An Explicit Solution for Multivariate Ridge Regression

Min Woong Shin* & Sung H. Park**

ABSTRACT

We propose that, in order to control the inflation and general instability associated with the least squares estimates, we can use the ridge estimator

$$\hat{B}^* = (X'X + kI)^{-1}X'Y : k \geq 0$$

for the regression coefficients B in multivariate regression. Our hope is that by accepting some bias, we can achieve a larger reduction in variance. We show that such a k always exists and we derive the formula obtaining k in multivariate ridge regression.

1. Introduction

Let $y_1, \cdots, y_p$ be $N \times 1$ vectors representing N independent observations on each of p correlated dependent random variables.

Assume the linear model

$$(1.1) \qquad y_j = X\beta_j + u_j, \quad j = 1, \cdots, p,$$

where X is an $(N \times q)$ matrix of known form and may be thought of as arising either as a "functional" or a conditional" regressor matrix; β_j is a $(q \times 1)$ vector of parameters; u_j is a $(N \times 1)$ vector of errors and $E(u_j) = 0$, $\text{Var}(u_j) = I\sigma^2$, so the elements of u_j are uncorrelated. For a given j, (1.1) is a univariate regression.

The basic model equation may be written in a more compact form in the following way. Define

$$Y \equiv (y_1, \cdots, y_p), \quad U \equiv (u_1, \cdots, u_p), \quad B \equiv (\beta_1, \cdots, \beta_p)$$

Then (1.1) becomes

$$(1.2) \qquad \underset{(N \times p)}{Y} = \underset{(N \times q)}{X} \; \underset{(q \times p)}{B} + \underset{(N \times p)}{U}.$$

* Department of Mathematics, Hankuk University of Foreign Studies.
** Department of Computer Science and Statistics, Seoul National University.

To define the multivariate regression model completely, impose the following assumptions and constraints on the quantities in (1.2)

(1.3) $$p+q \leq N$$

(1.4) $$\text{rank}(X)=q$$

Define the rows of U by

$$U \equiv (v_1, \cdots, v_N)'$$

where v_j is a $(p \times 1)$ vector, $j=1, \cdots, N$.

(1.5) $$E(v_j)=0, \quad \text{Var}(v_j)=\Sigma \equiv (\sigma_{ij}) \quad \text{and} \quad \Sigma > 0 \text{ (positive definite).}$$

An alternative form of this assumption, which is obtained by stringing out the columns of U into a long $Np \times 1$ vector u, where $u' \equiv (u_1', \cdots, u_p')$. Then

(1.6) $$E(u)=0, \quad \text{Var}(u)=\Sigma \otimes I_N \quad (\otimes \text{ is the direct product})$$

(1.7) $$\mathcal{L}(v_j)=N(0, \Sigma), \quad j=1, \cdots, N. \quad (\mathcal{L} \text{ means probability law}).$$

2. Multivariate Ridge Regression

From the notation and assumptions we know that

$$\hat{B}=(X'X)^{-1}X'Y$$

as an estimate of B and this gives the total minimum sum of squares of the residuals:

$$\phi(\hat{B})=\sum_{j=1}^{p}(y_j - X\hat{\beta}_j)' \; (y_j - X\hat{\beta}_j)$$

where X is $N \times q$ matrix of the known independent variables.

(2.1) $$\text{Var}(\hat{\beta})=\Sigma \otimes (X'X)^{-1}$$

where $\hat{\beta}' \equiv (\hat{\beta}_1', \cdots, \hat{\beta}_p')$

(2.2) $$L_1^2 \equiv \sum_{j=1}^{p}(\hat{\beta}_j - \beta_j)' \; (\hat{\beta}_j - \beta_j)$$

where L_1 is distance from $\hat{B}$ to B.

