Journal of the Korean
Data & Information Science Society
2000, Vol. 11, No. 2, pp. 201 ~ 215

The Identification Of Multiple Outliers [1]

Jin-Pyo Park [2]

Abstract

The classical method for regression analysis is the least squares method. However, if the data contain significant outliers, the least squares estimator can be broken down by outliers. To remedy this problem, the robust methods are important complement to the least squares method. Robust methods down weighs or completely ignore the outliers. This is not always best because the outliers can contain some very important information about the population. If they can be detected, the outliers can be further inspected and appropriate action can be taken based on the results. In this paper, I propose a sequential outlier test to identify outliers. It is based on the nonrobust estimate and the robust estimate of scatter of a robust regression residuals and is applied in forward procedure, removing the most extreme data at each step, until the test fails to detect outliers. Unlike other forward procedures, the present one is unaffected by swamping or masking effects because the statistics is based on the robust regression residuals. I show the asymptotic distribution of the test statistics and apply the test to several real data and simulated data for the test to be shown to perform fairly well.

Key Words and Phrases: MM-estimates, Outliers test, forward sequential test

I. INTRODUCTION

It is well known that outliers can have an extreme effect on least squares estimate. Intuitively, an outlier is an observation $(x_{i1}, \cdots x_{ip}, y_i)$ which deviate from the linear relation followed by the majority of the data. In regression model,

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

the outliers are classified into two categories, the outliers in the y-direction and the outliers in the x-direction. Especially the outliers in the x-direction are called

[1]The research was partially supported by Kyungnam University Sabbatical year, 1999.
[2]Professor, Division of information & communication engineering, Kyungnam University

leverage points. The non-outlying data will be referred to as the good data. It is assumed that the good data contains more data than 50% of the observations in the sample.

In lower dimensions, graphical techniques can be used to detect the outliers. When the regression model has less than three independent variables, the outliers can be detected by scatterplots and spin plots. But the degree of outlyingness is based on the judgement of the researcher.

However, once the independent variable is more than two, it is difficult to detect the outliers by graphical tool. We have to resort to other methods.

There are two general approaches to identify the outliers, diagnostics and robust methods. Each proceeds the same problem from opposite side. Diagnostics identifies the outliers and allows the researcher to decide how the outliers should be dealt with. But it is affected by the masking and swamping effects.

Robust procedures are immune to the masking and swamping effects. Robust approach usually fits a regression that does justice to the majority of the data and is immune to the masking and swamping effects. But it usually lacks efficiency. Since each approach is the advantage which the other is the disadvantage, we should combine the two methods to propose a diagnostic test that is unaffected by the masking and swamping effects.

In this paper, I will propose test statistics to test the outliers in regression context. This test is based on the sensitive estimate and the robust estimate of scatter of robust regression residuals. This test is applied sequentially in forward procedure to not only identify the outliers but also to indicate the number present as well. Furthermore, the test should be applied until it fails to detect the presence of an outlier because it should not be fooled by masked outliers. The robust regression residuals can be measured in several ways. Huber(1973) introduced M-estimates with monotone psi-function and Mallow(1975) proposed GM-estimate. But these estimates had breakdown point 0%. Siegel(1982), Rousseeuw(1984), Rousseeuw and Yohai(1984) suggested several estimates with high breakdown point 0.5 such as the repeated median estimate, the least median squares, the least trimmed squares and S-estimate. However these estimates are highly inefficient when all the observations satisfy the regression model with normal error.

Yohai(1987) proposed MM-estimate having simultaneously high breakdown point 0.5 and high efficiency under normal error. I use MM-estimate in this paper. The properties of the proposed test statistics is investigated through Monte Carlo simulations. These provide evidence about the asymptotic distribution, power and other properties of the test. Finally this test is applied to several real and artificial data sets in order that this is shown to perform fairly well.

II. PROPOSED OUTLIER TEST STATISTICS

In this paper, I propose the test statistics based on the ratio of two estimate of scatter of a robust regression residuals. I use the MM-estimate proposed by Yohai(1985) as robust regression residuals, because this has simultaneously high breakdown point and high efficiency under normal error. First I recall the definition of MM-estimate. The MM-estimate is defined in three stages. In the first stage a high breakdown estimate $\boldsymbol{\beta}_0^*$ is calculated. Then compute residuals

$$r_i(\boldsymbol{\beta}_0^*) = y_i - \boldsymbol{\beta}_0^* \mathbf{x_i} \qquad 1 \le i \le n$$

and compute the M-estimate of scale $s_n = s(r_i(\boldsymbol{\beta}_0^*))$. Finally the MM-estimate, $\widehat{\boldsymbol{\beta}_{MM}}$, is defined as any solution of

$$\sum_{i=1}^{n} \Psi(r_i(\widehat{\boldsymbol{\beta}})/\mathbf{s_n})\mathbf{x_i} = \mathbf{0},$$

which satisfies

$$S(\widehat{\boldsymbol{\beta}}) \le S(\boldsymbol{\beta}_0^*)$$

where

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} \rho(r_i(\boldsymbol{\beta})/s_n)$$

