[1] This work was supported by NSF Grants GU-2059 and GU-19568 and by U.S. Air Force Grant No. AFOSR-68-1415.

[2] On leave from Punjab Agricultural University (India).

PARTIAL RIDGE REGRESSION[1]

by

D. Raghavarao[2] and K.J.C. Smith

Department of Statistics
University of North Carolina at Chapel Hill

Institute of Statistics Mimeo Series No. 863

February, 1973

PARTIAL RIDGE REGRESSION[1]

by

D. Raghavarao[2] and K.J.C. Smith
Department of Statistics
University of North Carolina at Chapel Hill

ABSTRACT

A partial ridge estimator is proposed as a modification of the Hoerl and Kennard ridge regression estimator. It is shown that the proposed estimator has certain advantages over the ridge estimator. The problem of taking an additional observation to meet certain optimality criteria is also discussed.

---

[1] This work was supported by NSF Grants GU-2059 and GU-19568 and by U.S. Air Force Grant No. AFOSR-68-1415.

[2] On leave from Punjab Agricultural University (India).

1. Introduction. Consider the problem of fitting a linear model $\underline{y} = X\underline{\beta} + \underline{\varepsilon}$, where $\underline{y}' = (y_1, y_2, \ldots, y_n)$ is a vector of n observations on the dependent variable; $X = (x_{ij})$ is an $n \times p$ matrix of rank p, $\underline{x}_i' = (x_{i1}, x_{i2}, \ldots, x_{ip})$ being the vector of i-th observations on the independent variables ($i = 1,2,\ldots,n$); $\underline{\beta}' = (\beta_1, \beta_2, \ldots, \beta_p)$ is the vector of parameters to be estimated; and $\underline{\varepsilon}'$ is an n-dimensional vector of random errors assumed to be distributed with mean vector $\underline{0}'$ and dispersion matrix $\sigma^2 I_n$, $\underline{0}$ being a zero vector and I_n the identity matrix of order n. Without loss of generality we assume that the dependent and independent variables are standardized so that $X'X$ is a correlation matrix.

Let $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$ be the eigenvalues of $X'X$ and let $\underline{\xi}_1, \underline{\xi}_2, \ldots, \underline{\xi}_p$ be a set of ortho-normal eigenvectors associated with the eigenvalues λ_i ($i = 1,2,\ldots,p$). Let $\alpha_i = \underline{\xi}_i' \underline{\beta}$ for $i = 1,2,\ldots,p$.

The usual least squares estimator of $\underline{\beta}$ is given by

$$\hat{\underline{\beta}} = (X'X)^{-1} X'\underline{y} \quad , \tag{1.1}$$

and has the unsatisfactory property, when $X'X$ differs substantially from an identity matrix, that the mean squared error or expected squared distance from $\hat{\underline{\beta}}$ to $\underline{\beta}$ tends to be large compared to that of an orthogonal system.

Often an investigator is interested in obtaining a variance balanced design in which each parameter β_i is estimated with equal precision. The departure of a design from variance balancedness increases the more $X'X$ differs from an identity matrix.

The ridge regression method proposed by Hoerl and Kennerd (1970) estimates $\underline{\beta}$ by the ridge estimator given by

$$\hat{\underline{\beta}}^* = (X'X + kI_p)^{-1} X'\underline{y} \quad , \tag{1.2}$$

where k is a positive real number satisfying

$$k \leq \sigma^2/\alpha_{max}^2 \quad , \tag{1.3}$$

α_{max} being the maximum of $\alpha_i (i = 1,2,\ldots,p)$. The estimator $\hat{\underline{\beta}}^*$ is a biased estimator of $\underline{\beta}$ but has a smaller mean squared error than the least squares estimator $\hat{\underline{\beta}}$.

We propose here as an alternative to the ridge estimator $\hat{\underline{\beta}}^*$, the estimator

$$\hat{\underline{\beta}}_p = (X'X + k_p \underline{\xi}_p \underline{\xi}_p')^{-1} X'\underline{y} \quad , \tag{1.4}$$

where

$$k_p = \sigma^2/\alpha_p^2 \quad . \tag{1.5}$$

This estimator may be called the partial ridge estimator of $\underline{\beta}$. We show in Section 2 that the partial ridge estimator estimates $\underline{\xi}_p'\underline{\beta}$ (with minimum mean squared error and estimates the $\underline{\xi}_i'\underline{\beta}$) $(i = 1,2,\ldots,p-1)$ unbiasedly. In Section 3 we consider the problem of taking an additional observation so as to remove the bias of the partial ridge estimator and to attain certain optimality criteria.

