THIS REPORT HAS BEEN DELIMITED AND CLEARED FOR PUBLIC RELEASE UNDER DOD DIRECTIVE 5200.20 AND NO RESTRICTIONS ARE IMPOSED UPON ITS USE AND DISCLOSURE.

DISTRIBUTION STATEMENT A

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.
NOTICE: When government or other drawings, specifications or other data are used for any purpose other than in connection with a definitely related government procurement operation, the U. S. Government thereby incurs no responsibility, nor any obligation whatsoever; and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data is not to be regarded by implication or otherwise as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use or sell any patented invention that may in any way be related thereto.
LINEAR PROGRAMMING TECHNIQUES
FOR REGRESSION ANALYSIS

BY
HARVEY M. WAGNER

TECHNICAL REPORT NO. 51

PREPARED UNDER CONTRACT N6onr-25133
(NR-047-004)
FOR
OFFICE OF NAVAL RESEARCH

DEPARTMENT OF ECONOMICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

MAY 5, 1958
LINEAR PROGRAMMING TECHNIQUES FOR REGRESSION ANALYSIS

BY

HARVEY M. WAGNER

TECHNICAL REPORT NO. 51

PREPARED UNDER CONTRACT N6onr-25133
(NR-047-004)
FOR
OFFICE OF NAVAL RESEARCH

REPRODUCTION IN WHOLE OR IN PART IS PERMITTED FOR
ANY PURPOSE OF THE UNITED STATES GOVERNMENT

DEPARTMENT OF ECONOMICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

MAY 5, 1958
1. Introduction

Karst [5] has recently suggested an iterative procedure "for finding a straight line of best fit to a set of two dimensional points such that the sum of the absolute values of the vertical deviations of the points from the line is a minimum." It is well known that the general \( p + 1 \) dimensional version of this problem may be exactly formulated as a linear programming model consisting of \( n \) equations, where \( n \) is the number of observations. By employing the fundamental dual theorem [1, 6, 8] in linear programming, we shall show how the problem can be solved by a \( p \) equation linear programming model with bounded variables [2, 3, 9]. Secondly we shall demonstrate how a regular \( p + 1 \) equation linear programming model can be utilized to find a line of best fit according to a Chebyshev criterion [4], i.e., a line (or hyperplane) which minimizes the maximum of the vertical deviations from the sample points.

2. Minimizing the Sum of Absolute Deviations

Let \( X \) denote an \( n \times p \) dimensional matrix, where the columns consist of \( n \) observational measurements on \( p \) "independent" variables, and \( Y \) denote an \( n \)-dimensional column vector of measurements on the "dependent" variable.
We wish to find a p-dimensional column vector \( b \) such that

\[
Xb + Ie_1 - Ie_2 = Y, \quad e_1, e_2 \geq 0
\]

minimize \( E = (1 1 \ldots 1) \begin{pmatrix} e_1 \\ e_2 \end{pmatrix} \),

where \( I \) is an \( n \times n \) identity matrix. We interpret \( e_1 \) and \( e_2 \) as \( n \)-dimensional column vectors of vertical deviations "below" and "above" the fitted line; i.e., \( (e_1 + e_2) \) is the vector of absolute deviations between the fit \( Xb \) and \( Y \) (by the nature of the model, it is clear that the \( j \)-th components of \( e_1 \) and \( e_2 \) cannot both be strictly positive in an optimal solution). The solution to our problem yields the regression equation

\[
\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_p \end{pmatrix} = xb = y. \tag{2}
\]

Note that if we wish the left hand side of (2) to include a coefficient for the intercept of the \( y \) axis to be determined by the linear fit, then we can let \( x_p \equiv 1 \), and the \( p \)-th column of \( X \) be a vector of one's. We may force the fitted line to pass through some point, the usual example being the set of sample means, either by adding to (1) the linear restriction

\[
(\bar{x}_1 \ \bar{x}_2 \ \ldots \ \bar{x}_p) b = \bar{y} \tag{3}
\]
or by the usual least squares approach of subtracting each coordinate of the point, in our example the sample mean for each variable, from all the corresponding observations (including \( y \)) and then by fitting (1) without a \( y \)-intercept coefficient; the latter approach simply consists of shifting the origin of the axes in a \( p \)-dimensional space to the selected point, and then of fitting the line (hyperplane) through the new origin.

If it is desirable, the linear programming model (1) can be restricted further to permit only non-negative values for some or all of the components of \( b \), and to force \( b \) to satisfy additional linear constraints. It is noteworthy that collinearity in \( X \) (even to the degree that two columns of \( X \) are identical) will not cause a failure in the algorithm for (1). One drawback of the model is evident: when the number of observations \( n \) is sizeable, (1) becomes computationally unwieldy.

