The Korean Communications
in Statistics Vol. 5, No. 1, 1998
pp. 19-24

Scree Diagram for Detecting Multicollinearity and Estimating Ridge Constant in Linear Regression Model

Dae-Heung Jang[1]

Abstract

When multicollinearity appears in linear regression model, we can use ridge regression for stabilizing the regression coefficient estimates. We propose the scree diagram as a graphical method for detecting multicollinearity and estimating ridge constant in linear regression model.

1. Introduction

Linear regression model can be expressed in matrix notation as

$$\underset{\sim}{y} = X\underset{\sim}{\beta} + \underset{\sim}{\varepsilon} \tag{1}$$

where $\underset{\sim}{y} = (y_1, y_2, \cdots, y_n)'$ is the vector of observed response, X is the $n\times p(p\leq n)$ model matrix, $\underset{\sim}{\beta}$ is the the $p\times 1$ vector of parameters which appear in the chosen model, and $\underset{\sim}{\varepsilon} = (\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n)'$ is the vector of random errors associated with $\underset{\sim}{y}$. Here, p is the number of parameters in the model.

The vector of unknown parameters is estimated using ordinary least squares methods by

$$\underset{\sim}{b} = (X'X)^{-1}X'\underset{\sim}{y}. \tag{2}$$

The variance-covariance matrix of the estimated coefficients under the assumption that $\underset{\sim}{\varepsilon} \sim (\underset{\sim}{0}, \sigma^2 I)$ is

$$Var(\underset{\sim}{b}) = \sigma^2(X'X)^{-1}. \tag{3}$$

[1] Professor, Department of Applied Mathematics, Pukyong National University, Pusan, 608-737, Korea.

Multicollinearity is a condition among the set of p regressor variables in the model. If there exists an approximate linear dependence between the columns of X, then we have the condition usually identified as multicollinearity. When multicollinearity is present, $(X'X)^{-1}$ exists but is ill-conditioned. Therefore, the presence of unstable regression coefficient estimates arises. When multicollinearity appears in linear regression model, we can use ridge regression as a means for stabilizing the regression coefficient estimates in the fitted model.

The purpose of this paper is to suggest the scree diagram as a graphical method for detecting multicollinearity and estimating ridge constant. In Section 2, we propose the scree diagram as a graphical method for detecting multicollinearity and estimating ridge constant. In Section 3, we give a numerical example. In Section 4, we draw conclusion.

2. Scree diagram for detecting multicollinearity and estimating ridge constant

Various techniques for detecting and repairing multicollinearity have been proposed. A prevailing technique for repairing multicollinearity is ridge regression although there are some criticisms(For example, see Draper and Smith (1981).). We can use ridge regression as a mean for stabilizing the regression coefficient estimates in the fitted model. Hoerl and Kennard (1970 a, b) and, Marquardt (1970) have suggested problems associated with a ridge regression estimator. Until quite recently, there are being many researches for ridge regression. The ridge regression estimators for the parameters in the first order and in the second order polynomial models are calculated using the formula

$$\underline{b}(k) = (X'X + kI)^{-1}X'\underline{y} \tag{4}$$

where k is a constant and usually $0 < k < 1$.

From (4), the variance-covariance matrix of ridge regression estimator $\underline{b}(k)$, is

$$Var[\underline{b}(k)] = \sigma^2(X'X + kI)^{-1}X'X(X'X + kI)^{-1} \tag{5}$$

Let

$$V_k = \frac{Var[\underline{b}(k)]}{\sigma^2} = (X'X + kI)^{-1}X'X(X'X + kI)^{-1} \tag{6}$$

Then, from (6), we can execute the following spectral decomposition of V_k ;

$$V_k = P\Lambda P' \tag{7}$$

where Λ is the diagonal matrix of eigenvalues of V_k and P is the matrix of eigenvectors of V_k.

There are several techniques to estimate k. First, Hoerl and Kennard (1970b), Hoerl, Kennard and Baldwin (1975), McDonald and Galarneau (1975), Hocking, Speed and Lynn (1976), Lawless and Wang (1976), Wahba, Golub and Heath (1979), and Myers (1986) proposed stochastic methods. These stochastic methods have all exploited response data. Therefore, k is a random variable. Second, Tripp (1983, See Myers (1986).), St. John (1984), and Jang and Yoon (1997) proposed nonstochastic methods. Jang and Yoon (1997) proposed the prediction trace and the prediction variance trace as graphical tools for evaluating ridge regression estimator in mixture experiments. These nonstochastic methods do not have exploited response data. Therefore, k is not a random variable.