The influence function for MM-estimates is given by

$$IF_{MM}(\mathbf{X}, \mathbf{e}) = \frac{\sigma}{E(\boldsymbol{\Psi}'(e/\sigma))}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Psi},$$

where $\boldsymbol{\Psi} = (\Psi(e_1/\sigma), \Psi(e_2/\sigma), \cdots, \Psi(e_n/\sigma))'$.

The asymptotic representation of MM-estimate is given by

$$\widehat{\boldsymbol{\beta}_{\mathbf{MM}}} = \boldsymbol{\beta} + IF_{MM}(\mathbf{X}, \mathbf{e}) + O_p(n^{-1/2}),$$

and the residuals for robust fit is given by

$$r_{MMi} = y_i - \mathbf{x_i'}\widehat{\boldsymbol{\beta}_{\mathbf{MM}}}.$$

Let S_1 and S_2 be two estimates of scatter of r_{MMi}. Here S_2 is a standard deviation of r_{MMi} and S_1 is a median absolute deviation of r_{MMi}. The test statistics is defined as $R = S_2/S_1$. The test tests the following hypothesis

H_0 : no outliers in data $(x_{i1}, x_{i2}, \cdots, x_{ip}, y_i) \qquad i = 1, 2, \cdots, n$

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

The null hypothesis is rejected for large of R. However, if the test rejects, there is no indication of how many or which points are outliers. This problem is solved by applying the test in a forward sequential fashion. If the test rejects the null hypothesis then the point with the largest $D = |sort(r_{MMi}) - Med(r_{MMi})|$ is removed, where $sort(r_{MMi})$ is the sort of r_{MMi} and $Med(r_{MMi})$ is the median of r_{MMi}. The test is re-applied to the remaining data. This procedure is repeated until the test does not reject the null hypothesis. The median absolute deviation of r_{MMi} 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 masking effects and swamping effects of several outliers.

III. PROPERTIES OF THE TEST STATISTICS

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

$$y_i = x_{i1} + x_{i2} + \cdots + x_{ip} + e_i, \qquad i = 1, 2, \cdots, n,$$

in which $e_i \sim N(0, 1)$ and explanatory variables are generated as $x_{ij} \sim N(0, 100)$ for $j = 1, 2, \cdots, p$. Using 1000 replicates for each sampling situation, I compute the critical values for the test statistics. A summary of my results for $p = 1, 2, 3, 4$ and sample size up to 50 are presented in the Table 1.

Next, I consider the asymptotic distribution of the test statistics. This is Monte Carlo simulation of 1000 replications under the null hypothesis. For various sample sizes and number of explanatory variables, Q-Q plots of the test statistics is approximately laid on a straight line. Q-Q plots for a sample size of 50 in $p = 1, 2, 3, 4$ are shown in Figure 1. The plots to comparison cumulative distribution function are shown in Figure 2. In Figure 2, the solid line is the empirical distribution function of the test statistics and the dot line is distribution function of normal distribution. The solid line is approximately laid on dot line. The parameters for the normal distribution are estimated from 1000 test statistics. All of them appear to follow the normal distribution quite approximately.

Finally, I consider the power of the proposed test statistics for various situation. For this purpose, First, I 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, I 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, I compute the power of the proposed test statistics. A summary of our results for a single outlier, various magnitude of outliers, $\mu = 20, 30, 40, 50, 60, 70, 80, 90, 100$, and sample sizes 25, 40 are presented in the table 2 and 3. A summary for more than two outliers are not presented in this paper but the power of the test are high.