We shall now transform (1) into a more manageable dual problem which will yield the optimal \( b \) as a byproduct. To start, we shall assume we have added to (1) the restriction \( b \geq 0 \). The fundamental dual relationship in linear programming [1, 6, 8] asserts a solution to (1) can be found by considering the linear programming model

\[
X'd \leq \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \quad \quad (\text{4a})
\]

\[
Id \leq \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \quad \quad (\text{4b})
\]
where $X'$ is the transpose of $X$, $Y'$ the transpose of $Y$, and $d$ an n-dimensional column vector of "dual variables" which are unrestricted in sign (because (1) consists of a set of equations). Model (4), as it appears, is even a larger problem than (1), since it consists of $p + 2n$ relations. To reduce the problem to a model in $p$ equations and $n$ bounded variables we let

$$
(f) = \begin{pmatrix}
  f_1 \\
  f_2 \\
  \vdots \\
  f_n
\end{pmatrix} = (d) + \begin{pmatrix}
  1 \\
  1 \\
  \vdots \\
  1
\end{pmatrix} \quad (5)
$$

Then (4) is equivalent to

$$
X'f \leq X' \begin{pmatrix}
  1 \\
  1 \\
  \vdots \\
  1
\end{pmatrix}, \quad (6a)
$$

$$
0 \leq f_j \leq 2 \quad j = 1, 2, \ldots, n \quad (6b)
$$

$$
\max G^* = Y'f - Y' \begin{pmatrix}
  1 \\
  1 \\
  \vdots \\
  1
\end{pmatrix} \quad (6c)
$$
Upon appending a set of slack variables to (6a) and dropping the constant on the right side of (6c), we may solve (6) by one of the simplex algorithms for bounded variables [2, 3, 9]. If X and Y are deviations of sample values from their means, then the right hand side of (6a) is a vector of zero's, and the constant in (6c) is zero. Denoting the basis of the optimal solution of (6) by B (which may include slack vectors), and the associated coefficients in (6c) by \( r'_B \), we have

\[
b = (B^{-1})'r_B.
\]  

(7)

No extra computations are needed to find (7). In the original simplex method \( b \) appears in the \(( z_j - c_j )\) row of the final simplex tableau under the columns for the slack vectors [1, 8]; in the revised simplex method (7) is the "shadow price" vector for the optimal solution [7]. The optimal value of \( G^* \) is the minimized sum of absolute deviations.

When we drop our assumption that \( b \) be non-negative and allow the components of \( b \) to take on any sign, we modify (6) to

\[
X'f = X' \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \quad (6a')
\]

and introduce a set of artificial variables having an arbitrarily high cost to initiate one of the simplex algorithms. The optimal \( b \) remains (7), i.e., the shadow price vector in the revised simplex method or \( z_j \) of the final simplex tableau under the columns for the artificial vectors [1].
In summary, we can solve for \( b \) in (1) by applying a simplex algorithm for bounded variables to the \( p \) equation model (6). Although the mathematical manipulation underlying the transformation of problems appears involved, the computational procedure required to solve (6) is relatively straightforward, but somewhat laborious.

3. Minimizing the Maximum Absolute Deviation

The most bothersome aspect of the approach in the previous section is the requirement of a linear programming algorithm for bounded variables, as such techniques are (slightly) more difficult to perform than the standard simplex algorithm. We may eliminate the drawback if we are willing to accept a Chebyshev criterion for best fit. Our model in this case is to find a \( p \) dimensional column vector \( b \) such that

\[
Xb - Y \leq \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} e \quad (8a)
\]

\[
-Xb + Y \leq \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} e \quad (8b)
\]

minimize \( e \geq 0 \) \quad (8c)

Examination of (8) will reveal that \( e \) is the maximum absolute deviation. The equations (8) are reminiscent of a linear programming formulation for the minimax problem in two-person zero-sum games, and we shall use a similar approach for the solution. An equivalent expression for (8) is
where \( \vec{1} \) is an \( n \)-dimensional column vector of one's. Our previous remarks concerning additional linear constraints on \( b \) apply here equally as well.

