Journal of Korean
Data & Information Science Society
2005, Vol. 16, No. 4, pp. 1053~1066

The Forward Sequential Procedure for the Identifying Multiple Outliers in Linear Regression[1)]

Jinpyo Park[2)]

Abstract

In this paper we consider the problem of identifying and testing outliers in linear regression. First we consider the use of the so-called scale ratio tests for testing the null hypothesis of no outliers. This test is based on the ratio of two residual scale estimates. We show the asymptotic distribution of the test statistics and investigate its properties. Next we consider the problem of identifying the outliers. A forward sequential procedure using the suggested test is proposed. The new method is compared with classical procedure in the real data example. Unlike other forward procedures, the present one is unaffected by masking and swamping effects because the test statistic is based on robust scale estimate.

Keywords : Forward sequential procedure, Optimal weight function, Outliers detection, Outliers test,

1. INTRODUCTION

Consider the linear regression model,

$$y_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{ip}\beta_p + e_i, \quad i = 1, 2, \cdots, n \tag{1}$$

where the β_i's are unknown parameters and the error e_i's are independent normal random variables with mean zero and variance σ^2. The classical estimators of parameters are the least squares estimator. However, if outliers are present in the data, the classical estimates can be very inaccurate. Intuitively, an

1) Research funded by Kyungnam University, 2005.
2) Professor, Division of Computer engineering and science, Kyungnam University.
E-mail : jppark@kyungnam.ac.kr

outlier is an observation $(x_{i1}, x_{i2}, \cdots, x_{ip}, y_i)$ which deviate from the linear relation followed by the majority of the data. The non-outlying data will be referred to as the good observations. It is assumed that the good data contains more data than 50% of the observations in the sample. The problem of detecting of outliers has been around for many years. Some techniques have been developed to remedy this problem.

In lower dimension, graphical techniques can be used to detect outliers. When the regression model has less than three independent variables, outliers can be detected by scatter plots and spin plots. But the degree of outlyingness is based on the judgement of the researcher. Unfortunately, when the dimension is greater than three, it is difficult to detect the outliers by graphical tool. We have to resort to other methods.

There are two general approaches to the problem of detecting outliers, regression diagnostics and robust methods of analysis.

The advantage of a regression diagnostics is that it identifies the outliers and allows the researcher how the outliers should dealt with. However, it is difficult to detect outliers when there are several of them because the masking and swamping effects that outliers can have on the diagnosing procedures.

Robust procedures are devised for the case of several outliers. They are not affected by the masking and swamping effects. They usually result in good estimates for the good data when the sample is contaminated but usually lack efficiency when the sample is uncontaminated. It is the main problem that robust estimation down weighs or completely ignores the outliers. This is not always best because any information contained in the outliers is lost.

They attack the problem from opposite points of view and, oddly enough, the advantages of one method tend to be the disadvantages of the other. So the two approaches to the problem of identifying outliers should be combined to produce a diagnostic test that which is not affected by masking and swamping effects. Furthermore this test should be applied sequentially in a forward fashion to not only detect the outliers but to indicate the number present as well. Furthermore, the test have to applied until that it fails to identify the presence of an outlier because it should not be fooled by masking and swamping effects.

In this paper, we propose a robust diagnostic tool for detecting and testing outliers in a linear regression. This tool is based on the ratio of a robust estimate of scale and a non robust one. And then we propose the following forward sequential procedure for detecting the outliers. If the null hypothesis is rejected then the most extreme observation is removed and the test is applied again to the $n-1$ remaining observations. This procedure is applied iteratively and stops when the test is no longer significant.

The remaining of the paper is organized as follows. In Section 2 we introduce the outliers tests based on the ratio of two different estimate of scale and the forward sequential procedure. In Section 3 we derive that the asymptotic

distribution of the test statistics under the null hypothesis is normal distribution. And we calculate the critical values and powers of test by Monte Carlo simulation. In Section 4 the new method is applied to several real data sets and artificial data sets in order to show their performances. Section 5 contains some concluding remarks.

