Journal of Modern Applied Statistical Methods

Volume 9 | Issue 2 Article 16

11-1-2010

A Comparison between Unbiased Ridge and Least Squares Regression Methods Using Simulation Technique

Mowafaq M. Al-Kassab
Al-al Bayt University, Mafraq, Jordan, omar_qwaider_81@yahoo.com

Omar Q. Qwaider
Al-al Bayt University, Mafraq, Jordan, mowafaq2002@yahoo.co.uk

Recommended Citation
Al-Kassab, Mowafaq M. and Qwaider, Omar Q. (2010) "A Comparison between Unbiased Ridge and Least Squares Regression Methods Using Simulation Technique," *Journal of Modern Applied Statistical Methods*: Vol. 9 : Iss. 2 , Article 16.
DOI: 10.22237/jmasm/1288584900

Journal of Modern Applied Statistical Methods
November 2010, Vol. 9, No. 2, 488-494

1538 – 9472/10/$95.00

A Comparison between Unbiased Ridge and Least Squares Regression Methods Using Simulation Technique

Mowafaq M. Al-Kassab Omar Q. Qwaider
Al-al Bayt University,
Mafraq, Jordan

The parameters of the multiple linear regression are estimated using least squares ($\hat{B}_{LS}$) and unbiased ridge regression methods ($\hat{B}(KI,J)$). Data was created for fourteen independent variables with four different values of correlation between these variables using Monte Carlo techniques. The above methods were compared using the mean squares error criterion. Results show that the unbiased ridge method is preferable to the least squares method.

Key words: Least squares, prior information, unbiased ridge estimation, mean squares error.

Introduction

Consider the linear regression model:

$$Y^* = X^* B + U \qquad (1.1)$$

where X^* is a $(n\times(p+1))$ matrix of predictor variables of full rank, Y^* is a $(n\times 1)$ response vector, B is a $((p+1)\times 1)$ vector of parameters and U is a $(n\times 1)$ vector of errors with $E(U)=0$ and $Cov(U)=\sigma^2 I$. When multicollinearity exists, the least squares estimate $\hat{\mathrm{B}}_{\mathrm{LS}}=\left(\mathrm{X}^{*\mathrm{T}}\mathrm{X}^{*}\right)^{-1}\mathrm{X}^{*\mathrm{T}}\mathrm{Y}^{*}$ is unstable, and many different methods have been proposed to control multicollinearity (Hoerl & Kennard, 1970).

M. M. T. Al-Kassab is a Professor of Mathematical Statistics and Deputy Dean of the College of Science. His research interests include optimum stratum boundaries and biased and unbiased methods in regression. Email: mowafaq2002@yahoo.co.uk. Omar Q. Qwaider has his M.Sc. in statistics. His research interests are in regression estimation methods. Email: omar_qwaider_81@yahoo.com.

An alternative to the linear regression method is the unbiased ridge estimate

$$\hat{B}(KI,J)=\left(X^{*T}X^{*}+KI_P\right)^{-1}\left(X^{*T}Y^{*}+KJ\right)$$

where

$$J=\frac{\sum_{i=1}^{P}\hat{B}_{ils}}{P}$$

and

$$\hat{K}=\frac{P\sigma^2}{\left(\hat{B}-J\right)^T\left(\hat{B}-J\right)-\sigma^2 tr\left(X^TX\right)^{-1}}.$$

The unbiased ridge estimate regression, $\hat{B}(KI,J)$, has advantages and disadvantages. It is effective in practice but it is a complicated function of K, thus it is necessary to use rather complicated equations when employing some popular methods such as the Crouse, Jin and Hanumare (1995) criterion to select K (Swindel, 1976).

The General Multiple Linear Regression Model

The general multiple linear regression model is

$$Y_i^* = B_0 + B_1 X_{i1}^* + B_2 X_{i2}^* + ... + B_p X_{ip}^* + U_i \quad i = 1,2,...,n \qquad (2.1)$$

where $B_0, B_1, B_2, ..., B_P$ are the regression coefficients and $U_i \sim N(0,\sigma^2)$ is the random error associated with the observations. In matrix notation model (2-1) can be written as

$$\begin{bmatrix} y_1^* \\ y_2^* \\ \vdots \\ y_n^* \end{bmatrix} = \begin{bmatrix} 1 & x_{11}^* & x_{12}^* & \dots & x_{1p}^* \\ 1 & x_{21}^* & x_{22}^* & \dots & x_{2p}^* \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1}^* & x_{n2}^* & \dots & x_{np}^* \end{bmatrix} \begin{bmatrix} B_0 \\ B_1 \\ \vdots \\ B_p \end{bmatrix} + \begin{bmatrix} U_1 \\ U_2 \\ \vdots \\ U_n \end{bmatrix}$$

$$Y^* = X^* B + U,$$

where Y^* is a $(n \times 1)$ column vector of observations on the dependent variable, X^* is a $((p+1)\times 1)$ matrix resulting from n observations on P explanatory variables $X_1^*, X_2^*, ..., X_p^*$ where the first column of 1's represent the intercept term, that is, $X_0^* = 1$, and $U \sim N(0,\sigma^2)$ is $(n \times 1)$ column vector of errors.