(2.3) $$E(L_1^2) = \sum_{j=1}^{p} \text{Tr}(X'X)^{-1}\sigma_j{}^2$$

If the eigenvalues of $X'X$ are denoted by

(2.4) $$\lambda_{\max}=\lambda_1 \geq \cdots \geq \lambda_p = \lambda_{\min} > 0,$$

then the average value of the squared distance from $\hat{B}$ to B is given by

(2.5) $$E(L_1^2) = \sum_{j=1}^{p} \sigma_j^2 (\sum_{i=1}^{q}(1/\lambda_i))$$

and the variance when the error is normally distributed is given by

(2.6) $$\mathrm{Var}(L_1^2) = \sum_{j=1}^{p} (2\sigma_j^4(\sum_{i=1}^{q}(1/\lambda_i))^2)$$

Hence, if the shape of the factor space is such that reasonable data collection results in an $X'X$ with one or more small eigenvalues, the distance from $\hat{B}$ to B will tend to be large. In order to control the inflation and general instability associated with the least squares estimators, we might use a ridge estimator,

(2.7) $$\hat{B}^* = (X'X + kI)^{-1}X'Y \ : \ k \geq 0$$

in multivariate regression.

The relationship of a ridge estimate to an ordinary estimate is given by the alternative form

(2.8) $$\begin{aligned} \hat{B}^* &= (X'X + kI)^{-1}X'Y \\ &= (X'X + kI)^{-1}X'X\hat{B} \\ &= (I + k(X'X)^{-1})^{-1}\hat{B} \\ &= Z\hat{B} \end{aligned}$$

Let $\bar{B}$ be any estimate of the vector B. Then the residual sum of sruares can be written as

$$\begin{aligned} \phi &= \sum_{j=1}^{p}(y_j - X\bar{\beta}_j)'(y_j - X\bar{\beta}_j) \\ &= \sum_{j=1}^{p}(y_j - X\hat{\beta}_j)' \ (y_j - X\hat{\beta}_j) + \sum_{j=1}^{p}(\bar{\beta}_j - \hat{\beta}_j)'X'X(\bar{\beta}_j - \hat{\beta}_j) \\ &= \phi_{\min} + \phi(\bar{B}) \end{aligned}$$

The ridge regression coefficient matrix $\hat{B}^*$ is the single value of B which is the one with minimum length for a fixed ϕ.

This can be stated precisely as follows:

$$\text{Minimize} \quad \sum_j \bar{\beta}_j'\bar{\beta}_j$$

(2.9) $$\text{subject to} \quad \sum_j (\bar{\beta}_j - \hat{\beta}_j)'X'X(\bar{\beta}_j - \hat{\beta}_j) = \phi_0$$

As a Lagrangian problem this is

$$\text{minimize} \quad F=\sum_j \bar{\beta}_j'\bar{\beta}_j+(1/k)(\sum_j(\bar{\beta}_j-\hat{\beta}_j)'X'X(\bar{\beta}_j-\hat{\beta}_j)-\phi_0)$$

where $(1/k)$ is the multiplier.

Then

$$\frac{\partial F}{\partial \bar{\beta}_j}=2\bar{\beta}_j+(1/k)(2(X'X)\bar{\beta}_j-2(X'X)\hat{\beta}_j)=0, \quad j=1,\cdots,p.$$

This reduces to

$$\bar{\beta}_j = \hat{\beta}_j^* = (X'X+kI)^{-1}X'y_j, \; j = 1,\cdots,p.$$

That is,

$$\bar{B} = \hat{B}^* = (X'X + kI)^{-1}X'Y$$

where k is chosen to satisfy the restraint (2.9). This is the ridge estimator.

To look at $\hat{B}^*$ from the point of view of mean square error it is necessary to obtain an expression for $E(L_1^2(k))$.

$$E(L_1^2(k))=E(\sum_j(\hat{\beta}_j^*-\beta_j)' \; (\hat{\beta}_j^*-\beta))$$

$$=\sum_j(\sigma_j^2\sum_{i=1}^q \lambda_i/(\lambda_i+k)^2)+(\sum_j k^2\beta_j'(X'X+kI)^{-2}\beta_j$$

$$=\gamma_1(k)+\gamma_2(k)$$

where λ_i are the eigen-values of $X'X$.