Using scree diagram (See Krzanowski (1988).), we propose a graphical method for detecting multicollinearity and estimating ridge constant in linear regression model. Scree diagram is the plot of ith eigenvalue λ_i of V_k against $i(i=1,2,\cdots,p)$. Through scree diagram with several ridge constants, we can detect multicollinearity and decide the value of k. If the condition number of $(X'X)^{-1}$ is very large, we suspect severe multicollinearity. We can see approximately the degree of multicollinearity through secree diagram when k is zero. Condition number (See Cornell(1990).) is the ratio of the largest to the smallest eigenvalue, $\frac{\lambda_1}{\lambda_p}$. As k is increasing sequentially, eigenvalues are decreasing gradually. When this decreasing trend becomes weakened, we can decide the value of k. This method is a method for choosing k as a function of only the regressor data. Therefore, the choice of k is determined by the nature of the multicollinearity itself and k is not a random variable.

3. Numerical example

In mixture experiments, the measured response is assumed to depend only on the relative proportions of the components present in the mixture. For mixture experiments, if we let x_i represent the proportion of the ith component in the mixture where the number of components is q, then

$$\sum_{i=1}^{q} = 1,$$

where $0 \le x_i \le 1, i=1,2,\cdots,q$.

The experimental region is a regular $(q-1)$-dimensional simplex. When additional constraints are imposed on the proportions in the form of lower and upper bounds

$$0 < L_i \le x_i \le U_i < 1, \quad \text{where } i = 1, 2, \cdots, q, \tag{8}$$

the experimental region becomes a subregion of the simplex.

Typically, mixture models are of the Scheffé type where the first order model is

$$y = \sum_{i=1}^{q} \beta_i x_i + \varepsilon$$

and the second order model is

$$y = \sum_{i=1}^{q} \beta_i x_i + \sum\sum_{i<j}^{q} \beta_{ij} x_i x_j + \varepsilon$$

where y is observed response and ε is random error. When the component proportions in mixture experiments are restricted by lower and upper bounds, multicollinearity appears all too frequently, Thus, we can use ridge regression.

Our example is taken from McLean and Anderson (1966). The purpose of the experiment was to find the combination of the proportions of magnesium (x_1), sodium nitrate (x_2), strontium nitrate (x_3), and binder (x_4) for producing flare with maximum illumination. McLean and Anderson (1966) suggested the 15-point extreme vertices design consisting of the eight extreme vertices, the centroids of the six faces and the overall centroid of the region along with the flare illumination data. A second order polynomial was fit to the data. The component ranges are $0.40 \le x_1 \le 0.60$, $0.10 \le x_2 \le 0.50$, $0.10 \le x_3 \le 0.50$, and $0.03 \le x_4 \le 0.08$. We can use ridge regression because of multicollinearity in this example. St. John (1984) showed that $k = 0.005$ is appropriate by means of the VIF's and the ridge trace. Also, we can calculate $k = 0.004$ by the method proposed by Hoerl, Kennard, and Baldwin (1975).

Figure 1 shows the scree diagram when $k = 0.000$. From Figure 1, we can find that because the condition number is very large, severe multicollinearity arises in extreme vertices design. Figure 2 shows the scree diagram with several ks. From Figure 2, we can ascertain that as the value of k is changed from zero to 0.001, eigenvalues are decreasing dramatically, and that as the value of k is increasing sequentially from 0.001 to 0.007, eigenvalues are decreasing gradually, but that when k is greater than 0.005, this decreasing trend becomes weakened. Table 1 shows the difference of several sequential λ_1s. Therefore, we can conclude that $k = 0.005 \sim 0.006$ is appropriate. As other examples, in case of 15-point D-optimal design (Cornell(1990)) and 18-point Snee's design (Snee(1975)), using the scree diagram, we have the similar results as the above 15-point extreme vertices design, namely, we can conclude that

$k = 0.005 \sim 0.006$ is appropriate for D-optimal design and $k = 0.004 \sim 0.005$ is appropriate for Snee's design.

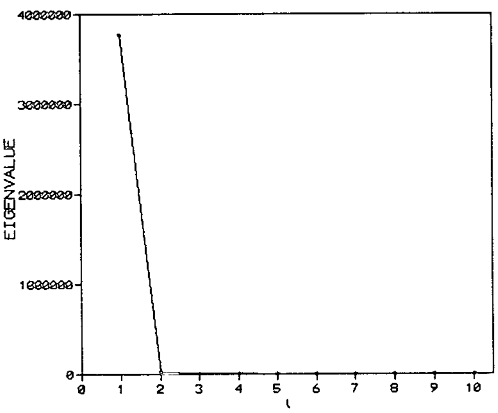

FIGURE 1. SCREE DIAGRAM IN EXTREME VERTICES DESIGN (K=0.000)

