Dhaka Univ. J. Sci. 58(2): 183-186, 2010 (July)

Median-Product Regression Estimators: New Robust Estimators for Bivariate Data

A. Z. M. Shafiullah and Jafar A. Khan

Department of Statistics, Biostatistics & Informatics, Dhaka University, Dhaka 1000, Bangladesh

Received on 16.05.2009. Accepted for Publication on 13.09.2009.

Abstract

In this article we consider the problem of calculating the linear regression estimates from bivariate data containing a fraction of outliers. Classical estimates of slope and intercept are affected by outliers, while robust estimates are computationally inefficient. In order to achieve robustness and computational efficiency at the same time, we propose new robust estimators of regression parameters. We call our estimators Median-Product (MP) regression estimators. To construct the proposed MP estimators, we replaced the non-robust building blocks of classical OLS estimators by their robust counterparts. Thus, we developed robust regression estimators that do not use iterative algorithm. Our simulation studies and real data application show that the proposed MP estimators give better results in the contaminated data compared to the classical estimators. The performance of our estimators is similar to that of the existing robust estimators. The advantage of our estimators is that they require less computing time than the existing robust estimators.

Keywords: Simple linear regression; Robustness; Computational efficiency; Sampling distribution; Standard error.

Abbreviations and acronyms: OLS - ordinary least squares; MP - median product; MAD - median absolute deviation from median; SS - sum of squares.

I. Introduction

Real datasets usually contain a fraction of outliers and other contaminations. The classical regression estimates, i.e., Ordinary Least Squares (OLS) estimates $\hat{\alpha}$ and $\hat{\beta}$ of the parameters of a simple linear regression model are much affected by these outliers and often give misleading results. Robust methods are designed to consider the *majority* of the data rather than *all* the data. Therefore, robust methods give reasonable results even when data contain a fraction of outliers. However, a major drawback of existing robust methods is that they are not computationally suitable, because fitting a robust model is a nonlinear optimization problem. Existing robust regression estimates such as S estimates (Huber, 1964), MM estimates (Yohai, 1987), and τ estimates (Yohai and Zamar, 1988) are also computationally inefficient. We propose new robust regression estimators for bivariate data that are resistant to outliers as well as computationally efficient. These estimators are obtained through the robustification of classical OLS estimators $\hat{\alpha}$ and $\hat{\beta}$. We call the proposed estimates the Median-Product (MP) estimates denoted by $\hat{\alpha}_{MP}$ and $\hat{\beta}_{MP}$ respectively. These estimates achieve robustness and computational efficiency at the same time.

The rest of the paper is organized as follows. In section II, we present our new robust estimators. In section III, we show the results of simulation study to compare the performance of our slope estimator with classical OLS and robust MM estimators. Section IV includes a real data application. Section V contains our conclusion.

II. Median Product Regression Estimates

Our MP regression estimators are based on a robust correlation estimator called MP correlation estimator proposed by Shafiullah and Khan (2009). We present this estimator below.

MP correlation estimator

Shafiullah and Khan (2009) argue that classical correlation estimator can be expressed as

$$r = \frac{\frac{1}{n}\sum_i (x_i - \bar{x})(y_i - \bar{y})}{s_x s_y} = \frac{1}{n}\sum_i \left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right),$$

i.e., $r = mean(Z_x \times Z_y)$, where $Z_x = \dfrac{x - \bar{x}}{s_x}$ and $Z_y = \dfrac{y - \bar{y}}{s_y}$. The authors proposed an initial robust correlation estimate (r_M) defined as:

$r_M = Median(Q_x \times Q_y)$, where

$Q_x = \dfrac{x - median(x)}{MAD(x)}$ and

$Q_y = \dfrac{y - median(y)}{MAD(y)}$ are the robust standardized variables. $MAD(x)$ and $MAD(y)$ are the median absolute deviations of X and Y, respectively. Since $-0.4549 \le r_M \le 0.4549$ (where 0.4549 is the median of χ_1^2 random variable), the authors denoted the asymptotic value of r_M by ρ_M and explored the relationship between ρ and ρ_M through a numerical study. The final MP estimate (denoted by $\hat{\rho}_{MP}$) that satisfies $-1 \le \hat{\rho}_{MP} \le 1$ is obtained using a transformation based on the numerical study. Shafiullah and Khan (2009) is referred to for details.