Theorem (Existence Theorem). There always exists a $k>0$ such that

$$E(L_1^2(k))<E(L_1^2(0))=\sum_j(\sigma_j^2\sum_{i=1}^q(1/\lambda_i))$$

Proof : $E[L_1^2(k)]=\sum_j[\sigma_j^2 \sum_i \lambda_i/(\lambda_i+k)^2]+\sum_j k^2\beta_j'(X'X+kI)^{-2}\beta_j$

$$\equiv\sum_j\gamma_{1j}+\sum_j\gamma_{2j}$$

If Λ is the matrix of eigenvectors of $X'X$ and p is the orthogonal transformation such that $X'X=P'\Lambda P$, then

$$\gamma_{2j}(k)=k^2\sum_i\alpha_{ji}^2/(\lambda_i+k), \text{ where } \alpha_j=P\beta_j$$

$$\frac{d}{dk}E(L_1^2(k))=\frac{d}{dk}\gamma_1(k)+\frac{d}{dk}\gamma_2(k)$$

$$=\sum_j(-2\sigma_j^2)\sum_i \lambda_i/(\lambda_i+k)^3+2k\sum_i\lambda_i\alpha_{ji}/(\lambda_i+k)^3$$

Let $E[L_{1j}^2(k)]=\gamma_{1j}(k)+\gamma_{2j}(k)$.

Then $E[L_1^2(k)]=\sum_j E[L_{1j}^2(k)]$.

It is known that

$$E[L_{1j}^2(k)]<E[L_{1j}^2(0)]=\sigma_j^2 \sum_i(1/\lambda_i)$$

for a $k<\sigma_j^2/\max[\sigma_{j1}^2,\cdots, \sigma_{jq}^2]$ and for each j, $j=1,\cdots,p$.

Hence

$$E[L_1^2(k)]=\sum_j E[L_{1j}^2(k)]<\sum_j E[L_{1j}^2(0)]=\sum_j[\sigma_j^2 \sum_i(1/\lambda_i)]$$

for a $k<\min\left[\frac{\alpha_1^2}{\max[\alpha_{11}^2,\cdots,\alpha_{1q}^2]},\cdots,\frac{\alpha_p^2}{\max[\alpha_{p1},\cdots,\alpha_{pq}^2]}\right]$.

3. Derivation of an explicit solution

The ridge regression estimators $\hat{B}^*$, for a fixed $k>0$, satisfy

(3.1) $$(X'X+kI)\hat{B}^*=X'Y$$

so that

(3.2) $$\hat{B}^*=(X'X+kI)^{-1}X'Y$$

The general form of ridge regression reduces $X'X$ to a diagonal matrix by applying an orthogonal transformation p.

We have that

$$P(X'X)P'=\Lambda$$

where P is a $q\times q$ orthogonal matrix and Λ is a diagonal matrix whose diagonal elements $\lambda_1,\cdots,\lambda_q$ are the characteristic roots of $X'X$. If we write

$$X^*=XP'$$

and

$$A=PB$$

then the model (1.2) may be written as

$$Y = X^* A + U$$

where $(X^*)'(X^*) = \Lambda$

The general ridge estimation procedure is then defined as

$$\hat{A}^* = [(X^*)'(X^*) + K]^{-1}(X^*)'Y \tag{3.3}$$

$$\equiv (\hat{\alpha}_1^*, \cdots, \hat{\alpha}_p^*)$$

where K is a diagonal matrix with nonnegative diagonal elements $k_1, \cdots, k_q$.

Optimal values for the k's in (3.3) can be considered to be those k_i's that minimize

$$Q = E[\sum_{j=1}^{p} (\hat{\alpha}_j^* - \alpha_j)'(\hat{\alpha}_j^* - \alpha_j)] \tag{3.4}$$

With a certain amount of algebra, (3.4) may be expressed as

$$Q = \sum_{j=1}^{p} [\sum_{i=1}^{q} (\sigma_j^2 \lambda_i + \alpha_{ji}^2 k_i)/(\lambda_i + k_i)^2] \tag{3.5}$$

and differentiation of (3.5) with respect to the k's yields the minimization equations

$$\frac{\partial Q}{\partial k_i} = \sum_{j=1}^{p} 2\lambda_i(\lambda_i + k_i)(k_i \alpha_{ji}^2 - \sigma_j^2)/(\lambda_i + k_i)^4 = 0, \quad i = 1, \cdots, q \tag{3.6}$$

$$\sum_{j=1}^{p} (k_i \alpha_{ji}^2 - \sigma_j^2) = 0.$$

From the full rank assumption on $X'X$ we have that $\lambda_i > 0$ for all i. Restricting the k_i's to be non-negative yields the solution

$$k_i = \sum_{j=1}^{p} [\sigma_j^2 / (\sum_j \alpha_j^2)], \quad i = 1, \cdots, q. \tag{3.7}$$

