Journal of Korean
Data & Information Science Society
2002, Vol. 13, No.2 pp. 271 284

Robust inference for linear regression model based on weighted least squares[1)]

Jin-Pyo Park[2)]

Abstract

In this paper we consider the robust inference for the parameter of linear regression model based on weighted least squares. First we consider the sequential test of multiple outliers. Next we suggest the way to assign a weight to each observation (x_i, y_i) and recommend the robust inference for linear model. Finally, to check the performance of confidence interval for the slope β using proposed method, we conducted a Monte Carlo simulation and presented some numerical results and examples.

Key Words and Phrases : *Least median of squares, Outliers test, Weighted least squares*

1. INTRODUCTION

We consider the robust inference for the parameters of linear regression model

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

where the error e_i is assumed to be normal distribution with mean zero and variance σ^2. Least squares estimate of β, $\hat{\beta}$ and standard error of $\hat{\beta}$, SE($\hat{\beta}$) are sensitive to the outliers. Hence inferences for the parameters of linear regression model using $\hat{\beta}$ and SE($\hat{\beta}$) are affected by outliers. To remedy this problem, many statistical method have been developed.

In this paper, we consider the tool of identifying and testing the outliers in linear regression model. This tool is based on the ratio of a robust scale estimate

1. Research funded Kyungnam University, 2002
2. Professor, Division of information & communication engineering, Kyungnam University.

and non robust scale estimate. And then we propose the forward sequential procedure for identifying the outliers. Next we consider the method to assign a weight to each observation (x_i, y_i). We make use of non increasing function of test statistics as a weight to each observation. Finally, we apply a weighted least squares analysis to introduce robust inference for linear regression model. The weighted least squares has a high breakdown point and is efficient in a statistical sense. Hence, inferences for the parameters of linear regression model using weighted least squares are not affected by outliers

The remaining of paper is organized as follows. In Section 2 we introduce the tool of identifying and testing the outliers in linear regression model. We suggest the method to assign a weight to each observation (x_i, y_i). We propose the robust inference for the parameters of linear regression model. In section 3 we consider the coverage and median length of confidence interval for the slope β by means of Monte Carlo simulation. In section 4 we apply the proposed method to several real data to check the performance of that. Section 5 contains some concluding remarks.

2. The Robust inference for the parameters of the linear regression model

We suggest the weighted least squares regression based on the sequential outliers test proposed by Jinpyo Park and Heechang Park(2001). First we recall the definition of the sequential outliers test. The test statistics is defined as follow. Least median of squares proposed by Rousseeuw(1984) minimizes the median of the squared residuals. Least median of squares regression has a very high breakdown point of almost 50%. The least median of squares estimator $\hat{\beta}_{LMS}$ is given by

$$\underset{\hat{\beta}_J}{\text{Minimize}} \quad \underset{i}{\text{med}}\, r_i^2 \tag{2}$$

where $r_i = y_i - x_i \hat{\beta}_J$, $\hat{\beta}_J = (X_J^T X_J)^{-1} X^T Y_J$, $x_i = (x_{i1}, x_{i2}, \cdots, x_{ip})$ and $J = \{i_1, i_2, \cdots, i_p\}$ is a subset of $\{1, 2, \cdots, n\}$ containing p indices. The residual is given by

$$r_{LMS_i} = y_i - x_i \hat{\beta}_{LMS}. \tag{3}$$

The initial scale estimate s_0 for the least median squares regression is given by

$$s_0 = 1.4826(1 + 5/(n - p - 1))\sqrt{\text{med}_i (r_{LMS_i})^2}. \tag{4}$$

The initial scale estimate is then used to determine a weight w_i for the ith observation, namely

$$w_i = \begin{cases} 1 & \text{if } c \le r_{LMS_i}/s_0 \le d \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

where $[c, d]$ is the inner fence of boxplot of r_{LMS_i}/s_0.

By means of these weights, the final scale estimate s for the least median squares

regression is given by

$$s = \sqrt{\sum_{i=1}^{n} w_i (r_{LMS_i})^2 / (\sum_{i=1}^{n} w_i - p - 1)}. \tag{6}$$

s also has a breakdown point 0.5, the highest possible value.

