An Application of Ordinary Least Squares and Maximum Likelihood Type Estimation in Roust Diagnostic Regression Analysis

by

Maw Maw Khin [1]

Abstract

This study shows that the OLS method is quite sensitive to outlier whereas maximum likelihood type estimation (*M*-estimation) methods resist outliers. The iterativated reweighted least squares (IRLS) method based on the Huber and the Bisquare ψ -functions clearly detect outliers that are given to less weight. The findings show that maximum likelihood type estimation based on the mean squares error (MSE) criterion can provide predicted values very close to actual values.

Keywords: Robust Regression, Ordinary Least Squares, Iterated Reweighted Least Squares

1. Introduction

Regression analysis is an important tool for any quantitative research. It explores the relationship between dependent and explanatory variables. The ordinary least squares (OLS) method is mostly applied in regression analysis. The application of this method requires a number of assumptions. A researcher should be aware of the fact that the OLS method performs poorly if the underlying assumptions are not fulfilled.

In the last two centuries, various strategies were introduced to test whether the model assumptions are fulfilled or not. Besides, more general regression techniques are available based on less stringent conditions. Until the mid-20th century, violations of the model assumptions were treated independently from any common error source. But, in particular, outlying observations within the data set can cause violations of model assumptions and thereby it can have a huge impact on regression results.

Robust regression analyses have been developed as an improvement to OLS estimation in the presence of outliers and provide information about what a valid observation is and whether this should be thrown out. The primary purpose of robust regression analysis is to fit a model which represents the information in the majority of the data. In this context, robust regression is to employ a fitting criterion that is not as vulnerable as OLS to unusual data. One remedy is to remove influential observations before using the OLS fit.

Robust regression analysis provides an alternative to an OLS regression model when fundamental assumptions are not fulfilled by the nature of data. Sometimes, the variables can be transformed to confirm the assumptions. Often, however, a transformation will not eliminate or satisfy the leverage of influential outliers that bias the prediction and distort the significance of parameter estimates. Under these

[1] Associate Professor, Dr., Department of Statistics, Yangon Institute of Economics

circumstances, to the best of the present researcher's knowledge, robust regression that is resistant to the influence of outliers may be the only reasonable remedy. This paper focuses on the (robustness-) performance of estimators if outliers occur within the data set.

2. Data and Methods

The cross-sectional data are used to study the effect of outlier in regression analysis. The human development indicators of 85 countries were obtained from the Human Development Report (2009) published by the UNDP. In order to assess the predicting performance of the model, the first eighty countries were used for model construction and rest of the five countries was used for validation of the predicted values. In this study, the two methods, OLS and *M*- estimation were used to analyze.

(a) Ordinary least squares (OLS) method

The OLS method computes the parameters $(\hat{\beta}_0,...,\hat{\beta}_p)$ that minimize the sum of squares of residuals. Formally, it can be written as

$$Q = \min_{(\hat{\beta}_0,...,\hat{\beta}_p)} \sum_{i=1}^{n} e_i^2. \quad (1)$$

Q is the sum of the squared vertical deviations from the hyperplane $H = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$. Taking the derivative of (1) with respect to $\beta = (\beta_0, \beta_1,..., \beta_p)$, obtain the normal equations

$$X^T X \hat{\beta} = X^T Y$$

and solving these equations give the least squares estimator of β

$$\hat{\beta} = (X^T X)^{-1} X^T Y. \quad (2)$$

The vector of predicted or fitted values is $\hat{Y} = X\hat{\beta} = HY$ where $H = X(X^T X)^{-1} X^T$ is called the hat matrix. The i th entry of $\hat{Y}$ is the i th fitted value (or predicted value) $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i,1} + \cdots + \hat{\beta}_p X_{i,p} = x_i^T \hat{\beta}$ for observation Y_i while the i th residual is $e_i = Y_i - \hat{Y}_i$. The vector of residuals is $\boldsymbol{e}$ = (**I** −**H**)**Y**.

(b) Maximum likelihood type estimation (M-estimation) method

The *M*-estimation method minimizes the objective function

$$\min \sum_{i=1}^{n} \rho(y_i - x_i' \hat{\beta}) = \min \sum_{i=1}^{n} \rho(e_i) \quad (3)$$

where the function ρ gives the contribution of each residual to the objective function. A reasonable ρ should have the following properties.

$$\rho(e) \geq 0$$
$$\rho(0) = 0$$
$$\rho(e) = \rho(-e)$$
$$\rho(e_i) \geq \rho(e_{i'}) \; for \; |e_i| > |e_{i'}|$$

The solution obtained from Equation (3) is not scale equivariant, and thus the residuals must be standardized by a robust estimate of their scale $\hat{\sigma}$, which is estimated simultaneously.

