MEAN SQUARE ERROR BEHAVIOR FOR PREDICTION
IN LINEAR REGRESSION MODELS

BY

ALAN E. GELFAND

TECHNICAL REPORT NO. 427
MARCH 7, 1990

PREPARED UNDER CONTRACT
N00014-89-J-1627 (NR-042-267)
FOR THE OFFICE OF NAVAL RESEARCH

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

MEAN SQUARE ERROR BEHAVIOR FOR PREDICTION
IN LINEAR REGRESSION MODELS

BY

ALAN E. GELFAND

TECHNICAL REPORT NO. 427
MARCH 7, 1990

Prepared Under Contract
N00014-89-J-1627 (NR-042-267)
For the Office of Naval Research

Herbert Solomon, Project Director

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

MEAN SQUARE ERROR BEHAVIOR FOR PREDICTION IN LINEAR REGRESSION MODELS

Alan E. Gelfand

ABSTRACT

For the problem of individual prediction in linear regression models, that is, estimation of a linear combination of regression coefficients, mean square error behavior of a general class of adaptive predictors is examined.

1. INTRODUCTION

Suppose the usual linear regression model with fixed regressors, $Y = X\beta + \varepsilon$, $Y_{n\times 1}$, $X_{n\times p}$ full rank, $\beta_{p\times 1}$ and $\varepsilon_{n\times 1} \sim (0,\sigma^2 I)$. Let $\hat{\beta}_{LS} = (X^TX)^{-1}X^TY$ denote the ordinary least squares estimator of β. At a new vector of predictor values, X_0, we seek to estimate $X_0^T\beta$. Using mean square error as a criterion, results of Cohen (1965) show that if ε is normally distributed, $\alpha X_0^T\hat{\beta}_{LS}$ is an admissible estimator of $X_0^T\beta$ for $0 \le \alpha \le 1$, e.g., the UMVU predictor is admissible. In fact, a predictor of the form ℓ^TY is admissible for $X_0^T\beta$ iff $(2\ell - X(X^TX)^{-1}X_0)^T(2\ell - X(X^TX)^{-1}X_0 \le X_0^T(X^TX)^{-1}X_0$.

In the sequel, we study the MSE under normality of predictors of the form $X_0^T\hat{\beta}_C$ where

$$\hat{\beta}_C = C\hat{\beta}_{LS} + (I - C)\beta^* \qquad (1)$$

C a matrix usually data dependent and β^* a specified vector. Such $\hat{\beta}_C$ include most alternatives to $\hat{\beta}_{LS}$ discussed in the literature. Earlier work in this direction appears in Baranchik (1964) and Radhakrishnan (1970).

2. NOTATION AND MOTIVATION

To simplify matters, we convert to canonical form. Let $\hat{\alpha} = P\hat{\beta}_{LS}$, P orthogonal such that $P(X^TX)^{-1}P^T = D^{-1}$, D diagonal with diagonal elements d_i. Define $\alpha = P\beta$, $\ell = P\underset{\sim}{X}_0$ and for convenience set $\beta^* = 0$. For the moment assume σ^2 known. Our problem now is to estimate $\theta = \ell^T\alpha$ given $\hat{\alpha} \sim N(\alpha, \sigma^2 D^{-1})$ wishing to do well near $\theta = 0$. Let $U = \ell^T\hat{\alpha}$, $Z = \hat{\alpha}^T D\hat{\alpha}$, $q = \ell^T D^{-1}\ell$, $V = Z - U^2/q$, $\lambda = \alpha^T D\alpha$ and $\phi = \lambda - \theta^2/q$. Then, U, V are independent, $U \sim N(\theta, \sigma^2 q)$, $V \sim \sigma^2\chi^2_{p-1}(\phi/\sigma^2)$.

