Journal of Korean
Data & Information Science Society
2004, Vol. 15, No. 4, pp. 921 ~934

The Detection and Testing of Multiple Outliers in Linear Regression[1]

Jin-Pyo Park[2] · Ruben H. Zamar[3]

Abstract

We consider the problem of identifying and testing outliers in linear regression. First, we consider the scale-ratio tests 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 test statistics and investigate the properties of the test. Next we consider the problem of identifying the outliers. A forward procedure based on the suggested test is proposed and shown to perform fairly well. The forward procedure is unaffected by masking and swamping effects because the test statistics used a robust scale estimate.

Keywords : Forward sequential test, Outliers test, Scale-ratio test

1. Introduction

Consider the linear regression model,

$$y_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{ip}\beta_p + e_{i\,,} \qquad i = 1, 2, \cdots, n \tag{1.1}$$

where the error e_i is assumed to be normally distribution with mean zero and variance σ^2. The aim of multiple regression is to estimate $\beta = (\beta_{0,} \beta_1, \cdots, \beta_p)^T$ from the data $(x_{i1,} x_{i2,} \cdots, x_{ip,} y_i)$. The most popular estimate $\widehat{\beta}$ is least squares estimate. However, it is well known that the outliers can have an extreme effect on the estimate.

[1] The research was supported by Kyungnam University, 2004.

[2] First Author : Professor, Division of computer engineering, Kyungnam University.
E-mail : jppark@eros.kyungnam.ac.kr

[3] Professor, Dept of statistics, British Columbia University, Canada.

An outlier is an observation $(x_{i1}, x_{i2}, \cdots, x_{ip}, y_i)$ which deviates from the pattern of majority of the data.

In lower dimension, graphical technique can be used to detect the outliers. Outliers can be hard to detect, when p exceeds 2, because we can no longer rely on graphical tool. Therefore we have to resort to other methods.

There are two general approaches to dealing with the outliers in regression analysis, outlier diagnostics test and robust methods. Each proceeds the same problem from opposite side. Since the advantages of one method tend to be the disadvantages of the other, we should combine two methods to propose a diagnostic test that is unaffected by masking effects.

In this paper, we propose a robust diagnostic tool to detect and test the outliers in regression context. This tool, which we call scale-ratio tests, is based on the ratio of robust scale estimates and non-robust scale estimate. And then we propose the following forward sequential procedure for identifying 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 estimate of scale, one expects that this procedure will not be affected by masking effects. This is confirmed by numerical examples.

The remaining of the paper is organized as follows. In Section 2 we introduce the scale-ratio test and the forward sequential procedure. In Section 3 we derive the asymptotic distribution of the scale-ratio test under the null hypothesis and calculate the critical values and powers of proposed test. In Section 4 the proposed test and the forward sequential procedure are applied to several real data sets and artificial data sets in order to show their performances. Section 5 contains some concluding remarks.

2. Scale-Ratio Test

The scale-ratio test was proposed to test the effects in 2^k factorial design without replicates by Le and Zamer(1992). They derived the asymptotic distribution of the scale-ratio test as well as condition that a pitman efficient test is obtained.

The scale-ratio test proposed for testing outliers in linear regression is defined as follows. Let ρ_1 be the bisquare function with tuning constants c,

$$\rho_1(x) = \begin{cases} (\frac{x}{c})^6 - 3(\frac{x}{c})^4 + 3(\frac{x}{c})^2 & \text{if } |x| < c \\ 1 & \text{if } |x| \geq c \end{cases}, \tag{2.1}$$

and $\psi_1(x) = \rho_1'(x)$.

$$\psi_1(x)=\begin{cases} \frac{6}{c}\left(\frac{x}{c}\right)^5-\frac{12}{c}\left(\frac{x}{c}\right)^3+\frac{6}{c}\left(\frac{x}{c}\right) & \text{if } |x|<c \\ 0 & \text{if } |x|\geq c \end{cases}. \tag{2.2}$$

For any $\widehat{\beta}$, let $s(\beta)$ be the solution of

$$\frac{1}{n}\sum\rho_1\left(\frac{y_i-x_i\widehat{\beta}}{s}\right)=\frac{1}{2}, \tag{2.3}$$

where $\widehat{\beta}=\underset{\beta}{\arg\min}\; s(\beta)$.

Let ρ_2 be the unbounded function,

$$\rho_2(x)=x^2 \tag{2.4}$$

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

$$\psi_2(x)=2x. \tag{2.5}$$

For any β^*, Let $\sigma(\beta)$ be solution of

$$\frac{1}{n}\sum\rho_2\left(\frac{y_i-x_i\beta^*}{\sigma}\right)=1, \tag{2.6}$$

where $\beta^*=\underset{\beta}{\min}\;\sigma(\beta)$.