$$\min \sum_{i=1}^{n} \rho\left(\frac{e_i}{\sigma}\right) \tag{4}$$

As in the case of *M*-estimates of location, the median absolute deviation (MAD) is often used.

$$\hat{\sigma} = 1.4826 \times MAD \tag{5}$$

$$\text{MAD} = \text{median}\,(\,|\,e_i - \text{median}\,\{e_i\}\,|\,)$$

Taking the derivative of Equation (4) and solving produces the score function

$$\sum_{i=1}^{n} \psi\left((y_i - x_i'\hat{\beta})/\hat{\sigma}\right) x_{ik} = \sum_{i=1}^{n} \psi\left(e_i/\hat{\sigma}\right) x_i = 0 \tag{6}$$

with $\psi = \rho'$ which is called the influence function. There is now a system of $p + 1$ equations, for which ψ is replaced by appropriate weights that decrease as the size of the residual increases. Define the weight function $w(e) = \psi(e)/e$, and let $w_i = w(e_i)$. Then the Equations (6) becomes

$$\sum_{i=1}^{n} w_i\left(e_i/\hat{\sigma}\right) x_i = 0 \quad for \quad i = 1,2,...,n.$$

This is exactly the system of equations that can be solved by using the iterated reweighted least squares (IRLS) procedure.

There are various *M*- estimators according to the choice of ψ-functions. In this study *M*-estimator based on Huber ψ-function and *M*-estimator based on Bisquare ψ-function are chosen to apply to the real data.

3. Results and Discussion

The maternal mortality ratio (per 100,000 live births) was studied as a function of the contraceptive prevalence rate (% of married women aged 15-49), births attended by skilled health personnel (%), total fertility rate (birth per women), life expectancy at birth (years), physicians (per 100,000 people), female labor force participating rate and public expenditure on health (% of GDP). Concerning these, a model was constructed that

Table (1) Summary of OLS Regression Model Fitted to the Maternal Mortality Data

Coefficients	Value of Coefficients	Standard Errors of Coefficients	*t* Statistics	Significance of *t*	Collinearity Statistics	
					Tolerance	VIF
Constant	11.391***	1.218	9.354	0.000	-	-
CPR	-0.001	0.005	-0.182	0.856	0.379	2.639
BABSHP	-0.013***	0.005	-2.697	0.008	0.400	2.503
TFR	0.061	0.125	0.486	0.628	0.230	4.353
LE	-0.067***	0.013	-5.282	0.000	0.348	2.876
PHY	-0.033***	0.009	-3.883	0.000	0.513	1.949
FLFPR	-0.002	0.005	-0.424	0.673	0.937	1.068
PEOH	-0.153***	0.043	-3.552	0.001	0.824	1.214

Note: (1) Dependent variable: lnMMR
(2) ***, **, and * : Significant at 1%, 5% and 10% respectively

Data Source: Human Development Report (2009)

Thus, a widely used procedure, "search" called stepwise regression was applied for exploring the regressors. Table (2) presents the results obtained from the SPSS STEPWISE procedure. According to these results, the three variables (CPR, TFR, and FLFPR) were dropped out from the model (7) and the new model under consideration thus contains four explanatory variables: BABSHP, LE, PHY and PEOH. The new model can be seen as follows:

$$\ln MMR_i = \beta_0 + \beta_1 BABSHP_i + \beta_2 LE_i + \beta_3 PHY_i + \beta_4 PEOH_i + \varepsilon_i. \qquad (8)$$

Table (2) Summary of New Model Fitted to the Maternal Mortality Data

Coefficients	Value of Coefficients	Standard Errors of Coefficients	*t* Statistics	Significance of *t*	Collinearity Statistics	
					Tolerance	VIF
Constant	11.794***	0.562	20.983	0.000	-	-
BABSHP	-0.015***	0.004	-3.373	0.001	0.498	2.008
LE	-0.072***	0.010	-7.170	0.000	0.555	1.802
PHY	-0.035***	0.008	-4.173	0.000	0.537	1.864
PEOH	-0.153***	0.042	-3.685	0.000	0.865	1.156

Note: (1) Dependent variable: lnMMR
(2) ***, **, and * : Significant at 1%, 5% and 10% respectively

Data Source: Human Development Report (2009)

In Table (2) the stepwise method shows a statistically significant negative effect of each of the explanatory variables (BABSHP, LE, PHY and PEOH) and suggests that when the CPR, TFR and FLFPR variables are removed, a slight change in the values of coefficients of the remaining variables is found out. As expected, the coefficients of BABSHP, LE, PHY and PEOH are negative. The intercept value slightly increases from 11.391 to 11.974. In addition, the slope coefficients of the BABSHP, LE, PHY and PEOH slightly change from -0.013, -0.067, -0.033 and -0.153 to -0.015, -0.072, -0.035 and -0.153 respectively. Moreover, the value of F increases from 51.346 to 92.064 which is also shown in Table (3).