Table1 Critical values for the test statistics

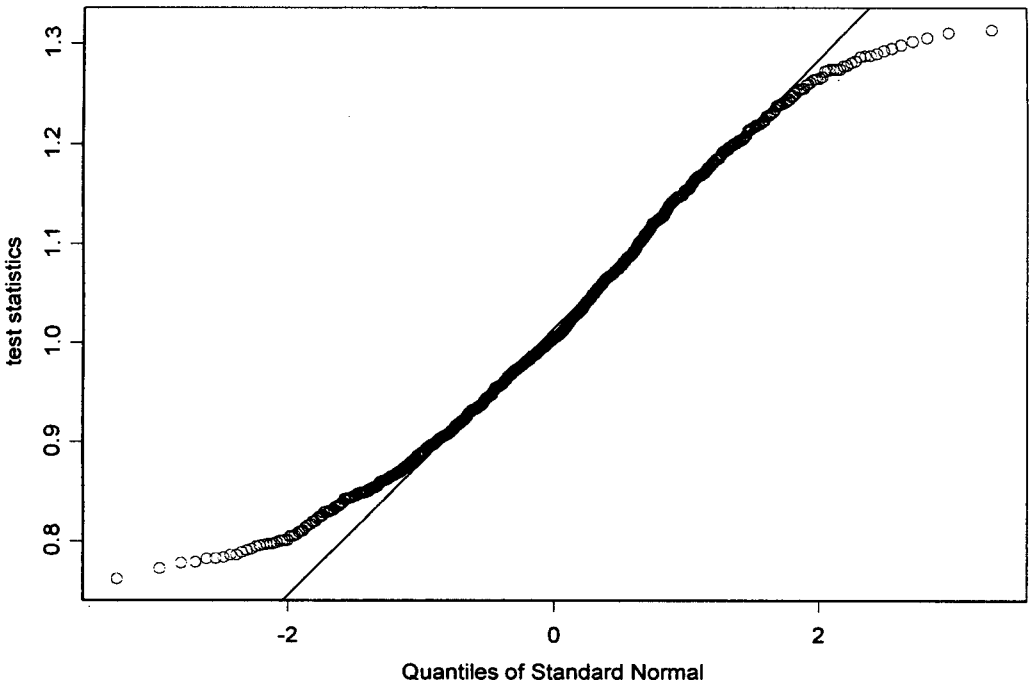
sample sizes	Number of explanatory variable											
	1			2			3			4		
	α level			α level			α level			α level		
	0.01	0.05	0.10	0.01	0.05	0.10	0.01	0.05	0.10	0.01	0.05	0.10
15	1.9583	1.7460	1.5901	1.7926	1.6790	1.5564	1.8671	1.7492	1.6376	1.8646	1.7511	1.6295
20	1.7992	1.5699	1.4645	1.7347	1.6077	1.5044	1.6827	1.5826	1.4980	1.6484	1.5592	1.4908
25	1.6435	1.4799	1.3553	1.6528	1.5078	1.4138	1.5545	1.4731	1.3910	1.5212	1.4480	1.3834
30	1.5990	1.4429	1.3179	1.5278	1.4300	1.3552	1.4565	1.3836	1.3237	1.4295	1.3783	1.3377
35	1.5696	1.3721	1.3027	1.4150	1.3526	1.2852	1.3448	1.3061	1.2594	1.3446	1.3052	1.2701
40	1.4334	1.3247	1.2499	1.3655	1.2983	1.2428	1.3110	1.2775	1.2363	1.2561	1.2319	1.2022
45	1.3462	1.2901	1.2144	1.3437	1.2820	1.2258	1.2877	1.2540	1.2084	1.2541	1.2198	1.1913
50	1.2953	1.2401	1.1960	1.2941	1.2468	1.1974	1.2518	1.2245	1.1952	1.2193	1.1935	1.1672

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

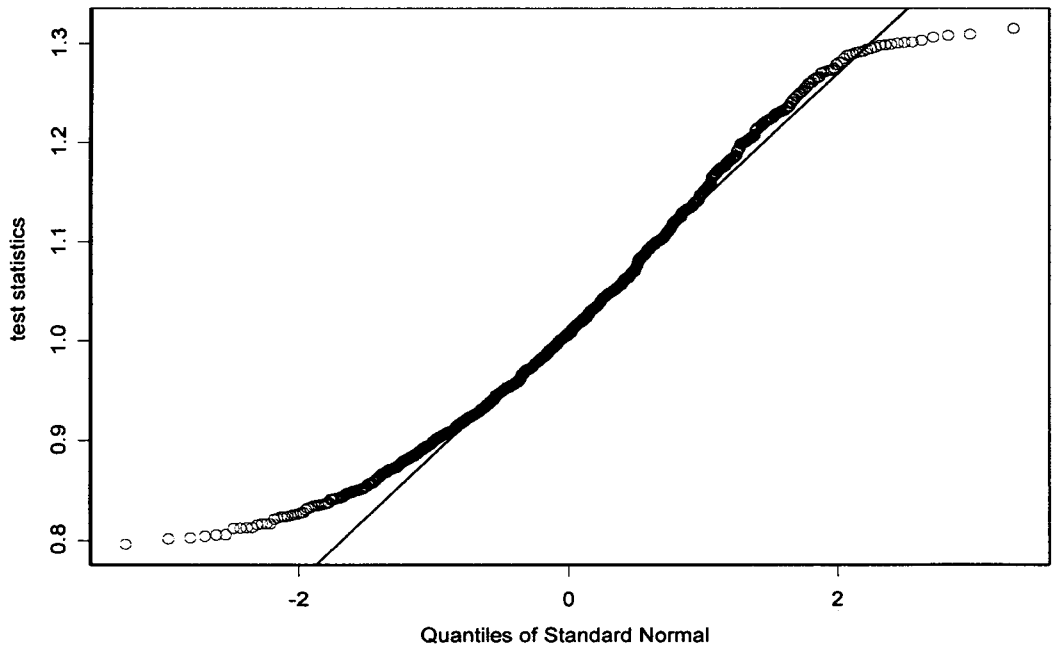
significant level	magnitude of outliers								
	20	30	40	50	60	70	80	90	100
0.1	0.955	0.989	0.998	1.00	1.00	1.00	1.00	1.00	1.00
0.05	0.949	0.986	0.997	1.00	1.00	1.00	1.00	1.00	1.00
0.01	0.943	0.975	0.995	0.998	1.00	1.00	1.00	1.00	1.00

