L_1-*Statistical Procedures and Related Topics*
IMS Lecture Notes – Monograph Series (1997) Volume 31

Notes on the early history of elemental set methods

Richard William Farebrother

Victoria University of Manchester, UK

Abstract: In this paper we outline the early history of traditional estimation procedures which are based on the use of elemental sets. There are two distinct classes of such procedures associated with the minimum values of the sum of absolute errors and the largest absolute error criteria respectively. As a matter of historical necessity, our study will concentrate on estimation procedures of the first type. However we shall also discuss some recent work on the least median of squared errors procedure which in principle involves elemental sets of the second type.

Key words: Least median of squared errors, least sum of absolute errors, least sum of squared errors, linear programming, minimax absolute error.

AMS subject classification: 01A50, 01A55, 62J05.

1 Introduction

In his study of the practical value of elemental set approximations to robust estimation procedures, Hawkins (1993, p.580) has summarised the history of such methods in the following terms:

Elemental set methods have their origins in the single-predictor proposal by Theil (1950). The extension of the idea to handling outlier problems in multiple regression was made independently by Rousseeuw (1984) and by Hawkins, Bradu and Kass (1984). They also arise naturally in the expression of the OLS multiple regression in terms of weighted U-statistics (as sketched in the technical appendix to Hawkins *et al.*, 1984).

Although this brief statement may well have been sufficient for the purposes of Hawkins's paper, it cannot be regarded as an adequate summary of the history of elemental set methods as it fails to mention that meth-

ods of this type had been used to estimate the parameters of linear and nonlinear models with two or more predictors for more than 240 years, or that the elemental set characterisation of the ordinary least squares (OLS) estimator has been known for more than 150 years. The purpose of the present paper is to provide a more comprehensive survey of the history of this subject. We shall also take the opportunity to mention some recent work on the least median of squared errors procedure.

2 Nature of the fitting problem

Some readers of this paper may be prompted to further their historical studies by consulting some of the original sources cited in it. Such readers will soon discover that the traditional notation and nomenclature of the Calculus of Observations is quite distinct from that of modern Mathematical Statistics. To help the interested reader through this difficulty we shall employ a variant of this traditional usage in our history.

In traditional notation the familiar curve or surface fitting problem may be expressed in the following terms: We are given a linear or nonlinear function $f(.)$ which is characterised by a set of p observed (variable) quantities $a, b, c, etc.$ and a set of q unobserved fixed quantities (we would call them parameters) $x, y, z, etc.$ We suppose that the variable quantities $a, b, c, etc.$ are observed without error but that the value of the function $f(.)$ is subject to an additive error. Denoting the observed value of the function by m and the corresponding value of the additive error by v, we find that we have a system of n equations:

$$m_i = f(a_i, b_i, c_i, etc.; x, y, z, etc.) + v_i \quad i = 1, 2, \ldots, n$$

which describes the relationship between the observed values of the variable quantities and the corresponding observed values of the function. In this context we have to choose values for the q unknown quantities $x, y, z, etc.$ in such a way that the observed errors (we would say residuals) $v_1, v_2, \ldots, v_n$ are as small as possible in some sense.

Although this general statement of the problem admits the possibility of nonlinear functions, we shall be largely concerned with functions $f(.)$ which are linear in the unknown constants. The major exceptions to this rule being found in the final paragraph of Section 3.

In passing we note that this traditional notation does not distinguish between the true and fitted values of the errors or between the true and fitted values of the unknown constants. This imprecision is of no immediate consequence as we shall only be concerned with fitted values in the remainder of this paper.

3 Elemental set methods

The most obvious solution to the general fitting problem outlined in Section 2 is to discard $n - q$ of the available equations and to use the remaining q equations to determine a set of values for the q unknown constants in such a way that these q equations are exactly satisfied with zero errors. This fitting procedure is variously known as the method of selected points, the subset selection method, and the method of elemental sets.

Now there are ${}^nC_q$ distinct ways of choosing q equations from a set of n equations, so that practitioners have either to choose a single set of q equations at random or according to some specified rule, or they have to select a greater number of sets and attempt to reconcile the discordant results obtained from their selection.