Assuming for the moment that we wish to impose the restriction \( b \geq 0 \), we convert the 2n equation model (9) into its dual form, which contains only \( p + 1 \) equations

\[
\begin{array}{c}
\begin{pmatrix}
x' & x'
\end{pmatrix}
\begin{pmatrix}
h_1 \\
h_2
\end{pmatrix} \leq \begin{pmatrix} 0 \\
1
\end{pmatrix}
\end{array}
\tag{10a}
\]

\[
\begin{array}{c}
\begin{pmatrix}
1' & 1'
\end{pmatrix}
\begin{pmatrix}
h_1 \\
h_2
\end{pmatrix} \leq 1
\end{array}
\tag{10b}
\]

\[
h_1, \ h_2 \geq 0,
\]

\[
\text{maximize } M = \begin{pmatrix}
-Y' & Y'
\end{pmatrix}
\begin{pmatrix}
h_1 \\
h_2
\end{pmatrix},
\tag{10c}
\]

where \( \vec{0} \) is a \( p \)-dimensional column vector of zero's. The vectors \( h_1 \) and \( h_2 \) are \( n \)-dimensional columns; if a component of \( h_1 \) (\( h_2 \)) is positive in the optimal solution of (10), then the maximum deviation occurs at the corresponding point or equation in (8), and this point will lie "below" ("above") the fitted line. Analogous to our result in (7),
where $B$ denotes the optimal basis for (10), and $r_B^\prime$ the coefficients in (10c) corresponding to the variables in $B$; and exactly as before, the solution (11) is a byproduct of the simplex method.

If we drop the assumption that $b$ be non-negative, we need only change (10a) to equalities, and results analogous to those in the previous section continue to hold.

4. A Numerical Example

Karst [5] examines the following data

$$X' = [-12.5 -8.5 -6.5 -3.5 -2.5 -1.5 -0.5 2.5 4.5 8.5 8.5 11.5]$$

$$Y' = [-8.4 -5.4 3.6 -2.4 -4.4 1.6 -0.4 -0.4 -2.4 3.6 5.6 9.6],$$

which comprise deviations of the original data about their sample means.

He finds the least squares fit to be

$$y = .539 \times$$

and the fit for the minimized sum of absolute deviations to be

$$y = .659 \times$$

As we have argued, (14) can be obtained by (6), where specifically we would find

$$b = (B^{-1})^\prime r_B = (1/8.5) 5.6 = .659$$

(7')
The solution by model (10) yields
\[
B = \begin{pmatrix} -6.5 & 11.5 \\ 1 & 1 \end{pmatrix}, \quad r_B' = [3.6 \quad 9.6] \tag{15}
\]
and consequently
\[
\begin{pmatrix} b \\ e \end{pmatrix} = \begin{pmatrix} .333 \\ 5.767 \end{pmatrix}. \tag{11'}
\]
That is, the Chebyshev fit is
\[
y = .333 x; \tag{16}
\]
since the vectors in B correspond to variables in \( h_2 \), the third and last sample point in (12) will lie above the fitted line and assume the maximum vertical deviation from it of 5.767.
REFERENCES


| Office of Naval Research Branch Office          | U.S.A.F. Air University Library | 1 |
| Office of Naval Research Branch Office          | Maxwell Air Force Base, Alabama | 1 |
| 346 Broadway, New York 13, N.Y.                  | U.S. Naval Supply Research and Development Facility | 1 |
| Office of Naval Research Branch Office          | Naval Supply Depot Bayonne, New Jersey | 1 |
| 1030 E. Green St., Pasadena 1, California       | Weapons Systems Evaluation Group | 1 |
| Office of Naval Research Branch Office          | Pentagon Bldg. Washington 25, D.C. | 1 |
| 1000 Geary St., San Francisco 9, California     | Ames Aeronautical Laboratory | 1 |
| Office of Naval Research Navy No. 100           | Moffett Field, California Attn: Technical Library | 1 |
| Fleet Post Office New York, N.Y.                | Armed Services Technical Info. Agency | 1 |
| Office of Naval Research Logistics Branch, Code 436 | Arlington Hall Station Arlington 12, Va. | 5 |
| Dept. of the Navy T3-Bldg. Washington 25, D.C.  | The Director Naval Research Laboratory | 1 |
| Air Controller's Office Headquarters, U.S. Air Force | Attn: Tech. Information Office | 1 |
| Washington 25, D.C.                             | Chief, Bureau of Supplies and Accounts | 1 |
| Office of Technical Services Dept. of Commerce  | Advanced Supply System R and D Division (384) Room 2434 Arlington Annex Washington 25, D.C. | 1 |
| Washington 25, D.C.                             | Naval War College Logistics Dept., Luce Hall Newport, Rhode Island | 1 |
| Operations Research Office The Johns Hopkins University 6410 Connecticut Ave. Chevy Chase, Maryland | Director National Science Foundation Washington 25, D.C. | 1 |
NACA for Aeronautics
1724 F Street, N.W.
Washington 25, D.C.
Attn: Chief, Office of
Aeronautical Engineering

Director
Operations Evaluation Group
Office of Chief
of Naval Operations (OP-03EG)
Navy Dept.
Washington 25, D.C.