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

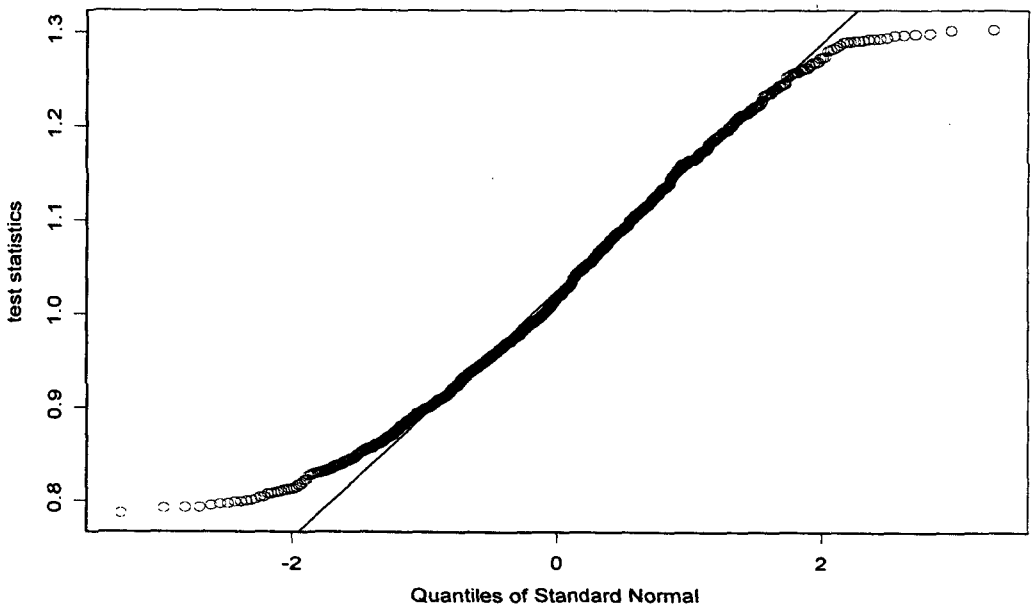
significant level	magnitude of outliers								
	20	30	40	50	60	70	80	90	100
0.1	0.965	0.992	0.999	1.00	1.00	1.00	1.00	1.00	1.00
0.05	0.959	0.990	0.998	1.00	1.00	1.00	1.00	1.00	1.00
0.01	0.947	0.980	0.996	0.998	1.00	1.00	1.00	1.00	1.00


