Intercontinental Geoinformation Days

http://igd.mersin.edu.tr/2020/

Detecting Outliers With The Least Trimmed Squares

Hasan Dilmaç*[1], Yasemin Şişman[2]

[1]*Ondokuz Mayis University, Faculty of Engineering, Department of Geomatics Engineering, Samsun, Turkey*

Keywords
The Least Squares
Outliers
Robust Estimation
The least trimmed squares

ABSTRACT

Classical outlier tests made classically based on the least-squares (LS) have significant disadvantages in some situations. The results of adjustment computation and classical outlier tests performed with classical methods are deteriorated when observations are not distributed independently and this distribution is not normal. To detect outliers that do not have a normal distribution, the robust techniques that are not sensitive to outliers have been developed. The least trimmed squares (LTS) known as having a high-breakdown point have been dealt with in this study. Adjustment computation has been carried out based on the least-squares (LS) and the least trimmed squares (LTS). A certain polynomial with arbitrary values has been used. In this way, the performances of these techniques have been investigated.

1. INTRODUCTION

Various observations are done in geodesy. Physical and geometric quantities, such as angles, distances, heights, and gravity are measured and processed. In this case, a great number of data appears (Fan 1977). Since the accuracy of data is always questioned, it is preferred that the number of observations is bigger than the number of unknowns. A quantity is always measured differently from each other even though it is measured many times under the same conditions (Ingram 1911). It is clear that observations are never exact, and they always contain error, however careful they are performed. Thus, an adjustment computation is applied to get unique solutions from these redundant measurements (Ghilani 2017; Mikhail and Ackermann 1982).

There are a lot of adjustment methods. The least-squares (LS) is a frequently used method. LS is a sort of regression that examines and models the relationship between data (usually obtained from observations). It is one of the most adopted methods because of its tradition and ease of computation (Cizek and Visek 2005; Rousseeuw and Leroy 1987). But, it has turned out with time that outliers (observations with different distribution compared to the distribution of majority) affect the LS method negatively.

Outliers in observations are encountered very often in applications (Rousseeuw and Leroy 1987). The results of adjustment with classical methods such as LS, which should meet some conditions like normal distribution are deteriorated. So, these outliers must be detected and eliminated from observations. There are outliers tests based on classical methods, especially LS. These outliers tests can be contaminated. Therefore, new statistical methods have been sought instead of LS sensitive to outliers. (Yetkin and Berber 2013).

The robust statistics deals with developing estimators insensitive to discrepancies from basic assumptions in classical models (Fabozzi et al. 2014). To overcome effects of outliers, robust methods aim to find results that are closest to adjustment results that would be found without outliers. Then, outliers can be detected through their residuals. (Rousseeuw and Hubert 2018). Many robust techniques have been developed. These techniques can be divided into classes with some concepts like a high-breakdown point, influence function, etc. The least trimmed squares (LTS) is a high-breakdown point estimator.

In this study, adjustment computations and outlier analysis have been performed according to LS and LTS method in different scenarios. Then, the results of LS and LTS have been compared with each other.

* **Corresponding Author**

*(hasan.dilmac@omu.edu.tr) ORCID ID 0000-0001-6877-8730
(ysisman@omu.edu.tr) ORCID ID 0000-0002-6600-0623

Cite this study

Dilmac H & Sisman Y (2020). Detecting Outliers with The Least Trimmed Squares. Intercontinental Geoinformation Days (IGD), 186-189, Mersin, Turkey

2. METHOD

2.1. Adjustment Computation

When there is the redundant observation in a problem, adjustment computation is performed to get unique values for the unknowns (Ogundare 2018). Adjustment is only meaningful when observations are more than the unknown number (Mikhail and Ackermann 1976). In this case, the problem is solved according to an objective function. For the solution of the problem, a mathematical model (model briefly) that represents the mathematical relationship of observations and unknowns is established (Schaffrin 2019). The mathematical model accounts for an essential part of adjustment computation, and it is usually composed of two parts: a functional model and a stochastic model. When observations are made, a functional model is typically chosen to represent the physical situation. The stochastic model determines variances and covariances of observations (Ghilani 2017; Mikhail and Ackermann 1976; Ogundare 2018). In the classical Gauss-Markov model, the functional and stochastic model can be expressed as below:

$$v = Ax - l$$
$$P = Q_{ll}^{-1} = \sigma_0^2 C_{ll}^{-1}$$

where v, A, x, l, P, σ_0^2 and C_{ll} are the residual or correction vector, the coefficient matrix, unknown vector, the observation vector, the weight matrix, a priori variance, and the covariance matrix, respectively.

