The Detection and Testing of Multiple Outliers in Linear Regression

Jinpyo Park

Kyungnam University, Division of information and communication engineering
449 Wolyoung-dong
Masan, Korea 631-701
jppark@eros.kyungnam.ac.kr

Ruben H. Zamar
University of British Columbia, Department of Statistics
333-6356 Agricultural Road
Vancouver, B.C., Canada V6T 1Z2
ruben@stat.ubc.ca

1. Introduction

It is now well known that outliers can have an extremely potent effect on estimation and analysis of linear regression. For this reason , we propose the new method that can identify and test multiple outliers without suffering from masking and swamping effects. This method, which we call the scale ratio tests, is based on the ratio of two estimates of scale. We show the asymptotic distribution of the test, and then we investigate the properties(critical values and powers) of the test through several Monte Carlo simulations. Furthermore, we propose the forward sequential procedure for identifying the outliers. Finally the proposed method is applied to several real and artificial data sets in order to show their performance.

2. Scale Ratio Tests

Suppose that we want to test the null hypothesis of no outliers in data set. Let s_1 and s_2 be two estimates of scale corresponding to the following ρ-functions, Tukey's bisquare functions $\rho_1(\cdot\,;\,c)$ and $\rho_2 = x^2$. Here, s_1 is robust estimate of scale with a breakdown point 0.5 and s_2 is the non-robust one of scale since ρ_2 is unbounded. The scale ratio test statistic is defined as $R = s_2/s_1$ and the null hypothesis is rejected for large value of R. However, if the null hypothesis is rejected, there is no indication of how many or which points are outliers. This problem is solved by applying the scale ratio test in a forward sequential procedure. If the test rejects null hypothesis then the point with the largest D=|sort(r_i)-med(r_i)| is removed and the test is applied again to the remaining points, where sort(r_i) is the sort of r_i; med(r_i) is the median of r_i; $r_i = y_i - x_i\hat{\beta}$; and $\hat{\beta}$ is S-estimate of regression coefficient β. This procedure is applied iteratively and stop when the test is no longer significant. We derive that the asymptotic distribution of the scale ratio test under the null hypothesis is normal. Next, the critical values are calculated for some sample size up to 50, number of explanatory variable =1,2,3,4 and significant level = 0,1,0.05,0.01. For larger sample sizes, the asymptotic

approximation $C_a = 1 + 0.6539 n^{-1/2} Z_a$ can be used, where Z_a is 100(1-α)th percentile of standard normal distribution and n is the sample size used to compute the test. The asymptotic approximation can also be used to calculate approximate $p-$values. Finally, we consider the power of the scale ratio test. The test is applied 1000 times to random samples of various sizes, some dimensions of explanatory variable, several magnitude of outliers, and significant level 0.1, 0.05, 0.01 to calculate the power. The results show that the performance of the scale ratio test are very good. The power increases with sample size and magnitude of outliers.

3. Applications of the Scale Ratio tests and Concluding Remarks

The scale ratio test is applied to several data sets for the purpose of detecting and testing outliers. The application for outliers detection come from the Brownlee(1965). The data is well-known stackloss data set. This consists of 21 observations in four dimension (one response and three explanatory). We select this data set because it is a set of real data and it is examined by many statisticians. Most people conclude that observations 1, 3, 4, and 21 were outliers. Some people reported that observation 2 was a outlier. The results for the test is presented in the table1.

Table 1. scale ratio test applied to the stackloss data

Sample size	Points selected	Scale ratio statistics	Critical Values			Sample size	Points selected	Scale ratio statistics	Critical Values		
			0.1	0.05	0.01				0.1	0.05	0.01
21	21	1.766	1.297	1.345	1.403	18	3	1.606	1.336	1.404	1.493
20	4	1.546	1.306	1.355	1.409	17	2	1.236	1.336	1.404	1.497
19	1	1.472	1.316	1.355	1.438						

The test identify 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. Additional applications not reported here show that the performance of the proposed method is very good and this is unaffected by masking and swamping effects.

The numerical example and Monte Carlo results presented in this paper suggest that the proposed method provides powerful method for detecting and testing outliers in linear regression. An important feature of our method is that the results can be objectively interpreted.

REFERENCE

[1] Brownlee, K. A. (1965), *Statistical Theory and Methodology in Science and Engineering*, 2nded, John Wiley & Sons, New York

[2] Rousseeuw, P.J., and Leroy, A.M..(1987*), Robust Regression and Outlier dection*, John Wiley & Sons, New York.

RESUME

In this paper we consider the problem of identifying and testing the outliers in linear regression. First we consider the scale ratio test for testing the null hypothesis of no outliers in data set. The test is based on the ratio of two estimate of scale. We show the asymptotic distribution of test statistics and investigate the properties of the test. Next we propose a forward sequential procedure of identifying the ourliers. Finally the proposed method is applied to several real and artificial data sets in order to show their performance. The proposed method is unaffected by masking and swamping effects.