Ⅱ. DETECTION AND TESTING OUTLIERS

We recall the s-estimate method introduced by Rousseeuw and Yohai(1984). Let ρ be a particular optimal symmetric bounded loss function and $\hat{\beta}$ be a robust estimate, the corresponding robust scale estimate, $\hat{s}(\beta)$, is a solution of the equation

$$\frac{1}{n}\sum_{i=1}^{n}\rho\left(\frac{y_i-\boldsymbol{x}_i\hat{\beta}}{s}\right)=\frac{1}{2}, \tag{2}$$

where $\hat{\beta}=\arg\min_{\beta} s(\beta)$.

We want to test the hypothesis

$$\begin{aligned} H_0&: \text{ no outlier in data } (x_{i1}, x_{i2}, \cdots, x_{ip}, y_i),\ i=1,2,\cdots,n \\ H_1&: \text{ some outliers in data } (x_{i1}, x_{i2}, \cdots, x_{ip}, y_i),\ i=1,2,\cdots,n \end{aligned} \tag{3}$$

in the linear regression.

The scale ratio test proposed for testing hypothesis is defined as followings. Let s_1 and s_2 be two estimate of scale corresponding to the following ρ-functions. Let ρ_1 be an optimal weight function introduced in Yohai and Zamar(1998). $\rho_1(\cdot;c)$ and derivative of $\rho_1(\cdot;c)$, $\psi_1(\cdot;c)$ are as follows:

$$\rho_1(x;c)=\begin{cases} 3.25c^2 & \text{if } |\frac{x}{c}|>3 \\ c^2[1.792-0.972(\frac{x}{c})^2+0.432(\frac{x}{c})^4-0.052(\frac{x}{c})^6+0.002(\frac{x}{c})^8 & \text{if } 2<|\frac{x}{c}|\le 3 \\ \frac{x^2}{2} & \text{if } |\frac{x}{c}|\le 2 \end{cases} \tag{4}$$

and

$$\psi_1(x;c)=\begin{cases}0 & \text{if } |\frac{x}{c}|>3\\ c[-1.944(\frac{x}{c})+1.728(\frac{x}{c})^3-0.312(\frac{x}{c})^5+0.016(\frac{x}{c})^7 & \text{if } 2<|\frac{x}{c}|\le 3\\ x & \text{if } |\frac{x}{c}|\le 2\end{cases}. \tag{5}$$

They showed that the function given above are optimal in following highly desirable sense: the final M-estimate has a breakdown point of one-half, and minimizes the maximum bias under contamination distributions, subject to achieving a desired efficiency when the data is Gaussian.

Let $\hat{\beta}$ be a robust estimate, the corresponding robust scale estimate, $s_1(\beta)$, is a solution of the equation:

$$\frac{1}{n}\sum_{i=1}^{n}\rho_1\left(\frac{y_i-\boldsymbol{x}_i\hat{\beta}}{s_1}\right)=\frac{1}{2}. \tag{6}$$

Let ρ_2 be the unbounded function. $\rho_2(\cdot)$ and $\psi_2(\cdot)$ are as follows:

$$\rho_2(x)=x^2 \tag{7}$$

and $\psi_2(x)=\rho_2'(x)=2x$.

Let $\tilde{\beta}$ be a non-robust estimate, the corresponding robust scale estimate, $s_2(\beta)$, is a solution of the equation:

$$\frac{1}{n}\sum_{i=1}^{n}\rho_2\left(\frac{y_i-\boldsymbol{x}_i\tilde{\beta}}{s_2}\right)=1. \tag{8}$$

Here, s_1 is the robust estimate of residuals scale with a breakdown point against outliers of about 50% and s_2 is the non-robust one of residuals scale since ρ_2 is unbounded. The scale ratio test statistic is defined as

$$R=s_2/s_1. \tag{9}$$

The null hypothesis is rejected for large value of R. The critical values approximated by Monte Carlo simulation using 1000 replicates are presented in Table 1, for sample size up to 30. For large sample sizes, the asymptotic approximation,

$$R_\alpha=1+0.7045\,n^{-1/2}Z_\alpha \tag{10}$$

can be used where Z_α is $100(1-\alpha)^{-}$ th percentile of standard normal distribution and n is the sample size used to compute the test statistics. The facts given above are detailed in next section.

