International Journal of Applied Mathematics
& Statistical Sciences (IJAMSS)
ISSN(P): 2319- 3972; ISSN(E): 2319-3980
Vol. 4, Issue 6, Oct -Nov 2015, 25-34

International Academy of Science,
Engineering and Technology
Connecting Researchers; Nurturing Innovations

BOOTSTRAPPING IN DETERMINATION OF RIDGE PARAMETER

OKEKE EVELYN NKIRUKA & OKEKE JOSEPH UCHENNA

Department of Mathematics and Statistics, Federal University Wukari, Nigeria

ABSTRACT

This article studied the application of ridge regression on multicollinear data whose ridge parameter was determined using bootstrap samples. Mean squared error of the samples and some arbitrary values were used to determine the ridge parameter that will give the minimum residual. The result of the study revealed that both the mean squared error and the smallest eigenvalue of the predictor variables of the original data play vital role in determining the ridge parameter of ridge regression.

KEYWORDS: Ridge Parameter, Least Squares, Bootstrapping, Multicollinearity, Penalized Residual Sum of Squares, Mean Squared Error and Smallest Eigenvalue

1.0 INTRODUCTION

The observations in many experiments of physical and medical sciences are often ill-conditioned. Their distribution can be highly skewed, they can have tail thicker than that of normal distribution, and random samples often have outliers and correlated variables. Outliers, correlated variables and heavy-tailed distribution are serious problems because they inflate the standard error of the estimates, causing them to have relatively low power. In regression analysis, when the predictor variables $X=(X_1, \dots, X_p)$ have $X'X$ that is of full rank least squares estimators are usually unbiased, The Gauss-Markov property assures us that the LS estimator has minimum variance in the class of unbiased linear estimators. Ordinary least squares regression is usually affected by outliers, multicollinearity, as well as skewed or heavy tailed distributions. When the method of least squares is applied to nonothogonal variables, very poor estimates of the regression coefficients are usually obtained. The variance of the estimates of the regression coefficients may be considerably inflated, and the length of the vector of coefficients too long on the average and very unstable. One of the common techniques to overcome the difficulty of least squares when the data is ill-conditioned is to drop the basic assumption of regression analysis that the estimators of regression coefficients be unbiased. Ridge regression is one of the biased estimators of regression coefficients that is applied to data whose predictor variables have $X'X$ matrix that is ill-conditioned (near singular) or even singular (has zero eigenvalues). In this article ridge regression will be applied to multicollinear real data whose ridge parameter will be determined using bootstrap samples. j

2.0 RIDGE REGRESSION

A method of regression analysis that is effective in the presence of multicollinearity was proposed by Hoerl and Kennard (1970) and is called ridged regression. Assuming that the data (X and Y) have been standardized, they suggested that some constant values will be added to diagonal elements of the $X'X$ matrix of the predictor variables to have regression coefficient that can be estimated from the modified normal equations, hereafter called ridged equation

$$(X'X + kI)b(k) = X'y \qquad 1$$

from which the regression coefficients can be estimated.

Making $b(k)$ the subject of the formula in (1) gives

$$b(k) = (X'X + kI)^{-1}X'y \quad 2$$

with $k \geq 0$, the nonstochastic quantity, being the ridge (or control) parameter. Of course $b(0) = b$ is the OLS estimate. $b(k) = [b_1(k), \dots, b_p(k)]'$ contain estimates of the parameters in the non-intercept part of the model.

Multiplying both sides of (2) by $(X'X)^{-1}$ we have that

$$b(k) = (I + k(X'X)^{-1})\hat{\beta} \quad 3$$

where $\hat{\beta}$ denotes the LS estimate in standard form. The equation (3) shows that the ridge estimator is biased and the amount of bias depends on the ridge (or the control) parameter k.