Here, s is s-estimate of scale for residuals with a breakdown point 0.5 and σ is the non-robust estimate of scale for residuals since ρ_2 is unbounded.

The scale-ratio test statistics is defined as $R=\sigma/s$. The scale-ratio test tests the hypothesis,

$$\begin{aligned} &H_0\text{: no outliers in data } (x_{i1},x_{i2},\cdots,x_{ip},y_i),\quad i=1,2,\cdots,n \\ &H_1\text{: outliers in data } (x_{i1},x_{i2},\cdots,x_{ip},y_i),\quad i=1,2,\cdots,n. \end{aligned} \tag{2.7}$$

The null hypothesis is rejected for large value of R. However, if the test rejects null hypothesis, 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 fashion to identify the outliers. If the test rejects null hypothesis then the point with the largest $|r_i|$, where $r_i = y_i - \widehat{\beta} x_i$ and $\widehat{\beta}$ is s-estimate of regression coefficients β, is removed and the test is applied again to the $n-1$ remaining data. This procedure is applied iteratively and stops when the test is no longer significant.

The s-estimate in the denominator is required to ensure that the test statistics is sensitive to outliers and that the forward procedure is not affected by possible effects of several outliers.

3. Properties of the Scale-Ratio Test

In this section we consider the properties of the scale-ratio test. First we derive that the asymptotic distribution of the scale-ratio test under the null hypothesis is $N(0, \tau^2)$ where

$$\tau^2 = \frac{\int_{-\infty}^{\infty} [\rho_1(y) - E_{\phi}(\rho_1(y))]^2 \phi(y)\,dy}{\left\{\int_{-\infty}^{\infty} \rho_1'(y) \cdot y\phi(y)\,dy\right\}^2} - \frac{\left\{\int_{-\infty}^{\infty} \rho_1(y) \cdot y^2 \cdot \phi(y)\,dy - \int_{-\infty}^{\infty} \rho_1(y) \cdot y \cdot \phi(y)\,dy\right\}}{\int_{-\infty}^{\infty} \rho_1'(y) \cdot y \cdot \phi(y)\,dy} + \frac{1}{2}, \tag{3.1}$$

and $\phi(\cdot)$ is the probability density function of the standard normal. The proof is sketched in the Appendix.

Next, we calculate the critical values for the scale-ratio test. For this purpose, we generate samples for various sample sizes up to 50 in the following situation,

$$y_i = x_{i1} + x_{i2} + \cdots + x_{ip} + e_i, \tag{3.2}$$

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 scale-ratio test. A summary of our results for $p = 1, 2, 3, 4$ and sample size up to 50 is presented in Table 1.

<Table 1> Critical values for the scale-ratio 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

For large sample size, the asymptotic approximate,

$$C_{\alpha} = 1 + 0.6539 n^{-1/2} Z\alpha \tag{3.3}$$

can be used. Where Z_{α} is 100(1- α)- th percentile of standard normal distribution and n is the sample size used to compute the test statistics. When n equals 50, $C_{0.01} = 1.215$, $C_{0.025} = 1.181$, $C_{0.05} = 1.152$ and $C_{0.01} = 1.119$. The asymptotic approximation can also be used to calculate approximate p-values.

Finally, we consider the power of the scale-ratio test for various situation. For this purpose, first, we generate sample as $e_i \sim N(0,1)$ and $x_{ij} \sim N(0,100)$. 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 $x_{ij} \sim N(\mu, 100)$. 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 $x_{ij} \sim N(0,100)$. Using 1000 replicates for each sampling situation, we compute the power of the scale-ratio 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, and sample sizes 25 and 40, is 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 Table 4. The power of the scale ratio test increases with sample size and magnitude of outliers.