When the null hypothesis is rejected, there is no indication of how many or which points are outliers. To solve this problem, we suggest to use this test sequentially in forward sequential procedure to identify the outliers. If the test rejects the null hypothesis then the point with the largest $D=|sort(r_i)-Med(r_i)|$ is defined as an outlier. Where $r_i=y_i-\hat{\beta}x_i$ and $\hat{\beta}$ is a robust estimate of regression coefficients β and $sort(r_i)$ is the sort of r_i and $Med(r_i)$ is the median of r_i. The observation detected as an outlier is removed and the test is applied again to the n-1 remaining observations. The procedure is repeated and stops when the test is no longer significant.

Ⅲ. PROPERTIES OF THE TEST STATISTIC

In this section we consider the properties of the proposed test. First we obtain the limiting distribution of test statistics under the null hypotheses. Observe that the test statistics $n^{1/2}\{(s_2/s_1)-1\}$ and $\sqrt{n}(s_2-s_1)$ are equivalent under null hypothesis. The Taylor expansion of equation (6) about $\hat{\beta}=\beta_0=0$ and $s_1=s_0=1$ gives,

$$\frac{1}{2}=\frac{1}{n}\sum\rho_1(y_i)-\hat{\beta}\frac{1}{n}\sum\rho_1'(y_i)x_i-(s_1-1)\frac{1}{n}\sum\rho_1'(y_i)(y_i)+\cdots \tag{11}$$

Since $n^{1/2}\hat{\beta}$ is asymptotically normal, the law of large number implies

$$n^{-\frac{1}{2}}\hat{\beta}\sum\rho_1'(y_i)x_i\rightarrow 0 \quad \text{in probability as} \quad n\rightarrow\infty \tag{12}$$

and

$$\frac{1}{n}\sum\rho_1'(y_i)(y_i)\rightarrow E\rho_1'(y)(y) \text{ almost surely as } n\rightarrow\infty. \tag{13}$$

Thus, using equation (11) - (13) we have the following asymptotic equivalence

$$\sqrt{n}(s_1-1)\approx\frac{\sqrt{n}\left[\frac{1}{n}\sum\rho_1(y_i)-\frac{1}{2}\right]}{E\rho_1'(y)\cdot y}. \tag{14}$$

By central limit theorem

$$\sqrt{n}(s_1-1) \rightarrow N(0, V) \tag{15}$$

where

$$V=\frac{\int_{-\infty}^{\infty}[\rho_1(y)-E_{\phi}(\rho_1(y))]^2\phi(y)dy}{\left\{\int_{-\infty}^{\infty}\rho_1'(y)\cdot y\phi(y)dy\right\}^2}.$$

Similarly

$$\sqrt{n}(s_2-1) \rightarrow N(0, \frac{1}{2}) \ . \tag{16}$$

So, under the null hypothesis

$$n^{\frac{1}{2}}\{(s_2/s_1)-1\} \rightarrow N(0, \tau^2) \tag{17}$$

where

$$\tau^2=V \ -\frac{\left\{\int_{-\infty}^{\infty}\rho_1(y)\cdot y^2\cdot\phi(y)dy-\int_{-\infty}^{\infty}\rho_1(y)\cdot y\cdot\phi(y)dy\right\}}{\int_{-\infty}^{\infty}\rho_1'(y)\cdot y\cdot\phi(y)dy}+\frac{1}{2},$$

and $\phi(\cdot)$ is probability density function of standard normal.