By contrast, the least squares estimator $\hat{\beta}_{LS}$ minimizes

$$\sum_{i=1}^{n} r^2{}_i. \tag{7}$$

The breakdown point of least squares estimator is 0. The residual is given by

$$r_{LS_i} = y_i - \boldsymbol{x_i}^T \hat{\boldsymbol{\beta}}_{LS}. \tag{8}$$

It is well known that outliers can have an extreme effect on the least squares estimator.

The scale estimate for the least squares regression is given by

$$\sigma = \sqrt{\sum_{i=1}^{n} (r_{LS_i})^2 / (n - p - 1)}. \tag{9}$$

The test statistics for testing the outliers is defined as

$$R = \sigma / s. \tag{10}$$

It tests the following hypothesis

$$H_0 : \text{no outlier in data } (x_{i1}, x_{i2}, \cdots, x_{ip}, y_i),\ i = 1, 2, \cdots, n$$
$$H_1 : \text{some outliers in data } (x_{i1}, x_{i2}, \cdots, x_{ip}, y_i),$$
$$i = 1, 2, \cdots, n. \tag{11}$$

The null hypothesis is rejected for large R. However, if the null hypothesis is rejected, there is no indication of how many or which points are outliers. To solve this problem, we 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 $D = |sort(r_{LMS_i}) - Med(r_{LMS_i})|$ is defined as an outlier, where $sort(r_{LMS_i})$ is the sort of r_{LMS_i} and $Med(r_{LMS_i})$ is the median of r_{LMS_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 critical values for the test(approximated by Monte Carlo simulation using 1000 replicates) are presented in the Table 1.

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		
	0.1	0.05	0.01	0.1	0.05	0.01	0.1	0.05	0.01	0.1	0.05	0.01
15	1.725	1.894	2.072	2.107	2.223	2.386	2.469	2.622	2.756	2.807	2.895	2.992
20	1.484	1.637	1.849	1.850	1.978	2.084	2.121	2.246	2.334	2.323	2.407	2.580
25	1.493	1.605	1.759	1.682	1.793	1.853	1.950	2.044	2.200	2.164	2.282	2.388
30	1.461	1.570	1.717	1.552	1.638	1.752	1.824	1.921	2.065	1.982	2.150	2.333
35	1.395	1.475	1.623	1.496	1.578	1.688	1.650	1.793	1.925	1.786	1.910	2.103
40	1.326	1.403	1.493	1.417	1.487	1.580	1.573	1.666	1.774	1.654	1.769	1.882
45	1.276	1.337	1.435	1.393	1.473	1.570	1.456	1.548	1.655	1.575	1.688	1.812
50	1.266	1.338	1.403	1.351	1.425	1.515	1.471	1.492	1.575	1.466	1.540	1.631

Next we suggest the way to assign a weight w_i to each observation (x_i, y_i). For this purpose, we can use several types of functions of the test statistics R. The first kind of weight function that we consider here is of the form

$$w(R_i) = \begin{cases} 1 & \text{if } R_i \le c_1 \\ 0 & \text{otherwise} \end{cases}. \tag{12}$$

where c_1 is a critical value for test statistics when significant level is 0.1. This weight function, yielding only binary weight, produces a clear distinction between accepted and rejected point. But this function is radical. So we introduce weight function that is less extreme. It consists of introducing a linear part that smooths the transition from weight 1 to weight 0.

In that way, extreme outliers disappear entirely and intermediate cases are gradually down-weighted. In the general formula

$$w(R_i) = \begin{cases} 1 & \text{if } R_i \leq c_1 \\ \dfrac{(c_2 - R_i)}{(c_2 - c_1)} & \text{if } c_1 \leq R_i \leq c_2 \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

Where c_2 is a critical value for test statistics when significant level is 0.01.

