Journal of Korean
Data & Information Science Society
2003, Vol. 14, No.3 pp. 673~685

The Scale Ratio Testing of 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 problem for testing the null hypothesis of no outliers. A test based on the ratio of two residual scale estimates is proposed. We show the asymptotic distribution of the test statistics by Monte Carlo simulation and investigate its properties. Next we consider the problem of identifying the outliers. A forward sequential procedure using the suggested test is proposed and shown to perform fairly well. 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 and Phrases : Outliers test, Forward sequential procedure, Optimal weight function.

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. It well known that outliers can have an extreme effect on the least squares estimate. Therefor the outlier problem has been around for many years. Intuitively, an 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

1) Research funded by Kyungnam University, 2003
2) Professor, Division of information & communication engineering, Kyungnam University. jppark@eros.kyungnam.ac.kr

observations. It is assumed that the good data contains more data than 50% of the observations in the sample.

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, once 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 dealing with outliers, diagnostics test and robust methods of analysis. They attack the problem from opposite points of view. Since the advantages of one method tend to be the disadvantages of the other, the two approaches to the outlier problem should be combined to produce a diagnostic test that which is not affected by masking and swamping effects. This test could then 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 scale estimate and a non robust scale estimate. 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. Since it is based on a robust scale estimate, one expects that this procedure will not be affected by masking and swamping effects. This is confirmed by numerical examples.

The remaining of the paper is organized as follows. In Section 2 we introduce the test statistic and the forward sequential procedure. In Section 3 we derive that the asymptotic distribution of the test statistics under the null hypothesis and calculate the critical values and powers of proposed test by Monte Carlo simulation. In Section 4 the proposed test and the forward sequential procedure 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

To test the hypothesis

$$H_0: \text{ no outlier in data } (x_{i1}, x_{i2}, \cdots, x_{ip}, y_i), \quad i=1,2,\cdots,n \tag{2}$$

$$H_1: \text{ some outliers in data } (x_{i1}, x_{i2}, \cdots, x_{ip}, y_i), \quad i=1,2,\cdots,n$$

in the linear regression, we propose a test statistic. The test statistic is a ratio of two scale estimates, s_1, which is sensitive to outliers, and s_2, which is highly robust.

The s_1 and s_2 are defined as follows. Let ρ_1 be an optimal weight function introduced in Yohai and Zamar(1988). 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.

The Yohai and Zamar's optimal functions $\rho_1(\cdot; c)$ is as follows:

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

and $\psi_1(x; c) = \rho_1'(x; c)$

$$\psi_1(x; c) = \begin{cases} 0 & \text{if } |\frac{x}{c}|>3 \\ c[-1.944(\frac{x}{c})+1.728(\frac{x}{c})^3-0.312(\frac{x}{c})^5+0.016(\frac{x}{c})^7 & \text{if } 2<|\frac{x}{c}|\le 3 \\ x & \text{if } |\frac{x}{c}|\le 2 \end{cases} \tag{4}$$

For any $\hat{\beta}$, let $s_2(\beta)$ be the solution of

$$\frac{1}{n}\sum\rho_1\left(\frac{y_i-x_i\hat{\beta}}{s_2}\right)=\frac{1}{2}, \tag{5}$$

where $\hat{\beta}=\underset{\beta}{\arg\min}\ s_2(\beta)$.

Let ρ_2 be the unbounded function,

$$\rho_2(x)=x^2 \tag{6}$$

and $\psi_2(x)=\rho_2'(x)$,

$$\psi_2(x)=2x \tag{7}$$

For any β^*, let $s_1(\beta)$ be solution of

$$\frac{1}{n}\sum\rho_2\left(\frac{y_i-x_i\beta^*}{s_1}\right)=1, \tag{8}$$

where $\beta^*=\min_{\beta} s_1(\beta)$.