Next, we calculate the critical values for the test. For this purpose, we generate samples for various situations in the following situation,

$$y_i=x_{i1}+x_{i2}+\cdots+x_{ip}+e_i, \tag{18}$$

in which $e_i\sim N(0,1)$ and the explanatory variables are generated as $x_{ij}\sim N(0,100)$ for $j=1,2,\cdots,p$. Using 1000 replicates for each sampling situation we compute the critical values for the test. A summary of our results for explanatory variables up to 4 and sample size up to 30 is presented in the Table 1. When n equals 50, $R_{0.01}=1.231$, $R_{0.05}=1.164$, and $R_{0.1}=1.127$. Finally, we consider the power of the test for various situation. First, we generate a sample as $e_i\sim N(0,1)$ and $x_{ij}\sim N(0,100)$. Second, to construct outliers in the independent variables space, we generate samples where $(1-\alpha)\times 100\%$ of the

cases are as in the first situation. $\alpha\times100\%$ are generated as $e_i\sim N(0,1)$ and $x_{ij}\sim N(\mu,100)$. Finally, we make the outliers in response variable space. For this purpose, $(1-\alpha)\times100\%$ of the cases are again as in the first situation. $\alpha\times100\%$ are generated as $e_i\sim N(\mu,1)$ and $x_{ij}\sim N(0,100)$.

Using 1000 replicates for each sampling situation, we compute the power of the test. The results for two outliers, various magnitude of outlier, number of explanatory variables 1 and sample size 25 are presented in the Table 2. The power of the test increases with sample size and magnitude of outliers.

Table 1. Critical values for the proposed test

Sample sizes	Number of explanatory variable											
	1			2			3			4		
	α level			α level			α level			α level		
	0.01	0.05	0.10	0.01	0.05	0.10	0.01	0.05	0.10	0.01	0.05	0.10
5	2.334	1.871	1.543	2.474	1.909	1.597	2.594	1.929	1.618	2.619	1.945	1.639
6	2.259	1.637	1.431	2.366	1.788	1.483	2.484	1.809	1.503	2.603	1.827	1.523
7	2.064	1.577	1.421	2.364	1.755	1.456	2.385	1.777	1.479	2.415	1.795	1.495
8	2.060	1.557	1.413	2.292	1.734	1.452	2.313	1.756	1.477	2.324	1.756	1490
9	1.939	1.555	1.413	2.138	1.612	1.421	2.159	1.634	1.441	2.179	1.654	1.463
10	1.784	1.531	1.405	1.888	1.558	1.423	2.090	1.604	1.426	2.186	1.627	1.445
11	1.748	1.531	1.380	1.834	1.471	1.392	1.954	1.591	1.412	1.976	1.612	1.432
12	1.745	1.522	1.378	1.811	1.462	1.381	1.931	1.582	1.402	1.956	1.603	1.422
13	1.725	1.498	1.369	1.802	1.443	1.371	1.822	1.562	1.394	1.846	1.582	1.412
14	1.679	1.472	1.367	1.785	1.433	1.370	1.705	1.554	1.390	1.728	1.578	1.410
15	1.650	1.464	1.348	1.721	1.424	1.354	1.643	1.523	1.376	1.691	1543	1.410
16	1.546	1.412	1.315	1.600	1.385	1.351	1.520	1.459	1.371	1.681	1.525	1.398
17	1.468	1.375	1.308	1.473	1.365	1.346	1.505	1.426	1.368	1.528	1.447	1.396
18	1.441	1.357	1.297	1.449	1.362	1.334	1.493	1.405	1.346	1.504	1.435	1.386
19	1.417	1.333	1.265	1.438	1.355	1.316	1.448	1.385	1.336	1.484	1.405	1.365
20	1.392	1.300	1.237	1.400	1.345	1.262	1.429	1.375	1.326	1.474	1.405	1.355
21	1.369	1.298	1.233	1.367	1.332	1.254	1.419	1.365	1.306	1.454	1.378	1.336
22	1.347	1.294	1.229	1.362	1.310	1.250	1.413	1.358	1.303	1.449	1.370	1.329
23	1.334	1.286	1.220	1.357	1.307	1.243	1.406	1.347	1.300	1.446	1.365	1.324
24	1.327	1.267	1.217	1.348	1.307	1.240	1.402	1.342	1.295	1.431	1.364	1.324
25	1.305	1.241	1.205	1.342	1.305	1.237	1.399	1.340	1.287	1.429	1.359	1.320
26	1.291	1.245	1.196	1.335	1.294	1.235	1.387	1.340	1.275	1.421	1.351	1.317
27	1.285	1.235	1.194	1.326	1.287	1.230	1.374	1.334	1.274	1.415	1.350	1.310
28	1.277	1.227	1.182	1.311	1.265	1.228	1.370	1.330	1.271	1.409	1.347	1.306
29	1.260	1.222	1.178	1.301	1.261	1.221	1.365	1.327	1.271	1.391	1.347	1.301
30	1.256	1.219	1.172	1.266	1.231	1.185	1.362	1.321	1.269	1.388	1.344	1.298

