OUTLIERS AND THEIR EFFECT ON PARAMETERS ESTIMATIONS IN REGRESSION ANALYSIS

Dr. Maw Maw Khin[14]

ABSTRACT

This study attempts to investigate the effect of outliers on estimation of parameters in regression analysis.The results about outlier robustness point out that the robust and classical methods both worked well data with no outliers indicating that their mean squares error (MSE) are quite close to each other. If there are outliers in the data, the robust methods perform better than the classical method. The OLS estimates provide poor estimates of true parameters of the regression model. As expected, OLS is a less efficient estimator whatever the type of outliers present in the data.

Keywords: RobustEstimators, Maximum Likelihood, Additive Outlier

1. Introduction

Outliers play an important role in regression. Outliers in the response variable represent model failure. Such observations are called outliers. Outliers with respect to the predictors are called leverage points. They can affect the regression model, too. Their response variables need not be outliers. Observation whose inclusion or exclusion results in substantial changes in the fitted model (coefficients, fitted values) is said to be influential.For this, about the types of outliers that can be found in regression analysis, their effects on regression coefficients and outliers detection were discussed in following.

Outliers can be thought of as observations in a data set that cause surprise in relation to the majority of the data. For example, surprising or extreme observations might be unusually large or unusually small values compared to the remaining data. Outliers are a common occurrence in data. They may be the result of an error in measurement of recording or transmission errors of exceptional phenomena such as earthquakes or strikes, or they may be due to the samples not being entirely from the same population. Apparent outliers may also be due to the values being the same, but nonnormal (in particular, heavy-tailed) distribution.

Outliers should be investigated carefully. Often they contain valuable information about the process under investigation or the data gathering and recording process. Before considering the possible elimination of these points from the data, one should try to understand why they appeared.

---

[14] Professor and Head, Department of Statistics, Yangon University of Economics

Outliers can be classified in statistics as outlying observations in linear regression, time series analysis, survey, directional and contingency table data (Barnett and Lewis, 1978). In the regression context, outliers are classified as y- and x-outliers. They always entail both theoretical and practical problems. Usually, depending on our goal(s), we need one or more procedures that are robust, to protect against and detect outlying observations in the data. For instance, in the case of a forecasting model, it is of utmost importance to be able to detect, estimate the effects of, and interpret outliers. In some cases, outliers in a residual series may indicate omission of an explanatory variable from the model. Furthermore, the robust regression estimates are less biased than OLS and provide estimates of outliers that are more strikingly seen in residual series.

Outliers may appear in data due to (i) gross errors, (ii) wrong classification of the data (outlying observations may not belong to the model followed by the bulk of the data), (iii) grouping, and (iv) correlation in the data (Hampel et al., 1986).

Gross errors often show themselves as outliers, but not all outliers are gross errors. Gross errors or outliers are data severely deviating from the pattern set by the majority of the data. This type of error usually occurs due to mistakes in copying or computation. They can also be due to part of the data not fitting the same model, as in the case of data with multiple clusters. Gross errors are often the most dangerous type of errors. In fact, a single outlier can completely spoil the least squares estimate, causing it to break down. Consequently, the estimators may not be efficient estimators. Some outliers are genuine and may be the most important observations of the sample. Rounding and grouping errors result from the inherent inaccuracy in collecting and recording data which are usually rounded, grouped, or even roughly classified. The departure from an assumed model means that real data can deviate from the assumed distribution. The departure from the normal distribution can manifest itself in many ways, for instance, in the form of skewed (asymmetric) or longer-tailed distributions.

