A SAS MACRO FOR RIDGE ANALYSIS OF MULTIVARIATE GENERAL LINEAR MODELS

Agam N. Sinha, Kenneth A. Hardy
Institute for Research in Social Science
University of North Carolina
Chapel Hill, N.C.

ABSTRACT

Ordinary least squares (OLS) fails to provide satisfactory estimates for General Linear Model parameters when the prediction vectors are not orthogonal, i.e., when the data are ill-conditioned. The technique of ridge regression provides better minimum variance estimates in such situations. This paper reviews the general theory of ridge regression analysis for the multivariate general linear model and presents a SAS macro which enables the user to obtain both OLS and ridge estimates for purposes of comparing the two techniques and visualizing the effects of non-orthogonality.

INTRODUCTION

The ordinary least squares (OLS) estimation procedure fails to provide satisfactory estimates of the parameters of the multiple regression model when the data are ill-conditioned due to multicollinearity of the regressors. Multicollinearity arises whenever there exists linear dependency among the predictor variables which is indicated by singularity in the matrix of their correlations. For a detailed discussion on ill-conditioning due to multicollinearity, the reader is referred to Farrar and Glauber (1967), Haitovsky (1969), Maquet and Snee (1975), and Franke and Kennard (1970). Heuer and Kennard (1970a, 1970b) have developed the technique of ridge regression to deal with this type of data ill-conditioning for the multiple regression model involving a single dependent variable. The purpose of the present paper is three-fold, viz., (a) to extend the method of ridge analysis to the case of general linear models involving more than one dependent variable, (b) to present a discussion of testing general linear hypotheses involving ridge estimates, and (c) to present a user oriented SAS macro to perform ridge analyses.

THE MULTICOLLINEARITY PROBLEM

Consider the general linear model

\[ Y = X \theta + \varepsilon, \]

where \( Y \) is an \( n \times p \) matrix of \( p \) dependent variables with \( n \) observations on each, \( X \) is an \( n \times q \) matrix of \( q \) predictor variables corresponding to each of \( n \) observations on \( Y \), \( \theta \) is a \( q \times p \) matrix of unknown model parameters, and \( \varepsilon \) is an \( n \times p \) matrix of non-observable disturbances associated with each of the \( n \) observations on \( p \) dependent variables.

The OLS solution for \( \theta \) assuming full rank of \( X \), is given by

\[ \hat{\theta} = (X'X)^{-1}X'Y, \]

where \( X'X \) is the variance-covariance matrix of each row of \( Y \) so that \( V(Y) = I_p \). This implies that the effective rank of \( X'X \) is less than \( q \). Furthermore, when \( V(Y) \) is near zero, indicating that the effective rank of \( X'X \) is less than \( q \), then it is referred to as "ill-conditioning," \( X'X \) is nearly singular. For seriously ill-conditioned data, the smallest eigenvalue of \( X'X \) is very near zero, indicating that the effective rank of \( X'X \) is less than \( q \). This implies that the variance inflation factors (VIF's) which are the diagonal elements of the variance-covariance inflation matrix (VIM) = \( (X'X)^{-1} \) are very large and so the OLS estimates \( \hat{\theta} \) are also very large, thereby disturbing the predictive efficiency of the model and, therefore, compromising property of the OLS estimator \( \hat{\theta} \). Furthermore, when \( X'X \) is effectively not of full rank, the \( \hat{\theta} \) are not MVUE of its unbiased estimates and the expected squared length of \( \hat{\theta} \) is not minimum. This is because (5) may be rewritten as

\[ E(D) = \text{Tr} [(\hat{\theta} - \theta)'(\hat{\theta} - \theta)] \]

(6)

where \( \theta \) is the variance-covariance matrix of each row of \( Y \) so that \( V(Y) = I_p \). Since \( E(D) \) is affected by the VIF's, the elements of \( E(D) \) are also very large in the presence of multicollinearity. The degree of inflation of the standard errors of \( \hat{\theta} \) may be estimated by replacing \( \theta \) in (6) with either of its unbiased estimates.

(1) \( \hat{\theta} \) minimizes the residual sum of squares, i.e., \( RSS = (Y - XB)'(Y - XB) \) when \( X'X \) is ill-conditioned.

(2) \( \hat{\theta} \) is BLUE and MVUE of \( \theta \).