Table 2. Estimated powers of the proposed test(n=25, p=1, two outliers)

magnitude of outliers	magnitude of outliers											
	20			30			40			50		
	significant level			significant level			significant level			significant level		
	0.1	0.05	0.01	0.1	0.05	0.01	0.1	0.05	0.01	0.1	0.05	0.01
20	0.965	0.943	0.938	0.995	0.993	0.988	0.998	0.997	0.996	1.00	1.00	1.00
30	0.966	0.955	0.944	0.996	0.994	0.99	0.999	0.998	0.997	1.00	1.00	1.00
40	0.969	0.957	0.950	1.00	0.999	0.995	1.00	1.00	1.00	1.00	1.00	1.00
50	0.970	0.962	0.954	1.00	0.999	0.995	1.00	1.00	1.00	1.00	1.00	1.00
60	0.973	0.966	0.958	1.00	1.00	0.996	1.00	1.00	1.00	1.00	1.00	1.00
70	0.975	0.973	0.963	1.00	1.00	0.997	1.00	1.00	1.00	1.00	1.00	1.00
80	0.977	0.975	0.967	1.00	1.00	0.998	1.00	1.00	1.00	1.00	1.00	1.00
90	0.983	0.977	0.972	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
100	0.991	0.985	0.983	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

continue(Table 2)

magnitude of outliers	magnitude of outliers											
	60			70			80			90		
	significant level			significant level			significant level			significant level		
	0.1	0.05	0.01	0.1	0.05	0.01	0.1	0.05	0.01	0.1	0.05	0.01
20	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
30	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
40	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
50	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
60	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.00	1.000	1.000	1.000
70	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
80	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
90	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
100	1.000	1.000	1.000	1.000	1.000	1.00	1.000	1.000	1.000	1.000	1.000	1.000

Ⅳ. APPLICATIONS OF THE PROPOSED TEST

In this section, the proposed test is applied to several data sets for the purpose of testing and detecting outliers.

Example 1 (Annual Rates of Growth of Prices in China)

The application begins by applying the test to the annual rates of growth of prices in China data given by Simkin(1978). Rousseew and Leroy(1987) used these data to illustrate the need for robust regression technique. For instance, in 1940 prices went up 1.62% as compared to the previous year. But a huge jump occurred in 1948. The data appear in the Table 3. The results applied the forward sequential procedure and robust method to this data are in Table 4.

Table 3. Annual Rates of Growth of Prices in China

index	Year(x)	Growth of Prices
1	40	1.62
2	41	1.63
3	42	1.90
4	43	2.64
5	44	2.05
6	45	2.13
7	46	1.94
8	47	15.50
9	48	364.00

Table 4. The results for the forward sequential procedure and robust method

sample size	observation selected	proposed test statistics	critical values			Robust(r_i/s)
			0.01	0.05	0.1	
9	9	299.089	1.939	1.555	1.413	2706.337
8	8	11.664	2.060	1.557	1.413	100.022
7	4	2.089	2.064	1.577	1.421	5.684
6	7	1.136	2.259	1.637	1.431	0.942

In the Table 4, the test is highly significant for observation 9 followed by observation 8, and observation 4, This test identifies observation 9, 8, and 4 as outliers. When the test is applied to the remaining 6 observations, null hypothesis is not rejected. This result confirms the conclusion drawn from the standardized LMS(least median squares) residual.

Example 2 (Number of Fire in 1976-1980)

This data set shows the trend from 1976 to 1980 of number of reported claims of Belgian fire insurance companies. One notices a slight upward trend the years. Rousseew and Leroy(1987) used these data to illustrate the need for robust regression technique. The data appear in the Table 5.

