THE EARLY USE OF MATRIX DIAGONAL INCREMENTS IN STATISTICAL PROBLEMS

BU-833-M

Revised November, 1985

Walter W. Piegorsch* and George Casella†

ABSTRACT

The early motivation for and development of diagonal increments to ease matrix inversion in least squares (LS) problems is discussed. It is noted that this diagonal incrementation evolved from three major directions: modification of existing methodology in non-linear LS, utilization of additional information in linear regression, and the improvement of the numerical condition of a matrix. The interplay among these factors, and the advent of ridge regression are considered in an historical and comparative framework.

Key Words: Matrix Inversion, Matrix Ill-conditioning, Non-linear Least Squares

RUNNING TITLE: Matrix Diagonal Increments

* Biometry and Risk Assessment Program, National Institute of Environmental Health Sciences, Research Triangle Park, NC, 27709.
† Biometrics Unit, Cornell University, Ithaca, NY, 14853

1. INTRODUCTION

The problem of matrix inversion with minimum computation and high accuracy has a long and rich history. Hotelling (1943) gave an early overview of the methodology. He included such currently-burgeoning topics as eigenvalue use and the nature of error analysis (cf. Peters and Wilkinson, 1979). Two years later, Waugh and Dwyer (1945) published a similar summary, concentrating on the more compact and efficient methods. An extensive review by von Neumann and Goldstine (1947) discussed the steps involved and accuracies of the then-available methods, becoming a popular early reference.

A problem many early authors recognized was that, for an ill-conditioned matrix, inversion beomes particularly difficult. The resulting inverse may only be approximately equal to the true inverse, and when the inverse is being used to solve the system of equations

$$A\beta = Z \quad , \qquad (1.1)$$

the solution, $A^{-1}Z$, suffers from unnatural variability. This can bring about problems in interpretation and use of the results.

To invert an ill-conditioned matrix, some early authors attempted to slowly work the ill-conditioning out of the inversion process. For example, Guttman (1946), and, later, Herzberger (1949) based an early method on the construction of sucessively larger sub-matrices. Another approach, known as preconditioning, involved linearly transforming the system to improve its condition (see Jennings and Ajiz, 1984). By far the more popular attempt has been to correct A slightly in order to make it easier to invert using a standard method. This correction comes in the form of the addition of a small positive quantity to the diagonal elements of A; a <u>diagonal incrementation</u>, $A + kI$. Very early on it was recognized that $(A + kI)^{-1}$ would be very close to A^{-1}: Duncan (1944) and also Guttman (1946), gave the relationship

$$(A - UD^{-1}V)^{-1} = A^{-1} + A^{-1}U(D - VA^{-1}U)VA^{-1} \qquad (1.2)$$

When $D = I = V$ and $U = -kI$, (1.2) gives

$$(A + kI)^{-1} = A^{-1} - k^2A^{-1}(I + A^{-1}/k)^{-1}A \quad . \qquad (1.3)$$

For very small k the second term in (1.3) is negligible, suggesting that $(A + kI)^{-1}$ will closely approximate A^{-1}. [Henderson and Searle (1981) give an interesting account of the development of (1.2) and of some associated quantities.]

In a wide variety of statistical applications, such a matrix inversion has received increasing attention over the past two decades. DiPillo (1976) introduced diagonal increments to a classification procedure in discriminant analysis. Reduced variance and improved performance resulted. Bhapkar (1973) explored their use in developing an alternative to the usual comparison of proportions in matched samples. Khare and Federer (1981) substituted $(A + rI)^{-1}$ for A^{-1} to obtain inter-block solutions for the treatment effects in an incomplete block design. Their increment was a ratio of experimental to inter-block variance, $r=\sigma_\varepsilon^2/\sigma_B^2$.