Table (3) Performance of Models Fitted to the Maternal Mortality Data

Model	Adjusted R^2	Standard Error of Estimates	F- Value	Significance of F	D.W.
Original[a]	0.788	0.734	51.346	0.000	1.818
New[b]	0.793	0.724	92.064	0.000	1.826

Note: (1) Dependent variable: lnMMR

(2)a: Predictors:(Constant),CPR, BABSHP, TFR, LE, PHY,FLFPR, PEOH

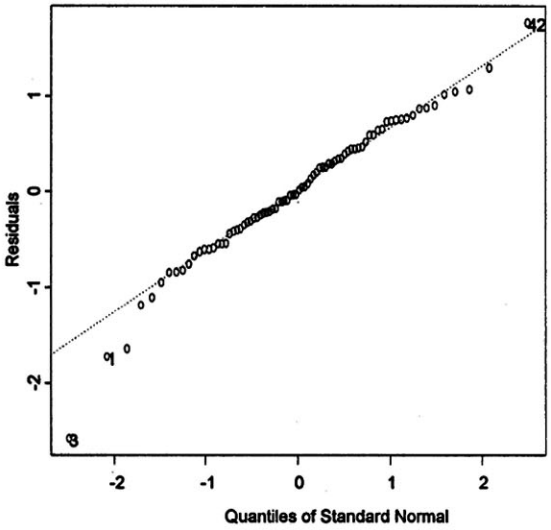
b: Predictors:(Constant), BABSHP,LE, PHY, PEOH

Data Source: Human Development Report (2009)

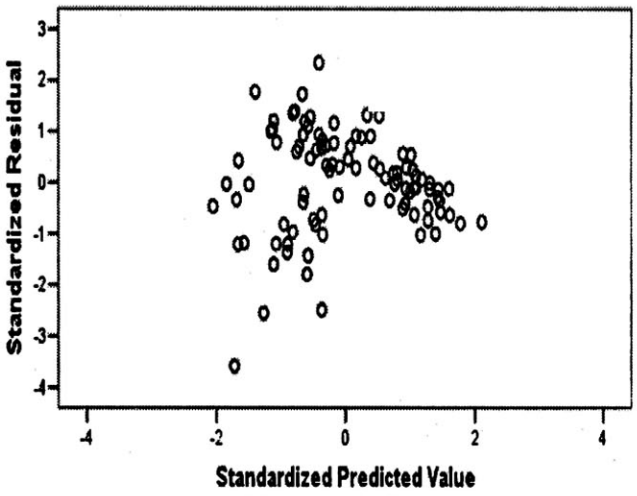
After the explanatory variables to be included in the model have been selected, a residual analysis was used to evaluate the aptness of the fitted model. Thus, the diagnostic plots which are shown in part (a) and (b) of Figure 1 are used for the study. Figure 1(a) suggests that the residuals of the fitted model (8) do not follow the normality assumption because some of the points do not fall in a straight line. It can be observed from the part (b) of Figure 1 that there is no apparent pattern between the standardized residual and predicted value. The residuals appeared to be evenly spread above and below the mean value for the predicted value. According to the part (a) of Figure 1, the new model violates the normality assumption. This result implies that, the data contain some outliers. Thus, the alternative procedure is used to achieve the robustness properties.

Before the application of robust methods, the types of unusual observations were investigated using a robust diagnostic plot. This plot is shown in Figure 2(a). Figure 2 reveals that observations 1, 3, 13 and 42 are vertical outliers and twelve observations are good leverage points. Therefore, according to the robust diagnostic plot, data contain four vertical outliers and twelve good leverage points. The Huber and bisquare M-estimation methods were applied to the same data set. The results are shown in Table (4).

Figure 1: (a) quantiles standard normal plot, and (b) standardized residuals versus predicted value of new model



(a)



(b)

Source: Based on Calculation

Table (4) OLS and *M*- Regression Models Fitted to the Maternal Mortality Data

Estimation Methods	β_0	β_1	β_2	β_3	β_4	MSE
OLS	11.794*** (20.983)	-0.015*** (3.373)	-0.072*** (7.170)	-0.035*** (4.173)	-0.153*** (3.685)	0.525
M-Huber	11.460*** (19.382)	-0.018*** (4.073)	-0.064*** (6.197)	-0.030*** (3.518)	-0.139*** (3.230)	0.399
M-Bisquare	10.703*** (18.590)	-0.015*** (3.515)	-0.054*** (5.391)	-0.035*** (4.212)	-0.123*** (2.940)	0.183

Note: (1) Absolute values of *t* statistics in parentheses

(2) ***, **, and * : Significant at 1%, 5% and 10% respectively

Source: Human Development Report (2009)