The proposed MP regression estimators

Let us consider a simple linear regression model

$$y_i = \alpha + \beta x_i + \varepsilon_i. \tag{2.1}$$

The ordinary least squares estimator of α and β can be expressed as $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$ and $\hat{\beta} = r\dfrac{s_y}{s_x}$. (2.2)

The non-robust ingredients of (2.2) are the means and standard deviations of *X* and *Y* and the Pearson's product moment correlation estimate *r*. We use the robust counterpart of these estimates to derive the proposed MP estimators. We used median and MAD as the robust counterparts of mean and standard deviation. MP correlation estimator $\hat{\rho}_{MP}$ is a robust counterpart of *r*. Thus following (2.2), the MP estimator of slope parameter β is:

$$\hat{\beta}_{MP} = \hat{\rho}_{MP} \frac{MAD(y)}{MAD(x)} \qquad (2.3)$$

and the MP estimator of intercept parameter α is:

$$\hat{\alpha}_{MP} = Median(y) - \hat{\beta}_{MP} Median(x) \qquad (2.4)$$

III. Simulations

First, we generated a single bivariate data and fitted the OLS line. We then contaminated the data by a single outlier and fitted the OLS line again along with the MP line. The results are presented in Figure 1.

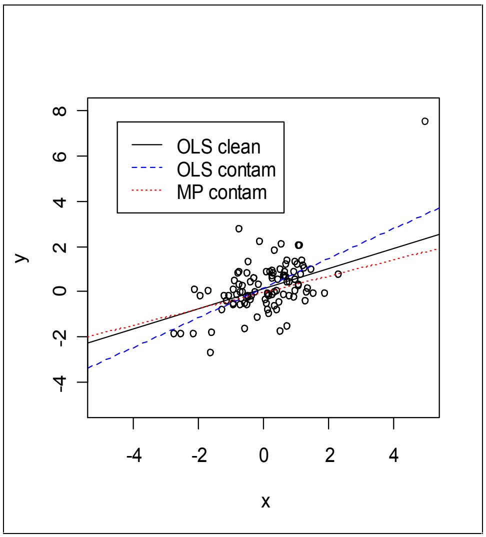

Fig.1. MP estimation fits the contaminated data better than OLS estimation.

This plot reveals that the OLS line for contaminated data has much larger slope than the OLS line for the clean data. The MP line, on the other hand, is close to the OLS line for the clean data. This means that the MP line is not affected by the outlier. The error SS for the three approaches are 71.06 (OLS clean), 85.72 (OLS contaminated) and 71.17 (MP contaminated) respectively.

We then conducted extensive simulation studies to examine the performance of our slope estimator $\hat{\beta}_{MP}$ and compare it with that of the classical and existing robust estimators.

Robustness of MP slope estimator $\hat{\beta}_{MP}$

We used R to carry out the simulation study. We considered the model (2.1) to generate 200 datasets each of size *n*=1000. The values of X and ε were generated from $N(0,1)$, and then Y values were generated using model (2.1) with $\alpha = 0$ and $\beta = 1$. Thus, the true value of the slope parameter is 1.

For each dataset, we calculated MP regression estimate $\hat{\beta}_{MP}$ along with classical estimate $\hat{\beta}$ and the existing robust MM estimate $\hat{\beta}_{MM}$. Then the data are contaminated by replacing a fraction of observations on X and Y by randomly generated extreme values that follow a different pattern from the majority of the data and are considered as outliers. To consider 5% outliers, each observation of a variable is assigned probability 0.025 of being replaced by a large number. [Therefore, the probability that any particular row of the dataset will be contaminated is $1\text{-}(1\text{-}0.025)^2$, which means that approximately 5% of the rows will be contaminated.] We then calculated the three estimates again from the contaminated data.

For the clean data, all the three estimators give similar results[1] not included in this paper. However, for contaminated data the classical estimate $\hat{\beta}$ took the value 0.581 as the average over 200 repeated trials. Thus OLS estimator of β failed to give precise estimate due to the presence of outliers. The average of 200 MM-estimates from 200 contaminated datasets was 0.773. The MP estimate attains remarkable robustness in comparison with MM estimate. While estimating β=1.0 our proposed estimator, applied to contaminated data took the value 0.905 as the average over 200 estimates. Therefore, the proposed MP estimator was found to be highly precise compared to OLS and even the existing robust MM estimate.