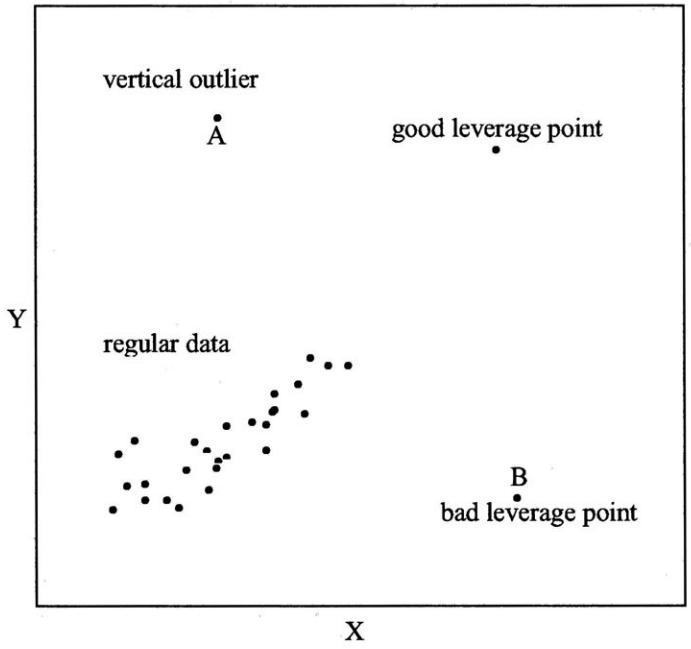
Types of Outliers in Regression

According to Rousseeuw and Van Zomeren(1990), there are several kinds of outliers. They proposed vertical outlier, good leverage point and bad leverage point. A point (x_i, y_i) which does not follow the linear pattern of the majority of the data but whose x_i is not outlying is called a vertical outlier. A point (x_i, y_i) whose x_i is outlying is called a good leverage point which follows the pattern of the majority, and a bad leverage point otherwise. To summarize, a data set can contain four types of points: regular observations, vertical outliers, good leverage points, and bad leverage points. Of course, most data sets do not have all four types. These types of outliers are shown in diagrammatic form.

Figure (1) shows these four types in simple regression. Point A clearly deviates from the typical linear relationship between the dependent (Y) and the independent (X)variable. Such 'vertical'

outlier is characterized by an unusual observation in the dependent variable. The impact of vertical outliers on the estimation of regression coefficients is usually small and mainly affects the regression intercept. If unusual observations occur in the set of independent variables, these outliers are called leverage points. If such leverage point deviates from the linear relationship described by the majority of observations it is called 'bad leverage point' such as Point *B* in Figure 1. Due to the exposed position of the outlier it has a leverage effect on the coefficient estimation. In contrast, a leverage point is called 'good leverage point' if it does not deviate from the typical relationship. Good leverage points are no outliers and even improve the regression inference as these points reduce standard errors of coefficient estimates.

Figure (1) Simple Regression Data with Points of All Four Types



Rousseeuw and Van Zomeren(1990) pointed out that high leverages can affect the estimated slope of the regression line in OLS, thus they may cause more serious problems than other outliers which might only affect the estimated intercept term. Moreover, their occurrence in regression models may move to some low leverage as well as high leverage and it can turn in vice versa. These two concepts are called masking and swamping in linear regression (Rousseeuw and Leory, 1987). Furthermore, the range of explanatory variables increases when they exist in regression analysis. Thus, the multiple coefficient determination statistics (R^2) which is a well known and popular measure of goodness-of-fit in the regression models will increase even by any changes of a single x variable (Ryan, 1997). In addition, high leverages may be the prime source of collinearity-influential observations whose presence can make collinearity and can destroy the existing collinearity pattern among the x variables (Hadi,

1992). In this respect, the identification of high leverage points to prevent their effect on linear regression becomes necessary.

2. Data and Methods

This paper focuses on the effect of the outlier on parameters estimation in regression by using the OLS method and robust methods. The required data sets are generated by using multiple linear regression models with three explanatory variables. Then, these data sets are transformed into outlier contaminated data sets. After that, the performances are compared in terms of bias and MSE criteria and then the most suitable estimation method is chosen.The statistical software packages namely S-PLUS 2000, STATA 10, and SPSS 13.0 were used to obtain the desire estimates throughout the analyses.

3. Results and Discussion

In this section, the performances of OLS and robust estimators were analyzed by simulations.First, the required data were generated from a multiple regression model.Then, simulated data were used to show that the robust methods outperform the classical method in presence of outliers. The data sets were generated from the following model:

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + e_i,\ i = 1,2,\ldots,n \qquad (1)$$

where all regression coefficients are fixed $\beta_0 = 5$ and $\beta_j = 1$, for each $i = 1,2,\ldots,n$ and $j = 1,2,\ldots,p$. The explanatory variables were randomly generated from a normal distribution with mean 0 and unit variance. The errors were assumed to be i.i.d. with $N(0,0.5)$.The data sets were generated under three regressors ($p = 3$) and the sample sizes were ($n = 30$ and $n = 40$) respectively. The true y's were calculated from the Equation (1).