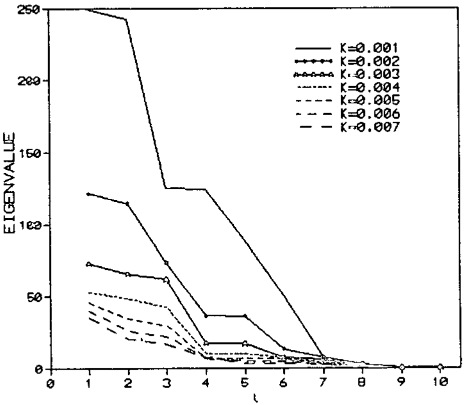

FIGURE 2. SCREE DIAGRAM IN EXTREME VERTICES DESIGN WITH SEVERAL RIDGE CONSTANTS

Table 1. difference of several sequential λ_1s

k	difference
0.000 vs. 0.001	3770010.60
0.001 vs. 0.002	127.91
0.002 vs. 0.003	48.69
0.003 vs. 0.004	19.91
0.004 vs. 0.005	7.21
0.005 vs. 0.006	5.83
0.006 vs. 0.007	4.77

4. Conclusion

In this thesis, as a graphical method for detecting multicallinearity and estimating ridge constant, the scree diagram have been proposed. The advantages of the scree diagram can be stated as follows;

(1) The scree diagram can be used as a tool for detecting multicollinearity and estimating ridge constant k.

(2) The scree diagram is nonstochastic methods.

(3) Through the scree diagram, we can evaluate the effect of ridge regression estimator.

References

[1] Cornell, J. A. (1990). *Experiments with Mixture: Designs, Models, and the Analysis of Mixture Data,* 2nd ed., New York : John Wiley & Sons, Inc.

[2] Draper, N. R. and Smith, H. (1981). *Applied Regression Analysis,* 2nd ed., New York : John Wiley & Sons, Inc., 319-325.

[3] Hocking, R. R., Speed, F. M., and Lynn, M. J. (1976). A Class of Biased Estimators in Linear Regression, *Technometrics*, 18, 425-437.

[4] Hoerl, A. E. and Kennard, R. V. (1970a). Ridge Regression : Biased Estimation for Nonorthogonal Problems, *Technometrics*, 12, 55-67.

[5] Hoerl, A. E. and Kennard, R. V. (1970b). Ridge Regression : Applications to Nonorthogonal Problems, *Technometrics*, 12, 69-82.

[6] Hoerl, A. E., Kennard, R. W., and Baldwin, K. F. (1975). Ridge Regression : Some Simulation, *Communications in Statistics - Theory and Methods*, 4, 105-123.

[7] Jang, Dae-Heung and Yoon, Min (1997). Graphical Methods for Evaluating Ridge Regression Estimator in Mixture Experiments, *Communications in Statistics - Simulation and Computation*, 26, 1049-1061.

[8] Krzanowski, W. J. (1988). *Principles of Multivariate Analysis,* Oxford University, Oxford, Press.

[9] Lawless, J. F. and Wang, P. (1976). A Simulation Study of Ridge and Other Regression Estimators, *Communications in Statistics - Theory and Methods*, 5, 307-323.

[10] Marquardt, D. W. (1970). Generalized Inverse, Ridge Regression, Biased Linear Estimation, and Nonlinear Estimation, *Technometrics*, 12, 591-612.

[11] McDonald, G. C. and Galarneau, D. I. (1975). A Monte Carlo Evaluation of Some Ridge-type Estimators, *Journal of American Statistical Association*, 70, 407-416.

[12] McLean, R. A. and Anderson, V. L.(1966). Extreme Vertices Design of Mixture Experiments, *Technometrics*, 8, 447-454.

[13] Myers, R. H. (1986). *Classical and Modern Regression with Applications*, Duxbury, Boston, Press.

[14] Snee, R. D. (1975). Experimental Designs for Quadratic Models in Constrained Mixture Spaces, *Technometrics*, 17, 149-159.

[15] St. John, R. C. (1984). Experiments with Mixtures, Ill-Conditioning, and Ridge Regression, *Journal of Quality Technology*, 16, 81-96.

[16] Tripp, R. E. (1983). Non-stochastic Ridge Regression and Effective Rank of the Regressors Matrix, Unpublished doctoral dissertation.

[17] Vining, G. G., Cornell, I. A., and Myers, R. H. (1993). A Graphical Approach for Evaluating Mixture Designs, *Applied Statistics*, 42, 127-138.

[18] Wahba, G., Golub, G. H. and Heath, C. G. (1979). Generalized Cross Validation as a Method for Choosing a Good Ridge Parameter, *Technometrics*, 21, 215-223.