In this case, both the functional model and the stochastic model must be correct if adjustment computation is to give real results (Ghilani 2017). After a mathematical model is formed, an optimization is made according to chosen objective function. Optimization means minimization or maximization of function (Grafarend and Sanso 2012).

2.2. The Least Squares

The Least Squares (LS) is a method used in adjustment computation by minimizing the sum of the squared weighted differences to get unique values with redundant measurements (Amiri-Simkooei 2003; Mikhail and Ackermann 1982; Wells and Krakiwsky 1971). The objective function of LS can be given the following:

$$v^T P v = \sum_{i=1}^{n} p_i v_i^2 \rightarrow min$$

The main problem of LS is that even one outlier might severely affect the LS method (Muhlbauer et al. 2009). LS can propagate errors in one observation to another observation. Therefore, masking and swamping effects occur. A bad observation could seem like a good one because of the propagation of errors; this is called a masking effect. On the contrary, the good observation could seem bad; this is called the swamping effect (Hekimoglu 2005). As a result, test for outliers like Baarda test (Data-snooping, W-test) and Pope test (Tau test) based on LS also can be affected negatively.

2.3. Robust Estimation and Determination of Outliers

Real data sets frequently contain outliers (Rousseeuw 1990). Therefore, methods that cannot be affected easily by outliers should be developed. These are the methods named as robust methods. Robustness usually means insensitivity to outliers (Huber 1981).

There are many robust methods. L_1-norm is the oldest method of these robust methods. Then, M-estimators, R-Estimators, and L-Estimators appeared. To compare the robustness of these methods, the 'breakdown point' has been used. The breakdown point means the smallest number of outliers, which may affect an estimator negatively (Hofmann et al. 2010). These methods above have a low-breakdown point (Rousseeuw and Leroy 1987). Because of this, generalised M-Estimators was developed. Then, Repeated Median, The Least Median Squared (LMS) (1984), S-Estimators, MM-Estimators, and The Least Trimmed Squares (LTS) were developed respectively (Hubert et al. 2008; Staudte and Sheather 2011; Toka and Cetin 2011).

2.4. The Least Trimmed Squares

The least trimmed squares (LTS) was developed by Rousseeuw. This method is quite similar to LS except that the largest squared residuals are removed from the data (Knight and Wang 2009). The objective function of LTS can be given the following:

$$Min \sum_{i=1}^{h} P v_i^2$$

where, h is the number of residuals (or corrections) after data removing

There are different criteria to determine the number of residuals in LTS that will be included in the sum. The h= n/2 (n, number of observations) should be taken for maximum robustness since LTS can give satisfactory results until %50 contamination (Cizek 2005). LTS problem requires dealing with finding the minimum one from $\binom{n}{h}$ LS solutions (Hofmann et al. 2010).

3. RESULTS

In numerical applications, a linear regression model such as $y=a_1x+a_2$ was used. Regression coefficients were taken as 2 and 0.5, respectively. y values were calculated according to x values that were chosen arbitrarily for 10 observations. In the first application, both LS and LTS methods were performed using x values and y values with random errors. In LTS, h was taken as 8 and $\binom{10}{8}$ solutions were made. Then, for the second application, gross errors were added to

some y values, and LS and LTS methods were performed again. Regression results are shown in Table 1 and Table 2. Distribution of x and y values and regression lines in Application 2 are shown in Figure 1. It was expected that LTS could detect precisely y values with gross errors.

The results of Application 1 have shown that the observations (measurements) are typically distributed, and they have only random errors. Thus, the results of LS and LTS are close to each other (Table 1). But, it is shown that the results of LTS are closer to real values than LS. In Application 2, it is clear that the products of LS are quite contaminated, and the sum of residuals squared has increased very much (Table 2). The coefficient a_2 of LS in Application 2 is quite different from expected. Compared to LS in Application 2, the results of LTS in Application 2 are more correct, and the sum of residuals squared is relatively much smaller.