In this simulation study, two types of outlier namely vertical outlier and bad leverage point were studied because they give different effects in the estimation of parameters of the regression model. After generating the data sets, two scenarios were considered in the following manners. They were seen as follows

(i) outliers in the independent variable: 10% of the y observations set to be vertical outliers by multiplying constant number 5 and keeping the others.

(ii) outliers in both y and x: 10% of both y and x observations were modified to be vertical outlier and bad leverage points and the remaining were unchanged. The vertical outlier was

obtained by multiplying 5 to its y value and the bad leverage point was obtained by adding 10 to its x value.

All simulations were done with 100 replications.To measure the robustness, the bias (that is the average of the estimated parameters minus the true value) and the mean squared errors (that is the variance of the estimated parameters plus the square of the bias) were used. For the first scenario, among the robust methods the LAV, *M* and *MM*-estimators were applied to this simulated data set because these estimators are robust subject to the vertical outliers. Then, this procedure was repeated 100 times and each time the parameters of OLS, LAV, *M*- (using Huber and Turkey) and *MM*-estimators (with a 70%, an 85% and a 95% efficiency) were estimated. On the basis of all the estimated parameters, the bias and the MSE were computed and the results were presented in Table (1). Figure (2) summarizes the results of simulations where $n = 30$ and $n = 40$ observations and three predictors. Bars represent bias and MSE for each estimator.

It is seen in Table (1) and Figure (2) that in the presence of vertical outliers, both the bias and MSE obtained from the *MM*-estimators (with a 70%, an 85% and a 95% efficiency), Huber and Turkey-*M*, and LAV are much close to each other but inferior to the OLS estimator. Their patterns shown in Figure 2(a) to (d) are intermingled and so no methods have a preferable bias and MSE in this case.

In the case of second scenario, the LAV, *M*, *MM*, LTS and LMS estimators were applied to this simulated data set. The results are shown in Table (2) and Figure(3). According to Figure3, the bias and MSE obtained from theHuber and Turkey-*M* are the smallest, followed by the *MM*-estimators (with a 70%, an 85% and a 95% efficiency) and LTS estimator in presence of vertical outlier and bad leverage points. In this case, the LMS behaves differently but just slightly, and have a bias and an MSE comparable to that of Huber and Turkey-*M* and *MM*-estimators. The OLS method also indicated in Figure 3(a) to (d) performs much worst in these situations. Therefore, the low bias and MSE values of theHuber and Turkey-*M* and *MM*-estimators are in line with the asymptotic robustness properties. As expected, OLS is a relatively less efficient estimator whatever the type of outliers occurred in the data.

4. Conclusion

In order to analyze the effect of outliers on the estimation of parameters in regression model, the classical and the robust estimation techniques are used. In this study, the multiple linear regression with three explanatory variables is used to generate the data sets. These clean data sets are transformed into outlier contaminated data sets. In this simulation study, two scenarios

are analyzed. According to the findings of the first scenario, it is shown that the *MM*-estimates (with a 70%, an 85% and a 95% efficiency), Huber and Turkey*M*- estimates, and LAV estimates are more resistant and efficient in the presence of vertical outliers. The OLS estimates provide poor estimates of true parameters of the regression model. Similarly, theHuber and Turkey*M*-estimates and *MM*-estimates are in line with the asymptotic robustness properties in the presence of both vertical and bad leverage points. As expected, OLS is a less efficient estimator whatever the type of outliers present in the data.

Table (1) Bias and MSE for OLS and Robust Methods of Simulated Data with Vertical Outliers