Consider a general adaptive predictor $\delta(\hat{\alpha})$ of the form

$$\delta(\hat{\alpha}) = \Sigma h_i(\hat{\alpha})\ell_i\hat{\alpha}_i. \tag{2}$$

Most predictors of θ discussed in the literature are special cases of (2). Apart from the LS predictor, U, we have:

i) A class of predictors given in Thompson (1968)

$$T_m = \frac{U^2}{U^2+m\sigma^2 q}U,\ \text{m a known constant, i.e., } h_i(\hat{\alpha}) = \frac{(\hat{\alpha}^T\ell)^2}{(\hat{\alpha}^T\ell)^2+m\sigma^2 q}.$$

ii) A class of predictors given in Mehta and Srivastava (1971)

$$MS_{b_1,b_2} = (1-b_1e^{-b_2U^2/\sigma^2 q})U,\ 0 < b_1 < 1,\ b_2 > 0,\ b_1, b_2 \text{ known,}$$

$$\text{i.e., } h_i(\hat{\alpha}) = 1 - b_1\exp(-b_2(\hat{\alpha}^T\ell)^2/\sigma^2 q).$$

iii) A predictor arising from the James-Stein estimator adapted for unequal variances (Sclove 1968)

$$JS_c = (1 - \frac{c\sigma^2}{Z})U,\ \text{c known usually taken equal to p - 2.}$$

A positive part adjustment should be applied so that $h_i(\hat{\alpha}) = [1 - c\sigma^2(\hat{\alpha}^T D\hat{\alpha})^{-1}]^+$.

iv) Predictors arising from (simple) ridge estimators

$$R_{k_t} = \Sigma \ell_i \frac{d_i}{d_i + k_t} \hat{\alpha}_i$$

where k_t is based on the data, i.e., $h_i(\hat{\alpha}) = d_i/(d_i + k_t(\hat{\alpha}))$. k's discussed include:

$k_1(\hat{\alpha}) = \sigma^2 p(\hat{\alpha}^T\hat{\alpha})^{-1}$ (Hoerl, Kennard, and Baldwin 1975),

$k_2(\hat{\alpha}) = \sigma^2 p Z^{-1}$ (Lawless and Wang 1976),

$k_3(\hat{\alpha})$, the solution to $\Sigma\hat{\alpha}_i^2 d_i^2 (d_i + k_3)^{-2} = \Sigma\hat{\alpha}_i^2 - \sigma^2 \Sigma d_i^{-1}$

(McDonald and Galarneau 1975),

$k_4(\hat{\alpha})$, the solution to $\Sigma\hat{\alpha}_i^2 d_i (d_i + k_4)^{-1} = \sigma^2 p$

(the RIDGM estimator of Dempster, Schatzoff and Wermuth 1977).

A subclass of (2) which includes (i), (ii), (iii), and R_{k_2} has the form

$$\delta(\hat{\alpha}) = \Sigma h_i(U,Z)\ell_i\hat{\alpha}_i \ . \tag{3}$$

A further subclass which still includes (i), (ii), and (iii) is

$$\delta(\hat{\alpha}) = h(U,Z) \cdot U. \tag{4}$$

When D = I, all of the aforementioned estimators belong to (4).

Taking another point of view (see e.g. Thompson (1968)), if h_i in (3) is constant, the optimal h_i to minimize the MSE are easily obtained:

$$h_i^* = \frac{\theta}{\sigma^2 + \lambda} \frac{\alpha_i}{\ell_i} \ . \tag{5}$$

An estimator of h_i^* would be of the form $c_i(\hat{\alpha},\hat{\sigma^2})$ leading to a predictor belonging to (2). If (5) was estimated by $c(U,Z,\sigma^2) \cdot \hat{\alpha}_i/\ell_i$ the class (4) results.

Suppose we take a Bayesian approach using a prior which centers θ at 0, where we want to do well. More precisely, let Q be an orthogonal matrix such that $QD^{\frac{1}{2}}\alpha = \binom{\theta/\sqrt{q}}{\eta}$ where η is $(p - 1) \times 1$ and $\eta^T\eta = \phi$. If we take as our prior

$$\binom{\theta/\sqrt{q}}{\eta} \sim N(0,\begin{pmatrix}\gamma & 0\\ 0 & \rho\gamma I_{p-1}\end{pmatrix}),\ \rho \text{ known},$$

then under squared error loss, the Bayes estimate of θ is $(\gamma + \sigma^2)^{-1} \cdot \gamma U$. Since (U,Z) is sufficient under the marginal distribution of $\omega = QD^{\frac{1}{2}}\hat{\alpha}$ an "empirical Bayes" estimator of θ takes the form in (4).

