac{\partial}{\partial\varepsilon}S\left((1-\varepsilon)F_\sigma+\varepsilon\Delta_{r_i}\right)|_{\varepsilon=0}\\
&=\sigma\frac{\partial}{\partial\varepsilon}S\left((1-\varepsilon)F_0+\varepsilon\Delta_{r_i/\sigma}\right)|_{\varepsilon=0}\\
&=\sigma IF\left(r_i/\sigma,S,F_0\right)
\end{aligned} \quad (12)$$

where, F_σ is the error distribution and by using scale equivariance S.

In a similar way, we can show that:

$$\frac{\partial}{\partial\varepsilon}S_0(H_\varepsilon)|_{\varepsilon=0}=\sigma_y IF\left(y-\mu_y/\sigma_y,S,F_0\right) \quad (13)$$

Inserting Equation 12 and 13 into Equation 11 yields:

$$IF\left((x,y),R_S^2,H\right)=\frac{2\sigma}{\sigma_y^3}\left(IF(\frac{y-\mu_y}{\sigma_y},S,F_0)-IF\left(r_i/\sigma,S,F_0\right)\right)$$

It is apparent that the influence function is bounded if the $IF\left(\frac{y-\mu_y}{\sigma_y},S,F_0\right)-IF\left(r_i/\sigma,S,F_0\right)$ function is bounded; and if $S = LS$ then R_{LS}^2 is non-robust, because it has an unbounded influence function (as previously noted by Romanazzi (1992)). Note that as Δy or ΔX increases, the value $\frac{\hat{S}^2(\Delta y-\alpha-\Delta X^T\beta)}{\hat{S}^2(\Delta y-\alpha_0)}$ approaches 1 then the influence function becomes an unbounded (decreasing) negative function. *M*-estimation is reputed to be robust with respect to *y*-space outlier. Note that large zone outliers in *y* have zero influence. In contrast, using the generally accepted expression of the influence function of an *LTS* of scale, the influence function of the R_{LTS}^2 is given by Equation 14:

$$\begin{aligned}
&IF\left((X,y),R_{LTS}^2,H\right)\\
&=\frac{2\sigma}{\sigma_y^3E_{F_0}\left[\dot{\rho}(\varepsilon)\varepsilon\right]}\begin{pmatrix}\max\left(\left(\frac{y-\mu_y}{\sigma_y}\right)^2,c\right)\\-\max\left(\left(r_i/\sigma\right)^2,c\right)\end{pmatrix}
\end{aligned} \quad (14)$$

Note that the influence function of R_{LTS}^2 is bounded and has zero influence in both direction (*X* and *y*) and the relative gross-error sensitivity at Gaussian distributions of *LTS* equals 1.543. In next section, the R_{LTS}^2 performance will be compared to the performance of the other criteria.

5. EXAMPLES

5.1. Example 1 (Simulation Study)

The setting for the simulation study is as follows. We generated 50 independent replicates of three independent uniformly random variables on [-1,1] of -x_{i1},x_{i2} and x_{i3}- and 50 independently normally distributed errors $\varepsilon_i \sim N(0,9)$. Then we define the true model $y_i = x_{i1}+x_{i2}+\varepsilon_i$, for $i = 1,\ldots,50$ using two variables x_{i1} and x_{i2}.

We compare three different *R*-square versions in this simulation study: Classical R^2_{LS} based on *LS* estimation, robust R^2_M based on *M*-estimation and robust R^2_{LTS} based on *LTS* estimation. We trimmed 10% observation then α = 0.1 and hence $h = [n(0.9)]+1$. To compute the robust R^2_M and R^2_{LTS}, we used, respectively, the function rlm(), ltsreg() from the *R* libraries MASS. For investigating how robust the methods perform against outliers, we apply them to three situations:

- Vertical outliers (outliers in y only)
- Good leverage points (outliers in y and *X*)
- Bad leverage points (outliers in some of *X* only)

For case (i) we randomly generated 10% of outliers from $N(50,0.1^2)$. For case (ii) we considered 10% of outliers on the variables x_{i1}, x_{i2} and x_{i3} are generated from $N(100,0.5^2)$ distribution, then generated y to get good leverage points. For case (iii) 10% of outlier on the variables x_{i1} and x_{i2} are generated from a $N(100,0.5^2)$ distrbution. For each of these setting we simulated 1000 samples.

We summarize the simulation results by reporting the percentage of selected models that are:

- Correct fit-the true model only
- Over fit-models containing all the variables in the true model plus some more that are actually redundant
- Under fit-models with only a strict of the variables in true model
- Wrong fit-all models that are not over fit, not a correct fit nor under fit

Table 1 shows detailed simulation results. We see that the classical R^2_{LS} selects a large proportion of over fit models for the data with and without outliers; on the other R^2_ρ ignoring some of the important variables in the model and a higher proportion of under fit are selected in all cases, while work better than others in that it gives the correct model as a good model for the all the cases.

A main message to be learned from this simulation study is thatit seems valid to use *R*-squared based on *LTS* estimation using expression (11) with different level of outliers. This is because we can control the level of trimming; based on the simulation results, R^2_ρ based on *M*-estimation selected a large proportion of under fit.