Using the abbreviation

$$G_k = (X'X + kI)^{-1} \quad 4$$

$E(b(k)), \text{Bias}(b(k), \beta)$ and $V(b(k))$ can be expressed as follows

$E(b(k))$

$$E(b(k)) = E((X'X + kI)^{-1}X'y)\backslash$$

$$= (X'X + kI)^{-1}X'E(y)$$

$$= (X'X + kI)^{-1}X'X\beta$$

$$= G_kX'X\beta \quad 5$$

$\text{Bias}(b(k), \beta)$

$$\text{Bias}(b(k), \beta) = E(b(k) - \beta)$$

$$= G_kX'X\beta - \beta$$

$$= (X'X + kI)^{-1}X'X\beta - \beta$$

$$= \frac{X'X\beta}{(X'X+kI)} - \beta$$

$$= \frac{X'X\beta - \beta(X'X+kI)}{(X'X+kI)}$$

$$= -k\beta G_k \quad 6$$

$V(b(k))$

$$V(b(k)) = \text{var}((X'X + kI)^{-1}X'y)$$

$$= \text{var}(G_kX'y)$$

$$= G_kX'\text{var}(y)XG_k$$

$$= \sigma^2 G_k X'X G_k \quad (7)$$

Hence the Mean Dispersion Error (MDE) matrix is

$$M(b(k), \beta) = E(b(k) - \beta)(b(k) - \beta)'$$

$$= E\left[\left(b(k) - E(b(k))\right) + \left(E(b(k)) - \beta\right)\right]\left[\left(b(k) - E(b(k))\right) + \left(E(b(k)) - \beta\right)\right]'$$

$$= E\Big[\{b(k) - E(b(k))\}\{b(k) - E(b(k))\}' + \{b(k) - E(b(k))\}\{E(b(k)) - \beta\}' + \{E(b(k)) - \beta\}\{b(k) - E(b(k))\}' + \{E(b(k)) - \beta\}\{E(b(k)) - \beta\}'\Big]$$

$$= V(b(k)) + Bias(b(k), \beta)Bias(b(k), \beta)'$$

$$= \sigma^2 G_k X'X G_k + (-k\beta G_k)(-k\beta G_k)'$$

$$= G_k(\sigma^2 X'X + k^2 \beta\beta')G_k$$

$$= \frac{\sigma^2 X'X + k^2\beta\beta'}{(X'X+kI)^2}$$

From the spectral decomposition of the symmetric matrix $X'X$ we have that

$$X'X = P\Lambda P' = \Lambda$$

$${G_k}^{-1} = X'X + kI = P\Lambda P' + kPP' = (\Lambda + k)$$

Note that $PP' = I$

$$G_k = (\Lambda + k)^{-1}$$

Therefore

$$M(b(k), \beta) = \frac{\sigma^2\Lambda + k^2\beta\beta'}{(\Lambda+kI)^2} \quad (8)$$

$$trM(b(k), \beta) = \sum_{i=1}^{k} \frac{\sigma^2\lambda_i + k^2\beta^2}{(\lambda_i+k)^2} \quad (9)$$

The scalar MDE of $b(k)$ for fixed σ^2 and a fixed vector β is a function of ridge parameter k, which starts at $\sum_{i=1}^{k} \frac{\sigma^2}{\lambda_i} = tr(V(b))$ for $k = 0$, takes its minimum for $k = k_{opt}$ and then increases monotonically, provided that $k_{opt} < \infty$ (Rao and Toutenburg 1995)

We now transform $M(b, \beta) = M(b) = \sigma^2(X'X)^{-1}$ as follows

$$M(b) = \sigma^2 G_k\left({G_k}^{-1}(X'X)^{-1}{G_k}^{-1}\right)G_k$$

$$= \sigma^2 G_k[(X'X + kI)(X'X)^{-1}(X'X + kI)]G_k$$

$$= \sigma^2 G_k[(X'X)(X'X)^{-1}(X'X) + (X'X)(X'X)^{-1}kI + kI(X'X)^{-1}(X'X) + kI(X'X)^{-1}kI]G_k$$

$$= \sigma^2 G_k\left((X'X) + k^2(X'X)^{-1} + 2kI\right)G_k \quad (10)$$

Definition 1: Let $\hat{\beta}_1$ and $\hat{\beta}_2$ be two estimators of β. Then $\hat{\beta}_2$ is called MDE-superior to $\hat{\beta}_1$ (or $\hat{\beta}_2$ is called MDE-improvement to $\hat{\beta}_1$) if the difference of their MDE matrices is nonnegative definite, that is, if

$$\Delta(\hat{\beta}_1, \hat{\beta}_2) = M(\hat{\beta}_1, \beta) - M(\hat{\beta}_2, \beta) \geq 0$$

From definition 1 we obtain the interval $0 < k < k^*$ in which the ridge estimator is MDE-superior to the OLS $b = (X'X)^{-1}X'y$.