2. Partial Ridge Estimator. To control the mean squared error of the estimator of the coefficient vector $\underline{\beta}$ in the model $\underline{y} = X\underline{\beta} + \underline{\varepsilon}$, Hoerl and Kennard (1970) proposed a ridge estimator, $\hat{\underline{\beta}}^*$, defined by (1.2) and showed that the mean squared error of $\hat{\underline{\beta}}^*$ was less than that of the least squares estimator $\hat{\underline{\beta}}$ of $\underline{\beta}$. Specifically, the mean squared error of $\hat{\underline{\beta}}^*$ is

$$E[(\hat{\underline{\beta}}^* - \underline{\beta})'(\hat{\underline{\beta}}^* - \underline{\beta})] = \sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i+k)^2} + \sum_{i=1}^{p} \frac{k^2\alpha_i^2}{(\lambda_i+k)^2} \tag{2.1}$$

$$= \gamma_1(k) + \gamma_2(k) \quad , \text{ say,}$$

where $E[\ ]$ denotes the expected value of the term in braces. The term $\gamma_1(k)$ is the sum of the variances of the components of $\hat{\beta}^*$ and the term $\gamma_2(k)$ is the bias component of the mean squared error. When $k = 0$, the ridge estimator coincides with the least squares estimator.

We propose as an alternative to the ridge estimator of $\underline{\beta}$ a partial ridge estimator of $\underline{\beta}$, denoted by $\hat{\underline{\beta}}_p$, defined by (1.4). The partial ridge estimator has the following property:

Theorem 2.1. The partial ridge estimator $\hat{\underline{\beta}}_p = (X'X + k_p \underline{\xi}_p \underline{\xi}_p')^{-1} X'\underline{y}$, where $k_p = \sigma^2/\alpha_p^2$, is such that $\underline{\xi}_p' \hat{\underline{\beta}}_p$ is the linear estimator of $\underline{\xi}_p' \underline{\beta}$ with minimum mean squared error and $\underline{\xi}_i' \hat{\underline{\beta}}_p$ is the best linear unbiased estimator of $\underline{\xi}_i' \underline{\beta}$ $(i = 1,2,\ldots,p-1)$.

Proof. Since $\underline{\xi}_1, \underline{\xi}_2, \ldots, \underline{\xi}_p$ are a set of ortho-normal eigenvectors associated with the eigen values $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$ of $X'X$; $X\underline{\xi}_i$ will be eigenvectors associated with the eigenvalues λ_i of XX' $(i = 1,2,\ldots,p)$.

Let $\underline{\eta}_1, \underline{\eta}_2, \ldots, \underline{\eta}_{n-p}$ be a set of orthogonal eigenvectors associated with the zero eigen value of multiplicity $n-p$ of XX'. The vectors $X\underline{\xi}_i$ $(i = 1,2,\ldots,p)$ and $\underline{\eta}_j$ $(j = 1,2,\ldots,n-p)$ form a basis of an n-dimensional vector space. Without loss of generality any linear estimator of $\ell = \underline{\xi}_p'\underline{\beta}$ can be taken to be

$$\tilde{\ell} = \sum_{i=1}^{p} c_i \underline{\xi}_i' X' \underline{y} + \sum_{j=1}^{n-p} d_j \underline{\eta}_j' \underline{y}, \tag{2.2}$$

where c_i and d_j are scalars. The mean squared error of $\tilde{\ell}$ as an estimator of ℓ can be shown to be

$$E[(\tilde{\ell}-\ell)^2] = \sum_{i=1}^{p-1} c_i^2(\lambda_i^2 \alpha_i^2 + \sigma^2\lambda_i) + (c_p \lambda_p - 1)^2\alpha_p^2 + c_p^2\lambda_p\sigma^2 + \sum_{j=1}^{n-p} d_j^2 \sigma^2 . \tag{2.3}$$