5.2. Example 2 (Stack Loss Data)

Stack Loss data was presented by Brownlee (1965). This data set consists of 21 observations on 3 independent variables and it contains 4 outliers (cases 1, 3, 4 and 21) and high leverage points (cases 1, 2, 3 and 21). The data are given in **Table 2**.

We applied the classical and robust versions of R^2 methods on the data. **Table 3** shows that the classical method selects the full model, robust R^2_ρ method ignored one of the important variables (x_2) and, robust R^2_{LTS} method agreed on the importance of the two variables x_1 and x_2.

Table 1. Proportion of select models from classical R^2_{LS}, robust R^2_ρ and robust R^2_{LTS} with different 10% outliers

		R^2_{LS}	R^2_ρ	R^2_{LTS}
Zero outliers	Correct fit over fit under fit wrong fit	0.6%	43.1%	59%
		99.4%	56.9%	41%
		0%	0%	0%
		0%	0%	0%
Vertical outliers	Correct fit over fit under fit wrong fit	0.1%	45.3%	68.8%
		99.7%	54.7%	31.2%
		0%	0%	0%
		0.2%	0%	0%
Bad leverage points	Correct fit over fit under fit wrong fit	0%	48.3%	68.8%
		99.9%	51.7%	31.2%
		0%	0%	0%
		0.1%	0%	0%
Good leverage points	Correct fit over fit under fit wrong fit	4.4%	45.6%	56.9%
		95.6%	54.4%	43.1%
		0%	0%	0%
		0%	0%	0%

Table 2. Stack loss data set

x_1	x_2	x_3	y
80	27	89	42
80	27	88	37
75	25	90	37
62	24	87	28
62	22	87	18
62	23	87	18
62	24	93	19
62	24	93	20
58	23	87	15
58	18	80	14
58	18	89	14
58	17	88	13
58	18	82	11
58	19	93	12
50	18	89	8
50	18	86	7
50	19	72	8
50	19	79	8
50	20	80	9
56	20	82	15
70	20	91	15

Table 3. Result ariable selection of Stack Loss data

Selected variables	R^2	R^2_{ρ}	R^2_{LTS}
x_1	0.85	0.9	0.73
x_2	0.77	0.21	0.32
x_3	0.16	0.55	0.32
x_1,x_2	0.90	0.83	0.77
x_1,x_3	0.85	0.79	0.72
x_2,x_3	0.77	0.22	0.34
x_1,x_2,x_3	0.91	0.83	0.76

6. CONCLUSION

Since the classical R^2 shows extreme sensitivity to outliers, we studied the R^2_{LTS} statistic in relation to Least Trimmed Squares (LTS) regression estimator of scale. Through calculating the influence functions of R^2_{LTS} we show that it is bounded and not sensitive to outliers or bad leverage point. We evaluated our method using both simulated and real data sets and compared its performance withthe classical method as well as R^2_M proposed by Croux and Dehon (2003). According to this study, the performance of R^2_{LTS} is better and more stable than the other methods. This study focused on the R^2 variable selection criteria; one might be interested to extend other robust model selection criteria to advanced robust breakdown point estimation methods, such as, Mallows'Cp criterion.

7. ACKNOWLEDGEMENT

I thank Dr. Rossita M. Yunus for helpful comments on an earlier version of this study and the editor, associate editor.

8. REFERENCES

Anderson-Sprecher, R., 1994. Model comparisons and R^2. Am. Stat., 48: 113-117.

Brownlee, K.A., 1965. Statistical Theory and Methodology in Science and Engineering. 2nd Edn., Wiley, New York, pp: 590.

Croux, C. and C. Dehon, 2003. Estimators of the multiple correlation coefficient: Local robustness and confidence intervals. Statistical, 44: 315-334. DOI: 10.1007/s00362-003-0158-7

Hahn, G.J., 1973. Coefficient of determination exposed. Chemische Technik.

Hampel, F.R., E.M. Ronchetti, P.J. Rousseeuw and W.A. Stahel, 2011. Robust Statistics: The Approach Based on Influence Functions. 1st Edn., J. Wiley Sons, New York, ISBN-10: 0471735779, pp: 536.

Kvålseth, T.O., 1985. Cautionary note about R^2. Am. Stat., 4: 279-285.

Leroy, A.M. and P.J. Rousseeuw, 1987. Robust Regression and Outlier Detection. 1st Edn., John Wiley and Sons, New York.

McKean, J.W. and G.L. Sievers, 1987. Coefficients of determination for least absolute deviation analysis. Stat. Probability lett., 5: 49-54. DOI: 10.1016/0167-7152(87)90026-5

Romanazzi, M., 1992. Influence in canonical correlation analysis. Psychometrika, 57: 237-259. DOI: 10.1007/BF02294507

Willett, J.B. and J.D. Singer, 1988. Another cautionary note about R^2: Its use in weighted least-squares regression analysis. Am. Stat., 42: 236-238. DOI: 10.2307/2685031

Yohai, V.J., 1987. High breakdown-point and high efficiency robust estimates for regression. Annals Stat., 15: 642-656. DOI: 10.1214/aos/1176350366