Anyway, we then apply weighted least squares defined by

$$\text{Minimize}_{\hat{\beta}} \sum_{i=1}^{n} w(R_i) r_i^2 \tag{14}$$

The weighted least squares estimator is given by

$$\boldsymbol{\beta}^* = (\boldsymbol{X}^T \boldsymbol{W}^T \boldsymbol{W} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{W}^T \boldsymbol{W} \boldsymbol{Y} \tag{15}$$

where $\boldsymbol{W} = diag(w_1^{\frac{1}{2}}, w_2^{\frac{1}{2}}, \cdots, w_n^{\frac{1}{2}})$.

Let $\boldsymbol{W}^T \boldsymbol{W} = diag(w_1, w_2, \cdots, w_n) = \boldsymbol{V}$, then

$$\begin{aligned} &\boldsymbol{\beta}^* = (\boldsymbol{X}^T \boldsymbol{V} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{V} \boldsymbol{Y} \\ &E(\boldsymbol{\beta}^*) = \boldsymbol{\beta} \\ \text{and} \quad &Var(\boldsymbol{\beta}^*) = (\boldsymbol{X}^T \boldsymbol{V} \boldsymbol{X})^{-1} \sigma^2. \end{aligned} \tag{16}$$

The standard error of i-th weighted least squares estimator is given by

$$\sqrt{\sigma^2 (\boldsymbol{X}^T \boldsymbol{V} \boldsymbol{X})_{ii}^{-1}}. \tag{17}$$

Where unknown σ^2 is estimated by $(s^*)^2 = \sum_{i=1}^{n} w(R_i) r_i^2 / (\sum_{i=1}^{n} w(R_i) - p)$.

To discuss robust inference for the linear regression model, we assume that the errors are independently and normally distributed with mean zero and variance σ^2. Under these conditions, it is well known that

$$\frac{\beta_i^* - \beta_i}{\sqrt{(s^*)^2(\boldsymbol{X}^T\boldsymbol{V}\boldsymbol{X})_{ii}^{-1}}}, \quad i = 1, 2, \cdots, p \tag{18}$$

has a Student t-distribution with $\sum_{i=1}^{n} w(R_i) - p$ degree of freedom. Let us denote the $1 - \alpha/2$ quantile of this distribution by $t_{\sum_{i=1}^{n} w(R_i) - p, 1 - \frac{\alpha}{2}}$. Then a $(1 - \alpha)100\%$ confidence interval for β_i is given by

$$[\beta_i^* - t_{\sum_{i=1}^{n} w(R_i) - p, 1 - \frac{\alpha}{2}} \sqrt{(s^*)^2(\boldsymbol{X}^T\boldsymbol{V}\boldsymbol{X})_{ii}^{-1}}, \ \beta_i^* + t_{\sum_{i=1}^{n} w(R_i) - p, 1 - \frac{\alpha}{2}} \sqrt{(s^*)^2(\boldsymbol{X}^T\boldsymbol{V}\boldsymbol{X})_{ii}^{-1}}], \quad i = 1, 2, \cdots, p. \tag{19}$$

To test the hypothesis

$$H_0 : \beta_i = 0$$
$$H_1 : \beta_i \neq 0, \tag{20}$$

We can use the following test statistics

$$\frac{\beta_i^*}{\sqrt{(s^*)^2(\boldsymbol{X}^T\boldsymbol{V}\boldsymbol{X})_{ii}^{-1}}}, \quad i = 1, 2, \cdots, p. \tag{21}$$

The robust coefficient of determination R_r^2 is calculated as follows:

$$R_r^2 = 1 - \frac{\sum_{i=1}^{n} w(R_i)(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} w(R_i)(y_i - \overline{y}_i)^2} \quad \text{in model with constant term} \tag{22}$$

and

$$R_r^2 = 1 - \frac{\sum_{i=1}^{n} w(R_i)(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} w(R_i) y_i^2} \quad \text{in model without constant term.} \tag{23}$$

Because the weighted least squares regression based on the sequential outliers test statistics R has the high breakdown point, inferences for the parameters of linear regression model using weighted least squares are not affected by outliers.