<Table 2> Estimated power of the scale-ratio test(n=25, p=1, one outlier)

Significant level	Magnitude of outliers								
	20	30	40	50	60	70	80	90	100
0.10	0.955	0.992	0.997	1.00	1.00	1.00	1.00	1.00	1.00
0.05	0.949	0.986	0.996	1.00	1.00	1.00	1.00	1.00	1.00
0.01	0.943	0.985	0.995	0.996	1.00	1.00	1.00	1.00	1.00

<Table 3> Estimated power of the scale-ratio test(n=40, p=1, one outlier)

Significant level	Magnitude of outliers								
	20	30	40	50	60	70	80	90	100
0.1	0.975	0.999	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.05	0.969	0.995	0.999	1.00	1.00	1.00	1.00	1.00	1.00
0.01	0.957	0.989	0.996	0.999	1.00	1.00	1.00	1.00	1.00

<Table 4> Estimated power of the scale-ratio 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	0.955	0.933	0.928	0.992	0.991	0.985	0.998	0.997	0.996	1.00	1.00	1.00
30	0.956	0.945	0.934	0.993	0.992	0.99	0.999	0.998	0.997	1.00	1.00	1.00
40	0.959	0.947	0.940	0.999	0.998	0.995	0.999	0.999	0.999	1.00	1.00	1.00
50	0.960	0.952	0.944	0.999	0.998	0.995	1.00	0.999	0.999	1.00	1.00	1.00
60	0.963	0.956	0.954	1.00	0.999	0.996	1.00	1.00	1.00	1.00	1.00	1.00
70	0.970	0.963	0.956	1.00	1.00	0.997	1.00	1.00	1.00	1.00	1.00	1.00
80	0.971	0.968	0.957	1.00	1.00	0.998	1.00	1.00	1.00	1.00	1.00	1.00
90	0.973	0.970	0.961	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
100	0.982	0.972	0.965	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

(continue)

<Table 4> Estimated power of the scale-ratio test(n=25, p=1, two outliers)(continue)

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

4. Applications of the Scale-Ratio test

In this section, the scale-ratio test is applied to several data sets for the purpose of outlier detection. The application begins by applying the scale-ratio test to the pilot-plant data given 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 is appeared in Table 5. The result for the scale-ratio test is in Table 6.

<Table 5> Pilot-Plant Data set

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

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

<Table 6> Scale-Ratio Test Applied to the contaminated pilot-plant data

Sample size	Observation selected	Scale-ratio test statistics	critical values		
			0.01	0.05	0.1
20	6	10.049	1.353	1.288	1.225
19	20	0.8496	1.373	1.298	1.238

In Table 6, the test is highly significant for observation 6 that is wrongly recorded. When the test is applied to the remaining 19, it is not rejected. For this example, the scale-ratio test yields a correct result.

The second application for outliers detection 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 observations 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 scale-ratio test appears in Table 8. In 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. But it does not detect observation 2 as outlier. This result is the same to conclusion that most people reported.

<Table 7> Stackloss Data

Obs	Rate(x1)	Temperature(x2)	Acid concentration(x3)	Stackloss(y)
1	80	27	89	42
2	80	27	88	37
3	75	25	90	37
4	62	24	87	28
5	62	22	87	18
6	62	23	87	18
7	62	24	93	19
8	62	24	93	20
9	58	23	87	15
10	58	18	80	14
11	58	18	89	14
12	58	17	88	13
13	58	18	82	11
14	58	19	93	12
15	50	18	89	8
16	50	18	86	7
17	50	19	72	8
18	50	19	79	8
19	50	20	80	9
20	56	20	82	15
21	70	20	91	15

<Table 8> Scale-Ratio Test Applied to the stackloss Data

Sample size	Observation selected	Scale ratio statistics	Critical Values		
			0.01	0.05	0.10
21	21	1.766	1.403	1.345	1.297
20	4	1.546	1.409	1.355	1.306
19	1	1.472	1.438	1.355	1.316
18	3	1.606	1.493	1.404	1.336
17	2	1.236	1.497	1.404	1.336

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 are the outliers that are not outlying in any of the individual variables.

The results for comparing the various diagnostic appear in Table 10. The contaminated data are shown in Table 9. We applied the scale-ratio test for the contaminated data. The result is listed in 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 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 Table 11, dbservation 19 is the most extreme followed by observation 6, observation 8, observation 4 and observation 5. But the test does not reject observation 5 at significant 0.01. This test identifies 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 least 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	1.783	1.484	1.415	1.365
19	6	1.948	1.518	1.445	1.395
18	8	2.068	1.547	1.472	1.412
17	4	2.635	1.577	1.492	1.433
16	5	1.227	1.671	1.522	1.463

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

5. 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 sets 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.