Assumptions of the standardized model are:

1. $E(U) = 0$
2. $Var(U) = E(UU^T) = \sigma^2 I$
3. Rank (X^*) = P where $p < n$

The ordinary least squares estimators are given by $\hat{B}_{LS} = \left(X^{*T}X^*\right)^{-1} X^{*T} Y^*$.

Properties of Ordinary Least Squares Estimators

1. Unbiasedness:
   An estimator, $\widehat{B}$, is said to be unbiased estimator of B if the expected value of $\widehat{B}$ equals B, that is, $E\left(\hat{B}_{LS}\right) = B$. (Casella & Berger, 2002)
2. Variance:
$$Var\left(\hat{B}_{LS}\right) = \sigma^2 \left(X^{*T} X^*\right)^{-1}$$
3. Mean squared error:

$$MSE\left(\hat{B}\right) = \sum_{i=1}^{P} Var\left(\hat{B}_i\right) + \sum_{i=1}^{P} \left(Bias(\hat{B}_i)\right)^2$$

$$\Rightarrow MSE\left(\hat{B}_{LS}\right) = \sum_{i=1}^{P} Var\left(\hat{B}_i\right) = \sigma^2 tr\left(X^T X\right)^{-1}$$

$$MSE\left(\hat{B}_{LS}\right) = \sigma^2 \sum_{i=1}^{P} \frac{1}{\ell_i} \qquad (2.2)$$

Unbiased Ridge Estimator

Ridge regression, which was proposed by Horel and Kennard (1970), suggests the use of $X^T X + K$, where K is a diagonal matrix rather than $X^T X$, so that the resulting estimators of B are known as the ridge regression estimators and are given by:

$$\hat{B} = (X^T X + K)^{-1} X^T Y \qquad (3.1)$$

Horel and Kennard (1970) suggested two forms for K. First, if $K = kI_p$, $0 < k < 1$. Substituting this in equation (3.1), results in

$$\hat{B}(k) = (X^T X + kI_p)^{-1} X^T Y \qquad (3.2)$$

and, using eigenvalues and eigenvectors, $\hat{B}(k)$ can be expressed as

$$\hat{B}(k) = \sum_{j=1}^{p} (\ell_j + k)^{-1} V_j V_j^T X^T Y. \qquad (3.3)$$

Second, if $K = diag(k_i),\ k_i > 0\ i = 1,2,....,p$, then

$$\hat{B}(k)=\sum_{j=1}^{p}(\ell_j+k_j)^{-1}V_jV_j^TX^TY \quad (3.4)$$

Swindle (1976) illustrated a technique for combining prior information with ridge regression that extended Hoerl and Kennard's model as follows:

$$B(kI,J)=(X^TX+kI)^{-1}(X^TY+kJ) \quad (3.5)$$

with J being a fixed vector of prior estimate of B. Swindle showed that there exists a value k which gives a smaller MSE than the least squares estimator for any fixed prior information, J.

Definition (1): A prior mean J is said to be good if the difference $\mathrm{MSE}(\hat{\mathrm{B}}(\mathrm{K}))-\mathrm{MSE}(\hat{\mathrm{B}}(\mathrm{kI},\mathrm{J}))$ is positive for all positive values k when both $\hat{B}(k)$ and $\hat{B}(kI,J)$ are computed by using the same value of k (Pliskin, 1987).