Table 5. Number of Fire Claims in Belgium from 1976 to 1980

index	Year(x)	Number of Fires
1	76	16,694
2	77	12,271
3	78	12,904
4	79	14,036
5	80	13,874

The results applied the forward sequential procedure and robust method to this data are in Table 6.

Table 6. The results for the forward sequential procedure and robust method

sample size	observation selected	proposed test statistics	critical values			Robust(r_i/s)
			0.01	0.05	0.1	
5	1	2.353	2.334	1.871	1.543	9.987
4	4	0.729	2.560	1.957	1.613	0.754

In the Table 6, the test is highly significant for observation 1. This test identifies observation 1 as outliers. When the test is applied to the remaining 4 observations, null hypothesis is not rejected. This result confirms the conclusion drawn from the standardized LMS residual.

Example 3 (Stackloss Data)

The second application for testing and detecting outliers comes from the Brownlee(1965). The data is well-known stackloss data set. We have selected this example because it is a set of real data and it is examined by many statisticians. Most people concluded that observation 1,
3, 4, and 21 were outliers. Some people reported that observation 2 was outlier. The data are shown in the Table 7. The result applied the forward sequential procedure and robust method to this data appear in the Table 8. In the Table 8, observation 21 is the most extreme followed by observation 4, observation 1, observation 3 and observation 2. The test identifies observation 21, 4, 1, and 3 as outliers. When the test is applied to the remaining 17 observations, null hypothesis is not rejected. Hence observation 2 is not a outlier. This result confirms the conclusion drawn from the standardized LMS residual. And It is the same to conclusion that most people reported.

Table 7. Stackloss data

index	rate (x1)	temper-ature(x2)	acid concen-tration(x3)	stackloss (y)	index	rate (x1)	temper-ature(x2)	acid concen-tration(x3)	stackloss (y)
1	80	27	89	42	12	58	17	88	13
2	80	27	88	37	13	58	18	82	11
3	75	25	90	37	14	58	19	93	12
4	62	24	87	28	15	50	18	89	8
5	62	22	87	18	16	50	18	86	7
6	62	23	87	18	17	50	19	72	8
7	62	24	93	19	18	50	19	79	8
8	62	24	93	20	19	50	20	80	9
9	58	23	87	15	20	56	20	82	15
10	58	18	80	14	21	70	20	91	15
11	58	18	89	14					

Table 8. The result applied the forward sequential procedure and robust method to the stackloss data

Sample size	observation selected	proposed test statistics	Critical Values			Robust(r_i/s)
			0.01	0.05	0.10	
21	21	1.7655	1.416	1.365	1.306	6.832
20	4	1.5459	1.429	1.375	1.326	7.245
19	1	1.4720	1.448	1.385	1.336	6.417
18	3	1.6047	1.493	1.405	1.346	6.210
17	2	1.236	1.505	1.426	1.368	2.277

Example 4 (Wood Specific Gravity)

Let us look at a finally example containing multidimensional real data. These data came from Draper and Smith(1966) and were used to determine the influence of anatomical factors on wood specific gravity. Rousseeuw and Leroy(1987) used a contaminated version of these data to compare the various diagnostic. These contaminated data is the outliers that are not outlying in any of the individual variables.

The result for comparing the various diagnostic appear in the table 10. The contaminated data is shown in the table 9. We applied the forward sequential procedure to the contaminated data . The result is listed in the table 10.

In the table 11, diagnostics based on least squares estimate did not succeed in identifying the actual contaminated observations, because they are susceptible to masking effect. But the standardized LMS(least median of squares)residuals and the resistant diagnostic suggested by Rousseeuw and Leroy identify the contaminated data 4, 6, 8, and 19 as the outliers.

In the table 10, Observation 19 is the most extreme followed by observation 6,

observation 8, observation 4 and observation 13. Because the test does not reject null hypothesis at significant 0.01 observation 13 is not an outlier. This test identify observation 19, 6, 8 and 4 as outliers. This result confirms the conclusions drawn from the standardized LMS residuals and the resistant diagnostic.

The above some examples confirmed that the forward sequential procedure based on robust estimate of scale is not affected by masking effects.