In the earliest period of enquiry in this area, scientists employed a variant of the first procedure in which the selection criterion is obscure, if not entirely hidden. Without making any attempt to explain their reasons, these authors arranged that there should be as many equations as were required for the problem to have a unique solution. For example, in his analysis of the problem of determining the height of a tree on a remote hillside, Liu Hui (third century) assumed that the surveyor had taken observations on the elevation of the top of the tree from each of two locations in a horizontal reference plane and a third observation on the elevation of the base of the tree from one of these locations. These three observations together with the known horizontal distance between the two observation sites are sufficient to determine the three unknowns of the problem, see Li and Du (1987, pp.76–78) for details. However it is not explained why the surveyor should not have observed both elevations from both locations, or what was to be done if he had. It is intriguing to speculate on how Liu Hui would have responded if he had been challenged on this point.

By the middle of the eighteenth century, this implicit choice of a single set of q equations had been replaced by a more explicit procedure. Mayer (1750, p.150), for example, suggested that one should choose a set of q equations that are typical of the n given equations. In the particular case of his determination of the position of the lunar crater Manilius, he obtained a system of $n = 27$ equations in $q = 3$ unknowns of the form

$$m_i = x + y\sin(\theta_i) + z\cos(\theta_i) + v_i \quad i = 1, 2, \ldots, n$$

and suggested that the $q = 3$ typical equations should be chosen in such a way that two of the angles differ from the third by 90 and 180 degrees respectively. However, having advanced this basic suggestion, he observed that one needs more than a single set of typical equations if one wishes to

check the accuracy of the original solution. In the limiting case one thus obtains a total of $[n/q]$ distinct solutions from disjoint subsets of the data which then have to be reconciled with one another.

Unfortunately Mayer did not discuss this technique in sufficient detail for us to be sure how many disjoint subsets of q equations he would have used, or how he would have attempted to reconcile the corresponding discordant results. Nevertheless his lack of precision in this regard is entirely comprehensible as his exposition of this topic was as a brief aside to his main purpose of introducing a new fitting procedure known as the method of averages.

A few years later, Boscovich addressed a similar problem relating to the ellipsoidal figure of the Earth in a similar way. In his contribution to Maire and Boscovich (1755) and again in Boscovich (1757), Boscovich was concerned with the solution of a system of $n = 5$ linear equations in $q = 2$ unknowns. In his first approach to this problem he evaluated all ${}^5C_2 = 10$ pairwise determinations of the unknown constants before taking an unweighted arithmetic mean. He was not satisfied with the result and tentatively suggested a variant which discards the two determinations with the smallest denominators before again taking an unweighted average of the remaining eight values.

Boscovich was far from satisfied with the results he obtained from either variant of this procedure and, like Mayer before him, resolved the impasse by proposing an alternative fitting procedure. Boscovich's alternative procedure is to be found in his scientific notes to a poem in Latin hexameters by Stay (1760, pp.420–425), see Farebrother (1993). It chooses values for the unknown constants in such a way as to minimise the sum of the absolute values of the observed errors subject to the condition that the corresponding sum of the signed errors is zero. The relationship between this procedure and the method of elemental sets will be outlined in the following section.

Finally, in this connection, we must observe that the method of elemental sets was not entirely supplanted by more advanced methods (notably the method of least squares) until well into the present century. This statement is particularly true of nonlinear problems as Pearson (1902, p.298) mentions the possible use of the method of elemental sets to fit Makeham's law (a generalisation of the Gompertz function) to actuarial data, and Yule (1925, pp.49–50) notes that this method produces acceptable results when fitting a logistic function to sufficiently smooth demographic data. Indeed, the more elementary statistical textbooks of the 1950s and 1960s still recommended the method of elemental sets for the latter purpose, see Croxton and Cowden (1939; 1955, p.215).

4 Elemental set characterisations

Boscovich provided a geometrical algorithm as an integral part of his solution procedure. This algorithm was subsequently given an analytical form by Laplace (1793), see Farebrother (1993), Sheynin (1973), or Stigler (1986) for details. Some years later Gauss (1809, sec.186) generalised Boscovich's optimality criterion to any number of unknowns and suggested that the adding-up constraint could be deleted. (In passing we note that this adding-up constraint has no direct connection with the method of least squares which was developed by Gauss and Legendre some years after Boscovich's death.)

In his discussion of the unconstrained least sum of absolute errors problem, Gauss (1809, sec.186) notes that the optimal solution to this problem is characterised by a set of q zero errors and that the other $n - q$ errors only help to determine this optimal set. He gives no justification for this result but a proof which would have been accessible to Gauss and his contemporaries has been suggested by Waterhouse (1990).