Sample Size	Estimation Method		β_0	β_1	β_2	β_3
n = 30	OLS	Bias	1.9748	0.9043	1.0060	1.1331
		MSE	4.4833	1.7279	1.6046	2.0718
	LAV	Bias	0.3237	0.3173	0.2570	0.2460
		MSE	0.3378	0.3391	0.1148	0.1133
	M-H	Bias	0.3151	0.2832	0.2277	0.2381
		MSE	0.3284	0.3018	0.0916	0.0958
	M-T	Bias	0.2354	0.2419	0.1813	0.1787
		MSE	0.2809	0.2519	0.0625	0.0574
	MM-(0.70)	Bias	0.2481	0.2794	0.2143	0.2081
		MSE	0.2905	0.2940	0.0934	0.0762
	MM(0.85)	Bias	0.2403	0.2513	0.1914	0.1892
		MSE	0.2800	0.2744	0.0690	0.0647
	MM-(0.95)	Bias	0.2387	0.2506	0.1932	0.1813
		MSE	0.2827	0.2744	0.0729	0.0593
n = 40	OLS	Bias	2.0424	0.9613	0.7830	0.8337
		MSE	5.2361	1.4872	1.0443	1.3226
	LAV	Bias	0.3744	0.1748	0.2479	0.2070
		MSE	0.9587	0.0454	0.1463	0.1095
	M-H	Bias	0.3800	0.1810	0.2233	0.1918
		MSE	1.0227	0.0544	0.1283	0.1095
	M-T	Bias	0.2994	0.1446	0.2092	0.1579
		MSE	0.8796	0.0346	0.1100	0.0643
	MM-(0.70)	Bias	0.3171	0.1702	0.2202	0.1726
		MSE	0.8723	0.0455	0.1151	0.0696
	MM(0.85)	Bias	0.3099	0.1543	0.2111	0.1599
		MSE	0.8772	0.0407	0.1081	0.0602
	MM-(0.95)	Bias	0.2999	0.1511	0.2153	0.1556
		MSE	0.8763	0.0380	0.1107	0.0585

Simulation setup: simulations = 100, contamination = 10%

Source: Calculations based on simulation data

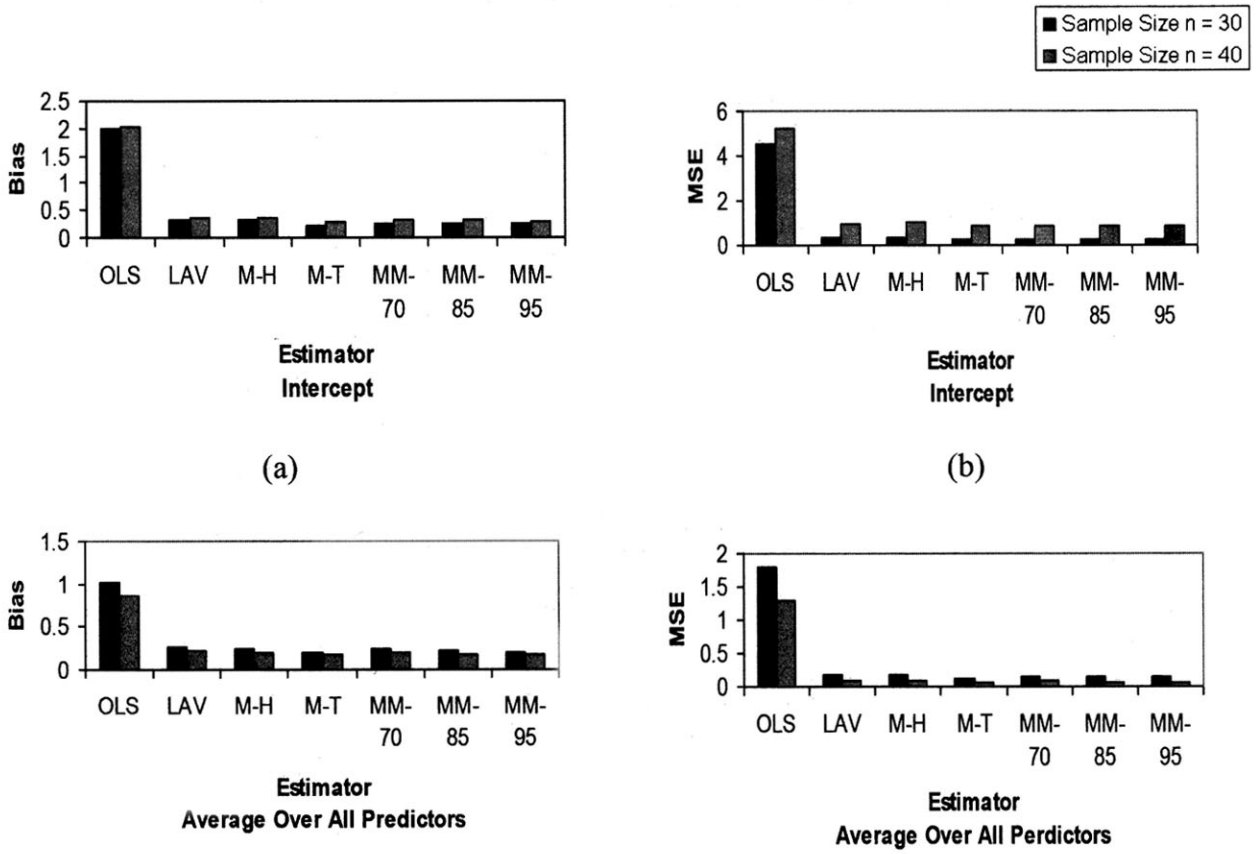
Table (2) Bias and MSE for OLS and Robust Methodsof Simulated Data with Vertical Outlier and Bad Leverage Points