Industrial College of the Armed Forces
Fort Lesley J. McNair
Washington 25, D.C.
Attn: Mr. L. L. Henkel

Commanding Officer
Office of Naval Research
Branch Office
86 E. Randolph St.
Chicago 1, Illinois

Superintendent
U.S. Naval Postgraduate School
Attn: Library
Monterey, California

Electronic Computer Division
Code 280
Bureau of Ships
Dept. of the Navy
Washington 25, D.C.

The RAND Corporation
1700 Main St.
Santa Monica, California

Mr. Adam Abruzzi
Applied Statistics Group
Dept. of Management
Stevens Institute of Technology
Roboken, New Jersey

Dr. S. G. Allen
SRI
Menlo Park, California

Mr. Lee Bach
Carnegie Institute of Technology
Pittsburgh 13, Pa.

Professor E.F. Beckenbach
Dept. of Mathematics
University of California
Los Angeles 24, California

Dr. Martin J. Beckmann
Box 2125
Yale Station
New Haven, Conn.

Dr. Richard Bellman
The RAND Corporation
1700 Main St.
Santa Monica, California

Prof. Max R. Bloom
School of Bus. Admin.
University of California
Berkeley 4, California

Dean L.M.K. Boelter
School of Engineering
University of California
Los Angeles 24, California

Prof. James N. Boles
University of California
Agricultural Experiment Sta.
Berkeley 4, California

Prof. S.S. Cairns, Head
Dept. of Mathematics
University of Illinois
Urbana, Illinois

Prof. A. Charnes
R. R. No. 10
Lafayette, Indiana

Prof. John S. Chipman
Dept. of Economics
University of Minnesota
Minneapolis 14, Minn.
Mr. Frank Hahn  
Dept. of Economics  
University of Birmingham  
Birmingham, England  

Dr. H. Heller  
Navy Management Office  
Washington 25, D. C.  

Dr. Theodore E. Harris  
The RAND Corporation  
1700 Main Street  
Santa Monica, California  

Prof. C. Lowell Harriss  
Dept. of Economics  
Columbia University  
New York 27, N.Y.  

Prof. M. R. Hestenes  
Dept. of Mathematics  
University of California  
Los Angeles 24, Calif.  

Prof. C. Hildreth  
Michigan State University  
East Lansing, Michigan  

Mr. C. J. Hitch  
The RAND Corporation  
1700 Main Street  
Santa Monica, Calif.  

Mr. Alan J. Hoffman  
General Electric Co.  
Management Consultation Services  
570 Lexington Ave.  
New York 22, N. Y.  

Dr. C. C. Holt  
Grad. Sch. of Industrial Admin.  
Carnegie Inst. of Technology  
Pittsburgh 13, Pa.  

Mr. John W. Hooper  
Econometric Institute  
Netherlands School of Economics  
Rotterdam, Netherlands  

Prof. H. Hotelling  
Dept. of Mathematical Statistics  
University of North Carolina  
Chapel Hill, N. C.  

Dr. H. M. Hughes  
Dept. of Biometrics  
School of Aviation Medicine  
U.S.A.F.  
Randolph Field, Texas  

Prof. L. Hurwicz  
School of Business Admin.  
University of Minnesota  
Minneapolis 14, Minn.  

Prof. W. Grant Ireson  
Dept. of Industrial Engineering  
Stanford University  

Prof. J. R. Jackson  
Management Sciences Res. Proj.  
University of California  
Los Angeles 24, Calif.  

Dr. E. H. Jacobson  
Survey Research Center  
Institute for Social Research  
University of Michigan  
Ann Arbor, Michigan  

Mr. Dale W. Jorgenson  
12A Wendell St.  
Cambridge 38, Mass.  

Mr. Sidney Kaplan  
RCA Princeton Laboratories  
Lab. 3  
Princeton, N. J.  

Cdr. Walter H. Keen, USN  
Aircraft Design Division  
Bureau of Aeronautics  
Navy Dept.  
Washington 25, D. C.  

Prof. T. C. Koopmans  
Cowles Foundation  
for Research in Econ.  
Box 2125, Yale Station  
New Haven, Conn.  