$$\Delta(b, b(k)) = M(b) - M(b(k), \beta)$$

$$= \sigma^2 G_k\big((X'X) + k^2(X'X)^{-1} + 2kI\big)G_k - G_k(\sigma^2 X'X + k^2\beta\beta')G_k$$

$$= \sigma^2 G_k(X'X)G_k + \sigma^2 G_k k^2(X'X)^{-1}G_k + \sigma^2 G_k 2kIG_k - G_k\sigma^2 X'XG_k + G_k k^2\beta\beta'G_k$$

$$= kG_k[\sigma^2(2I + k(X'X)^{-1}) - k\beta\beta']G_k \qquad 11$$

Since $G_k > 0$, we have that $\Delta(b, b(k)) \geq 0$ if and only if

$$\sigma^2(2I + k(X'X)^{-1}) - k\beta\beta' \geq 0$$

$$\sigma^2(2I + k(X'X)^{-1}) \geq k\beta\beta' \qquad 12$$

Dividing through by $k\beta\beta'$ gives

$$\frac{\sigma^2(2I + k(X'X)^{-1})}{k\beta\beta'} \geq 1$$

$$\frac{k\beta\beta'}{\sigma^2(2I + k(X'X)^{-1})} \leq 1$$

$$\sigma^{-2}k\beta'(2I + k(X'X)^{-1})^{-1}\beta \leq 1 \qquad 13$$

As a sufficient condition for (12), independent of the model matrix, we obtain

$$2\sigma^2 I - k\beta\beta' \geq 0$$

$$2\sigma^2 I \geq k\beta\beta'$$

$$k \leq \frac{2\sigma^2}{\beta\beta'} \qquad 14$$

The range of k, which ensures the MDE-1 superiority of $b(k)$ compared to b is dependent on $\sigma^{-1}\beta$ and hence unknown. If auxiliary information about the length (norm) of β is available in the form

$$\beta\beta' \leq r^2$$

then

$$k \leq \frac{2\sigma^2}{r^2} \qquad 15$$

is sufficient for (14) to be valid. Hence possible values for k, in which $b(k)$ is better than b, can be found by estimation of σ^2 or by specification of a lower limit or by a combined a priori estimation $\sigma^2\beta\beta' \leq \tilde{r}^2$

Swamy et al (1978) and Swamy and Mehta (1977) investigate the following problem

$$\min_{\beta}\{\sigma^{-2}(y-X\beta)'(y-X\beta)\beta'\beta \le r^2\}$$

The solution to this problem

$$\hat{\beta}(\mu) = (X'X + \sigma^2\mu I)^{-1}X'y \qquad 16$$

is once again a ridge estimate and $\hat{\beta}(\mu)'\hat{\beta}(\mu) = r^2$ is fulfilled. Replacing σ^2 by the estimate s^2 provided a practical solution for the estimator (16) but its properties can only be calculated approximately.

Hoerl and Kennard (1970) derived the ridge estimator by the following reasoning. Let $\hat{\beta}$ be any estimator and $b = (X'X)^{-1}X'y$ the OLS estimator. Then the error sum of squares estimated with $\hat{\beta}$ can be expressed, according to the property of optimality of b, as

$$S(\hat{\beta}) = (y - X\hat{\beta})'(y - X\hat{\beta}) \qquad 17$$

$$= \left((y - Xb) + X(b - \hat{\beta})\right)'\left((y - Xb) + X(b - \hat{\beta})\right)$$

$$= (y - Xb)'(y - Xb) + (b - \hat{\beta})'X'X(b - \hat{\beta}) + 2(y - Xb)'X(b - \hat{\beta})$$

$$= S(b) + \Phi(\hat{\beta}) \qquad 18$$

noting that the term

$$2(y - Xb)'X(b - \hat{\beta}) = 2(y - X(X'X)^{-1}X'y)'X(b - \hat{\beta})$$

$$= 2y(I - X(X'X)^{-1}X')X(b - \hat{\beta})$$

$$= 2MX(b - \hat{\beta}) = 0$$

since $MX = 0$

Let $\Phi_0 > 0$ be a fixed given value for the error sum of squares. Then a set $\{\hat{\beta}\}$ estimate exists that fulfill the condition $S(\hat{\beta}) = S(b) + \Phi_0$. In this set we look for the estimate with minimal length

$$\min_{\hat{\beta}}\left\{\hat{\beta}'\hat{\beta} + \frac{1}{k}\left[(b - \hat{\beta})'X'X(b - \hat{\beta}) - \Phi_0\right]\right\} \qquad 19$$

where $\frac{1}{k}$ is a Lagrangian multiplier. Differentiation of this function with respect to $\hat{\beta}$ and $\frac{1}{k}$ leads to the normal equations,

let

$$\min_{\hat{\beta}}\left\{\hat{\beta}'\hat{\beta} + \frac{1}{k}\left[(b - \hat{\beta})'X'X(b - \hat{\beta}) - \Phi_0\right]\right\} = L$$