Table 1. The regression results of Application 1

Methods	a_1	a_2	[VV]
LS	2.03	0.37	1.24
LTS	2.01	0.38	0.15

Table 2. The regression results of Application 2

Methods	a_1	a_2	[VV]
LS	2.14	1.76	208.38
LTS	2.07	0.09	0.96

It can be seen that the outliers have affected the results of LS regression in Application 2. Point 6 and, which is designed as outliers have drawn LS regression line towards themselves. However, LTS regression has not been affected by outliers (Figure 1).

The residuals of LS regression in Application 1 are small as expected (Figure 2a). The effects of outliers on the residuals for LS in Application 2 can be seen in Figure 2b. Also, it is seen that the gross errors added to Point 6 and 7 have been distributed to the other points in the LS method (Figure 2b).

LTS regression in Application 2 could determine outliers precisely and remove outliers (6 and 7. Points) from the observations. Also, LTS regression has not distributed outliers effects to the other points (Figure 3).

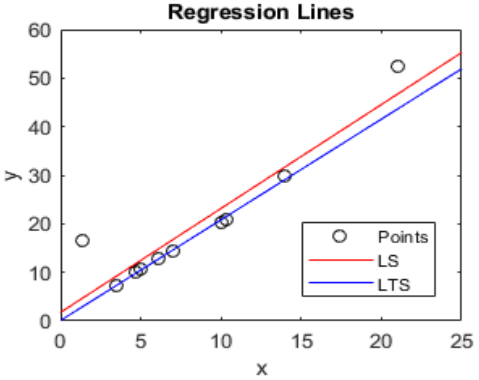

Figure 1. The regression lines of LS and LTS in Application 2

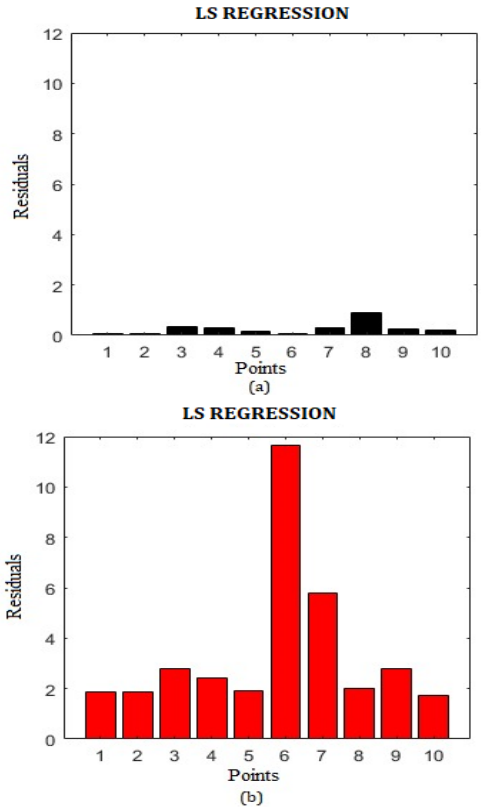

Figure 2. The residuals of LS regression in Application 1 (a). The residuals of LS regression in Application 2 (b)

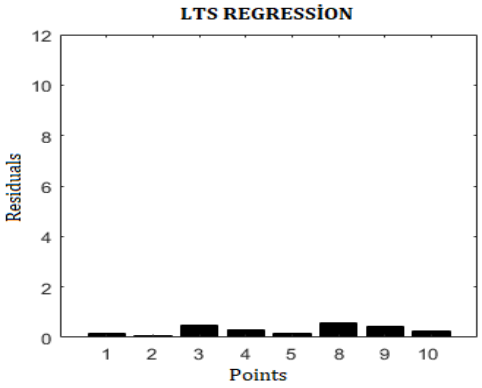

Figure 3. The residuals of LTS regression in Application 2

4. DISCUSSION AND CONCLUSION

A linear regression model was used in this study. The analysis were performed according to LS and LTS method using different scenarios. In the first application, points with only random errors were used. The LS and LTS methods gave good results. But, LTS had a little better results. In the second application where contaminated points were used, although LS results were affected badly from outliers, LTS results gave results close to ones in Application 1.

As a result, LTS results are as good as LS results when observations are normally distributed. On the other hand, LTS can give much better results than LS when observations have outliers. Also, the LS method can distribute the outlier effect to the other points.