Remark: The restriction $k>0$ is made because, if $k=0$ then

$$\hat{B}_{LS}=\hat{B}(k)=\hat{B}(kI,J)=(X^TX)^{-1}X^TY$$

for all J, thereby implying that all three estimators have the same risk. In this study, it was found that the vector of prior information J depends on the arithmetic mean of the least squares estimators multiplying by a vector whose elements are ones, that is

$$J=\left[\frac{\sum_{i=1}^{p}\hat{B}_{iLS}}{p}\right]I_{p\times 1} \quad (3.6)$$

Unbiasedness of Ridge Estimators: Theorem (1)

Consider the standard linear regression model (2.1), where U is normally distributed $N(0,\sigma^2I)$, and the least square estimator, $\hat{B}$ is normally distributed $N(B,\sigma^2(X^TX)^{-1})$. The prior information J is independent of $\hat{B}_{LS}$, and J is normally distributed $N(B,V)$. Also assume that V has full rank covariance matrix and that the convex estimator is $B(C,J)=C\hat{B}_{LS}+(I-C)J$, where I is the $P\times P$ identity matrix and C is a $P\times P$ matrix. The optimal C in terms of minimum MSE is then

$$\mathrm{C}=\mathrm{V}(\sigma^2(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}+\mathrm{V})^{-1} \quad (3.7)$$

Corollary (1): Suppose $\hat{B}$ is an estimator of B with mean B and covariance matrix $\sum$, and J is prior information with mean B and covariance matrix V. Further assume that if J is uncorrelated with $\hat{B}$, and V and $\sum$ are of full rank, then the convex estimator $B(C,J)$ has a minimum MSE of optimal value

$$C=V(V+\sum)^{-1} \quad (3.8)$$

Theorem (2): Unbiased Ridge Estimate of B (Crouse, et al., 1995)

Let $\hat{B}_{LS}$ have a distribution with mean B and covariance $\sigma^2(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}$, denoted by $N(B,\sigma^2(X^TX)^{-1})$, as in the linear model. Similarly, let J be distributed $\mathrm{N}(\mathrm{B},(\frac{\sigma^2}{\mathrm{k}})\mathrm{I})$ for $\mathrm{k}>0$, and define $\mathrm{B}(\mathrm{C},\mathrm{J})=\mathrm{C}\hat{\mathrm{B}}_{\mathrm{LS}}+(\mathrm{I}-\mathrm{C})\mathrm{J}$; then, for the optimal value C in terms of minimum MSE $\mathrm{B}(\mathrm{C},\mathrm{J})=\hat{\mathrm{B}}(\mathrm{kI},\mathrm{J})=(\mathrm{X}^{\mathrm{T}}\mathrm{X}+\mathrm{kI})^{-1}(\mathrm{X}^{\mathrm{T}}\mathrm{Y}+\mathrm{KJ})$, and $B(C,J)$ is an unbiased estimate of B.

Proof: Assuming that $J \sim N(B,(\frac{\sigma^2}{k})I)$ and, from corollary (1), $\hat{B}$ has a distribution with mean B and covariance $\sum = \sigma^2 (X^T X)^{-1}$, that is, $\hat{B} \sim N(B, \sum)$, it is found that J is distributed with mean B and covariance $V = (\frac{\sigma^2}{k})I$ denoted by $J \sim N(B, V)$. Substituting this into equation (3.8) results in

$$\begin{aligned}\hat{C} &= \frac{\sigma^2}{k}\left(\sigma^2\left(X^T X\right)^{-1} + \frac{\sigma^2}{k} I\right)^{-1} \\ &= \frac{I}{k}\left((X^T X)^{-1} + \left(\frac{I}{k}\right) I\right)^{-1} \\ &= \left[k(X^T X)^{-1} + I\right]^{-1}\end{aligned}$$

Substituting $B(C,J) = C\hat{B} + (I-C)J$, results in

$$\begin{aligned}B(C,J) = &\left(k\left(X^T X\right)^{-1} + I\right)^{-1}\left(X^T X\right)^{-1} X^T Y \\ &+ \left(I - \left(k\left(X^T X\right)^{-1} + I\right)^{-1}\right) J\end{aligned}$$

and

$$\begin{aligned}B(C,J) = &\left(\left(X^T X\right) + kI\right)^{-1} X^T Y \\ &+ \left(I - \left(k\left(X^T X\right)^{-1} + I\right)^{-1}\right) J .\end{aligned}$$

Multiplying $\left(K(X^T X)^{-1} + I\right)^{-1}$ by $X^T X (X^T X)^{-1}$, results in

$$\begin{aligned}B(C,J) = &\left(\left(X^T X\right) + kI\right)^{-1} X^T Y + \\ &\left(I - X^T X\left(X^T X\right)^{-1}\left(k\left(X^T X\right)^{-1} + I\right)^{-1}\right) J \\ = &\left(\left(X^T X\right) + kI\right)^{-1} X^T Y \\ &+ \left(I - X^T X\left((X^T X) + kI\right)^{-1}\right) J\end{aligned}$$