An explicit characterisation of the solution to the least squares problem as a weighted sum of elemental set determinations was established by Jacobi (1841). See Sheynin (1973) for an excellent description of this result which closely follows Jacobi's own derivation. This result was subsequently rediscovered by Glaisher (1879), Subrahmanyam (1972), Hawkins, Bradu and Kass (1984), and Ben-Tal and Teboulle (1990) amongst others. The long interlude between the publication of the papers by Glaisher and Subrahmanyam would seem to be due to the intervention of a clear statement of Jacobi's result in the popular textbook by Whittaker and Robinson (1924; 1944, pp.251–252).

This explicit characterisation of the solution to the least squares problem may be generalised to a weighted sum of least squares determinations from sets of $m > q$ equations, see Sheynin (1993) and Wu (1986) for details. As a further generalisation of this result, Ben-Tal and Teboule (1990) have shown that, for all members of a class of strictly isotone functions which includes the weighted sum of the kth powers of the absolute errors (for some fixed finite positive value of k), every set of values for the unknown constants which minimises the chosen function of the errors will lie within the convex hull of the elemental set determinations. In addition they have shown that, for all members of a class of isotone (but not strictly isotone) functions including the largest absolute error and the median absolute error functions, at least one of the sets of values for the unknown constants which minimise the chosen function of the errors must lie within the convex hull of the elemental set determinations of these values.

Thus, although the solutions to the least sum of absolute errors ($k = 1$) and the least sum of squared errors ($k = 2$) problems must lie in the convex hull of the elemental set determinations, the solutions of the minimax absolute errors and least median of (squared) absolute errors problems will necessarily be members of the convex hull only if these solutions are unique.

Further, although all the solutions of the weighted sum of kth powers (or other strictly isotone) problem must lie in a convex set for all sets of weights, the set of all such solutions need not itself be convex. Gilstein and Leamer (1983) have given a precise characterisation of the nonconvex set of solutions to the weighted least squares problem.

5 Minimax absolute error criterion

The class of (Gaussian) elemental set methods discussed in Sections 3 and 4 may be associated with the optimal value of the sum of absolute errors criterion. A second class of (Laplacian) elemental set methods may similarly be associated with the optimal value of the maximum absolute error criterion. As its name implies, the minimax absolute error procedure chooses values for the unknown constants in such a way as to minimise the largest in absolute value of the n observed errors. This fitting procedure was first discussed by Laplace (1786). Given a set of n linear equations in $q = 2$ unknowns, Laplace arbitrarily selected a set of $r = q + 1$ equations. Using any q of these equations to eliminate the q unknowns from the remaining equation, he obtained a single (reduced) equation with a linear function of the r observed errors on one side and a nonnegative constant on the other. Without further explanation, he asserted that the largest in absolute value of these r errors is minimised when all r errors take the same absolute value and their signs are given by the signs of the corresponding coefficients in the single reduced equation.

A determinantal formulation of this procedure was subsequently developed by de la Vallée Poussin (1911). In this alternative formulation of the problem one has to set the r selected errors proportional to the signs of the cofactors of $m_1, m_2, \ldots, m_r$ in the $r \times r$ determinant

$$\begin{vmatrix} a_1 & b_1 & c_1 & \ldots & m_1 \\ a_2 & b_2 & c_2 & \ldots & m_2 \\ \vdots & \vdots & \vdots & & \vdots \\ a_r & b_r & c_r & \ldots & m_r \end{vmatrix}$$

where, for notational simplicity, we assume that the trial solution is defined by the first r equations of the model. The values of the unknown constants and the common absolute value of the r errors are then obtained by applying

Cramer's Rule to the corresponding system of r equations in r unknowns.

Under either of these schemes we have to evaluate all n errors $v_1, v_2, \ldots, v_n$ and select a new set of r equations if one or more of these errors is larger in absolute value than the common value of the r errors in the current set. Laplace (1786) apparently proposed to choose this new set without reference to earlier selections. By contrast, de la Vallée Poussin (1911) introduced an automated selection procedure which Stiefel (1960) subsequently identified with his own (1959) procedure and both as equivalent to a standard simplex implementation of the linear programming dual formulation of the minimax problem.