test statistics
0.8
0.9
1.0
1.1
1.2
1.3
-2
0
2
Quantiles of Standard Normal

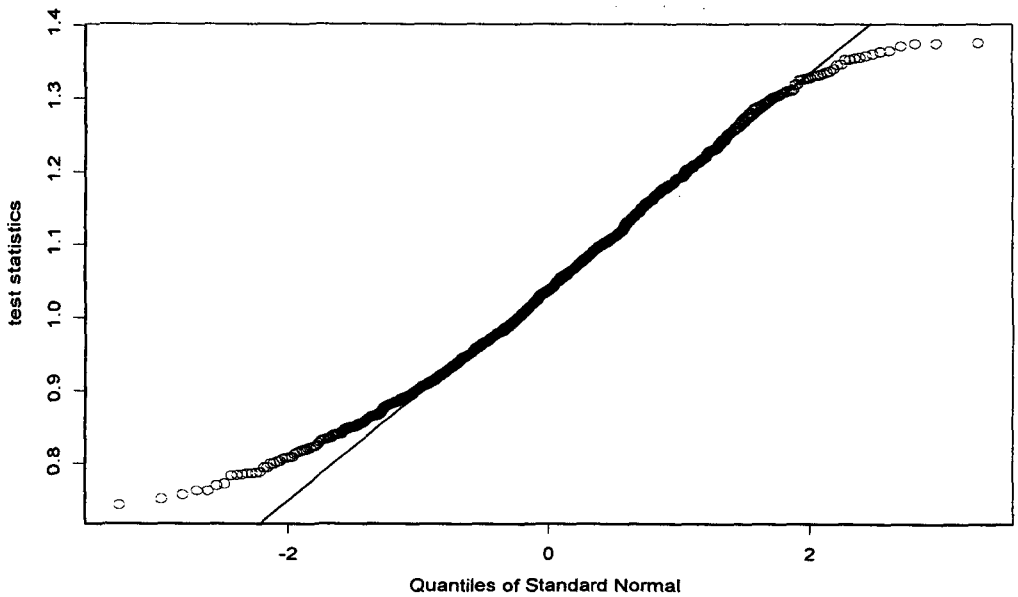
(a) p=1


test statistics
0.8
0.9
1.0
1.1
1.2
1.3
-2
0
2
Quantiles of Standard Normal

(b) p=2

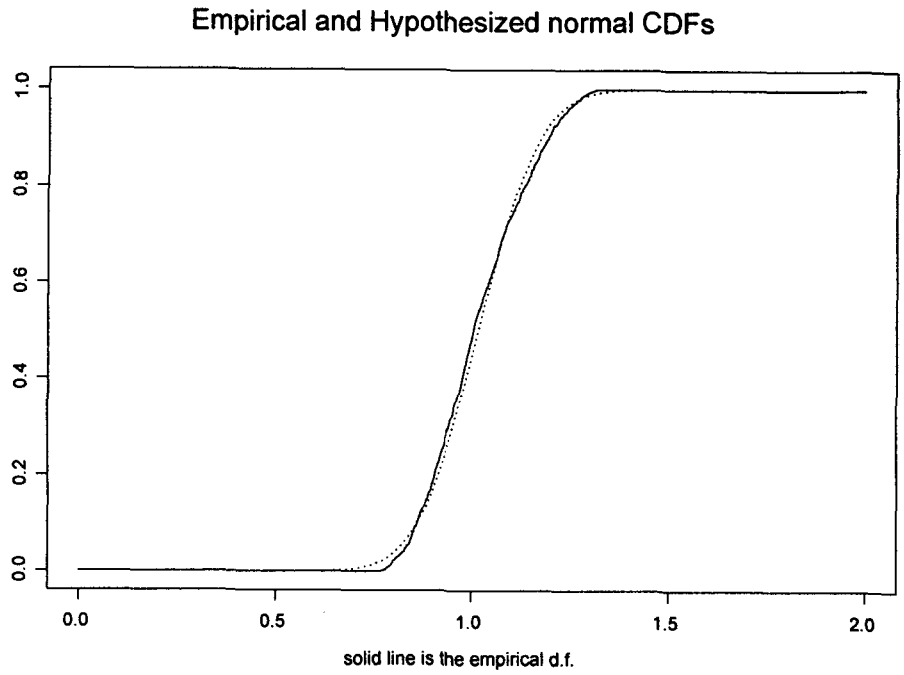


(c) p=3

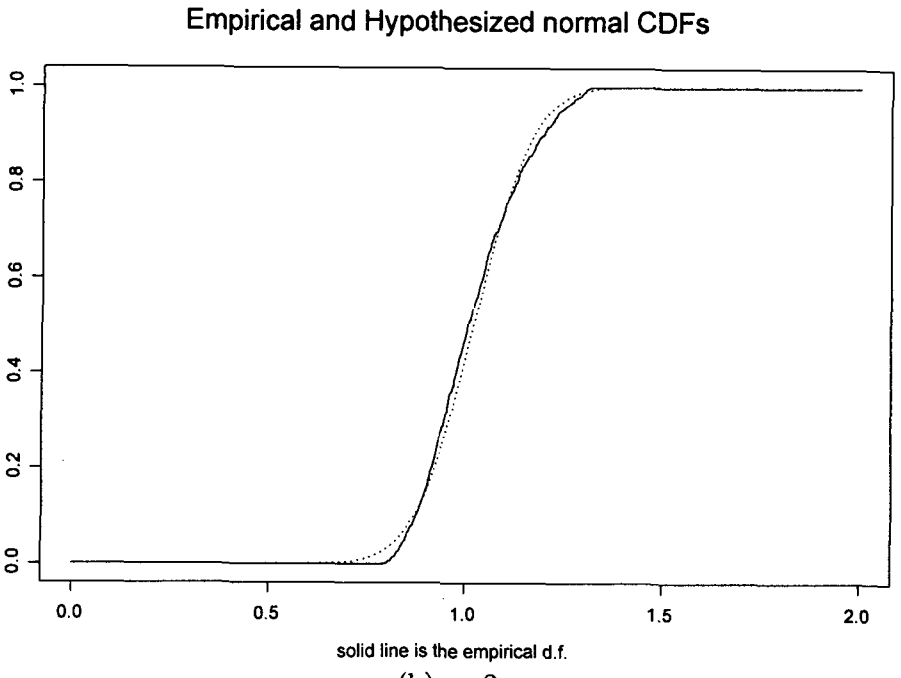