When the response and predictor variables are standardized, the model (1) reduces to

\[ y* = x^*\theta* + \varepsilon, \]

(3)

where \( y* \) and \( x* \) are standardized \( Y \) and \( X \) variables. Then the correlation based OLS estimator of \( \hat{\theta}^* \) is given by

\[ \hat{\theta}^* = (\text{CORX})^{-1}(\text{CORXY}), \]

(4)

where \( \text{CORX} = (x'^*x^*)/n \) is the correlation matrix of the predictor variables and \( \text{CORXY} = (x'^*y^*)/n \) is the correlation matrix of the predictor variables with the response variables.

It is useful to examine the correlation matrix \( \text{CORX} \) when studying multicollinearity. The reasons for this are given by Franke (1978) and Marquardt and Snee (1975). In the presence of "ill-conditioning," \( \text{CORX} \) (or \( X'X \)) is nearly singular. For seriously ill-conditioned data, the smallest eigenvalue of \( X'X \) is very near zero, indicating that the effective rank of \( X'X \) is less than \( q \) (Marquardt, 1970). This implies that the variance inflation factors (VIF's) which are the diagonal elements of the variance-covariance inflation matrix (VIM) = \( (X'X)^{-1} \) are very large and so the OLS estimates \( \hat{\theta} \) are also very large, thereby disturbing the predictive efficiency of the model and, therefore, compromising property of the OLS estimator \( \hat{\theta} \). Furthermore, when \( X'X \) is effectively not of full rank, the \( \hat{\theta} \) are not MVUE of \( \theta \) and the expected squared length of \( \hat{\theta} \) is not minimum. This is because (5) may be rewritten as

\[ E(D) = \text{Tr} [(\hat{\theta} - \theta)'(\hat{\theta} - \theta)] \]

(6)

where \( \theta \) is the variance-covariance matrix of each row of \( Y \) so that \( V(Y) = I_p \). Since \( E(D) \) is affected by the VIF's, the elements of \( E(D) \) are also very large in the presence of multicollinearity. The degree of inflation of the standard errors of \( \hat{\theta} \) may be estimated by replacing \( \theta \) in (6) with either of its unbiased estimates.

\[ \hat{\theta} \] is the solution which minimizes the residual sum of squares, i.e., \( RSS = (Y - XB)'(Y - XB) \) when \( X'X \) is ill-conditioned.

\[ \hat{\theta} \] is BLUE and MVUE of \( \theta \).

When the response and predictor variables are standardized, the model (1) reduces to

\[ y* = x^*\theta* + \varepsilon, \]

(3)

where \( y* \) and \( x* \) are standardized \( Y \) and \( X \) variables. Then the correlation based OLS estimator of \( \hat{\theta}^* \) is given by

\[ \hat{\theta}^* = (\text{CORX})^{-1}(\text{CORXY}), \]

(4)

where \( \text{CORX} = (x'^*x^*)/n \) is the correlation matrix of the predictor variables and \( \text{CORXY} = (x'^*y^*)/n \) is the correlation matrix of the predictor variables with the response variables.

It is useful to examine the correlation matrix \( \text{CORX} \) when studying multicollinearity. The reasons for this are given by Franke (1978) and Marquardt and Snee (1975). In the presence of "ill-conditioning," \( \text{CORX} \) (or \( X'X \)) is nearly singular. For seriously ill-conditioned data, the smallest eigenvalue of \( X'X \) is very near zero, indicating that the effective rank of \( X'X \) is less than \( q \) (Marquardt, 1970). This implies that the variance inflation factors (VIF's) which are the diagonal elements of the variance-covariance inflation matrix (VIM) = \( (X'X)^{-1} \) are very large and so the OLS estimates \( \hat{\theta} \) are also very large, thereby disturbing the predictive efficiency of the model and, therefore, compromising property of the OLS estimator \( \hat{\theta} \). Furthermore, when \( X'X \) is effectively not of full rank, the \( \hat{\theta} \) are not MVUE of \( \theta \) and the expected squared length of \( \hat{\theta} \) is not minimum. This is because (5) may be rewritten as

\[ E(D) = \text{Tr} [(\hat{\theta} - \theta)'(\hat{\theta} - \theta)] \]

(6)

where \( \theta \) is the variance-covariance matrix of each row of \( Y \) so that \( V(Y) = I_p \). Since \( E(D) \) is affected by the VIF's, the elements of \( E(D) \) are also very large in the presence of multicollinearity. The degree of inflation of the standard errors of \( \hat{\theta} \) may be estimated by replacing \( \theta \) in (6) with either of its unbiased estimates.