$$\frac{\partial L}{\partial \hat{\beta}} = 2\hat{\beta} - 2\frac{1}{k}(b - \hat{\beta})X'X = 0$$

$$2\hat{\beta} + 2\frac{1}{k}X'X(\hat{\beta} - b) = 0$$

Since $2 \neq 0$

$$\hat{\beta} + \frac{1}{k}X'X(\hat{\beta} - b) = 0$$

$$\hat{\beta} + \frac{1}{k}X'X\hat{\beta} = \frac{1}{k}X'Xb$$

$$\hat{\beta}\left(\frac{1}{k}X'X + 1\right) = \frac{1}{k}X'Xb$$

$$\hat{\beta}\left(\frac{X'X+kI}{k}\right) = \frac{1}{k}X'Xb$$

$$\hat{\beta} = (X'X + kI)^{-1}X'Xb$$

$$= G_k X'Xb \qquad 20$$

$$\frac{\partial L}{\partial \hat{\beta}} = \left((b - \hat{\beta})'X'X(b - \hat{\beta}) - \Phi_0\right) = 0$$

$$(b - \hat{\beta})'X'X(b - \hat{\beta}) - \Phi_0 = 0$$

$$\Phi_0 = (b - \hat{\beta})'X'X(b - \hat{\beta}) \qquad 21$$

Hence the solution of the equation (19) is the ridge estimator $\hat{\beta} = b(k)$ in (20).

Minimizing a penalized residual sum of squares the ridge estimator $b(k)$ is

$$b(k) = \text{argmin}_\beta \left\{\left(y_i - \sum_{j=1}^{p} X_{ij}\beta_j\right)^2 + k\sum_{j=1}^{p}\beta_j^2\right\} \qquad 22$$

The ridge parameter k is to be determined iteratively so that (21) is fulfilled.

Consider $\hat{y}(k) = Xb(k)$ to be estimated y.

$$\hat{y}(k) = X(X'X + kI)^{-1}X'y$$

$$= X(X'X)^{-1}\{I + k(X'X)^{-1}\}^{-1}X'y \qquad 23$$

The sum of the squares of the deviation of the y's from their fitted values is

$$(y - \hat{y}(k))'(y - \hat{y}(k)) \qquad 24$$

2.1 BOOTSTRAP

Bootstrapping describes how sample data can be handled to obtain reliable standard error, confidence interval and other measures of uncertainty for a wide range of problems. The key idea is to resample from the original data either directly or indirectly or via a fitted model-to create replicate dataset, from which variability of the quantity of interest can be assessed without longwinded and error analytical calculations. This process involved repeating the original data analysis procedure with many replicate sets of data. The initial reaction was that resampling from the original data is a fraud. But in fact it is not (Davision and Hinley 1985). It turns out that a wide range of statistical problems can be tackled this way, liberating the investigator from the need to oversimplify complex problems. Bootstrap methods are intended to help avoid tedious calculations based on questionable assumptions. But they cannot replace clear critical thought about the problem, appropriate design of the investigation, data analysis and incisive presentation of the conclusions. The methods can be applied when there is a well-defined probability model for the data and when there is not. There are four optional

resampling schemes under bootstrap-classical bootstrap, smooth bootstrap, wild bootstrap, and residual-based (Bayesian) bootstrap Hall and Mammen (1994). In this article we wish to make use of classical bootstrap in forming samples from which different values of mean square error will be generated.

The classical bootstrap may be thought of as rather a device for constructing a new data sequence having the same size as the original sample. All the member of the new sequence are drawn from the original sample, and are present in proportions which are determined by a uniform multinomial distribution on the original values. Of course, the later distribution is a consequence of the "random sampling with replacement" concept that underlies the classical bootstrap algorithm. Under classical bootstrap $\{(X_i^*, y_i^*)\}_{1\leq i\leq n}$ is taken at random from the original sample $\{X_i, y_i\}_{1\leq i\leq n}$.This resampling method goes back to the pioneering work of Efron (1979)

2.1.1 Statistical Error

The basic idea of bootstrapping is to approximate the quantity q(f) -such as var(β|f) by the estimate$q(\hat{f})$, where $\hat{f}$ is either a parametric or a nonparametric estimate of f based on the data $\{X_i, y_i\}_{1\leq i\leq n}$. The statistical error is then the difference between $q(\hat{f})$, and q(f). Bootstrap methods wish to minimize this error as far as possible or remove it entirely.