(d) p=4

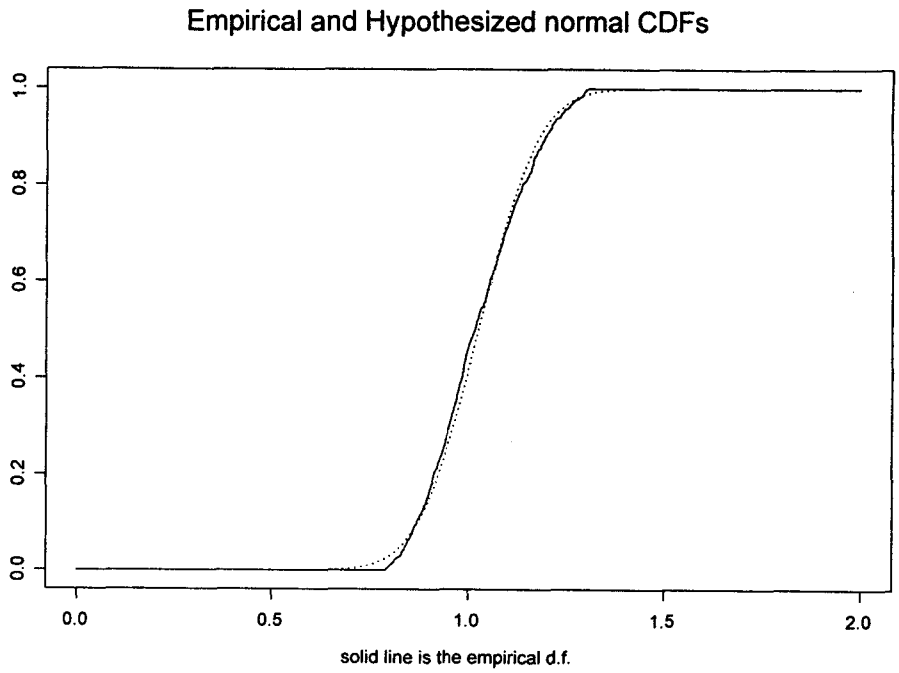
Figure 1 Normal probability plot of 1000 test statistics for p=1, 2, 3, 4


Empirical and Hypothesized normal CDFs
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.5
1.0
1.5
2.0
solid line is the empirical d.f.