3. Simulation and its Results

We focus on confidence interval for the slope parameter β. We consider the coverage and median length of confidence interval for the slope β. First, we generate samples in following situation,

$$y_i = x_{i1} + x_{i2} + \cdots + x_{ip} + e_i \quad i = 1, 2, \cdots, n \tag{24}$$

in which $e_i \sim N(0, 1)$ and the explanatory variables are generated as $x_{ij} \sim N(0, 100)$, for $j = 1, 2, \cdots, p$. Second, (1- α) $\times$ 100% of samples are generated as in the first situation and the remaining $\alpha \times 100\%$ are generated as $e_i \sim N(0, 1)$ and $x_i \sim N(\mu, 100)$. Finally (1- α) $\times$ 100% of samples are generated as in the first situation. The remaining $\alpha \times 100\%$ are generated as $e_i \sim N(\mu, 1)$ and $x_i \sim N(0, 100)$.

To check the performance of confidence interval for the slope β using equation 19, we conducted a Monte Carlo simulation using 1000 replicates of the following sampling situation: sample sizes n=20, 40, 60, 80, 100, μ=10, 50, 90, p=1, α = 0.05 and α = 0.1.

The mean coverage and median length(in parenthesis) of confidence interval for the slope β based on the least squares method are presented in the Table 2. The nominal confidence level in all cases is 0.95.

Table 2. mean coverage and median length of confidence interval for the slope β = 1 based on the least squares

% of contamination	Sample size	Mild Contamination (μ=20)	Medium Contamination (μ=50)	Strong Contamination (μ=90)
5%	20	0.75(0.41)	0.51(0.784)	0.47(1.231)
	40	0.71(0.40)	0.47(0.641)	0.33(0.951)
	60	0.65(0.34)	0.32(0.534)	0.26(0.823)
	80	0.61(0.32)	0.29(0.423)	0.23(0.741)
	100	0.47(0.29)	0.27(0.341)	0.21(0.647)
10%	20	0.65(0.35)	0.42(0.35)	0.30(0.92)
	40	0.47(0.24)	0.32(0.25)	0.25(0.83)
	60	0.27(0.21)	0.21(0.20)	0.11(0.76)
	80	0.15(0.19)	0.10(0.18)	0.07(0.43)
	100	0.13(0.17)	0.09(0.16)	0.05(0.28)

The computed percentages are below the nominal level, specially for larger sample sizes. This indicates that the low coverage levels are due to outliers in the data.

To overcome this problem, we use the weighted least squares based on the sequential outliers test statistics R. The mean coverage and median length(in parenthesis) of confidence interval for the slope β based on the weighted least squares method are presented in the Table 3. The nominal confidence level in all cases is 0.95.

Table 3. mean coverage and median length of confidence interval for the slope $\beta = 1$ based on the weighted least squares

% of contamination	Sample size	Mild Contamination (μ=20)	Medium Contamination (μ=50)	Strong Contamination (μ=90)
5%	20	0.96(0.098)	0.96(0.097)	0.93(0.112)
	40	0.96(0.096)	0.96(0.096)	0.94(0.0991)
	60	0.95(0.094)	0.94(0.095)	0.94(0.098)
	80	0.94(0.093)	0.94(0.094)	0.95(0.095)
	100	0.94(0.092)	0.94(0.093)	0.95(0.094)
10%	20	0.97(0.135)	0.98(0.137)	0.92(0.192)
	40	0.95(0.124)	0.97(0.125)	0.94(0.183)
	60	0.93(0.121)	0.98(0.120)	0.95(0.176)
	80	0.93(0.119)	0.97(0.18)	0.96(0.413)
	100	0.92(0.117)	0.95(0.16)	0.96(0.128)

Notice that the coverages are all above 92% and several exceed the nominal 95% level. And median length of confidence interval in the Table 3 is shorter than that in the Table 2.

4. Numerical Results

In this section, the proposed method is applied to several data sets to check the performance.

Example 1 (Pilot-Plant Data)

This data comes form 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 4. The results for the proposed method are in the Table 5.