In this context it is interesting to note that Farebrother (1985) has shown that the minimax absolute error procedure is closely related to the linear programming dual formulation of the least sum of absolute errors problem. Also see Farebrother (1997) for a detailed account of the early history of the minimax absolute error procedure with particular reference to the work of de Prony (1804) and Fourier (1827).

6 Median squared absolute error criterion

These results on the minimax absolute error procedure have recently come to prominence as Rousseeuw's (1984) so-called least median of squares procedure actually chooses values for the unknown constants in such a way as to minimise the median or middlemost value of any increasing function of the absolute values of the observed errors. For example, suppose that we wish to minimise the hth largest squared error where h is set close to $n/2$. Then, conditional on the choice of the $h-1$ equations which are to be ignored, the least median of squared errors problem may be expressed in the form of a minimax absolute error problem applied to the $n-h+1$ retained equations. The optimal solution to the least median of squared errors problem is thus also characterised by a (Laplacian) elemental set determination of the unknown constants. In principle, we may therefore determine the optimal values of these q constants by evaluating the median squared error function for a sufficiently large sample of the minimax absolute error determinations of the q unknowns from sets of $q+1$ equations.

However, it is clear that the minimax fitting of a system of $q+1$ equations in q unknowns is vastly more expensive than the direct solution of a set of q equations in q unknowns. Rousseeuw and Leroy (1987) have therefore suggested that a sufficiently accurate approximation to the exact least median of squares solution may be obtained by evaluating the median squared error function for a sufficiently large sample of Gaussian elemental set determinations. This conjecture was subsequently confirmed

by Stromberg (1993) whilst Hawkins (1993) has shown that this technique yields satisfactory results for a wide class of robust fitting procedures.

References

[1] Ben-Tal, A. and Teboulle, M. (1990). A geometric property of the least squares solution of linear equations. *Linear Algebra and Its Applications* **139** 165–170.

[2] Boscovich R.J. (1757). De litteraria expeditione per pontificiam ditionem et synopsis amplioris operis ..., *Bononiensi Scientiarum et Artum Instituto atque Academia Commentarii* **4** 353–396. Reprinted by the Institute for Higher Geodesy, Zagreb, 1961.

[3] Croxton, F.E. and D.J. Cowden. (1939). *Applied General Statistics.* Prentice–Hall Inc., Englewood Cliffs, New Jersey. Second Edition, Sir Isaac Pitman and Sons, London, 1955.

[4] De la Vallée Poussin, C.J. (1911). Sur la méthode de l'approximation minimum. *Annales de la Société Scientifique de Bruxelles* **35** 1–16.

[5] de Prony, G.C.F.M.(1804). *Recherches physico-mathématiques sur la théorie des eaux courantes.* Paris: Imprimerie Impériale.

[6] Farebrother, R.W. (1985). Unbiased L_1 and L_∞ estimation. *Commun. Statist. Theory and Method* **14** 1941–1962.

[7] Farebrother, R.W. (1987). The historical development of L_1 and L_2 estimation procedures 1793–1930. In *Statistical Data Analysis based on the L_1-norm and Related Methods,* Ed. Y. Dodge, pp. 37-63. Amsterdam: North-Holland Publishing Company.

[8] Farebrother, R.W. (1992). Geometrical foundations of a class of estimators which minimise sums of Euclidean distances and related quantities. In *L_1-Statistical Analysis and Related Methods,*Ed. Y. Dodge, pp. 337-349. Amsterdam: North-Holland Publishing Company.

[9] Farebrother, R.W. (1993). Boscovich's method for correcting discordant observations. In *R.J. Boscovich: His Life and Scientific Work,*Ed. P. Bursill–Hall, pp. 255-261. Rome: Istituto della Enciclopedia Italiana.

[10] Farebrother, R.W.(1997). The historical development of the linear minimax absolute residual estimation procedure 1786–1960. *Comput. Statist. Data Anal.* Forthcoming.

[11] Fourier, J.B.J. (1827). Second Extrait, *Histoires de l'Académie Royale des Sciences de Paris [pour 1824]*, p. 47. Reprinted in his *Oeuvres,*Vol 2. Gautier–Villars et Fils, Paris, 1890, pp. 325 –328.

[12] Gauss, C.F. (1809). *Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientium.*Hamburg: F. Perthes and I.H. Besser.

Reprinted in his *Werke,*Vol. 7. F Perthes, Gotha, 1871.

