Introduction

Yadolah Dodge

University of Neuchâtel, Switzerland

While the method of least squares (and its generalizations) have served statisticians well for a good many years (mainly because of mathematical convenience and ease of computation), and enjoys certain well known properties within strictly Gaussian parametric models, it is recognized that outliers, which arise from heavy-tailed distributions, have an unusually large influence on the estimates obtained by these methods. Indeed, one single outlier can have an arbitrary large effect on the estimate. Outlier diagnostics have been developed to detect observations with a large influence on the least squares estimation. For excellent books related to such diagnostics the reader is referred to Cook and Weisberg (1982, 1994) and Chatterjee and Hadi (1988).

Parallel to diagnostic techniques, robust methods with varying degrees of robustness and computational complexity have been developed to modify the LS method so that the outliers have less influence on the final estimates. Among others are the bounded influence estimators, the repeated median, the least median of squares and the regression quantile methods.

In 1964, Huber published what is now considered to be a classic paper on robust estimation of location parameter and subsequently extended to that linear model. The development of selected robustness concepts since their inception in the 1960's and their current status, is given by Huber (1995).

One of the simplest robust alternatives to LS is the least absolute value method. This method, which is the subject of this volume, is a widely recognized superior method especially well-suited to longer-tailed error distributions, such as the Laplace distribution.

Depending on the field of application, the least absolute value method has been studied in several contexts under a variety of names such as minimum, or least sums of absolute errors, deviations or values; and here we

refer to it as the L_1- norm method (for minimizing the L_1- norm of the vector of deviations). The L_1 method estimates the unknown parameters in a stochastic model so as to minimize the sum of the absolute deviations of a given set of observations from the values predicted by the model.

Historically, L_1 estimation is the oldest of all robust methods. The method of least absolute deviations was introduced almost 50 years before the method of least squares, in 1757 by Roger Joseph Boscovich (1711-1787). He devised the method as a way to reconcile inconsistent measurements for estimating the shape of the earth. After Pierre Simon, Laplace adopted the method 30 years later, it saw occasional use but was soon overshadowed by the method of least squares. The popularity of least squares was at least partly due to the relative simplicity of its computations and to the supporting theory that was developed by Gauss and Laplace. Laplace, in his second memoir on the Figure of the Earth in 1789, adopted Boscovich's two criteria for a line of best fit, and gave an algebraic formulation and derivation of Boscovich's algorithm.

After nearly seventy years following the publication of Laplace's second supplement to the Théorie Analytique des Probabilités (1818), Edgeworth (1887) presented a method for linear regression using L_1 method. But since the publication of Edgeworth's work , few attempts have been made to convince the statisticians and particularly the applied users to employ this method (see Turner, 1887; Rhodes, 1930; Singleton, 1940; Karst,1958). Reasons for such a long silence may be summarized as follows :

(1) Computational difficulties in producing the numeric values of the L_1 estimates in regression. (Lack of closed form formulae similar to that of least squares).

(2) Lack of an asymptotic theory for L_1 estimation in the regression model, and more generally the nonexistence of accompanying statistical inference procedures.

(3) Insufficient evidence to show the superiority of the small sample properties of L_1 estimation compared to the LS estimators when sampling from long tailed distributions.

Following the work of Charnes, Cooper and Ferguson (1955) a renewed interest in using L_1 estimation for regression problem was created. They showed the equivalence between the L_1 problem and a linear programming problem. Wagner (1959) suggested that the L_1 problem in a linear regression of the form $y_i = \theta_0 + \theta_1 x_{1i} + \cdots + \theta_p x_{pi} + \varepsilon_i$ or in matrix form $Y = X\theta + \varepsilon$ can be solved by solving the dual of the L_1 problem. He also observed that the dual problem can be reduced to a problem with a smaller basis but where the dual variables have upper-bound restrictions. Wagner's formulation of the problem is to restate the problem of minimizing $\sum |\varepsilon_i|$ with

respect to θ where ε_i is the deviation between the observed and predicted values of the i^{th} observation Y_i, as: minimize $\sum |\varepsilon_i|$, subject to $X\theta+\varepsilon = Y$, where θ, ε unrestricted in sign.

Noting the fact that $|\varepsilon_i| = \varepsilon_{1i}+\varepsilon_{2i}$ where $\varepsilon_1 = \varepsilon$ if $\varepsilon > 0$, 0 otherwise and $\varepsilon_2 = -\varepsilon$ if $\varepsilon < 0$, 0 otherwise, that both are nonnegative and $\varepsilon_i = \varepsilon_{1i} - \varepsilon_{2i}$, we can reformulate the problem as a linear programming problem: minimize $\sum \varepsilon_{1i} + \sum \varepsilon_{2i}$, subject to $X\theta + \varepsilon_1 - \varepsilon_2 = Y$, where θ unrestricted in sign, $\varepsilon_1, \varepsilon_2 \geq 0$.