In the Table 5, 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. In this example, all $w(R_i)$ are equal to 1, except for case 6. The inference results by means of least squares(LS) and weighted least

squares(WLS) for Pilot-Plant data with outlier are presented in the Table 6. The scatterplot for the pilot-plant data without outliers suggests a strong statistical relationship between the response and the explanatory variable. However, in the Table 6, we can conclude that the LS slope is not significantly differ from zero and the R^2 corresponding to LS is 0.141, on the other side, the WLS slope is significantly differ from zero and the R^2 corresponding to WLS is 0.994. And the length of the confidence interval for LS is longer than that for WLS. This result demonstrate the fact that WLS is unaffected but LS is affected by outliers. Therefore the inference using the weighted least squares method based on the function of the sequential test statistics is robust.

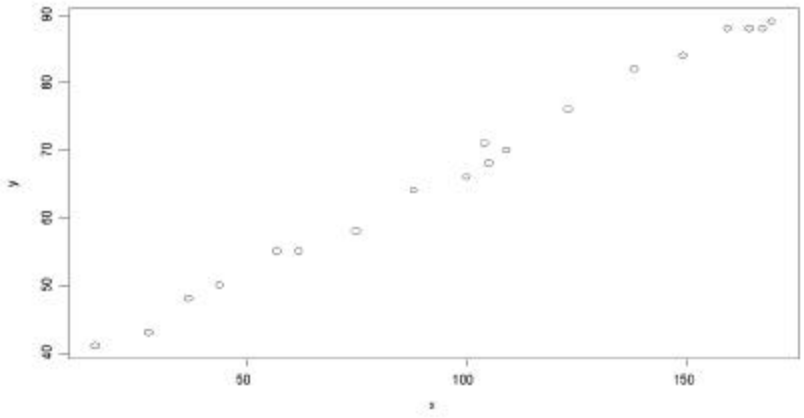

Figure 1. Scatterplot for Pilot-Plant without outliers

Table 4. Pilot-Plant data set with outlier

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 5. The proposed method applied to the contaminated pilot-plant data

sample size	observation selected	proposed test statistics	critical values			weight
			0.01	0.05	0.1	
20	6	11.703	1.849	1.637	1.484	0
19	11	0.941	1.858	1.651	1.495	1

Table 6. The inference results by means of LS and WLS for Pilot-Plant data with outlier

	coefficient		standard error	T-value	p-value	95% confidence interval		length	coefficient of determination
						lower	upper		
LS	x1	0.081	0.047	1.719	0.103	-0.018	0.179	0.197	0.141
WLS	x1	0.323	0.006	54.214	0.000	0.310	0.335	0.025	0.994

Example 2 (Stockloss Data)

The second example comes from the Brownlee(1965). 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 proposed method appear in the Table 8. In the Table 8, all $w(R_i)$ are equal to 1, except for observation 4, 21, 1, 3 and 2. The inference results by means of LS and WLS for stackloss data are presented in the Table 9.

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 method applied to the stackloss data

Sample size	observation selected	proposed test statistics	Critical Values			weight
			0.01	0.05	0.10	
21	4	2.685	2.304	2.204	2.101	0
20	21	2.895	2.334	2.246	2.121	0
19	1	2.367	2.359	2.296	2.231	0
18	3	2.9067	2.384	2.346	2.321	0
17	2	2.610	2.421	2.396	2.384	0
16	13	2.326	2.634	2.583	2.421	1

Table 9. The inference results by means of LS and WLS for stackloss data

	coefficient		standard error	T-value	p-value	95% confidence interval		length	coefficient of determination
						lower	upper		
LS	x1	0.716	0.135	5.307	0.000	0.431	1.000	0.569	0.914
	x2	1.295	0.368	3.520	0.003	0.519	2.072	1.553	
	x3	-0.152	0.156	-0.973	0.344	-0.482	0.178	0.66	
WLS	x1	0.686	0.088	7.834	0.000	0.495	0.877	0.382	0.942
	x2	0.567	0.153	3.702	0.003	0.233	0.901	0.668	
	x3	-0.017	0.063	-0.273	0.789	-0.155	0.120	0.275	

In the Table 9, the length of confidence interval of each regression coefficient for LS are longer than that for WLS. And the significance of the regression coefficients turns out to be different in the LS fit and the WLS fit.