[13] Gilstein, C.Z. and E.E. Leamer (1983). The set of weighted regression estimates. *J. Am. Statist. Assoc.* **78** 942–948.

[14] Glaisher, J.W.L. (1879). On the method of least squares. *Memoirs of the Royal Astronomical Society* **40** 600–614.

[15] Hawkins, D.M. (1993). The accuracy of elemental set approximations for regression. *J. Am. Statist. Assoc.* **88** 580–589.

[16] Hawkins, D.M., Bradu D. and C.V. Kass (1984). Location of several outliers in multiple regression data using elemental sets. *Technometrics* **19** 197–208.

[17] Jacobi, C.G.J. (1841). De formatione et proprietatibus determinatium. *Journal für die reine und angewandte Mathematik* **22** 285–318. Reprinted in his *Werke*, Vol.3. Georg Reimer, Berlin, 1884, pp.355–392.

[18] Laplace, P.S. (1786). Mémoire sur la figure de la Terre. *Mémoires de l'Académie Royale des Sciences de Paris [pour 1783],* pp.17–46. Reprinted in his *Oeuvres Complètes,* vol.11. Gauthier–Villars et Fils, Paris, 1895, pp.3–32.

[19] Laplace P.S. (1793). Sur quelques points du système du monde. *Mémoires de l'Académie Royale des Sciences de Paris [pour 1789],* pp.1–87. Reprinted in his *Oeuvres Complètes.* Vol. 11. Gauthier–villars et Fils, Paris, 1895, pp.447–558.

[20] Li, Y. and S. Du (1987). *Chinese Mathematics: A Concise History.*Translated by J.N. Crossley and A.W.C. Lun. Oxford University Press, Oxford.

[21] Maire, C. and R.J. Boscovich. (1755). *De Litteraria Expeditione per Pontificiam Ditionem.*Rome: N and M Palearini.

[22] Mayer, J.T. (1750). Abhandlung über die Umwalzung des Monds um seine Axe. *Kosmographische Nachrichten und Sammlungen [1748]* **1** 52–183.

[23] Pearson, K. (1902). On the systematic fitting of curves. *Biometrika* **1** 265–303.

[24] Rousseeuw, P.J. (1984). Least median of squares regression. *J. Am. Statist. Assoc.* **79** 871–880.

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

[26] Sheynin, O.B. (1973). R J Boscovich's work on probability. *Archive for History of Exact Sciences* **9** 306–324 and **28**, 173.

[27] Sheynin, O.B. (1993). On the history of the principle of least squares. *Archive for History of Exact Sciences* **46** 39–54.

[28] Stay, B. (1760). Annotated by R.J. Boscovich, *Philosophiae Recentioris... Versibus Traditae...*, *Tomus II.*Rome: N and M Palearini.

[29] Stiefel, E.L. (1959). Uber diskrete und lineaire Tschebyscheff Approximationen. *Numerische Mathematik* **1** 1–28.

[30] Stiefel, E.L. (1960). Note on Jordan elimination, linear programming and Tchebycheff approximation. *Numerische Mathematik* **2** 1–17.

[31] Stigler, S.M. (1986). *The History of Statistics: The Measurement of Uncertainty before 1900.*Harvard University Press, Cambridge, Massachusetts.

[32] Stromberg, A.J. (1993). Computing the exact value of the least median of squares estimate and stability diagnostics in multiple linear regression. *SIAM J. Scient. Comput.* **14** 1289–1299.

[33] Subrahmanyam, M (1972). A property of simple least squares estimates. *Sankhyā* B **34** 355–356.

[34] Theil, H (1950). A rank-invariant method for linear and polynomial regression analysis. *Koninklijke Nederlandse Akademie van Wetenschappen Proceedings* Part A **53** 386–392, 521–525 and 1397–1412.

[35] Waterhouse, W.C. (1990). Gauss's first argument for least squares. *Archive for History of Exact Sciences* **41**, 41–52.

[36] Whittaker, E.T. and G. Robinson (1924). *The Calculus of Observations.*Blackie and Son Ltd., London. Fourth Edition (1944). Reprinted by Dover Publications Inc., New York, 1967.

[37] Wu, C.F.J. (1986). Jacknife, Bootstrap and other resampling methods in regression analysis. *Ann. Statist.***14** 1261–1295.

[38] Yule, G.U. (1925). The growth of population and the factors which control it, *J. R. Statist. Soc.* **88** 1–62.