Minimizing (2.3) with respect to the coefficients c_i and d_j we have

$$(2.4)\quad c_1 = c_2 = \ldots = c_{p-1} = d_1 = d_2 = \ldots = d_{n-p} = 0,\ c_p = \frac{1}{\lambda_p + \frac{\sigma^2}{\alpha_p^2}} .$$

Choosing $k_p = \sigma^2/\alpha_p^2$, the linear estimator of $\underline{\xi}_p' \underline{\beta}$ with least mean squared error is given by $(\lambda_p + k_p)^{-1} \underline{\xi}_p' X'\underline{y}$. The best linear unbiased estimators of $\underline{\xi}_i' \underline{\beta}$ are the least squares estimators $\lambda_i^{-1} \underline{\xi}_i' X'\underline{y}$ $(i = 1,2,\ldots,p-1)$. Making a 1-1 correspondence of estimators of $\underline{\xi}_i' \underline{\beta}$ with estimators of β_i, the required estimator $\hat{\underline{\beta}}_p$ of $\underline{\beta}$ is given by

$$\hat{\underline{\beta}}_p = \left(\sum_{i=1}^{p-1} \frac{1}{\lambda_i} \underline{\xi}_i \underline{\xi}_i' + \frac{1}{\lambda_p + k_p} \underline{\xi}_p \underline{\xi}_p'\right) X'\underline{y}$$

$$= (X'X + k_p \underline{\xi}_p \underline{\xi}_p')^{-1} X'\underline{y} .$$

This completes the proof of Theorem 2.1.

The problem of estimating k_p can be solved either by graphical or iterative procedures as described by Hoerl and Kennard (1970).

From (2.1) and (2.3) we note that the bias component in the mean squared error of the partial ridge estimator is smaller than that of the ridge estimator.

3. Optimum choice of an additional observation. The equation (1.4) defining the partial ridge estimator suggests taking an additional observation y_{n+1} on the dependent variable corresponding to some choice of values of the independent variables. Let us assume without loss of generality that the design matrix with an additional observation is

$$(3.1)\quad X_1 = \begin{bmatrix} X \\ \sqrt{w}\ \underline{x}'_{n+1} \end{bmatrix} ,$$

where $\underline{x}'_{n+1}\ \underline{x}_{n+1} = 1$ and w is a non-zero scalar. The least squares estimator of $\underline{\beta}$ using the additional observation is

$$\underline{\hat{\beta}}_A = (X_1' X_1)^{-1} X_1' \begin{pmatrix} \underline{y} \\ y_{n+1} \end{pmatrix} \tag{3.2}$$

$$= (X'X + w\, \underline{x}_{n+1}\, \underline{x}'_{n+1})^{-1} X_1' \begin{pmatrix} \underline{y} \\ y_{n+1} \end{pmatrix} ,$$

which is an unbiased estimator of $\underline{\beta}$.

Before discussing the optimal choice of the additional observation, we shall introduce the following:

Definition 3.1 Let $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$ be the eigenvalues of $X'X$, where X is a $n \times p$ design matrix. The departure from variance balancedness of the design X is measured by

$$Q(X) = \sum_{i=1}^{p} (\lambda_i - \bar{\lambda})^2 , \tag{3.3}$$

where $\bar{\lambda} = (\sum_{i=1}^{p} \lambda_i)/p$.

An equivalent expression for $Q(X)$ is

$$Q(X) = \mathrm{tr}[(X'X)^2] - [\mathrm{tr}(X'X)]^2/p , \tag{3.4}$$

where $\mathrm{tr}[A]$ denotes the trace of the matrix A.

Definition 3.2. [Kiefer (1959)] Of the class of all $n \times p$ design matrices X, the design X is A-optimal if $\mathrm{tr}[(X'X)^{-1}]$ is minimum.

Definition 3.3. [Kiefer (1959)] Of the class of all $n \times p$ design matrices X, the design X is D-optimal if $\det[(X'X)^{-1}]$ is minimum, where $\det[\ ]$ denotes the determinant of the matrix in braces.