3.0 NUMERICAL APPLICATION

3.1 Data and its Description

We applied ridge regression equation to a real data set, and five bootstrap samples generated from the real data set. Some arbitrary values of k were also considered to enable us determined how well the ridge parameter of the ridge regression can be chosen. The real life data we used were obtained from unpublished B.SC research project presented at the Department of Statistics, Nnamdi Azikiwe University, Awka, by Iteire (2004). The data were from Nigeria Stock Exchange and is on their transaction for the period of 1991-2007. The data was chosen because it is ill-conditioned. The predictor variables studied as affecting the response variable(market capitalization) includes-share volume index, share value index, daily average volume, daily average value, number of listed securities, all share index, and number of listed companies. Below is the correlation matrix of the predictor variables.

Correlations: shar vol, shar val, D Av vol, D.Av.val, N.lis sec., ...

	shar vol	shar val	D Av vol	D.Av.val	N.lis sec.	A sha ind.
shar val	0.995					
D Av vol	0.999	0.998				
D.Av.val	0.993	1.000	0.997			
N.lis sec	0.605	0.585	0.596	0.581		
A sha ind.	0.923	0.893	0.910	0.883	0.636	
N.lis.comp	0.493	0.434	0.469	0.421	0.707	0.700

Observe that the presence of multicollinearity is highly pronounced in the correlation matrix above with two predictor variables being perfectly correlated. Two of the eigenvalues of $X'X$ matrix are zero and it is the matrix smallest eigenvalue.

4.0 RESULT

Table 1: Residual Analysis of Real life and Bootstrap Samples

Samples	Mean Squared Error from Unstandardized Variables	Residual	Mean Squared Error from Standardized Variables	Residual
Real Data	103202	0.009051	0.9532.	60063335
Bootstrap Sample1	101	0.011234	0.9532	12738352.5
Bootstrap Sample 2	63867	0.009522	0.7990	87408183.4
Bootstrap Sample 3	576	3.412094	0.2228	$1.26x10^{8}$
Bootstrap Sample 4	2154	0.988904	1.557	$2.24x10^{9}$
Bootstrap Sample 5	135	0.011232	0.2469	$1.27x17^{8}$

The sum of the squared of deviation of the y's from their fitted values stated in (24) were calculated using different value of k obtained from (15). The variances were obtained from the estimated mean squared error of the real life and bootstrap data, before and after the data (X and Y) have been standardized. The result of analysis is presented in Table 1 above. Table 1 shows that the biasing parameter is better estimated from the mean squared error of the original data as compared with different values obtained from the bootstrap samples when the data are not standardized. There is no definite order of choosing the appropriate values of k as one can observe from Table 1.

Table 2: Residual Analysis from Arbitrary Values of k

K Value	Residual
0	0.00832
0.00001	0.00832
0.006	0.01114
0.05	0.01173
1.0	0.52339

Table 2 shows, for some selected values of k, values of the residual sum of squares (17). In section 3.1, we made mention that the smallest eigenvalue of the $X'X$ is zero, then observing what we have in Table 2, one can see that as the value of the biased parameter k approaches zero (15) converges. Values of k that deviates much from the value of the smallest eigenvalue have large values of residual.

5.0 CONCLUSIONS

Comparing the residual of Tables 1 and 2, we can state that the biasing parameter of ridge regression may better be determined through the use of $X'X$ matrix. It is also observed that the smallest residual in Table 2 is very close to that of

Table 1. This is to say that both the mean squared error and the smallest eigenvalue of the predictor variables of the original data play vital role in determining the biased parameter of ridge regression.

REFERENCES

1. Davision, A. C, and Hinkley, D.V (1985). Bootstrap method and their Applications, Cambridge University Press, USA
2. Efron, B. (1979). Bootstrap methods: Another look at the Jackknife. Annals of Statistics, 7, 1-26.
3. Hall, P. and Mammen, E. (1994). On general resampling algorithms and performance in distribution estimation. Annals of Statistics 22,4, 2011-2031.
4. Hoerl, A. E. and Kennard, R.W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12, 59-84.
5. Iteire, I. A. (2004). Transaction of Nigeria Stock Exchange from 1991-2007. B.SC Project, Nnamdi Azikiwe University, Awka, Nigeria, 33-35.
6. Rao, C. R., and Toutenburg, H. (1995). Linear models: Least squares and alternatives. Springer, New York, 63-67.
7. Swamy, P. A. V. B. and Mehta, J. S. (1977). A note on minimum average risk estimators for coefficients in linear models. Comm. Statist. A, 6, 1181-1186.
8. Swamy, P. A. V. B., Mehta, J. S., and Rappoport, P. N. (1978). Two methods of evaluating Hoerl and Kennard's ridge regression. Comm. Statist. A, 12, 1133-1155.