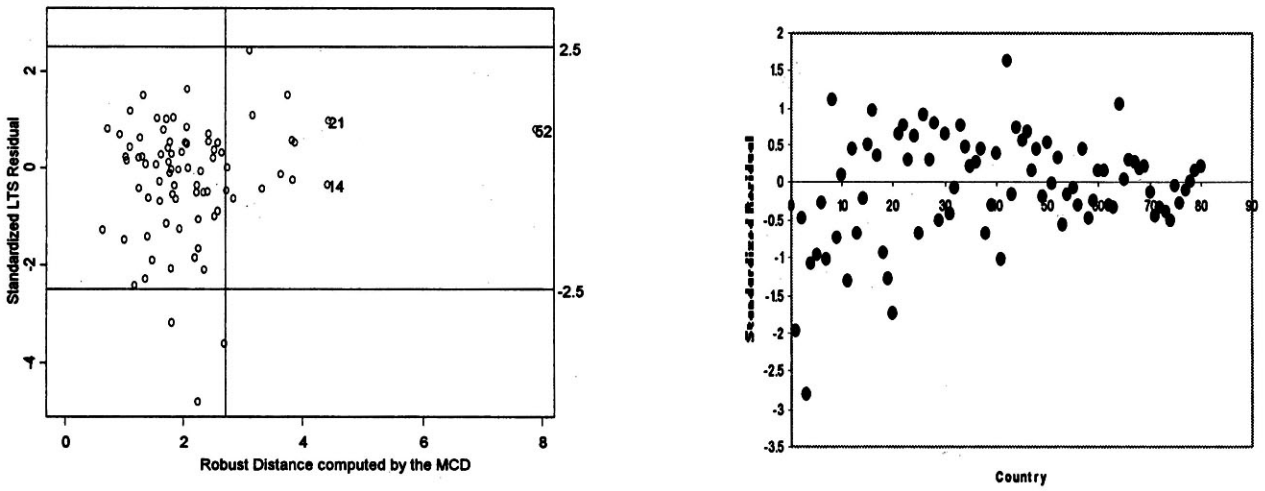
Table (4) gives the estimates from OLS and *M*-estimation regressions fitted to the maternal mortality data. The intercepts of *M*-estimates based on Huber and bisquare *M*-estimation are not too much different. Due to the vertical outliers, the mean squares error

(MSE) of OLS is the largest value among the others. The MSE of *M*-estimates is quite different from OLS method. The *M*-estimation methods detected the outliers for Australia, Ireland, Mauritius and Guyana and these are given relatively small weights. In this study, the optimal result was obtained by using the *M*-estimates based on bisquare ψ -function.The fitted regression model is given below:

$$\ln \hat{MMR} = 10.703 - 0.015\,\hat{BABSHP} - 0.054\,\hat{LE} - 0.035\,\hat{PHY} - 0.123\hat{PEOH}. \quad (9)$$

Based on the above fitted model (9), the maternal mortality ratios of some countries are estimated and the results are displayed in Table (5). From this table, it can be easily seen that the actual and estimated values are not very different.

Figure 2: Diagnostic Plots: (a) robust diagnostic plot, and (b) plot of the standardized residuals



(a) (b)

Source: Based on Calculation

Table (5) Actual and Estimated values of Maternal Mortality Ratio

Country	MMR		
	Actual	Estimated	Error
Netherlands	6	14	8
Canada	7	14	7
Syrian, Arab Republic	130	135	5
Tajikistan	170	147	23
Vietnam	150	120	30

Source: Calculated from Equation (9)

4. Conclusion

The data of the maternal mortality ratio for 80 countries are used to analyze the performance of ordinary least squares (OLS) and *M*-estimation methods in multiple linear regression model. The OLS method, which is sensitive to outliers, is used to estimate the regression parameters. It is found that, the OLS method does not meet the basic assumptions due to the vertical outliers and the mean squares error (MSE) is large. Therefore, the model is estimated by *M*-estimation method based on Huber and bisquare ψ-function. It is found that *M*-estimation based on bisquare ψ-function gives the actual and estimated values which are not very different.

Reference

1. Atkinson, A. C., and M. Riani (2000), *Robust Diagnostic Regression Analysis*, New York : Springer-Verlag.
2. Barnett, V., and T. Lewis (1978), *Outliers in Statistical Data*, New York: John Wiley and Sons.
3. Davies, L. (1993), Aspects of Robust Linear Regression, *The Annals of Statistics*, vol. 21, 1843-1899.
4. Draper, N. R., and H. Smith (1981), *Applied Regression Analysis*, New York: John Wiley.
5. Huber, P. J. (1981), *Robust Statistics*, New York: John Wiley and Sons.