Here, s_2 is S-estimate of residuals scale with a breakdown point 0.5 and s_1 is the non-robust estimate of residuals scale since ρ_2 is unbounded.

The test statistic is defined as

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

The null hypothesis is rejected when v is too large. However, when the null hypothesis is rejected, there is no indication of how many or which points are outliers. To solve this problem, we propose to apply the 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 S-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. The robust estimate of scale in the denominator is required to ensure that the test statistic is sensitive to outliers and that the forward sequantial procedure is not affected by possible masking and swamping effects of several outliers.

Ⅲ. PROPERTIES OF THE TEST STATISTIC

In this section we consider the properties of the proposed test. First we calculate the critical values for the test. For this purpose, we generate samples for various sample size up to 50 in the following situation,

$$y_i=x_{i1}+x_{i2}+\cdots+x_{ip}+e_i, \tag{10}$$

in which $e_i\sim N(0,1)$ and the explanatory variables are generated as $x_{ij}\sim N(0,49)$ 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 $p=1,2,3,4$ and sample

size up to 50 is presented in the Table 1. Next, we consider the power of the test for various situation. First, we generate a sample as $e_i \sim N(0,1)$ and $x_{ij} \sim N(0,49)$. Second, to construct outliers in the independent variables space, $(1-\alpha)\times 100\%$ of samples are as in the first. The remaining $\alpha\times 100\%$ are generated as $e_i \sim N(0,1)$ and $x_{ij} \sim N(\mu,49)$. Finally, we make the outliers in response variable space. For this purpose, $(1-\alpha)\times 100\%$ of the samples are as in the first. The remaining $\alpha\times 100\%$ are generated as $e_i \sim N(\mu,1)$ and $x_{ij} \sim N(0,49)$.

Using 1000 replicates for each sampling situation, we compute the power of the test. A summary of our results for a single outlier, various magnitude of outliers, $\mu = 10, 20, 30, 40, 50, 60, 70, 80, 90, 100$, p=1 and sample sizes 25 and 40, are presented in the Table 2 and 3. The results for two outliers, various magnitude of outlier, $\mu = 10, 20, 30, 40, 50, 60, 70, 80, 90, 100$, $p=1$ and sample size 25 are presented in the Table 4. 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
20	1.353	1.288	1.225	1.357	1.298	1.246	1.409	1.355	1.306	1.454	1.385	1.335
25	1.276	1.246	1.200	1.322	1.285	1.241	1.389	1.338	1.287	1.419	1.369	1.319
30	1.263	1.226	1.180	1.269	1.234	1.193	1.342	1.291	1.249	1.368	1.324	1.278
35	1.225	1.191	1.153	1.245	1.208	1.177	1.304	1.265	1.227	1.333	1.278	1.237
40	1.209	1.172	1.146	1.221	1.185	1.164	1.283	1.239	1.195	1.292	1.253	1.213
45	1.197	1.169	1.129	1.214	1.182	1.145	1.256	1.220	1.189	1.274	1.258	1.200
50	1.182	1.153	1.128	1.196	1.168	1.141	1.220	1.195	1.166	1.230	1.199	1.176

Table 2. Estimated power of the proposed test(n=25, p=1, one outlier)

significant level	magnitude of outliers								
	20	30	40	50	60	70	80	90	100
0.1	0.955	0.997	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.05	0.949	0.996	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.01	0.941	0.995	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Table 3. Estimated power of the proposed test(n=40, p=1.one outlier)

significant level	magnitude of outliers								
	20	30	40	50	60	70	80	90	100
0.1	0.974	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.05	0.972	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.01	0.965	0.999	1.00	0.999	1.00	1.00	1.00	1.00	1.00

Table 4. Estimated power 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 4)

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

Finally, we consider the asymptotic distribution of the test statistic. This is obtained by the result of Monte Carlo simulation of 1000 replications under the null hypothesis. For various sample sizes and the number of explanatory variables, Q-Q plots of the test statistics are similar. So a Q-Q plot of the test statistic for sample size 100 in $p=3$ is shown only in Figure1.
Though the extreme quantiles for the test statistic is the greater spread, all of them appear to follow the normal distribution approximately.