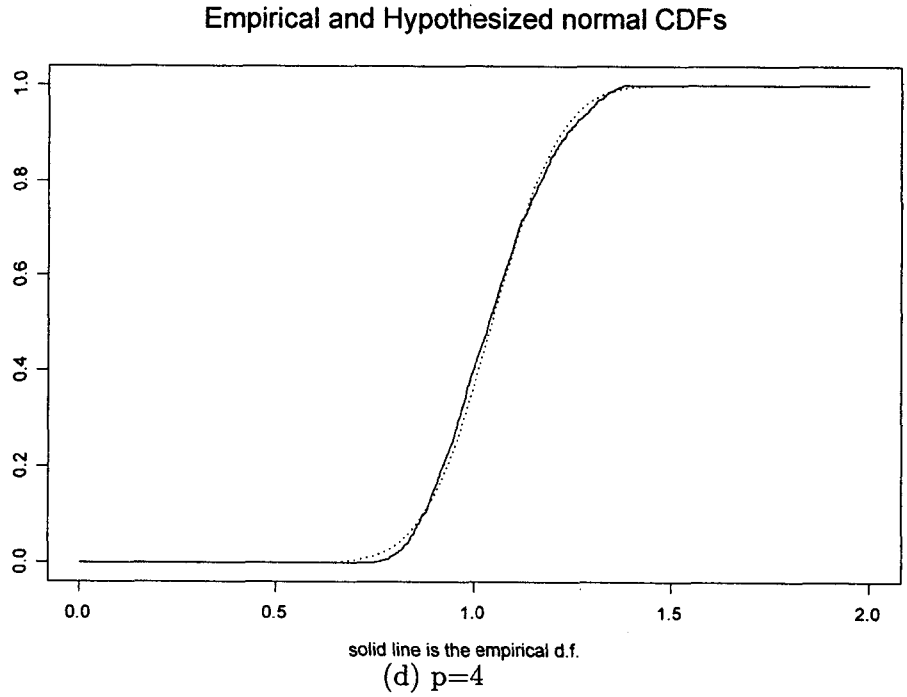
(a) p=1


Empirical and Hypothesized normal CDFs
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.5
1.0
1.5
2.0
solid line is the empirical d.f.

(b) p=2



(c) p=3



(d) p=4

Figure 2 Empirical CDFs of 1000 test statistics and Hypothesized Normal CDFs

IV. APPLICATIONS OF THE PROPOSED TEST

In this section, the proposed test is applied to several data sets to detect outliers. This includes an estimate of the number of outliers present as well as the identity of outliers. The application begins by applying the proposed test to pilot-plant data given Daniel and Wood(1971). Rousseeuw and Leroy(1987) used these data to illustrates the need for a robust regression technique. The data appear in the Table 4. Suppose now that one of the observations has been wrongly recorded. The x-value of the sixth observation have been recorded as 370 instead of 37. The results for the proposed test appear in the Table 5. In Table 5, observation 6 is the most extreme followed by observation 20. The test detects that observation 6 is outlier.

Table 4. Pilot-Plant Data

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

Table 5. Proposed test applied to the pilot-plant Data

Sample size	observation selected	scale ratio statistics	Critical Values		
			0.01	0.05	0.10
20	6	17.4901	1.7992	1.5699	1.4645
19	20	0.8496	1.8234	1.6254	1.5321

The next application for outliers detection comes from the Brownlee(1965). These data is well-known stackloss data set. It has been 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 is shown in the Table 6. The results for the proposed test appear in the Table 7. In the Table 7, observation 21 is the most extreme followed by observation 4, observation 1, observation 3 and observation 2. The test identifies observations 21, 4, 1, 3 as outliers. But it does not detect observation 2 as outlier. This result is the same to conclusion that most people reported.

Table 6. Stackloss Data

observation	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 7. The proposed test applied to the stackloss Data

Sample size	observation selected	scale ratio statistics	Critical Values		
			0.01	0.05	0.10
21	21	2.3451	1.6603	1.5617	1.4872
20	4	2.1115	1.6827	1.5826	1.4980
19	1	2.6016	1.7179	1.6159	1.5197
18	3	2.1626	1.7453	1.6259	1.5286
17	2	1.3214	1.7746	1.6521	1.5532

Let us look at a final 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 9. The contaminated data is shown in the table 8. I applied the proposed test for the contaminated data. The result is listed in the table 10.

Table 8. 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 9, 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 identifies 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 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 9. 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

	Based on Lesat squares method						
index	h_{ii}	MD_i^2	r_i/s	t_i	t(i)	$CD^2(i)$	DFFITS
i	0.600	11.07	2.50	2.50	2.50	1.00	1.095
1	0.278	4.327	-0.73	-0.85	-0.84	0.047	-0.524
2	0.132	1.552	0.05	0.05	0.05	0.000	0.019
3	0.220	3.224	1.24	1.41	1.46	0.093	0.776
4	0.258	3.959	0.35	0.41	0.40	0.010	0.236
5	0.223	3.277	1.00	1.14	1.15	0.062	0.615
6	0.259	3.974	-0.45	-0.53	-0.51	0.016	-0.302
7	0.530	9.124	0.91	1.32	1.36	0.329	1.448
8	0.289	4.536	-0.03	-0.04	-0.04	0.000	-0.025
9	0.348	5.665	-0.40	-0.49	-0.48	0.021	-0.348
10	0.449	7.588	-0.42	-0.56	-0.55	0.043	-0.492
11	0.317	5.075	1.99	2.40	3.02	0.447	2.059
12	0.410	6.833	-1.20	-1.56	-1.65	0.281	-1.376
13	0.287	4.506	-0.49	-0.58	-0.56	0.022	-0.356
14	0.129	1.500	-1.26	-1.35	-1.40	0.045	-0.537
15	0.152	1.945	-0.59	-0.64	-0.62	0.012	-0.264
16	0.526	9.049	0.52	0.76	0.75	0.107	0.789
17	0.289	4.548	-0.25	-0.30	-0.29	0.006	-0.187
18	0.294	4.637	0.28	0.34	0.33	0.008	0.211
19	0.292	4.599	-1.08	-1.29	-1.32	0.114	-0.849
20	0.318	5.084	0.55	0.66	0.65	0.034	0.441