Sample Size	Estimation Method		β_0	β_1	β_2	β_3
n = 30	OLS	Bias	0.7798	0.1761	0.1618	0.6959
		MSE	1.1732	0.0558	0.0464	0.8442
	LAV	Bias	0.3016	0.0832	0.0983	0.2025
		MSE	0.6038	0.0135	0.0166	0.0829
	M-H	Bias	0.2894	0.0796	0.0855	0.1934
		MSE	0.5636	0.0124	0.0120	0.0743
	M-T	Bias	0.2896	0.0926	0.0932	0.1814
		MSE	0.5688	0.0201	0.0180	0.0624
	MM-(0.70)	Bias	0.3057	0.1286	0.1261	0.2095
		MSE	0.6098	0.0397	0.0387	0.0852
	MM-(0.85)	Bias	0.2966	0.1175	0.1165	0.1872
		MSE	0.5638	0.0309	0.0315	0.0665
	MM-(0.95)	Bias	0.2905	0.1218	0.1112	0.1841
		MSE	0.5686	0.0387	0.0262	0.0674
	LMS	Bias	0.3150	0.0992	0.0999	0.2046
		MSE	0.5869	0.0233	0.0215	0.0762
	LTS	Bias	0.4382	0.2175	0.2680	0.3365
		MSE	0.7850	0.1027	0.1484	0.1802
n = 40	OLS	Bias	0.4899	0.1193	0.1449	0.1400
		MSE	0.2817	0.0240	0.0556	0.0357
	LAV	Bias	0.1614	0.0775	0.0855	0.0762
		MSE	0.0462	0.0097	0.0170	0.0094
	M-H	Bias	0.1408	0.0637	0.0714	0.0708
		MSE	0.0352	0.0067	0.0152	0.0082
	M-T	Bias	0.1432	0.0786	0.0804	0.0769
		MSE	0.0362	0.0163	0.0177	0.0099
	MM-(0.70)	Bias	0.1559	0.1138	0.1013	0.0944
		MSE	0.0444	0.0296	0.0242	0.0172
	MM(0.85)	Bias	0.1546	0.1016	0.0886	0.0905
		MSE	0.0407	0.0253	0.0182	0.0144
	MM-(0.95)	Bias	0.1462	0.0962	0.0887	0.0800
		MSE	0.0366	0.0252	0.0203	0.0107
	LMS	Bias	0.1701	0.1078	0.0953	0.0855
		MSE	0.0515	0.0290	0.0228	0.0129
	LTS	Bias	0.3212	0.2157	0.1847	0.1872
		MSE	0.1621	0.0978	0.0715	0.0750