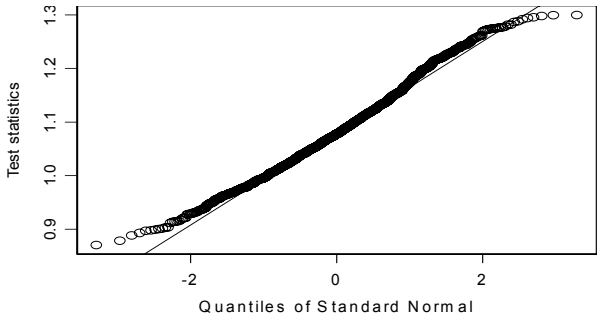

Figure1. Normal probability plot of 1000 test statistics for size 100 in p= 3

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 (Pilot-Plant Data)

The application begins by applying the test to the pilot-plant data given by Daniel and Wood(1971). Rousseew and Leroy(1987) used these data to illustrate the need for robust regression technique. Suppose now that one of the observations has been wrongly recorded. For example, the x-value of the sixth observation has been recorded as 370 instead of 37. This error produces an outlier in the independent variable space. The data appear in the Table 5. The results for the proposed test are in the Table 6.

Table 5. Pilot-Plant data set

index	Extraction(x)	Titration(y)	index	Extraction(x)	Titration(y)
1	123	76	11	138	82
2	109	70	12	105	68
3	62	55	13	159	88
4	104	71	14	75	58
5	57	55	15	88	64
6	370(37)	48	16	164	88
7	44	50	17	169	89
8	100	66	18	167	88
9	16	41	19	149	84
10	28	43	20	167	88

*(37) is original data of pilot-plant data set

Table 6. The proposed test applied to the contaminated pilot-plant data

sample size	observation selected	proposed test statistics	critical values		
			0.01	0.05	0.1
20	6	10.048	1.353	1.288	1.225
19	11	0.8567	1.373	1.298	1.238

In the Table 6, the test is highly significant for observation 6 that wrongly recorded. When the test is applied to the remaining 19 observations, null hypothesis is not rejected. For this example, the proposed test yields a correct result.

Example 2 (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 for the proposed test 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 is the same to conclusion that most people reported.

Table 7. Stackloss data

index	rate (x1)	temper-ature(x2)	acid concen-tration(x3)	stackless (y)	index	rate (x1)	temper-ature(x2)	acid concen-tration(x3)	stackless (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 proposed test applied to the stackloss data

Sample size	observation selected	proposed test statistics	Critical Values		
			0.01	0.05	0.10
21	21	1.7655	1.403	1.345	1.297
20	4	1.5459	1.409	1.355	1.306
19	1	1.4720	1.438	1.355	1.316
18	3	1.6047	1.493	1.404	1.336
17	2	1.236	1.497	1.404	1.336

Example 3 (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 scale-ratio test for the contaminated data. The result is listed in the table 11.

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

In the table 10, 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 11, 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.

Table 10. 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, and RD_i

index i	Based on Lesat squares method													Robust	
	h_{ii}	MD_i^2	r_i/s	t_i	$t(i)$	$CD^2(i)$	DFFITS	CFBETAS(0.447)						r_i/s	RD_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

Table 11. Scale-Ratio 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

The above example demonstrate the performance of the scale-ratio test and is unaffected by masking effects.

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. Daniel, C., and Wood, F. S.,(1971) *Fitting Equations to data,* John Wiley & Sons, New York.
3. Rousseeuw, P. J.,(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. Yohai, V.J. and Zamar(1998) Optimal locally robust M-estimates of regression, *Jour, of statist. Inf. and Planning*

[received date : Jun. 2003, accepted date : Jul. 2003]