The greatest statistical attention devoted to diagonal incrementation has involved parameter estimation in the linear model

$$E[Y_{n\times 1}] = X_{n\times p}\beta_{p\times 1} \qquad (1.4)$$

The least squares (LS) regression estimates of β,

$$b = (X'X)^{-1}X'Y \ , \qquad (1.5)$$

are critically based on $(X'X)^{-1}$. When X'X is ill-conditioned, as occurs frequently, for example, in polynomial regression (Bradley and Srivastava, 1979), the LS estimates become unstable. Small perturbations in the data, Y, can lead to large changes in the solution vector. In order to achieve some reduction in this variation, Hoerl (1962) suggested the use of X'X + kI in place of X'X.

This became known as ridge regression, and the procedure sparked a large literature (see Smith and Campbell, 1980).

Although statisticians have provided great motivation and use for diagonal incrementation, it is in the engineering sciences where the concept first arose. Prompted by problems in fitting non-linear equations to data, the method has been thriving there for some 40 years. We start, in Section 2, by considering this non-linear LS development. In Section 3 we follow this development into the linear model, and in Section 4 we continue through to overlaps with the Bayesian regression framework. Section 5 gives a short summary.

2. THE ORIGINS OF DIAGONAL INCREMENTATION

In December 1943, Kenneth Levenberg presented a paper at the American Mathematical Society's annual meetings in Chicago. Entitled "A Method for the Solution of Certain Non-linear Problems in Least Squares," the paper was published the following year (1944). It involved Levenberg's work at the War Department's Frankford Arsenal. There, he had noticed that the usual LS method for approximating a non-linear function, $E[Y]=F(X,\beta)$, did not always improve upon the initial estimates of the function's parameters. If the LS estimates strayed too far from their initial values, β^*, then the values of $\Delta\beta_j = b_j - \beta_j^*$ would be quite large. Denote the residuals by $f(X,b)$. Then, their first order Taylor approximation,

$$Y_i - F_i(X,b) = f_i(X,b) \approx f_i(X,\beta^*) + \sum_{j=1}^{p} \Delta\beta_j(\partial f_i/\partial\beta_j) \quad , \qquad (2.1)$$

would be greatly in error. This occurs because of the neglect of the higher order terms, $(\Delta\beta_j)^2$, $(\Delta\beta_j)^3$, etc., in the Taylor approximation. Levenberg's algorithm was designed to insure improvement of β^* by limiting, or "damping", the values of $|\Delta\beta_j|$, accomplishing this by minimizing a weighted sum of these differences:

$$w \sum_{i=1}^{n} \left[f_i(X,\beta^*) + \sum_{j=1}^{p} \Delta\beta_j(\partial f_i/\partial\beta_j) \right]^2 + \sum_{j=1}^{p} u_j(\Delta\beta_j)^2 \quad . \qquad (2.2)$$

The normal equations which resulted from this approach were the same as the ordinary ones except for the coefficients of the principal diagonal. These were incremented by quantities proportional to the weighting factors, u_j, on the parameter differences.

Levenberg went on to show that when the u_j were all equal, the directional derivative of the residual sum of squares (taken at w=0 along the new solution vector) would be a minimum. Without loss of generality, he took these values all equal to one. The diagonal increment then became a constant, equal to w^{-1}. Although designed for the solution of non-linear LS problems - literally a modification of the Taylor series method - this was the first presented use of diagonal incrementation.

As digital computer technology progressed in the 1950's, so did the ability of these computers to handle more and more complicated algorithms. This helped in the dissemination of Levenberg's procedure, which was rather tedious to work out by hand. The value of w was, of course, critical to the entire estimation process, and it became known as the Levenberg parameter (Wilde and

Beightler, 1967). However, the algorithm did not gain widespread attention very quickly. It was the independent development of a procedure very much like Levenberg's that led to its current-day notoriety (e.g. over two dozen citations in 1981).