Figure 2 shows the sampling distribution of the three estimators. We observe that the proposed MP estimator captures most of the values near the parameter $\beta = 1$ (observed mean of MP estimates is 0.94). Thus, MP estimator attains the desired property of robustness. Further details of $\hat{\beta}_{MP}$ are given in Table 1.

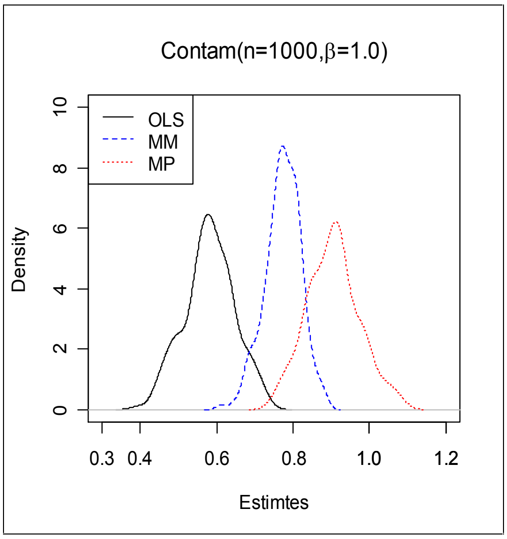

Fig. 2. Sampling distributions of OLS, MM and MP slope estimators for contaminated data.

From the above figure, we, therefore, argue that the OLS estimates provide misleading predictions of the dependent variable if the data contain even a small portion of outliers.

The proposed MP estimators are resistant to outliers and hence attain robustness to a great extent.

Computational efficiency of MP slope estimator $\hat{\beta}_{MP}$

We compared the proposed robust MP and existing robust MM slope estimator on the basis of average standard error, average magnitude of bias and total elapse computing time taken by CPU for 200 trials each comprising a sample with 5% outliers. The study considered small (*n*=25), moderately large (*n*=100) and large samples (*n*=400) and different values of the parameter (β= 0.5, 1.0, 1.5 and 2.0). The results are summarized in the following table.

Table. 1. Standard error, CPU time and magnitude of bias of MM and MP estimators of β in contaminated data.

Criteria	β	n=25		n=100		n=400	
		MM	MP	MM	MP	MM	MP
Standard error	0.5	0.184	0.347	0.116	0.178	0.055	0.084
	1.0	0.324	0.394	0.165	0.209	0.079	0.104
	1.5	0.240	0.486	0.101	0.279	0.080	0.126
	2.0	0.473	0.561	0.175	0.274	0.083	0.139
Elapse Computing time (second)	0.5	3.79	0.73	4.08	0.78	5.29	0.93
	1.0	4.38	0.70	5.19	1.05	8.52	1.12
	1.5	5.01	0.74	5.45	1.02	6.64	0.91
	2.0	5.49	0.66	5.27	0.99	6.60	1.14
Magnitude of Bias (Average).	0.5	0.276	0.119	0.186	0.035	0.202	0.053
	1.0	0.484	0.242	0.296	0.085	0.292	0.095
	1.5	0.59	0.248	0.312	0.141	0.306	0.15
	2.0	0.676	0.300	0.313	0.220	0.312	0.185

Table 1 reveals that, MP estimator has the average standard error greater than that of MM estimator for *n*=25 but as the sample size increases, the standard errors of $\hat{\beta}_{MP}$ decreases. The average amount of bias of MP estimator is always smaller than that of MM estimator for every sample size (the last 4 rows of Table 1). So MP estimates are more evenly distributed around the parameter than MM estimate.

Another reason for the preference of the proposed MP estimator is the less computing time required by it. The proposed MP estimator always takes less CPU time than the existing MM estimator and hence achieves remarkable computational efficiency.

Based on these extensive simulations we, therefore, argue that the new robust estimators, i.e., MP estimators of slope and intercept of simple linear regression model are preferable to the existing robust MM estimators if the sampled data contain a fraction of outliers.