Adding and subtracting kI to $X^T X$,

$$\begin{aligned}B(C,J) &= \begin{pmatrix}\left(X^T X + kI\right)^{-1} X^T Y + \\ \left(I - \left(X^T X + kI - kI\right)\left(X^T X + kI\right)^{-1}\right) J\end{pmatrix} \\ &= \begin{pmatrix}\left(X^T X + kI\right)^{-1} X^T Y + \\ \left(I + k\left(X^T X + kI\right)^{-1} - I\right) J\end{pmatrix}\end{aligned}$$

Simplifying the above results in:

$$\begin{aligned}B(C,J) &= \hat{B}(kI, J) \\ &= (X^T X + kI)^{-1}(X^T Y + kJ)\end{aligned} \tag{3.9}$$

Swindle (1976) did not propose a method for estimating the parameter k, however, Crouse, et al. (1995) proposed a procedure to estimate k, as follows:

$$\hat{k} = \begin{cases} \dfrac{P\sigma^2}{(\hat{B}-J)^T(\hat{B}-J) - \sigma^2 tr(X^T X)^{-1}}, \\ \quad \text{if} \quad (\hat{B}-J)^T(\hat{B}-J) - \sigma^2 tr(X^T X)^{-1} > 0 \\ \dfrac{P\sigma^2}{(\hat{B}-J)^T(\hat{B}-J)} \quad , \quad \text{o.w.} \end{cases} \tag{3.10}$$

If σ^2 is unknown, then σ^2 can be estimated by an unbiased estimator,

$$\hat{k}=\begin{cases}\dfrac{Ps^2}{(\hat{B}-J)^T(\hat{B}-J)-s^2tr(X^TX)^{-1}}, \\ \quad \text{if} \quad (\hat{B}-J)^T(\hat{B}-J)-\sigma^2tr(X^TX)^{-1}>0 \\ \dfrac{Ps^2}{(\hat{B}-J)^T(\hat{B}-J)} \quad , \quad \text{o.w.}\end{cases} \tag{3.11}$$

Properties of the Unbiased Ridge Estimators

1. Unbiasedness:

$$E\left(\hat{B}\left(kI,J\right)\right)=B$$

2. Variance:

$$Var\left(\hat{B}_i\left(kI,J\right)\right)=\sigma^2\frac{\ell_i}{\left(\ell_i+k_i\right)^2}$$

3. Mean Square's Error:

$$MSE\left(\hat{B}\left(kI,J\right)\right)= \sum_{i=1}^{P}\frac{I}{\left(\ell_i+k_i\right)^2}\left(\sum_{i=1}^{P}k_i\left(B_i-J\right)\right)^2+\sigma^2\sum_{i=1}^{P}\frac{\ell_i}{\left(\ell_i+k_i\right)^2} \tag{3.12}$$

Methodology

Model Description and Monte Carlo Simulation

This research used a Monte Carlo study to examine the properties of least squares and unbiased ridge methods. The properties were then compared in the sense of the MSE, which was evaluated using equations (2.2) and (3.12) respectively. Thirty observations (n=30) were generated for each of fourteen (p=14) explanatory variables; the explanatory variables were generated using the device:

$$\begin{cases}X_{ij}^*=(1-\alpha^2)^{1/2}Z_{ij}^*+\alpha Z_{i15}^* \ (j=1,2,...,m \ . \ i=1,2,...,30)\\ X_{ij}^*=Z_{ij}^* \ (j=m+1,m+2,...,14 \ . \ i=1,2,...,30)\end{cases}$$

Where Z_{ij} are independent standard normal pseudo-random numbers, Z_{i15} is the i^{th} element of the column vector of random error Z_{15}, α is specified so that the correlation between any two explanatory variables is given by α^2. The n observations for the dependent variable Y are determined by:

$$Y_i=\lambda_1X_{i1}+\lambda_2X_{i2}+.....+\lambda_{14}X_{i14}+U_i$$
$$i=1,2,....,30$$

where U_i are independent normal $(0,\sigma^2)$ pseudo-numbers evaluated by: $U_i=Z_{i15}-\overline{Z}_{15}$, and Y is standardized using unit length scale.

Results

The primary purpose of this research was to compare the MSE of the considered estimators, thus, the MSE for all estimators was evaluated. In addition, the efficiency of each estimator was evaluated. Thirteen experiments using Monte Carlo methods were conducted. The results of each experiment consist of five tables. The tables display the MSE of each estimator under one of five levels of correlation between explanatory variables. One set of experimental results is presented and consists of tables displaying the MSE of the least square and unbiased ridge methods for the desired correlation coefficients.

Conclusion

As shown in Tables 1-5, based on the thirteen experiments, it is concluded that the unbiased ridge method is preferable to the least square method because it results in smaller MSE values.