Table 9. Contaminated Data on Wood Specific Gravity

Index	x_1	x_2	x_3	x_4	x_5	y
1	0.5730	0.1059	0.4650	0.5380	0.8410	0.5340
2	0.6510	0.1356	0.5270	0.5450	0.8870	0.5350
3	0.6060	0.1273	0.4940	0.5210	0.9200	0.5700
4	0.4370	0.1591	0.4460	0.4230	0.9920	0.4500
5	0.5470	0.1135	0.5310	0.5190	0.9150	0.5480
6	0.4440	0.1628	0.4290	0.4110	0.9840	0.4310
7	0.4890	0.1231	0.5620	0.4550	0.8240	0.4810
8	0.4130	0.1673	0.4180	0.4300	0.9780	0.4230
9	0.5360	0.1182	0.5920	0.4640	0.8540	0.4750
10	0.6850	0.1564	0.6310	0.5640	0.9140	0.4860
11	0.6640	0.1588	0.5060	0.4810	0.8670	0.5540
12	0.7030	0.1335	0.5190	0.4840	0.8120	0.5190
13	0.6530	0.1395	0.6250	0.5190	0.8920	0.4290
14	0.5860	0.1114	0.5050	0.5650	0.8890	0.5170
15	0.5340	0.1143	0.5210	0.5700	0.8890	0.5020
16	0.5230	0.1320	0.5050	0.6120	0.9190	0.5080
17	0.5800	0.1249	0.5460	0.6080	0.9540	0.5200
18	0.4480	0.1028	0.5220	0.5340	0.9180	0.5060
19	0.4170	0.1687	0.4050	0.4150	0.9810	0.4010
20	0.5280	0.1057	0.4240	0.5660	0.9090	0.5680

Table 10. Forward sequential procedure Test for the Data in table 9

Sample size	observation selected	scale ratio statistics	Critical Values		
			0.01	0.05	0.10
20	19	1.867	1.584	1.515	1.465
19	6	2.144	1.718	1.645	1.595
18	8	2.525	1.847	1.772	1.712
17	4	2.772	1.977	1.892	1.833
16	13	1.557	1.981	1.922	1.863

Table 11. Diagnostics for the Data in Table 9[h_{ii} ; Squared Mahalanobis Distance; Standardized, Studentized, and Jackknifed Ls Residuals; $CD^2(i)$; DFFITS; DFBETAS; Standardized LMS Residuals, RD_i]