If the ridge estimation procedure is defined as $\hat{A}^* = [(X^*)'(X^*) + kI]^{-1}(X^*)'Y$, then the optimal value k that minimizes Q is

$$k = q \sum_j \sigma_j^2 / \sum_i \sum_j \alpha_{ji}^2 \tag{3.8}$$

In (3) the author suggests using an iterative procedure to estimate k_i. The procedure may be described by the formula

$$k_i(k) = \sum_{j=1}^{p} \sigma_j^2 / (\sum_j \alpha_{ji}^{2*}(k))^2, \quad i = 1, \cdots, q \tag{3.9}$$

where the bracketed k subscript is used to denote the kth iteration. As initial values we use

$$\hat{\alpha}^*_{ji(0)} = \hat{\alpha}_{ji}' \quad i=1,\cdots,q \tag{3.10}$$

where $\hat{\alpha}_{ji}$ is the ordinary least squares estimate of α_{ji}. The $k_i(k)$ values are used in equation (3.3) in order to obtain the next $\hat{\alpha}^*_{ji(k+1)}$ values for use in (3.9). Presumably, $\hat{\sigma}^2_j$ is the residual sum of squares for the model (1.2) divided by $(N-q)$, the ordinary least squares estimator for σ^2_j.

Hemmerle (3) shows that an explicit solution is available for the limiting $\hat{\alpha}^*_{i(k)}$ values so that it is not necessary to iterate in order to obtain these values.

It will be convenient to represent the p-vectors $(X^*)'y$ and $\alpha^*_i(k)$ as diagonal matrices. In this context let

$$B=\text{diag}(((X^*)'y)_1,\cdots,((X^*)'y)_p)$$

and

$$A_k=\text{diag}(\alpha_{j1(k)},\cdots,\alpha_{jp(k)}).$$

As a consequence we have that

$$A=\Lambda^{-1}B. \tag{3.11}$$

Furthemore

$$A_{k+1}=(\Lambda+\hat{\sigma}^2_j A^{-2})^{-1}\Lambda A_0 \tag{3.12}$$

and

$$A_{k+1}=(I+\sigma^2_j\Lambda^{-1}A_k{}^{-2})^{-1}A. \tag{3.13}$$

If we next let

$$D=\Lambda/\hat{\sigma}^2_j \tag{3.14}$$

and we let

$$E_k=D^{-1}A_k{}^{-2}, \tag{3.15}$$

the iterative procedure is reduced to the simple formula

$$E_{k+1}=E_0(I+E_k)^2. \tag{3.16}$$

Let us assume that $\alpha_{ji}\neq 0$ for all i and that the iterative procedure is convergent such that

$$\lim_{k\to\infty} E_k=E^*. \tag{3.17}$$

From (3.16) and (3.17) we must have the relationship

(3.18) $$E^*=E_0(I+E^*)^2$$

or

(3.19) $$(E^*)^2+(2I-E_0^{-1})E^*+I=0.$$

Now (3.19) consists of p equations of the form

(3.20) $$(e^*)^2+(2-1/e_0)e^*+1=0$$

where the e_0 and e^* are scalar. Solving (3.20) for e^* we obtain

(3.21) $$e^*=[(1-2e_0)\pm(1-4e_0)]/2e_0.$$

Hemmerle shows that the kth iterative process defined by (3.22) converges whenever $0<e_0<\frac{1}{4}$ and diverges for $e_0>\frac{1}{4}$. That is

(3.22) $$e_{k+1}=e_0(1+e_k)^2$$

Let

(3.23) $$e_i^*=\lim_{k\to\infty} e_i(k),\quad \hat{\alpha}_{ji}^*=\lim_{k\to\infty} \hat{\alpha}_{ji}^*(k)$$

where $e_{i(j)}$ denote the jth iterate of the ith equation.