3. EXAMINATION OF THE MSE

We can calculate the MSE for the general predictor in (2) in terms of the h_i, assuming σ^2 known.[1]

<u>Theorem 1</u>. If $E|\frac{\partial h_i}{\partial U} \cdot \hat{\alpha}_i| < \infty$, $i = 1,2,\ldots,p$,

$$\begin{aligned}\text{MSE}(\delta) = \sigma^2 q + E(\delta - U)^2 - 2\sigma^2 E\Sigma \ell_i^2(1 - h_i)\\ + 2\sigma^2 qE\Sigma \ell_i \hat{\alpha}_i \frac{\partial h_i}{\partial U}\ .\end{aligned} \tag{6}$$

<u>Proof</u>. By direct calculation

$$\text{MSE}(\delta) = \sigma^2 q + E(\delta - U)^2 - 2E\{r(\hat{\alpha})(U - \theta)\} \tag{7}$$

where $r(\hat{\alpha}) = \Sigma(1 - h_i)\ell_i\hat{\alpha}_i$. Stein's identity (Stein 1981, p. 1148) converts the right-most term of (7) to $\sigma^2 qE(\frac{\partial r(\hat{\alpha})}{\partial U})$. Simplification yields (6).

$\frac{\partial h_i}{\partial U}$ *would be calculated using the transformation* $\hat{\alpha} = D^{-\frac{1}{2}}Q^T\omega$ of the previous section. In the case of (3), it can be calculated directly writing h_i as a function of U and V. For predictors of the form (4), $\text{MSE}(\delta)$ depends only on θ and ϕ and is given as Corollary 1.

<u>Corollary 1</u>. For the predictors in (4), if $E|U \frac{\partial h}{\partial U}| < \infty$

$$\text{MSE}(\delta) = \sigma^2 q + E(1 - h)^2 U^2 + 2\sigma^2 qEU \frac{\partial h}{\partial U} - 2\sigma^2 qE(1 - h). \tag{8}$$

Under (4) choices of h in the literature are such that h is symmetric in U about 0 and restricted to $[0,1]$. Using essentially the argument of Efron and Morris (1976, p. 14) positive part restriction of h uniformly reduces risk. Restriction of $h \leq 1$ is less clear. Taking $h > 0$ the predictor $h^* \cdot U$ where $h^* = \min(h,1)$ does not necessarily dominate $h \cdot U$. For example, let $h(U,V) = \begin{cases} 1 + c, & a^2 < U^2 < b^2 \\ 1 & \text{, elsewhere} \end{cases}$. Then at each ϕ, for $|\theta|$ sufficiently large, MSE of $h(U,V)U$ is less than MSE of $h^*(U,V)$. Nonetheless, to improve in a neighborhood of a specified θ_0 requires convex combinations of U and θ_0. Theorem 2 details MSE properties of predictors in (4) relative to the MSE of U.

Theorem 2. For $\delta(\hat{\alpha})$ in (4) with $h \in [0,1]$, let h be symmetric in U about 0. Let $g = (1 - h)U$ with $\lim_{|U|\to\infty} \sup_V g = 0$ and assume $\frac{\partial g}{\partial U}$ exists for all U. Finally, assume that the Lebesgue measure of $A = \{(U,V) : h(U,V) < 1\}$ is greater than 0. Then,

(i) For each ϕ there is a neighborhood N_ϕ of $\theta = 0$ where $MSE(\delta;\theta,\phi) < \sigma^2 q$.

(ii) $MSE(\delta;\theta,\phi)$ is bounded and $\lim_{|\theta|\to\infty} MSE(\delta;\theta,\phi) = \sigma^2 q$.

(iii) $MSE(\delta;\theta,\phi)$ is symmetric in θ about 0 and $\left.\frac{\partial MSE(\delta;\theta,\phi)}{\partial\theta}\right|_{\theta=0} = 0$.

(iv) $g^2 - 2\frac{\partial g}{\partial U}$ changes sign at least once in $0 < U < \infty$. If $g^2 - 2\frac{\partial g}{\partial U}$ changes sign b times in $0 < U < \infty$, then for fixed ϕ, $MSE(\delta;\theta,\phi) - \sigma^2 q$ changes sign at most 2b times.