index	Based on Lesat squares method													Robust	
	h_{ii}	MD_i^2	r_i/s	t_i	$t(i)$	$CD^2(i)$	DFFITS	DFBETAS(0.447)						r_i/s	RD_i
i	0.600	11.07	2.50	2.50	2.50	1.00	1.095	β_1	β_2	β_3	β_4	β_5	Const.	2.50	2.50
1	0.278	4.327	-0.73	-0.85	-0.84	0.047	-0.524	-0.004	0.055	0.328	-0.052	0.215	-0.347	-0.16	0.798
2	0.132	1.552	0.05	0.05	0.05	0.000	0.019	0.009	0.002	-0.005	0.002	0.000	-0.003	0.00	0.701
3	0.220	3.224	1.24	1.41	1.46	0.093	0.776	-0.651	-0.523	-0.206	-0.429	0.549	-0.356	0.55	0.577
4	0.258	3.959	0.35	0.41	0.40	0.010	0.236	0.035	-0.049	0.015	-0.105	0.118	-0.074	-14.79	3.938
5	0.223	3.277	1.00	1.14	1.15	0.062	0.615	0.286	-0.517	0.164	-0.388	0.437	-0.244	1.75	0.605
6	0.259	3.974	-0.45	-0.53	-0.51	0.016	-0.302	-0.053	0.037	0.035	0.130	-0.113	0.050	-17.68	4.520
7	0.530	9.124	0.91	1.32	1.36	0.329	1.448	-0.956	0.424	0.521	0.133	-0.964	1.027	0.73	1.421
8	0.289	4.536	-0.03	-0.04	-0.04	0.000	-0.025	0.011	-0.012	0.005	-0.005	0.006	-0.005	-17.31	4.466
9	0.348	5.665	-0.40	-0.49	-0.48	0.021	-0.348	0.052	0.105	-0.224	0.161	0.007	-0.075	-0.73	1.243
10	0.449	7.588	-0.42	-0.56	-0.55	0.043	-0.492	-0.008	-0.198	-0.256	-0.137	-0.029	0.257	-0.40	1.267
11	0.317	5.075	1.99	2.40	3.02	0.447	2.059	0.425	0.970	0.748	0.198	-0.800	0.521	0.00	1.258
12	0.410	6.833	-1.20	-1.56	-1.65	0.281	-1.376	-0.597	0.013	0.556	0.359	0.368	-0.566	-1.88	1.030
13	0.287	4.506	-0.49	-0.58	-0.56	0.022	-0.356	-0.098	0.045	-0.251	0.106	-0.121	0.180	0.00	1.015
14	0.129	1.500	-1.26	-1.35	-1.40	0.045	-0.537	-0.169	0.228	0.178	-0.006	-0.103	0.021	-1.30	0.668
15	0.152	1.945	-0.59	-0.64	-0.62	0.012	-0.264	0.148	-0.061	-0.011	-0.162	0.108	-0.073	-0.34	0.465
16	0.526	9.049	0.52	0.76	0.75	0.107	0.789	-0.529	0.559	-0.052	0.745	-0.432	0.122	0.00	0.865
17	0.289	4.548	-0.25	-0.30	-0.29	0.006	-0.187	-0.019	0.019	-0.044	-0.055	-0.086	0.133	0.00	0.802
18	0.294	4.637	0.28	0.34	0.33	0.008	0.211	-0.062	-0.096	0.081	-0.024	0.045	-0.002	-0.21	0.985
19	0.292	4.599	-1.08	-1.29	-1.32	0.114	-0.849	0.195	-0.287	0.231	-0.024	0.079	-0.128	-20.84	5.201
20	0.318	5.084	0.55	0.66	0.65	0.034	0.441	0.092	-0.154	-0.305	0.037	0.046	0.064	0.00	0.816

V. CONCLUDING REMARKS

It is very important to test and detect the multiple outliers in linear regression. Several diagnostic measures based on the resulting from the least squares estimate have been proposed to identify the multiple outliers. However, the accuracy of diagnostic measures is very suspect because these can be severely affected by the masking and swamping effects. This inaccuracy can seriously affect their performance.

In this paper, we proposed the forward sequential test for testing and detecting the multiple outliers. This was founded on a robust estimate of scale.

In principle, the forward sequential test set up a natural simple approach for identifying the multiple outliers. However, if the forward sequential test is founded on the resulting from the least squares estimate, it can be seriously affected by the masking and swamping effects.

On the other hand, if the forward sequential test is founded on a robust estimate of scale, like the test proposed in this paper, the problem for the masking and swamping effects can be overcome.

We proved that the proposed forward sequential test was not affected by the masking and swamping effects through the Monte Carlo results and numerical examples. These suggest that the proposed test provides a conservative and fairly powerful method for the detection of the multiple outliers in linear regression.

References

1. Brownlee, K. A.(1965), *Statistical Theory and Methodology in Science and Engineering,* 2nd ed, John Wily & Sons, New York.
2. Draper, N. R., and Smith, H.(1966), *Applied Regression Analysis*, John Wily & Sons, New York.
3. Rousseeuw, P. J., and Yohai, V.(1984) Least median of squares regression, *J. Am. Stat. Assoc., 79, 871-884.*
4. Rousseeuw, P. J., and Leroy, A. M.,(1987) *Robust regression and outlier detection,* John Wiley & Sons, New York.
5. Simkin, C. G. F.(1978), Hyperinflation and Nationalist China, *Stability and Inflation, and A .Volume of Essays to Honour the memory of A. W. H. Phillips,* John Wiley & Sons, New York
6. Yohai, V.J. and Zamar(1998) Optimal locally robust M-estimates of regression, *Jour, of statist. Inf. and Planning*

[received date : Sep. 2005, accepted date : Oct. 2005]