Then since

(3.24) $$e_i(k)=\hat{\sigma}_j^2/\lambda_i(\hat{\alpha}_{ji}^*(k))^2$$

we have that

(3.25) $$\hat{\alpha}_{ji}^*(k)\underset{k\to\infty}{\to 0}\quad\text{for}\quad e_i(0)>\frac{1}{4}$$

whenever the procedure defined by (3.22) diverges for the ith equation. Thus we let

$$\hat{\alpha}_{ji}^*=0\quad\text{for}\quad e_j(0)>\frac{1}{4}.$$

When the procedure converges for the ith equation we have that

(3.26) $$\hat{\alpha}_{ji}^*=\frac{[(X^*)'y]_i}{\lambda_i+\lambda_i e_i^*}=\frac{\hat{\alpha}_{ji}}{(1+e_i^*)}\quad\text{for}\quad 0<e_{i(0)}\leq\frac{1}{4}$$

where e^* is evaluated using the formula (3.21).

4. Numerical Example

The data consist of

$$X=\begin{pmatrix} 3\sqrt{2}/10 & 4\sqrt{2}/10 \\ 4\sqrt{2}/10 & 3\sqrt{2}/10 \\ 5\sqrt{2}/10 & 5\sqrt{2}/10 \end{pmatrix}, \quad Y=\begin{pmatrix} 1 & 1 \\ 2 & 3 \\ 3 & 5 \end{pmatrix}.$$

After orthogonal reparametrization of these data, we obtain the following results.

$$X'X=\begin{pmatrix} 1 & 49/50 \\ 49/50 & 1 \end{pmatrix}, \quad X'Y=\begin{pmatrix} 26\sqrt{2}/10 & 40\sqrt{2}/10 \\ 25\sqrt{2}/10 & 38\sqrt{2}/10 \end{pmatrix},$$

$$A=\begin{pmatrix} 85/33 & 130/33 \\ 5 & 10 \end{pmatrix}.$$

For this example we see that

$$\hat{\sigma}_1^2=y_1'y_1-\hat{\sigma}_1^*(X^*)'y_1=12/33$$

$$\hat{\sigma}_2^2=75/33$$

Consequently,

$$e_{11(0)}=\frac{12/33}{(51/10)\ (85/33)}=\frac{8}{289}<\frac{1}{4} \quad \text{and}$$

$$e_{21(0)}=\frac{12/33}{(1/10)(5)}=\frac{24}{33}>\frac{1}{4} \quad \text{for } y_1.$$

Similarly $\quad e_{12(0)}=\frac{75}{1014}>\frac{1}{4}$ and $e_{22(0)}=\frac{75}{60}>\frac{1}{4}$ for y_2.

Using the explicit method developed in the previous sections we evaluate e^* by formula (3.21) and obtain

$$e_{11}^*=0.0293 \quad \text{and} \quad e_{12}^*=0.0875$$

Therefore

$$\hat{\alpha}_{11}^*=\frac{\hat{\alpha}_{11}}{(1+e_1^*)}=2.502, \quad \hat{\alpha}_{12}^*=\frac{\hat{\alpha}_{12}}{(1+e_2^*)}=3.623 \quad \text{and} \quad \hat{\alpha}_{21}^*=\hat{\alpha}_{22}^*=0.$$

We obtain k_1 and k_2 by formula (3.9), that is

$$k_1=0.9551, \quad k_2=0.7278.$$

The resulting solution is then given by

$$\hat{B}^*=\begin{pmatrix} 1.769 & 2.562 \\ 1.769 & 2.562 \end{pmatrix}.$$

(Received January 1982; Revised September 1982)

REFERENCES

(1) Anderson, T.W., (1956). "An introduction to multivariate statistical analysis," John Wiley and Sons, New York.

(2) Donald, F. Morrison, (1976). "Multivariate statistical methods," McGraw-Hill, NewYork.

(3) Hemmerle, W.J., (1975). "An explicit solution for generalized ridge regression," Technometrics, 17.

(4) Hocking, P.R., (1976). "The analysis and selection of variables in linear regression." Biometrics, 32.

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

(6) Kennard, R.W., (1971). "A note on the C_p-statistics," Technometrics, 13.

(7) Mallows, C.L., (1973). "Some comments on C_p," Technometrics, 15.

(8) S. James Press., (1977). "Applied multivariate analysis," Holt, New York.