REFERENCES

Amiri-Simkooei A (2003). Formulation of L1 Norm Minimization in Gauss-Markov Models. *Journal of Surveying Engineering*, 129:1, 37:43. doi: 10.1061/(ASCE)0733-9453(2003)129:1(37)

Čížek P & Víšek J Á (2005). Least Trimmed Squares. *XploRe®—Application Guide*,49-63. Springer, Berlin, Heidelberg.doi: 10.1007/978-3-642-57292-0_2

Cizek P (2005). Least trimmed squares in nonlinear regression under dependence. *Journal of Statistical Planning,* 136, 3967-3988, doi: 10.1016/j.jspi.2005.05.004

Fabozzi F J, Focardi S M, Rachev S T & Arshanapalli B G (2014). *The Basics of Financial Econometric: Tools, Concepts and Asset Management Applications*. John Wiley ve Sons, Inc.

Fan H (1993). *Theory of errors and least squares adjusment*. Royal Institute of Technology, 72, 100-44, Stockholm, Sweden.

Ghilani C D (2017). *Adjustment computations: Spatial data analysis* (Sixth edition). John Wiley ve Sons, Inc., Hoboken, New Jersey.

Grafarend E W & Sansò F (Editors) (2012). *Optimization* and *design of geodetic networks*. Springer Science & Business Media, Heidelberg, Berlin.

Hekimoglu S (2005). Do Robust Methods Identify Outliniers More Reliably Than Conventional Tests for Outliniers? *Zeitschrift für Vermessungwesen, 3, 174-180.*

Hofmann M, Gatu C & Kontoghiorghes E J (2010). An Exact Least Trimmed Squares Algorithm for a Range of Coverage Values*, Journal of Computational and Graphical Statistics,* 19:1, 191-204, doi: 10.1198/jcgs.2009.07091

Huber P J (1981). *Robust Statistics*. John Wiley and Sons, Inc.

Hubert M, Rousseeuw P J & Van Aelst S (2008). High-breakdown robust multivariate methods. *Statistical science*, 23:1, 92-119. doi: 10.1214/088342307000000087

Ingram E L (1911). *Geodetic surveying and the adjustment of observations* (methods of least squares). McGraw-Hill Book Company, Inc. 370 Seventh Avenue, New York.

Knight N L & Wang J (2009). A comparison of outlier detection procedures and robust estimation methods in GPS positioning. *The Journal of Navigation,* 62:4, 699-709 doi: 10.1017/S0373463309990142

Mikhail E M & Ackermann F E (1976). *Observations and least squares*. Thomas Y. Crowell Company, Inc. 666 Fifth Avenue, New York.

Montgomery D C, Peck E A & Vining G G (2012). *Introduction to linear regression analysis* (Fifth edition). John Wiley & Sons, Inc, 821, Hoboken, New Jersey.

Muhlbauer A, Spichtinger P & Lohmann U (2009). Application and comparison of robust linear regression methods for trend estimation. *Journal of Applied Meteorology and Climatology,* 48:9, 1961-1970 doi: 10.1175/2009JAMC1851.1

Ogundare J O (2018). *Understanding Least Squares Estimation and Geomatics Data Analysis*. John Wiley ve Sons, Inc, 111 River Street, Hoboken, NJ 07030, USA

Rousseeuw J R & Leroy A M (1987). *Robust Regression and Outlier Detection.* John Wiley ve Sons, Inc.

Rousseeuw J R (1990). Robust estimation and identifying outliers. *Handbook of statistical methods for engineers and scientists*, 16-1.

Rousseeuw P J & Huber M (2018). Anomaly detection by robust statistics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery,* 8:2, e1236. doi: 10.1002/widm.1236

Schaffrin B (2019). Notes On Adjustment Computations Part I.

Staudte R G & Sheather S J (2011). *Robust estimation and testing*. John Wiley ve Sons, Inc, 918.

Toka O & Cetin M (2011). The comparing of S-estimator and M-estimators in linear regression. *Gazi University Journal of Science,* 24:4, 747-752

Wells D & Krakiwsky E (1971). The Method of least squares. University of New Brunswick: Canada

Yetkin M & Berber M (2013). Application of the sign-constrained robust least-squares method to surveying networks. *Journal of Surveying Engineering,* 139:1, 59-65. doi: 10.1061/(ASCE)SU.1943-5428.0000088