Proof. The proof of (i) is clear since $MSE(\delta;0,\phi) < \sigma^2 q$. For (ii),

$$MSE(\delta;\theta,\phi) = \sigma^2 q + Eg^2 - 2E(U - \theta)g. \tag{9}$$

Given ε, $\exists\, u_0$ such that for all V, $U > u_0 \Rightarrow |g| < \varepsilon$ and $\exists\, \theta_0 > 0$ such that $|\theta| > \theta_0 \Rightarrow P(|U| > u_0) > 1 - \varepsilon$. Then the second term and the third term (using the Cauchy-Schwarz Inequality) in (9) can be made arbitrarily small as $|\theta| \to \infty$. It is clear that the r.h.s. of (9) is bounded. (iii) is obvious. The first part of (iv) follows since U is admissible. The second part follows from the sign change theorem of Karlin (1957) by noting that $MSE(\delta;\theta,\phi) - \sigma^2 q = E(g^2 - 2\frac{\partial g}{\partial U})$.

Remark 1. Predictors in (i), (ii), (iii) of Section 2 satisfy the conditions of Theorem 2.

Remark 2. Result (ii) is a simple case of the "tail minimaxity" notion of Berger (1976).

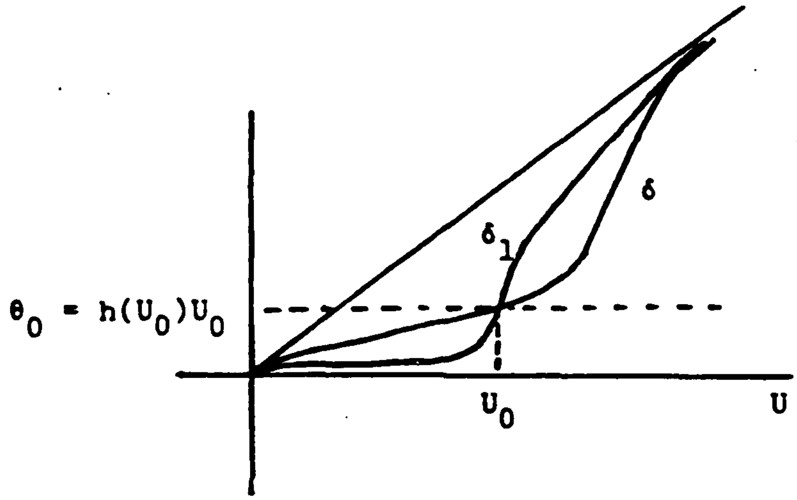
Remark 3. In (iii), $\inf_{\theta} MSE(\delta;\theta,\phi)$ need not occur at $\theta = 0$. If, however, $h(U,V)$ is increasing in $|U|$ it must as may be shown by establishing the result for h, a step function in U. An induction argument proves this.

Remark 4. If $b = 1$ in (iv), then a graph of $MSE(\delta;\theta,\phi)$ for $\theta \geq 0$ must start below $\sigma^2 q$ at $\theta = 0$, cross above $\sigma^2 q$ at some θ and then asymptotically return to $\sigma^2 q$ from above. Any predictor satisfying the conditions of Theorem 2 must necessarily perform worse for a set of θ's near 0 than for a set arbitrarily far away.

Remark 5. No immediate extension of Theorem 2 to $\delta(\hat{\alpha})$ as in (3) is available. For an arbitrary member of (3), MSE depends upon θ and η and, even if each h_i meets the "tail minimaxity" condition, need not approach $\sigma^2 q$ as $|\theta| \to \infty$ for fixed η.

Remark 6. Theorem 2 is readily extended to the comparison of any pair of predictors in (4).

We conclude with a comment on admissibility for the above predictors. Within the class of predictors based solely on U, i.e., $h(U)U$, those meeting the conditions of Theorem 2 will either be admissible or if not then improvement cannot be substantial. We employ ideas of Chow and Hwang (1984). Suppose $\delta_1(U)$ is to dominate $\delta_0 = h(U)U$ meeting the conditions of Theorem 2. We can write δ_1 as $h^*(U)U$, and assume $h^* \geq 0$. For δ_1 to dominate δ_0 requires, when $|U|$ is large, that generally h^* be closer to 1 than h and that, when $|U|$ is small, generally h^* be closer to 0 than h. A simplified picture of δ_0, δ_1 for $U > 0$ might look like



But, at $\theta = \theta_0$, it would be almost impossible for δ_1 to dominate. Thus, the simplest h* which realistically could dominate would have to have at least 3 sign changes for h - h* on U > 0. For such an h*, its form would be complicated, domination would be difficult to show, and improvement would be minimal.