Table 9. Continued

index	Based on Lesat squares method						Robust	
	CFBETAS(0.447)						r_i/s	RD_i
i	β_1	β_2	β_3	β_4	β_5	Const.	2.50	2.50
1	-0.004	0.055	0.328	-0.052	0.215	-0.347	-0.16	0.798
2	0.009	0.002	-0.005	0.002	0.000	-0.003	0.00	0.701
3	-0.651	-0.523	-0.206	-0.429	0.549	-0.356	0.55	0.577
4	0.035	-0.049	0.015	-0.105	0.118	-0.074	-14.79	3.938
5	0.286	-0.517	0.164	-0.388	0.437	-0.244	1.75	0.605
6	-0.053	0.037	0.035	0.130	-0.113	0.050	-17.68	4.520
7	-0.956	0.424	0.521	0.133	-0.964	1.027	0.73	1.421
8	0.011	-0.012	0.005	-0.005	0.006	-0.005	-17.31	4.466
9	0.052	0.105	-0.224	0.161	0.007	-0.075	-0.73	1.243
10	-0.008	-0.198	-0.256	-0.137	-0.029	0.257	-0.40	1.267
11	0.425	0.970	0.748	0.198	-0.800	0.521	0.00	1.258
12	-0.597	0.013	0.556	0.359	0.368	-0.566	-1.88	1.030
13	-0.098	0.045	-0.251	0.106	-0.121	0.180	0.00	1.015
14	-0.169	0.228	0.178	-0.006	-0.103	0.021	-1.30	0.668
15	0.148	-0.061	-0.011	-0.162	0.108	-0.073	-0.34	0.465
16	-0.529	0.559	-0.052	0.745	-0.432	0.122	0.00	0.865
17	-0.019	0.019	-0.044	-0.055	-0.086	0.133	0.00	0.802
18	-0.062	-0.096	0.081	-0.024	0.045	-0.002	-0.21	0.985
19	0.195	-0.287	0.231	-0.024	0.079	-0.128	-20.84	5.201
20	0.092	-0.154	-0.305	0.037	0.046	0.064	0.00	0.816

Table 10. The proposed test for the Data in table 8

Sample size	observation selected	scale ratio statistics	Critical Values		
			0.01	0.05	0.10
20	19	4.783	2.484	2.291	2.287
19	6	3.948	2.518	2.445	2.395
18	8	5.068	2.547	2.472	2.412
17	4	5.635	2.577	2.492	2.433
16	5	1.227	2.671	2.522	2.463

The above example demonstrate the performance of the proposed test and is unaffected by masking effects.

V. CONCLUDING REMARKS

The Monte Carle results and numerical examples shown in this paper suggest that the proposed test provides a conservative and fairly powerful method for the detection of the outlires in linear regression. An important feature of my method is that the results can be objectively interpreted. The forward sequential procedures constitute a natural and simple approach for identifying outliers. However if the forward sequential procedure is based on a classical test statistics, then it can be affected by masking effects or swamping effects, whereas if the forward sequential procedure is based on a robust estimate of scatter of robust residuals, like the proposed test, it is not affected by masking effects and swamping effects.

References

1. Huber, P. J.(1973). Robust regression: Asymptotics, conjectures and Monte Carlo, *The Annals of Statistics*, 1, 799-821

2. Mallows, C. L.(1975). On some topics in robustness, unpublished memorandum, Bell Telephone Laboratories, Murray Hill, NJ. [1.2]

3. Siegel, A. F.(1982). Robust regression using repeated median, *Biometrika*, 69, 242-244

4. Rousseeuw, P. J.(1984). Least median of squares regression, *Journal of the American Statistical Association* , 79, 871-8804.

5. Rousseeuw, P. J., and Yohai, V.(1984). Robust regression by means of S-estimators, Lecture Notes in Statistics No. 26, Springer Verlag, New York, 256-272.

6. Yohai, V.J.(1985). High breakdown-point and high efficiency robust estimates for regression, *The Annals of Statistics*

7. Daniel, C., and Wood, F. S.(1971). *Fitting Equations to data,* John Wiley & Sons, New York.

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

9. Brownlee, K. A., (1965). *Statistical Theory and Methodology in Science and Engineering,* 2nd ed., John Wiley & Sons, New York

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