While developing computer algorithms and associated procedures at duPont's Engineering Labs during the 1950's and 1960's, Donald Marquardt made discoveries very similar to Levenberg's. Just as Levenberg had noticed problems with the Taylor series approach, Marquardt recognized a disparity in the other computerized non-linear LS approach, known as the steepest descent, or gradient, approach. There, proper convergence from the initial values was not always assured, and the procedure sometimes lead to nonsensical results. Marquardt explains,

> "At first by plotting and later by algebraic calculation, I had observed that the gradient and the Taylor-series methods invariably give correction vectors whose included angle . . . is nearly a right angle. Recognition of the orientation of these vectors in the sum-of-squares contours explained for the first time the apparently anomalous behaviors of the [two] methods" (1979).

These observations led Marquardt to reconcile these two earlier approaches. This was done in an algorithm which displayed some of the better properties of both predecessors, while avoiding some of their limitations. The work was published in 1963. Critical to it was the development of a Lagrange parameter, λ, which varied monotonically over $(0,\infty)$ (the Taylor series and gradient method correspond to the two extremes for λ). The parameter was used to control the iterative solution of the non-linear normal equations. At each iteration, equations of the form

$$(X'X + \lambda I)\beta = X'Y \qquad (2.3)$$

were solved so that the iterative residual sum of squares was always

decreasing. The resulting algorithm had the ability to converge quickly from a wide range of initial estimates (Marquardt, 1963). It became an important tool in the estimation of non-linear parameters, with, e.g., almost 1000 citations between 1963 and 1977 (Marquardt, 1979).

As can be seen, the motivation, development, structure, and optimality of Levenberg's and Marquardt's algorithms are almost identical. Indeed, both authors are now referred to as its progenitors (Kennedy and Gentle, 1980), and $\lambda=w^{-1}$ is now called the Levenberg-Marquardt parameter (Morê, 1977). However, Marquardt was not aware of Levenberg's work throughout much of his research period. He was only informed of it, by H. O. Hartley, just before the 1963 paper went to press. The best Marquardt could do was to comment on Levenberg's paper in his Acknowledgments (the 1944 paper is referenced last, out of alphabetical order) and thank Hartley for bringing it to his attention.

3. LINEAR MODELS AND RIDGE REGRESSION

One very interesting comment Marquardt makes is to highlight the "corollary numerical benefit" of adding λ to the diagonal of X'X, in that it helps improve the condition of the matrix (1963, p.439). This was precisely Arthur Hoerl's observation of the previous year; he reported that incrementing the diagonal of X'X by some small positive quantity was a helpful way to correct for any ill-conditioning. Later works (Hoerl and Kennard, 1970; Marquardt, 1970) developed this into a formal approach to estimating β, and the problem of ill-conditioned regression has since generated a great deal of interesting work (Bradley and Srivastava, 1979; Hocking, 1983; Wold, et al., 1984).

Much of the early justification for this ridge regression procedure was more heuristic than theoretic. Indeed, finding a theoretically optimal basis

for the ridge procedure has been a lengthy process (cf. Rolph, 1976; Strawderman, 1978; Casella, 1980), and is still not fully developed. Still, the observation that X'X+kI can be numerically easier to invert than X'X is very true, and the correspondence and timing of Hoerl's (1962) and Marquardt's (1963) observations was no coincidence; both men were involved in statistical research with the duPont group. Throughout the the 1960's and 1970's, Hoerl, Kennard, and Marquardt worked at improving and developing their results on diagonal incrementation. The Wilmington, Deleware area was where much of the early research on ridge regression was performed, and the works of all three men were critically intertwined.

An important, positive aspect of the ridge method was the improvement in conditioning the diagonal incrementation provided. It is a strange anomaly then, that in the history of ridge's development, neither Hoerl nor Marquardt was the first to note it. The first indication of the usefulness of digonal incrementation - indeed, the first use of the matrix notation A + kI - came from James Riley (1955). Riley's approach was of a slightly different nature than that of Marquardt's and Levenberg's, but did bear some resemblance to that of Hoerl and Kennard. Instead of starting with a non-linear problem and developing the diagonal increment k, Riley simply proposed the use of the increment and then examined its usefulness; again, a more heuristic approach.