Standardization is done by subtracting respective means from the observed values and then dividing by the corresponding standard deviations.
The ridge estimate of $B^*$ is obtained by minimizing (6) by adding a constant $k(>0)$ to the diagonal elements of $(X^*X^*)$. Denoting the ridge estimate of $B^*$ by $\hat{B}_k$ for some $k(>0)$, we have

$$\hat{B}_k = [X^*X^* + kI_q]^{-1}X^*y^*;$$  

(8)

or in correlation form

$$\hat{B}_k = [nCORX + kI_q]^{-1}(nCORXY).$$  

(9)

The ridge method provides a biased estimate of $B^*$, the bias being given by

$$BIAS_k = B^* - \hat{B}_k = (I_q - Z_k)B^*,$$  

(10)

where

$$Z_k = [X^*X^* + kI_q]^{-1}(X^*X^*).$$

(11)

or in correlation form

$$Z_k = [CORX + k*I_q]^{-1}(CORXY),$$  

(12)

Furthermore, following Hoerl and Kennard (1970a), one can show that for $k \neq 0$,

$$\text{Tr} (\hat{B}_k^*\hat{B}_k^*) < \text{Tr} (\hat{B}_0^*\hat{B}_0^*).$$

(13)

Thus, while the estimates are biased, they are not as inflated as the OLS estimates (in presence of multicollinearity).

The residual sum of squares and product matrix using the ridge estimate $\hat{B}_k$ is given by

$$RSS_k = (Y^* - X^*\hat{B}_k)(Y^* - X^*\hat{B}_k)^T = Y^*Y^* - X^*\hat{B}_k\hat{B}_k^T$$

(14)

which is based on $n-q-p$ degrees of freedom. Note that the last expression for $RSS_k$ is used when one works with correlation matrices (CORX) and (CORXY). Hence a biased estimate of $\Sigma^*$ is given by

$$\hat{\Sigma}_k = \left\{ \frac{Y^*Y^* - n\hat{B}_k^T(CORXY)n\hat{B}_k}{n-q-p} \right\}^{-1}.$$

(15)

Choosing $k$ and $\hat{B}_k$ is equivalent to choosing $k^*(\hat{B}_k)$ and $\hat{B}_k$ when one uses the correlation matrices (CORX) and (CORXY) in place of $(X^*X^*)$ and $(X^*Y^*)$ respectively. The 'Ridge Trace' (see Hoerl and Kennard, 1970a and 1970b) provides a
framework for choosing \( k \) (or \( k^* \)), and also for examining which coefficients are sensitive to "ill-conditioning" in the data. The ridge trace is a plot of the estimate of each coefficient versus the value of \( k \) (or \( k^* \)) used to obtain it. The objective in choosing \( k \) (or \( k^* \)) is to get a stable set of coefficient estimates, \( \hat{b}_k \), which will do a good job in predicting future observations (Marquardt and Snee (1975)). By 'stable' it is meant that the coefficients are not sensitive to small changes in the data. A highly correlated pattern of predictor variables results in a rapidly changing coefficient estimates for small values of \( k \) (or \( k^* \)) but as \( k(k^*) \) increases, these estimates stabilize (i.e. narrow the interval within which they fluctuate). One should choose that value of \( k(k^*) \) at which the coefficients' estimates have stabilized, and the stabilized estimates should be taken to describe the model.

In the SAS macro presented here, we have implemented a heuristic, iterative procedure for selecting \( k^* \) which approximates the notion of examining the ridge trace for a point at which the coefficients' estimates "stabilize" in the above sense. The convergence criteria used in the iterative scheme in the macro is as follows. If \( \hat{b}_k \) is the ridge estimate at the \( k \)th iteration then the macro tests whether \( |\hat{b}_{k1} - \hat{b}_{k+11}| \leq C \). The default value of \( C \) in the macro is .001.

**TESTING GENERAL LINEAR HYPOTHESES USING RIDGE ESTIMATES**

Since computation of \( S* \) involves computation of \((X^*X^*)^{-1}\) or \((CORX)^{-1}\) where \((X^*X^*) \) or \((CORX)\) may be very near singular (because of multicollinearity), one might question using \( S* \) in obtaining the unbiased estimates of the variances of \( \hat{b}^* \). However, if \( S* \) is used instead of \( S* \), the estimates of the variances of \( \hat{b}^* \) will not be unbiased. Since the unbiasedness of an estimate is not as important in the ridge analysis as is the reduced variance of the estimate, one is faced with the problem of constructing inferential statistics similar to Student's \( t \) and Snedecor \( F \) for hypothesis testing. Under normal theory, if one is to construct such statistics, a basic requirement is that the numerator and denominator of the test statistics must be independent of each other. Thus, if one used \( S^* \) instead of \( S* \) as an estimate of \( \Sigma \), one cannot construct statistics satisfying this requirement. However, one can construct similar statistics for comparative and heuristic purposes which may be called Heuristic Statistics [Vinod (1976)]. Therefore, it is advisable to check the singularity of \((X^*X^*)\) or \((CORX)\) first (by Wilks-Bartlett's test or some other test, see e.g., Haltovsky, 1969) and then construct the classical test statistics of Heuristic Statistics depending on whether \((X^*X^*)\) or \((CORX)\) is non-singular or singular. In the event \((X^*X^*)\) is almost singular, it is not advisable to use Ridge analysis when the investigator's interest is inference oriented. However, if the investigator's aim is just model fitting and making meaningful future prediction, the ridge method is quite appropriate even if \((X^*X^*) \) (or \(CORX)\) is nearly singular.