IV. Application: Motorola VS. Market Data

The response variable (*Y*) is the difference between the monthly Motorola returns and the returns on 30-day US Treasury bills. The explanatory variable (*X*) is the difference between the monthly Market returns and the returns on 30-day US Treasury bills. The financial economists fit a straight line to this type of data. The slope measures the riskiness of the stock, the larger the slope the riskier the stock. We are interested to estimate the parameters of the regression equation $Y = \alpha + \beta X + \varepsilon$. The following table gives the results obtained from MP and OLS estimation. The estimate of the intercept was close to zero.

Table. 2. Different slope estimates for Motorola vs. Market data

Method of Estimation	$\hat{\beta}$		$R^2 (= r^2)$	
	“Clean” data	Contaminated data	“Clean” data	Contaminated data
OLS	0.918151	0.847862	0.3635	0.3496
MP	1.197368	1.184615	0.4225	0.4356

Table 2 shows that the OLS and MP estimate of β are 0.918 and 1.197, respectively for the "clean" data. The reason for this difference is that, there are some masked outliers even in the "clean" data, which are difficult to identify. Figure 3 contains the two fitted lines with a scatter plot for original data.

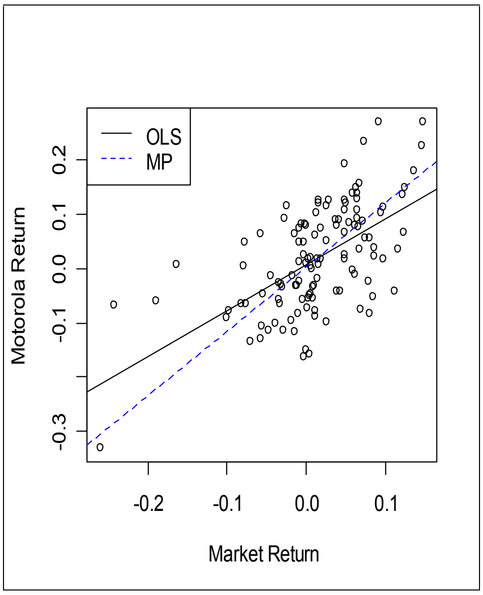

Fig. 3. A scatter plot of Motorola data and two different regression lines.

According to OLS estimate, Motorola's stocks are safer than the market. But according to MP estimate, Motorola's stocks are riskier than the market. As the outliers are eliminated, the OLS estimate of the slope increases. This indicates that for clean dataset, the OLS estimate tends to the MP estimate, which is not affected by the outlier.

V. Conclusion

In this study, we consider the problem of estimating the regression coefficients for bivariate data that may contain a fraction of outliers. Classical OLS estimates of regression parameters are much affected by the outliers. On the other hand, robust methods are computationally inefficient since they use iterative algorithms. Our goal was to achieve robustness and computational efficiency at the same time. We proposed new robust estimator of regression coefficients: Median-Product (MP) regression estimators. We performed simple robustification of OLS estimators $\hat{\alpha}$ and $\hat{\beta}$ by replacing the non-robust building blocks by their robust counterparts. We denoted our robust estimators by $\hat{\alpha}_{MP}$ and $\hat{\beta}_{MP}$. The new robust regression estimates have much better performance compared to classical estimates in the contaminated data. The performance of our estimators is comparable to the existing robust MM estimators. The advantage of our estimator is that they take less computing time.

When applied to a real dataset (Motorola vs. Market data), the proposed Median-Product regression estimates show better performance compared to the classical estimates.

----

1. Huber, P. J., 1964. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, **35:** 73-101.
2. Rousseeuw, P. J., and V. J. Yohai, 1984. Robust Regression by means of S- estimators. *Robust and Nonlinear Time Series Analysis* (J. Franle, W. Hardle, and R. D. Martin, *eds.*), Lecture notes in Statistics **26**, Springer Verlag, New York: 256-272.
3. Maronna, R. A., R. D. Martin, and V. J. Yohai, 2006. Robust Statistics, Theory and Methods. Wiley, England.
4. Shafiullah, A. Z. M., and J. A. Khan, 2009. Median-product correlation estimator: A new robust estimator for bivariate data. *Computational Statistics & Data Analysis*, submitted.
5. Yohai, V. J., 1987. High breakdown-point and high efficiency robust estimates for regression. *The Annals of Statistics*, **15:** 642-656.