From the computational point of view, the L_1 method is now extremely simple and it requires only a routine to fit the L_1 regression. There are several computer programs available for calculation of L_1 estimates. See for example Sadovski (1974) and Farebrother (1988). For the case of multiple regresssion we can use, for example, the modified simplex algorithm of Barrodale and Roberts (1973) that exists in the IMSL library under the name RLLAV. The L_1 estimation problem with additional linear restrictions (restricted L_1 problem) is considered along the same lines in Barrodale and Roberts (1974). Arthanari and Dodge (1993) devoted a complete chapter on computational aspects of the L_1 estimation. L_1 regression estimates are also obtainable from the function l1fit in the computer language S-Plus and from the robust regression package ROBSYS (Marazzi, 1993). Detection of outlying points in both dependent and independent variables in regression model are explained in Dodge (1997).

The major difficulty for applied researchers in using L_1 estimation for many years was the lack of accompanying statistical inference procedures. Such procedures would include methods for testing general linear hypothesis, obtaining confidence intervals, analysis of variance tables and for performing multiple comparison procedures.

Bassett and Koenker (1978) developed the asymptotic theory for L_1 estimators in the regression model. Their finding is considered to be a breakthrough for the problem. Their main result is that the sampling distribution of L_1 estimators will be asymptotically normal with a specified mean and variance. Under very general assumptions they confirmed that the L_1 estimator $\widehat{\theta}$ has a normal distribution with mean θ and covariance matrix $\tau^2(X'X)^{-1}$ where τ^2/n is the asymptotic variance of the sample median from random samples of size n taken from the error distribution with a continuous and positive derivative at the median. This result is remarkably similar to that for LS. Therefore the L_1 confidence intervals for an estimable function $\lambda'\widehat{\theta}$ is

$$\lambda'\widehat{\theta} \pm z_{\alpha/2} \cdot \widehat{\tau}\sqrt{\lambda'(X'X)^{-}\lambda}$$

where $(X'X)^{-}$ is the q-inverse of $X'X$ and $\widehat{\tau}$ is an estimate of τ given in

McKean and Schrader (1987).

Koenker and Bassett (1982) investigate the asymptotic distribution of three alternative L_1 test statistics of a linear hypothesis in the standard linear model. They showed that the three test statistics, which correspond to the Wald, likelihood ratio and Lagrange multiplier tests, under mild regularity conditions on the design and error distribution, have the same limiting chi-square behavior. For a complete treatment of L_1 regression the reader is referred to Chapter 4 of Birkes and Dodge (1993).

With the availability of many computationally efficient algorithms and developed inference procedures for testing general linear hypotheses, for obtaining confidence intervals, selection of variables, analysis of variance tables and multiple comparison, it is hoped that L_1 estimation methods will be employed more often by researchers in applied sciences than before.

Certainly, there are many other areas of statistical data analysis based on the L_1-norm (such as density estimation, time series analysis, multivariate analysis and classification methods) that could have been discussed here. But, unfortunately, limitation of space and time have caused many interesting and important lines of research to be treated lightly or not at all. While it is now evident that no single robust procedure is best by any criteria, it may be appropriate (or at least reasonable) to use adaptive convex combinations of L_1 with other methods rather than a single criterion to estimate the unknown parameters. However, for the error distributions for which the median is superior to the mean as an estimator of location, L_1 estimation is certainly preferred to least squares and strongly recommended for use in these cases.

Bloomfield and Steiger (1983), Devroye and Györfi (1985), and Gonin and Money (1989) are the only books entirely devoted to L_1 topic. The authors of these three texts had the courage to pull together a rich and diverse literature in this field. I hope that the proceedings contained in this volume and its predecessors, Dodge (1987, 1992), will encourage someone to write a fourth.

References

[1] Arthanari, T.S. and Dodge, Y. (1993). *Mathematical Programming in Statistics.* New York: Wiley, Classical Edition.

[2] Barrodale, I. and Roberts, F.D.K. (1973). An improved algorithm for discrete L_1 linear approximation. *SIAM J. Numer. Anal.* **10,** 839-848.

[3] Barrodale, I. and Roberts, F.D.K. (1974). Algorithm 478: Solution of an overdetermined system of equations in the L_1 norm. *Commun.*

Assoc. Comput. Mach. **17,** 319.

[4] Bassett, G.W. and Koenker, R. (1978). Asymptotic theory of least absolute error. *J. Am. Statist. Assoc.* **73,** 618-22.

[5] Birkes, D. and Dodge, Y (1993). *Alternative Methods of Regression.* New York: Wiley.

[6] Bloomfield, P. and Steiger, W.L. (1983). *Least Absolute Deviations: Theory, Applications, and Algorithms.* Boston: Birkäuser.

[7] Boscovich, R.J. (1757). De litteraria expeditione per pontificiam synopsis amplioris operis, ac habentur plura eius ex exempla ria etiam sensorum impressa. Bononiensi Scientiarum et Artium Instituto Atque Academia Commentarii 4, 353-96.