Dr. Howard Laitin  
2134 Homcrest Ave.  
Brooklyn 29, N. Y.
<table>
<thead>
<tr>
<th>Name</th>
<th>Position</th>
<th>Institution</th>
<th>Address</th>
<th>City, State</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dr. R. F. Lanzellotti</td>
<td>Dept. of Economics</td>
<td>Washington State College</td>
<td>Pullman, Wash.</td>
<td></td>
</tr>
<tr>
<td>Prof. C. E. Lemki</td>
<td>Dept. of Mathematics</td>
<td>Rensselaer Polytechnical Inst.</td>
<td>Troy, New York</td>
<td></td>
</tr>
<tr>
<td>Mr. Bernhardt Lieberman</td>
<td>Clinical Psychology Section</td>
<td>Boston Veterans Adm. Hospital</td>
<td>South Huntington Ave.</td>
<td>Boston, Massachusetts</td>
</tr>
<tr>
<td>Prof. S. B. Littauer</td>
<td>Dept. of Industrial Engineering</td>
<td>Columbia University</td>
<td>New York 27, N. Y.</td>
<td></td>
</tr>
<tr>
<td>Dr. R. Duncan Luce</td>
<td>Dept. of Social Relations</td>
<td>Harvard University</td>
<td>Cambridge 38, Mass.</td>
<td></td>
</tr>
<tr>
<td>Dr. Craig A. Magwire</td>
<td>Dept. of Mathematics</td>
<td>U.S. Naval Postgraduate School</td>
<td>Monterey, California</td>
<td></td>
</tr>
<tr>
<td>Prof. Julius Margolis</td>
<td>Dept. of Business Admin.</td>
<td>University of California</td>
<td>Berkeley 4, California</td>
<td></td>
</tr>
<tr>
<td>Prof. Jacob Marschak</td>
<td></td>
<td>Box 2125, Yale Station</td>
<td>New Haven, Conn.</td>
<td></td>
</tr>
<tr>
<td>Prof. Lionel M. McKenzie</td>
<td>Dept. of Economics</td>
<td>University of Rochester</td>
<td>River Campus Station</td>
<td>Rochester 20, N.Y.</td>
</tr>
<tr>
<td>Prof. Maurice McManus</td>
<td>School of Business Admin.</td>
<td>University of Minnesota</td>
<td>Minneapolis 14, Minn.</td>
<td></td>
</tr>
<tr>
<td>Dr. Richard A. Miller</td>
<td></td>
<td></td>
<td>4071 West 7th St.</td>
<td>Fort Worth 7, Texas</td>
</tr>
<tr>
<td>Prof. Franco Modigliani</td>
<td>Dept. of Economics</td>
<td>Carnegie Inst. of Technology</td>
<td>Pittsburgh 13, Pa.</td>
<td></td>
</tr>
<tr>
<td>Prof. O. Morgenstern</td>
<td>Dept. of Economics and Social Institutions</td>
<td>Princeton, New Jersey</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mr. M. L. Norden</td>
<td>Research Division</td>
<td>College of Engineering</td>
<td>New York University</td>
<td>New York 53, N.Y.</td>
</tr>
<tr>
<td>Prof. R. R. O'Neil</td>
<td>Dept. of Engineering</td>
<td>University of California</td>
<td>Los Angeles 24, Calif.</td>
<td></td>
</tr>
<tr>
<td>Prof. Stanley Reiter</td>
<td>Dept. of Economics</td>
<td>Purdue University</td>
<td>Lafayette, Indiana</td>
<td></td>
</tr>
<tr>
<td>Prof. D. Rosenblatt</td>
<td>Dept. of Statistics</td>
<td>The George Washington University</td>
<td>Washington 7, D.C.</td>
<td></td>
</tr>
</tbody>
</table>
Prof. A. W. Tucker  
Fine Hall, Box 708  
Princeton, N. J.  

Prof. D. F. Votaw, Jr.  
Dept. of Economics  
Yale University  
New Haven, Conn.  

Prof. W. A. Wallis  
207 Haskell Hall  
University of Chicago  
Chicago 37, Illinois  

Prof. J. L. Walsh  
Dept. of Mathematics  
Harvard University  
Cambridge 38, Mass.  

Dr. T. Whitin  
Office of Operations Analysis  
U.S. Atomic Energy Commission  
Washington 25, D.C.  

Mr. Philip Wolfe  
The RAND Corporation  
1700 Main Street  
Santa Monica, California  

Prof. J. Wolfowitz  
Dept. of Mathematics  
Cornell University  
Ithaca, N. Y.  

Mr. Marshall K. Wood, Chief  
Industrial Vulnerability Branch  
Office of Ass't. Sec. of Defense  
Washington 25, D.C.  

Prof. M. A. Woodbury  
Dept. of Mathematics  
New York University  
New York 53, N. Y.  

Additional copies for project leader and assistants and reserve for future requirements 50