We shall outline below the hypothesis testing procedure which involves using \( S^* \), the unbiased estimate of \( S* \).

(a) Test of Hypotheses \( H_0 : b^*_{1j} = 0 \).

For testing \( H_0 : b^*_{1j} = 0 \) vs. \( H_1 : b^*_{1j} \neq 0 \), \( j = 1, 2, \ldots, q \), \( j = 1, 2, \ldots, q \), we compute

\[
I_{1j}(b) = \frac{b^*_{1j}}{v_{b^*_{1j}}(b^*_j)},
\]

where \( v_{b^*_{1j}}(b^*_j) \) is taken from (17). The two-tailed test of size \( \alpha \) is then given by comparing \( I_{1j}(b) \) with \( t_{1-\alpha/2, n-q} \) the \( (1 - \alpha) \) 100 percent point of the central Student's \( t \) distribution with \( n-q \) degrees of freedom.

Using the biased estimate of the variance of \( b^*_{1j} \), one can compute the Heuristic Statistic

\[
H_{1j}(b) = \frac{\hat{b}_{1j}}{\hat{v}_{b_{1j}}(b^*_j)},
\]

where \( \hat{v}_{b_{1j}}(b^*_j) \) is taken from (18). This, of course, cannot be compared with the Student's \( t \) distribution.

(b) Test of hypotheses of secondary parameters, viz., \( H_0 : \theta = C\beta^*U \).

Consider any secondary parameter \( \theta = C\beta^*U \), where \( C \) and \( U \) are known matrices of constants. \( r \times q \) \( p \times s \)

The ridge estimate of \( \beta^*_k \) is then given by

\[
\hat{\theta}_k = C\hat{\beta}_k^*,
\]

The unbiased estimate of the variance-covariance matrix of \( \hat{\theta}_k \) is given by

\[
v(\hat{\theta}_k) = (C[VIM_k]\)' \otimes [U'S^*U],
\]

where \( S^* \) and \( VIM_k \) are given by (11) and (16) respectively. Therefore,

\[
v(\hat{\theta}_{k1j}) = [C[VIM_k]'C']_{1j} [U'S^*U]_{jj}, i = 1, 2, \ldots, r,
\]

\[
j = 1, 2, \ldots, q.
\]

Thus, \( \hat{\theta}_k = (\hat{\theta}_{k1j})_{i=1, \ldots, r} \) \( j=1, \ldots, q \)

for testing \( H_0 : \theta_{1j} = 0 \) where \( \theta = (\theta_{1j})_{i=1, \ldots, r} \) \( j=1, \ldots, q \)

is then given by

\[
T_{1j}(\theta) = \frac{\hat{\theta}_{1j}}{v(\hat{\theta}_{1j})}
\]

231
which can be compared with the $1 - \alpha$ 100 percent point of the central student’s $t$ distribution with $n-q$ degrees of freedom, for a two-sided test of size $\alpha$. Similarly, the biased estimate of $V(\theta_k)$ is given by

\[ v_B(\hat{\theta}_k) = [C(VM_k)C']_1 [U'SP_k]_{11}. \]  

where $S_k$ is given by (14), so that

\[ v_B(\hat{\theta}_{ij}) = [C(VM_k)C']_{11}[U'SP_k]_{11}. \]  

Hence the corresponding Heuristic statistic is

\[ H_{ij}(\alpha) = \frac{\hat{q}^k_{ij}}{v_B(\hat{\theta}_{ij})}, \]  

which, likewise, cannot be compared with the student's $t$ statistic for significance testing purposes.