This argument does not extend to the more general class (4). Though U and V are independent, conditioning on V in the above heuristic leads to θ_0 depending upon V. We, nonetheless, conjecture "approximate admissibility" for members of (4) meeting the conditions of Theorem 2.

FOOTNOTE

[1]When σ^2 is unknown, we customarily assume an estimator S^2 of σ^2 such that $\nu S^2 \sim \sigma^2\chi^2_\nu$ independent of $\hat{\alpha}$. In the foregoing predictors, σ^2 is replaced by cS^2. As Lawless (1981, pp. 463-464) notes, when $\nu \to \infty$ and even when ν is moderate, resulting MSE will differ little from that with σ^2 known.

BIBLIOGRAPHY

Baranchik, Alvin J. (1964). "Multiple Regression and Estimation of the Mean of a Multivariate Normal Distribution," Stanford University Tech. Report #51.

Berger, J.O. (1976). "Tail minimaxity in location vector problems and its applications," Annals of Statistics, 4, 33-50.

Chow, M.S. and Hwang, J.T. (1984). "The comparison of estimators for the noncentrality of a chi-square distribution," Cornell Statistical Center, Technical Report.

Cohen, A. (1965). "Estimates of linear combinations of the parameters in the mean vector of a multivariate distribution," Annals of Mathematical Statistics, 36, 78-87.

Dempster, A.P., Scharzoff, M. and Wermuth, N. (1977). "A simulation study of alternatives to ordinary least squares," Journal of the American Statistical Association, 72, 77-106.

Efron, B. and Morris, C. (1976). "Families of minimax estimators of the mean of a multivariate normal distribution," Annals of Statistics, 4, No. 1, 11-21.

Hoerl, A.E., Kennard, R.W. and Baldwin, K.F. (1975). "Ridge regression: Some simulations," Communications in Statistics, 4, 105-123.

Karlin, S. (1957). "Polya type distributions II," Annals of Mathematical Statistics, 28, No. 2, 281-308.

Lawless, J.F. (1981). "Mean squared error properties of generalized ridge estimators," Journal of the American Statistical Association, 76, 462-466.

Lawless, J.F. and Wang, P. (1976). "A simulation study of ridge and other regression estimators," Communications in Statistics, A, 5, 307-323.

McDonald, G.C. and Galarneau, D.L. (1975). "A Monte Carlo evaluation of some ridge type estimators," Journal of the American Statistical Association, 70, 407-416.

Mehta, J.S. and Srivastava, R. (1971). "Estimation of the mean by shrinking to a point," Journal of the American Statistical Association, 66, 86-91.

Radhakrishnan, R. (1970). "Some Estimation Problems in Multivariate Analysis," unpublished Ph.D. Thesis, Carnegie Mellon University.

Sclove, S.L. (1968). "Improved estimators for coefficients in linear regression," Journal of the American Statistical Association, 63, 596-606.

Stein, C. (1981). "Estimation of the mean of a multivariate normal distribution," Annals of Statistics, 9, No. 6, 1135-1151.

Thompson, J.R. (1968). "Some shrinkage techniques for estimating the mean," Journal of the American Statistical Association, 63, 113-122.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: 427	2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE (and Subtitle): Mean Square Error Behavior For Prediction In Linear Regression Models	5. TYPE OF REPORT & PERIOD COVERED: TECHNICAL REPORT
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): Alan E. Gelfand	8. CONTRACT OR GRANT NUMBER(s): N00014-89-J-1627
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Department of Statistics, Stanford University, Stanford, CA 94305	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: NR-042-267
11. CONTROLLING OFFICE NAME AND ADDRESS: Office of Naval Research, Statistics & Probability Program Code 1111	12. REPORT DATE: March 7, 1990
	13. NUMBER OF PAGES: 11
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)	15. SECURITY CLASS. (of this report): UNCLASSIFIED
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

prediction; adaptive estimators; linear regression; mean square error.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

For the problem of individual prediction in linear regression models, that is, estimation of a linear combination of regression coefficients, mean square error behavior of a general class of adaptive predictors is examined.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

S/N 0102-014-6601

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)