To solve $A\beta=Z$, Riley set $C=A+kI$ so that $A=C-kI$ and thus

$$A^{-1} = \sum_{m=0}^{\infty} k^m C^{-m-1} \quad . \tag{3.1}$$

Then, a solution is

$$\beta = A^{-1}Z = \sum_{m=0}^{\infty} (kC^{-1})^m C^{-1}Z \quad . \tag{3.2}$$

For k>0, when all of the eigenvalues of A are positive (e.g., if A is symmetric and positive definite), the eigenvalues of kC^{-1} are all contained in (0,1). Hence (3.2) converges (Riley, 1955, p.98). For very small values of k (Riley suggested 10^{2-M} where M is the number of decimal places being carried), terms involving k^m in (3.2) are negligible for m>1, and the resulting approximation is the ridge estimator; i.e. for A=X'X and Z=X'Y, we write

$$\sum_{m=0}^{\infty} k^m(X'X + kI)^{-m}(X'X + kI)^{-1}X'Y \approx (X'X + kI)^{-1}X'Y \quad . \qquad (3.3)$$

Riley then used a number of different measures of matrix condition to show that C is better conditioned than A, suggesting a sort of numerical improvement. In particular, he considered the ratio of largest to smallest (absolute) eigenvalues for the matrix. This is one form of the well-known condition number (cf. Marshall and Olkin, 1965, 1969; Longley, 1981; Casella, 1985). As Riley shows, the condition number of A + kI is always smaller than that of A (for k>0).

Except perhaps for the use of the term, Riley's work could be considered as an early example of ridge methodology. Unfortunately, Marquardt did not know of Riley's paper, while Hoerl and Kennard (1970) gave it only passing reference [when they utilized some of Riley's matrix manipulations to help show that their (non-stochastic) ridge estimator dominated the risk of the LS estimator over a portion of the parameter space]. However, even Riley fell victim to a similar oversight. His only reference to Levenberg's (1944) work occured late in his 1953 paper, in the last appendix. It too was done in passing.

4. BAYES APPROACH TO THE REGRESSION PROBLEM

Before Riley's (1955) paper appeared, James Durbin (1953) published a work, entitled "A Note on Regression when there is Extraneous Information About

one of the Coefficients." In it, he considered model (1.4) when there was some outside, unbiased estimator of the first regression coefficient, β_1. At issue was how to best use the information about β_1 in estimating β. By applying Aitken's (1935) extension of Gauss's LS theorem on best linear unbiased estimators, Durbin showed that the normal equations were only slightly modified by this additional information. The difference from the ordinary LS expressions was the addition of the ratio σ^2/σ_1^2 to the leading diagonal term in $\mathbf{X'X}$, where $\text{var}(Y_i) = \sigma^2$ and $\text{var}(b_1) = \sigma_1^2$. Later authors (Theil, 1963; Lee, et al., 1968; Havenner and Craine, 1981) successfully applied this approach to a number of statistical and mathematical problems. Of particular interest was a paper by Chipman (1964), which considered topics ranging from multicollinearity to problems of estimability in LS regression. It was well-written, paying close attention to the various historical perspectives, as well as to analytical and technical rigor. Durbin (1953) also considered estimation of the ratio σ^2/σ_1^2 when both variances were unknown and when there is outside information on β. The results were similar.

These results are also similar to the estimators produced when operating under a Bayesian framework. Hoerl and Kennard (1970) noted that, under the prior distributional assumption

$$\beta \sim N_p(\mathbf{0}, \sigma_\beta^2 \mathbf{I}) \tag{4.1}$$

the Bayes estimator, when $\mathbf{Y} \sim N_n(\mathbf{X}\beta, \sigma^2\mathbf{I})$, is

$$\mathbf{B} = (\mathbf{X'X} + k\mathbf{I})^{-1}\mathbf{X'Y} \tag{4.2}$$

where $k=\sigma^2/\sigma_\beta^2$ (notice the similarity to the Khare and Federer [1981] ratio mentioned in Section 1). The authors are quick to point out the link to ridge regression by noting the similarity of (4.2) to (3.3). This is an interesting property of the ridge estimator, showing that it is mathematically equivalent

to this Bayes estimator (of course, their motivations are substantively different; each solves a different statistical problem). Lindley and Smith (1972) go into greater detail on this Bayes regression problem, and the result can be traced back in the literature at least as far as Raiffa and Schlaifer (1961, Ch.13), although it was probably known long before this.