THE SAS MACRO GLMR

The GLMR macro is designed to perform either OLS or Ridge analysis or both at the user's options. The minimum core requirement for the macro is 220k. In the macro the validity of performing the ridge analysis is tested as follows:

(a) The statistic

\[ H = R^T \hat{\delta}_n(1 - |CORX|), \]

where

\[ R = \begin{bmatrix} [n - 1 - \frac{1}{6}(2q + 5)] \end{bmatrix} \]  

is computed, where $n$ is the total number of observations and $q$ is the number of predictor variables as defined in the beginning. The statistic $H$ is approximately distributed as $\chi^2$ with

\[ v = \frac{1}{2}n(n - 1), \]

degrees of freedom (see Bartlett, 1950, 1954, and Haladcky, 1969). As (CORX) approaches singularity, $\hat{\delta}_n(1 - |CORX|)$ approaches zero, so that a small value of $\chi^2$ indicates the existence of multicollinearity and its severity can be measured by the level of significance at which the null hypothesis $|CORX| = 0$ is accepted. Thus $P(H \geq \chi^2) = \alpha$, and so higher the $\alpha$, the greater is the severity of multicollinearity. The macro has the default value of $\delta = 1 - \alpha$ as .10. Even if $\delta < .10$, $\hat{\delta}$ is computed for sake of comparison.

(b) The smallest eigen value, $\lambda$ of (CORX) is computed and then the value of

\[ J = \sum_{q} \frac{1}{\lambda^q}, \]

is found. The macro has the default upper limit of $J$ as $10^4$. (i.e., if $J$ is less than or equal to $10^4$, ridge analysis is attempted, otherwise not). Generally, this upper limit of $J$ can be any value between $10^{-7}$ to $10^{-4}$. (This test is derived as a natural consequence to Marquardt's (1970, p. 596) discussions on computing effective rank of (CORX) when using the generalized inverse method."

If either (a) or (b) holds, then the macro proceeds to perform ridge analysis if the user uses the default option. If the ridge analysis is done the $\hat{\delta}$ and $\hat{\beta}$ are obtained using the formula given in (18) and (19) respectively. If (a) or (b) holds, then it is not feasible to rely on an unbiased estimate $\hat{\delta}$ of $\delta$. The various other quantities are also computed according to the discussions and formulas mentioned in the preceding section.

(A) When OLS analysis is performed, the following are printed as default or with particular options:

(i) OLS estimates $\hat{\beta}$.

(ii) $\hat{\beta}$, the ridge estimate for $\hat{\beta}$, when $\hat{\beta}$ starts getting stable, and the value of $\delta$.

(iii) $\hat{\beta}$, the transformed ridge estimate of $\hat{\beta}$.

(iv) The unbiased estimate of the standard error of $\hat{\beta}$ (if it is feasible) along with the student’s $t$ statistics and statistics of 'heuristic interest' based on biased estimate of the standard error of $\hat{\beta}$.

(v) Unbiased variance-covariance matrix of $\hat{\beta}$ (if it is feasible) along with the biased variance-covariance matrix of $\hat{\beta}$.

(vi) The VIF’s and maximum VIF.

(vii) Values of $k^*$ and corresponding values of $\hat{\delta}$ and maximum VIF.

(viii) Tests of hypotheses of the type $H_0$: $\delta = \chi^2 = 0$ at the user's option if $C$ and $U$ are provided.

(B) When Ridge analysis is performed the following are printed as default or with particular options:

(i) OLS estimates $\hat{\beta}$ (Correlation form).

(ii) $\hat{\beta}$, the ridge estimate for $\hat{\beta}$, when $\hat{\beta}$ starts getting stable, and the value of $\delta$.

(iii) $\hat{\beta}$, the transformed ridge estimate of $\hat{\beta}$.

(iv) The unbiased estimate of the standard error of $\hat{\beta}$ (if it is feasible) along with the student’s $t$ statistics and statistics of 'heuristic interest' based on biased estimate of the standard error of $\hat{\beta}$.

(v) Unbiased variance-covariance matrix of $\hat{\beta}$ (if it is feasible) along with the biased variance-covariance matrix of $\hat{\beta}$.

(vi) The VIF's and maximum VIF.

(vii) Values of $k^*$ and corresponding values of $\hat{\delta}$ and maximum VIF.

(viii) Tests of hypotheses of the type $H_0$: $\delta = \chi^2 = 0$ at the user's option based on unbiased estimate of $\hat{\delta}$, and also print the statistics of 'heuristic interests' based on the biased estimate of $\hat{\delta}$.

(ix) It plots the Ridge Trace, $k^*$ vs. maximum VIF, and $k^*$ vs. SSE (or maximum eigen value of $SSR$ for $p > 1$) for each dependent variable (if $p \geq 1$), at the user's option.

When asked to use both estimation methods, the macro can do everything listed at (A) and (B) except testing of hypotheses. This is done using the ridge estimates of the coefficients only.

Copies of the macro and examples of use are available from the authors.
REFERENCES