APPENDIX : A sketch of proof for the Asymptotic distribution of scale-ratio test

To derive the asymptotic distribution of the scale-ratio test under the null hypothsis, the Taylor expansion of (2.3) about $\widehat{\beta}=\beta_0$ and $s=s_0$ gives,

$$\frac{1}{2}=\frac{1}{n}\sum\rho_1\left(\frac{y_i-\beta_0^T x_i}{s_0}\right)-\left(\frac{1}{ns_0}\sum\rho_1'\left(\frac{y_i-\beta_0^T x_i}{s_0}\right)x_i\right)(\widehat{\beta}-\beta_0)$$
$$-\frac{1}{ns_0}\sum\rho_1'\left(\frac{y_i-\beta_0^T x_i}{s_0}\right)\left(\frac{y_i-\beta_0^T x_i}{s_0}\right)(s-s_0)+\cdots. \tag{A-1}$$

Since $\sqrt{n}(\widehat{\beta}-\beta_0)$ is asymptotically normal, the law of large number implies

$$n^{-\frac{1}{2}}\left[\frac{1}{s_0}\sum\rho_1'\left(\frac{y_i-\beta_0^T x_i}{s_0}\right)x_i\right)\Bigg](\widehat{\beta}-\beta_0)\rightarrow 0 \text{ in probability as } n\rightarrow\infty \tag{A-2}$$

and

$$\frac{1}{n}\sum\rho_1'\left(\frac{y_i-\beta_0^T x_i}{s_0}\right)\left(\frac{y_i-\beta_0^T x_i}{s_0}\right)\rightarrow E\rho_1'\left(\frac{y-\beta_0^T x}{s_0}\right)\left(\frac{y-\beta_0^T x}{s_0}\right)$$
$$\text{almost surely as } n\rightarrow\infty. \tag{A-3}$$

Thus, using (A-1) - (A-3) we have the following asymptotic equivalence

$$\sqrt{n}(s-s_0)\approx s_0\frac{\sqrt{n}\left[\frac{1}{n}\sum\rho_1\left(\frac{y_i-\beta^T x_i}{s_0}\right)-\frac{1}{2}\right]}{E\rho_1'\left(\frac{y-\beta^T x}{s_0}\right)\left(\frac{y-\beta^T x}{s_0}\right)}. \tag{A-4}$$

Without loss of generality we assume that $\beta_0^T=0$, $s_0=1$

$$\sqrt{n}(s-1)\approx\frac{\sqrt{n}\left[\frac{1}{n}\sum\rho_1(y_i)-\frac{1}{2}\right]}{E\rho_1'(y)\cdot y}. \tag{A-5}$$

By central limit theorem

$$\sqrt{n}(s-1) \rightarrow N(0,\, V), \tag{A-6}$$

where

$$V = \frac{\int [\rho_1(y) - E_{\Phi}(\rho_1(y))]^2 \phi(y)\,dy}{\left\{ \int \rho_1{}'(y) \cdot y\,\phi(y)\,dy \right\}^2}. \tag{A-7}$$

Similarly

$$\sqrt{n}(\sigma-1) \rightarrow N(0, \frac{1}{2}). \tag{A-8}$$

Thus, using (A-6) and (A-8) we have the following asymptotic distribution of $\sqrt{n}(s-\sigma)$

$$\sqrt{n}(s-\sigma) \rightarrow N(0, \tau^2), \tag{A-9}$$

where

$$\tau^2 = \frac{\int_{-\infty}^{\infty} [\rho_1(y) - E_{\Phi}(\rho_1(y))]^2 \phi(y)\,dy}{\left\{ \int_{-\infty}^{\infty} \rho_1'(y) \cdot y\phi(y)\,dy \right\}^2} - \frac{\left\{ \int_{-\infty}^{\infty} \rho_1(y) \cdot y^2 \cdot \phi(y)\,dy - \int_{-\infty}^{\infty} \rho_1(y) \cdot y \cdot \phi(y)\,dy \right\}}{\int_{-\infty}^{\infty} \rho_1{}'(y) \cdot y \cdot \phi(y)\,dy} + \frac{1}{2}. \tag{A-10}$$

Moreover, the test statistics $n^{\frac{1}{2}}\{(\sigma/s)-1\}$ and $\sqrt{n}(\sigma-s)$ are equivalent under null hypothesis. Hence we can have the following conclusion

$$n^{\frac{1}{2}}\{(\sigma/s)-1\} \rightarrow N(0, \tau^2). \tag{A-11}$$

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. Draper, N. R., And Smith, H.,(1966), *Applied Regression Analysis,* John Wiley & Sons, New York.

4. Nhu D. Le and Ruben H. Zammer,(1992), A Global Test for Effects in 2^k Factorial Design without Replicates, *J. Statist. Comput. Simul., 41, 41-54.*

5. Rousseeuw, P. J., and Leroy, A. M.,(1987), *Robust regression and outlier detection,* John Wiley & Sons, New York.

[received date : Jul. 2004, accepted date : Oct. 2004]