Unfortunately, the similarity of (4.2) to Durbin's eariler work has not been extensively discussed in the literature. Even in an excellent and extensive review by Draper and van Nostrand (1979), Durbin's work is left unmentioned (although both the Levenberg [1944] and Riley [1955] papers are properly described). It seems that Durbin's contribution to the diagonal increment problem was fated for early anonymity.

5. SUMMARY

There are three basic problems that led to the use of matrix diagonal increments. First, the improvement of a non-linear LS solution when the usual methods fail to provide acceptable estimates. This was first investigated by Levenberg (1944) and later by Marquardt (1963). Next, the utilization of additional information about a regression parameter by Dubin (1953), which was later developed into the Bayes approach by Lindley and Smith (1972), and many others (see Rolph, 1976). And third, the need to improve the condition of a matrix in order to solve a system of simultaneous equations with less difficulty and greater precision (Riley, 1955). On the surface, the ridge regression procedure would seem to derive much of its motivation from the last of these three research problems. Far deeper, however, one can find an interesting interplay among all three.

ACKNOWLEDGMENTS

The authors thank Drs. Donald Marquardt and William Provine for helpful discussions, and the Editor and referees for their suggestions, in the preparation of this work. This is paper no. BU-833-M in the Biometrics Series, Cornell University, Ithaca, New York.

REFERENCES

Aitken, A.C. (1935). On least squares and linear combination of observations. Proc. Roy. Soc. Edinburgh, **55**, 42-48.

Bhapkar, V.P. (1973). On the comparison of proportions in matched Samples. Sankhya, Ser. A, **35**, 341-356.

Bradley, R.A. and Srivastava, S.S. (1979). Correlation in polynomial regression. Amer. Statist., **35**, 11-14.

Casella, G. (1980). Minimax ridge regression estimation. Ann. Statist., **8**, 1036-1056.

Casella, G. (1985). Condition numbers and minimax ridge-regression estimators. J. Amer. Statist. Assoc., **80**, 753-758.

Chipman, J.S. (1964). On least squares wth insufficient observations. J. Amer. Statist. Assoc., **59**, 1078-1111.

DiPillo, P.J. (1976). "The application of bias to discriminant analysis. Commun. Statist. - Theor. Meth., **A5**, 843-854.

Draper, N.R. and van Nostrand, R.C. (1979). Ridge regression and James-Stein estimation: Review and Comments. Technometrics, **21**, 451-466.

Duncan, W.J. (1944). Some devices for the solution of large sets of simultaneous linear equations (with an appendix on the reciprocation of partitioned matrices). London, Edinburgh, Dublin Phil. Mag. J. Sci., Seventh Ser., **35**, 660-670.

Durbin, J. (1953). A note on regression when there is extraneous information about one of the coefficients. J. Amer. Statist. Assoc., **48**, 799-808.

Guttman, L. (1946). Enlargement methods for computing the inverse matrix. Ann. Math. Statist., **17**, 336-343.

Havenner, A. and Craine, R. (1981). Estimation analogies in control. J. Amer. Statist. Assoc., **76**, 850-859.

Henderson, H.V. and Searle, S.R. (1981). On deriving the inverse of a sum of matrices. SIAM Rev., **23**, 53-60.

Herzberger, M. (1949). The normal equations of the method of least squares and their solution. Q. Appl. Math., **7**, 217-223.

Hoerl, A.E. (1962). Application of ridge analysis to regression problems. Chem. Eng. Progress, **60**, 54-59.

Hoerl, A.E. and Kennard, R.W. (1970). Ridge regression: Biased estimation for non-orthogonal problems. Technometrics, **12**, 55-67.

Hocking, R.R. (1983). Developments in linear regression methodology: 1959-1982. Technometrics, **25**, 219-230.