[8] Charnes, A., Cooper, W.W. and Fergusson, R.O. (1955). Optimal estimation of executive compensation by linear programming. *Manage. Sci.* **1,** 138.

[9] Chatterjee, S. and Hadi, A.S. (1988). *Sensitivity Analysis in Linear Regression.* New York: Wiley.

[10] Cook, R.D. and Weisberg, S. (1982). *Residuals and Influence in Regression.* New York: Chapman and Hall.

[11] Cook, R.D. and Weisberg, S. (1994). *An Introduction to Regression Graphics.* New York: Wiley.

[12] Devroye, L. and Györfi, L. (1985). *Nonparametric Density Estimation, the L_1 View.* New York: Wiley.

[13] Dielman, T.E. and Rose, L.E. (1995). A bootstrap approach to hypothesis testing in least absolute value regression. *Comp. Statist. Data Anal.* **20,** 119-130.

[14] Dielman, T.E. and Rose, L.E. (1996). A note on the hypothesis testing in LAV multiple regression: A small sample comparison. *Comp. Statist. Data Anal.* **21,** 463-470.

[15] Dodge, Y. (1987). (Ed). *Statistical Data Analysis Based on the L_1-Norm and Related Methods.* Amsterdam: North-Holland.

[16] Dodge, Y. (1992). *L_1-Statistical Data Analysis and Related Methods.* Amsterdam: North Holland.

[17] Dodge, Y. (1997). LAD regression for detecting outliers in response and explanatory variables. *J. Multiv. Anal.* **61,** 144-158.

[18] Edgeworth, F.Y. (1887). On observations relating to several quantities. *Phil. Mag.* (5th Series) **24,** 222.

[19] Ellis, S.P. and Morgenthaler, S. (1992). Leverage and breakdown in L_1 regression. *J. Am. Statist. Assoc.* **87,** 143-148.

[20] Farebrother, R.W. (1988). A simple recursive procedure for the L_1 norm fitting of a straight line. *Appl. Statist.* **37,** 457-489.

[21] Gonin, R. and Money, A.H. (1989). *Nonlinear L_p-Norm Estimation.*

New York: Marcel Dekber.

[22] Huber, P.J. (1987). The place of the L_1-norm in robust estimation. In *Statistical Data Analysis - Based on L_1-Norm and Related Methods*, Ed. Y. Dodge, pp. 23-33. Amsterdam: North Holland.

[23] Huber, P.J. (1995). Robustness: Where are we now ? *Student* **1,** 75-86.

[24] Karst, O.J. (1958). Linear curve fitting using least deviations. *J. Am. Statist. Assoc.* **53,** 118-132.

[25] Koenker, R. and Bassett, G.W. (1982). Tests of hypotheses and L_1 estimation. *Econometrica* **50,** 1577-83.

[26] Koenker, R. (1987). A comparison of asymptotic testing methods for L_1-regression. In this volume.

[27] Laplace, P.S. (1789). Sur les degrés mesurés des méridiens, et sur les longueurs observées sur pendule. Histoire de l'Académie royale des inscriptions et belles lettres, avec les Memoires de littérature tirez des registres de cette académie, Année 1789. Paris. "Mémoires", 18-43.

[28] Laplace, P.S. (1793). Sur quelques points du système du monde. Mémoires de l'Académie Royale des Sciences de Paris, 1-87. Reprinted in Oeuvres Complètes de Laplace, Vol.11, pp. 477-558. Paris: Gauthier-Villars, 1895.

[29] Laplace, P.S. (1818). Deuxième supplément to Laplace (1812).

[30] Legendre, A.M. (1805). Nouvelles méthodes pour la détermination des orbites des comètes. Paris: Courcier. Appendice sur la méthode des moindres carrés, 72-80.

[31] Marazzi, A. (1993). *Algorithms, Routines, and S Functions for Robust Statistics.* Wadsworth and Brooks/Cole, California.

[32] McKean, J.W. and Schrader, R.M. (1987). Least absolute errors analysis of variance. In *Statistical Data Analysis Based on the L_1-Norm and Related Methods*, Ed. Y. Dodge, pp 297-321. Amsterdam: North Holland.

[33] Rhodes, E.C. (1930). Reducing observations by the methode of minimum deviations. *Phil. Mag.* (7th Series) **9,** 974.

[34] Rousseeuw, P.J. and Leroy, A.M. (1987). *Robust Regression and Outlier Detection.* New York: Wiley.

[35] Sadowski, A.N. (1974). Algorithm AS74. L_1-norm fit of a straight line. *Appl. Statist.* **23,** 244-248.

[36] Singleton, R.R. (1940). A method of minimizing the sum of absolute values of deviations. *Ann. Math. Statist.* **11,** 301-310.

[37] Turner, H.H. (1887). On Mr. Edgeworth's method of reducing observations relating to several quantities. *Phil. Mag.* (5th Series) **24,** 466-70.

[38] Wagner, H.M. (1959). Linear programming techniques for regression analysis. *J. Am. Statist. Assoc.* **54,** 206.