Example 3 These raw 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 contaminated data is shown in the Table 10. The result for proposed method appear in the Table 11. In the Table 11, all $w(R_i)$ are equal to 1, except for observation 19, 6, 8, and 4. The inference results by means of LS and WLS for modified data on wood specific gravity are presented in the Table 12.

Table 10. 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 11. The proposed method applied to modified data on wood specific gravity

Sample size	observation selected	scale ratio statistics	Critical Values			weight
			0.01	0.05	0.10	
20	19	1.783	1.484	1.415	1.365	0
19	6	1.948	1.518	1.445	1.395	0
18	8	2.068	1.547	1.472	1.412	0
17	4	2.635	1.577	1.492	1.433	0
16	5	1.227	1.671	1.522	1.463	1

Table 12. The inference results by means of LS and WLS for modified data on wood specific gravity

	coefficient		standard error	T-value	p-value	95% confidence interval for slop		length	coefficient of determination
						lower	upper		
LS	x1	0.441	0.117	3.770	0.002	0.190	0.691	0.501.	0.808
	x2	-1.475	0.487	-3.029	0.009	-2.519	-0.431	2.088	
	x3	-0.261	0.112	-2.332	0.035	-0.501	-0.021	0.48	
	x4	-0.021	0.161	0.129	0.899	-0.325	0.366	0.691	
	x5	0.171	0.203	0.840	0.415	-0.265	0.607	0.872	
WLS	x1	0.217	0.042	5.162	0.000	0.124	0.311	0.187	0.958
	x2	-0.085	0.198	-0.430	0.676	-0.526	0.356	0.882	
	x3	-0.564	0.043	-12.975	0.000	-0.661	-0.467	0.194	
	x4	-0.400	0.065	-6.118	0.000	-0.546	-0.255	0.291	
	x5	0.607	0.079	7.730	0.000	0.432	0.783	0.351	

In the Table 12, the significance of the regression coefficients turns out to be different in the LS fit and the WLS fit. The variables x4, x5 have LS regression coefficients that are not significantly different from zero for significant level 0.05. Only the variable x2 has WLS regression coefficient that is not significantly different from zero for significant level 0.05. Draper and Smith conclude that x2 could be removed from the model. The variable x2 has LS regression coefficient that is significantly different from zero for significant level 0.01 is only caused by outliers. And the length of confidence interval of each regression coefficient for LS are longer than that for WLS.

The many examples demonstrate the fact that the WLS is unaffected by masking and swamping effects.

5. Concluding Remarks

It is well known that outliers can have an extreme effect on the least squares estimation. Hence, it is important to detect the outliers and to manage how to deal with the detected outliers.

In this paper, we proposed the tool to identify outliers and the method to deal with the detected outliers. To detect the outliers, we suggested the forward sequential test. And to deal with the detected outliers, we recommended the weighted least squares method based on the function of the sequential test statistics.

We proved that the weighted least squares method based on the function of the sequential test statistics was not affected by the masking and swamping effects through the Monte Carlo results and numerical examples. These suggest that the newly proposed tool provides conservative and fairly powerful method for the analysis of the data from linear regression model

Reference

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 W`ley & Sons, New York.
3. Draper, N, R., and Smith, H.,(1966). *Applied Regression Analysis*, John Wiley & Sons, New York.
4. Jinpyo Park and Heechang Park,(2001). The sequential testing of multiple outliers in linear regression, The Korean Communication in Statistics Vol. 8, No.2, 337-346
5. Rousseeuw, P. J.,(1984) Least median of squares regression, *J. Am. Stat. Assoc., 79, 871-884.*
6. Rousseeuw, P. J., and Leroy, A. M.,(1987) *Robust regression and outlier detection*, John Wiley & Sons, New York.

[2002년 9월 접수, 2002년 10월 채택]