Hoteling, H. (1943). Some new methods in matrix calculation. Ann. Math. Statist., **14**, 1-43.

Jennings, A. and Ajiz, M.A. (1984). Incomplete methods for solving $A^TAx=b$. SIAM J. Sci. Statist. Comp., **5**, 978-987.

Kennedy, W.J. Jr. and Gentle, J.E. (1980). Statistical Computing. New York: Marcel Dekker.

Khare, M. and Federer, W.T. (1981). A simple construction procedure for resolvable incomplete block designs for any number of treatments. Biom. J., **23**, 121-132.

Lee, T.C., Judge, G.G., and Zellner, A. (1968). Maximum likelihood and Bayesian estimation of transition probabilities. J. Amer. Statist. Assoc., **63**, 1162-1179.

Levenberg, K. (1944). A method for the solution of certain non-linear problems in least squares. Q. Appl. Math., **2**, 164-168.

Lindley, C.V. and Smith, A.F.M. (1972). Bayes estimates for the linear model. J. Roy. Statist. Soc., Ser. B, **34**, 1-41 (with disc.).

Longley, J.W. (1981). Least squares computations and the condition of the matrix. Comm. Statist. - Simula. Computa., **B10**, 593-615.

Marquardt, D.W. (1963). An algorithm for least-squares estimation of non-linear parameters. J. Soc. Indust. Appl. Math., **11**, 431-441.

Marshall, A.W. and Olkin, I. (1965). Norms and inequalities for condition numbers. Pac. J. Math., **15**, 241-247.

Marshall, A.W. and Olkin, I. (1969). Norms and inequalities for condition numbers, II. Lin. Alg. Appl., **2**, 167-172.

Marquardt, D.W. (1970). Generalized inverses, ridge regression, biased linear estimation, and non-linear estimation. Technometrics, **12**, 591-612.

Marquardt, D.W. (1979). Week's citation classic. Curr. Contents Eng. Technol. Appl. Sci., July 12, 1979, 14.

Moré, J.J. (1977). The Levenberg-Marquardt algorithm: Implementation and theory, in Numerical Analysis, ed. G.A. Watson, Springer Ser. Math. **630**, New York:Springer-Verlag, 105-116.

Peters, G. and Wilkinson, J.H. (1979). Inverse iteration, ill-conditioned equations, and Newton's method. SIAM Rev., **21**, 339-360.

Raiffa, H. and Schlaifer, R. (1961). Applied Statistical Decision Theory. Boston: Harvard University.

Riley, J.D. (1955). Solving systems of linear equations with a positive definite, symmetric, but possibly ill-conditioned matrix. Math. Tables and Other Aids Comp., **9**, 96-101.

Rolph, J.E. (1976). Choosing shrinkage estimators for regression problems. Comm. Statist. - Theor. Meth., **A5**, 789-802.

Smith, G. and Campbell F. (1980). A critique of some ridge regression estimators. J. Amer. Statist. Assoc., **75**, 74-81 (with disc.).

Strawderman, W.E. (1978). Minimax adaptive generalized ridge regression analysis. J. Amer. Statist. Assoc., **73**, 623-627.

Theil, H. (1963). On the use of incomplete prior information in regression analysis. J. Amer. Statist. Assoc., **58**, 401-414.

von Neumann, J. and Goldstine, H.H. (1947). Numerical inverting of matrices of higher order. Bull. Amer. Math. Soc., **53**, 1021-1099.

Waugh, F.V. and Dwyer, P.S. (1945). Compact computation of the inverse of a matrix. Ann. Math. Statist., **16**, 259-271.

Wilde, D.J. and Beightler, C.S. (1967). Foundations of Optimization. Englewood Cliffs, NJ: Prentice Hall.

Wold, S., Ruhe, A., Wold, H., and Dunn, W.J. III (1984). The collinearity problem in linear regression. The partial least squares approach to generalized inverses. SIAM J. Sci. Statist. Comp., **5**, 735-743.