The following theorem gives the optimum choice of w and $\underline{x}_{n+1}$ for an additional observation:

Theorem 3.1. Given the $n \times p$ design matrix X, among possible choices of w and $\underline{x}'_{n+1}$ in (3.1), the design

$$X^* = \begin{bmatrix} X \\ \sqrt{\frac{\bar{\lambda}-\lambda_p}{1-\frac{1}{p}}}\ \underline{\xi}'_p \end{bmatrix} , \tag{3.5}$$

has the following properties:

(i) $$Q(X^*) < Q(X) .$$

(ii) Among the class of designs X_1 in (3.1),

$$Q(X^*) \leq Q(X_1) .$$

(iii) Among the class of designs X_1 in (3.1) and subject to $Q(X_1)$ minimum, X^* is A- and D-optimal..

Proof. For the design X_1 of (3.1),

$$Q(X_1) = \mathrm{tr}[(X_1' X_1)^2] - [\mathrm{tr}(X_1' X_1)]^2/p \tag{3.6}$$

$$= Q(X) + 2w\, \underline{x}'_{n+1} X'X\, \underline{x}_{n+1} + w^2(1 - \tfrac{1}{p}) - 2w\, \bar{\lambda} .$$

The quadratic form $\underline{x}'_{n+1}(X'X)\underline{x}_{n+1}$ is minimized when $\underline{x}'_{n+1} = \underline{\xi}'_p$ and the minimum value is λ_p. Substituting this least value of $\underline{x}'_{n+1}X'X\,\underline{x}_{n+1}$ in (3.6) and minimizing with respect to w, we obtain the stationary value of w to be

$$w = \frac{\bar{\lambda} - \lambda_p}{1 - \frac{1}{p}} . \tag{3.7}$$

Substituting into (3.6), the minimum value of $Q(X_1)$ is

$$Q_{\min}(X_1) = Q(X^*) = Q(X) - \frac{(\bar{\lambda} - \lambda_p)^2}{1 - \frac{1}{p}} \quad . \tag{3.8}$$

Thus $Q(X^*) < Q(X)$. Moreover $Q(X^*)$ is the minimum value of $Q(X_1)$.

Now

$$\begin{aligned} \det[(X_1' X_1)^{-1}] &= \det[(X'X + w\, \underline{x}_{n+1} \underline{x}_{n+1}')^{-1}] \\ &= \det[(X'X)^{-1}](1 + w\, \underline{x}_{n+1}' (X'X)^{-1} \underline{x}_{n+1})^{-1} \quad . \end{aligned} \tag{3.9}$$

The maximum value of $\underline{x}_{n+1}' (X'X)^{-1} \underline{x}_{n+1}$ is $1/\lambda_p$ for $\underline{x}_{n+1}' = \underline{\xi}_p'$. Hence $\det[(X_1' X_1)^{-1}]$ is minimized with respect to $\underline{x}_{n+1}$ when $\underline{x}_{n+1}' = \underline{\xi}_p$. In order that $Q(X_1)$ be least, w must be given by (3.7). Hence X^* is D-optimal among the class of designs X_1 with minimum $Q(X_1)$.

To prove the A-optimality of X^* among the class of designs X_1 with minimum $Q(X_1)$, we observe that

$$\operatorname{tr}[(X_1' X_1)^{-1}] = \operatorname{tr}[(X'X)^{-1}] - \frac{w\, \underline{x}_{n+1}' (X'X)^{-2} \underline{x}_{n+1}}{1 + w\, \underline{x}_{n+1}' (X'X)^{-1} \underline{x}_{n+1}} \quad . \tag{3.10}$$

The maximum value of the second term on the right hand side of (3.10) is the maximum of

$$\mu = \frac{1}{\lambda(\frac{\lambda}{w} + 1)} \quad , \tag{3.11}$$

where λ's are the eigenvalues of $X'X$. In order that $Q(X_1)$ is least, w is given by (3.7) and the maximum μ is attained when $\lambda = \lambda_p$ and $\underline{x}_{n+1}' = \underline{\xi}_p'$. Thus X^* is A-optimal among the class of X_1 matrices with minimum $Q(X_1)$.

References

Hoerl, Arthur E. and Kennard, Robert W. (1970). "Ridge Regression: Biased Estimation for Non-orthogonal Problems." *Technometrics*, 12, 55-67.

Kiefer, J. (1959). "Optimum Experimental Designs." *J. Roy. Statist. Soc.*, 21B, 272-304.