Simulation setup: simulations = 100, contamination = 10%

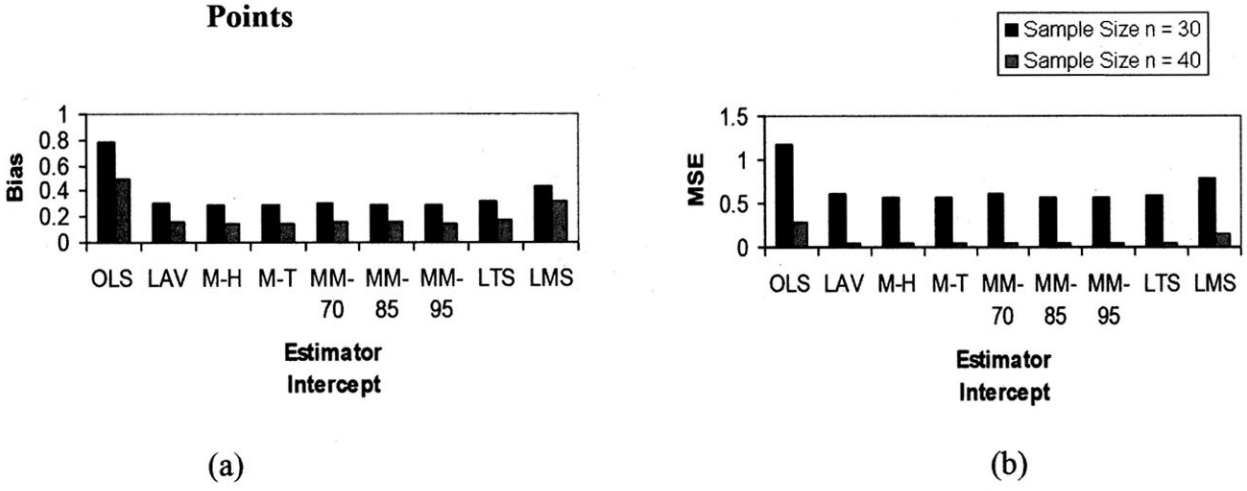
Source: Calculations based on simulation data

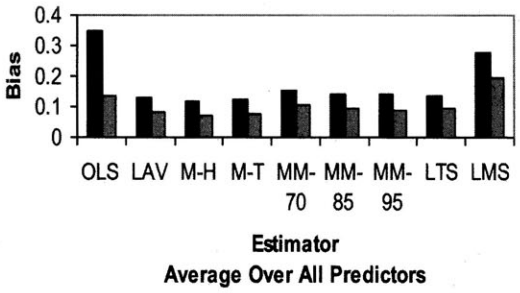
Figure (2) Bias and MSE of Simulated Data with Vertical Outliers



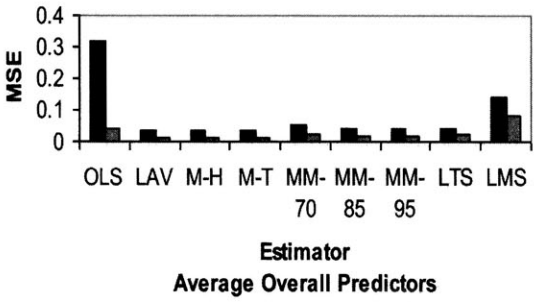
Source: Table (1)

Figure (3) Bias and MSE of Simulated Data with Vertical Outlier and Bad Leverage Points





(c)



(d)

Source: Table (2)

ACKNOWLEDGEMENT

I would like to express my gratitude to Professor Dr. Khin Naing Oo, Ractor, Yangon University of Economics, for giving me an opportunity to write this research paper.

REFERENCES

1. Andrews, D. F. (1974), A Robust Method for Multiple Linear Regression, *Technometrics*, vol.16, 523-531.
2. Atkinson, A. C., and M. Riani (2000), *Robust Diagnostic Regression Analysis*, New York: Springer-Verlag.
3. Barnett, V., and T. Lewis (1978), *Outliers in Statistical Data*, New York: John Wiley and Sons.
4. Davies, L. (1993), Aspects of Robust Linear Regression, *The Annals of Statistics*, vol.21, 1843-1899.
5. Draper, N. R., and H. Smith (1981), Applied *Regression Analysis*, New York: John Wiley.
6. Hadi, A. S. (1992), Identifying Multiple Outliers in Multivariate Data, *Journal of the Royal Statistical society, Series B*, vol. 54, 761-771.
7. Hampel, F. R., E. M. Ronchetti, P. J. Rousseeuw, and W.A. Stahel (1986), *Robust Statistics, The Approach Based on Influence Functions,* New York: John Wiley and Sons.
8. Huber, P. J. (1981), *Robust Statistics*, New York: John Wiley and Sons.
9. Rousseeuw, P. J., and V. J. Yohai (1984), Robust Regression by Means of S-Estimators. Robust and Nonlinear Time series Analysis, *Lecture Notes in Statistics*, vol. 26, 256-272.
10 Rousseeuw, P. J., and B. C. Van Zomeren (1990), Unmasking Multivariate Outliers and Leverage Points, *Journal of American Statistical Association,* vol.85, 633-639.
11. Ryan, T. P. (1997), Modern *Regression Methods*, New York: John Wiley and Sons.