Table 1: Correlation Coefficient $\alpha^2 = 0.35$

Correlation Between	MSE Using Least Squares	MSE Using Unbiased Ridge
X1,X2	25.4000	0.122822
X1-X3	23.1838	0.0039352
X1-X4	30.7029	0.0368124
X1-X5	36.5401	0.144270
X1-X6	28.5695	0.0714341
X1-X7	25.3975	0.0241636
X1-X8	36.4954	0.128423
X1-X9	46.5005	0.0045159
X1-X10	1.57386	0.0355173
X1-X11	27.4589	0.0231471
X1-X12	38.3113	0.0382758
X1-X13	39.3052	0.0080928
X1-X14	46.2861	0.0331327

Table 2: Correlation Coefficient $\alpha^2 = 0.51$

Correlation Between	MSE Using Least Squares	MSE Using Unbiased Ridge
X1,X2	26.2279	0.0444524
X1-X3	24.3146	0.0111884
X1-X4	33.2860	0.0029957
X1-X5	45.0645	0.006178
X1-X6	33.4187	0.0064171
X1-X7	29.1076	0.0451311
X1-X8	44.4291	0.0554930
X1-X9	61.7260	0.0508783
X1-X10	29.6791	0.0113255
X1-X11	33.2142	0.0075011
X1-X12	47.9239	0.0162912
X1-X13	49.6498	0.0027323
X1-X14	68.5781	0.0129765

Table 3: Correlation Coefficient $\alpha^2 = 0.67$

Correlation Between	MSE Using Least Squares	MSE Using Unbiased Ridge
X1,X2	28.3239	0.0295791
X1-X3	26.6763	0.0084244
X1-X4	38.7922	0.0066016
X1-X5	61.6469	0.015094
X1-X6	42.9897	0.0242642
X1-X7	37.0921	0.0174133
X1-X8	59.5534	0.0087105
X1-X9	91.8063	0.0000733
X1-X10	38.3784	0.0059245
X1-X11	44.8721	0.0023228
X1-X12	66.4587	0.0110501
X1-X13	69.5876	0.0075445
X1-X14	113.142	0.0031203

Table 4: Correlation Coefficient $\alpha^2 = 0.84$

Correlation Between	MSE Using Least Squares	MSE Using Unbiased Ridge
X1,X2	35.6903	0.0079587
X1-X3	34.3758	0.0061240
X1-X4	57.3043	0.0004140
X1-X5	115.0238	0.042486
X1-X6	74.0384	0.0159949
X1-X7	64.4310	0.0030147
X1-X8	107.380	0.0007958
X1-X9	189.863	0.0025083
X1-X10	68.1341	0.0057765
X1-X11	83.3222	0.0018858
X1-X12	125.433	0.0037824
X1-X13	132.940	0.0039350
X1-X14	259.746	0.0037324

Table 5: Correlation Coefficient $\alpha^2 = 0.99$

Correlation Between	MSE Using Least Squares	MSE Using Unbiased Ridge
X1,X2	253.630	0.0064327
X1-X3	250.785	0.0011331
X1-X4	606.225	0.0098181
X1-X5	1645.7712	0.026003
X1-X6	974.1748	0.0078642
X1-X7	900.5158	0.0048502
X1-X8	1461.07	0.0158666
X1-X9	3049.69	0.0170372
X1-X10	976.803	0.0026540
X1-X11	1218.541	0.0011670
X1-X12	1787.02	0.0043619
X1-X13	1908.30	0.0040964
X1-X14	4586.19	0.0038209

References

Casella, G., & Berger, R. (2002). *Statistical Inference, 2nd Ed*. USA: Duxbury.

Crouse, R., Chun, J., & Hanumara, R. C. (1995). Unbiased ridge estimation with prior information and ridge trace. *Communications in Statistics*, *24*(*9*), 2341-2354.

Wichern, D. W., & Churchill, G. A. (1978). A comparison of ridge estimators. *Technometrics*, *20*(*3*), 301-310.

Gunst, R. F., Webster, J. T., & Mason, R. L. (1977). Biased estimation in regression: An evaluation using mean squared error. *Journal of the American Statistical Association*, *72*(*356*), 616-628.

Horel, A. E., & Kennard, R. W. (1970a). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*, 55-83.

Mason, R. L., Gunst, R. F., & Webster, J. T. (1977). Regression analysis and problems of multicollinearty. *Communications in Statistics*, *4*(*3*), 279-292.

McDonald & Galarneau. (1975). A Monte Carlo evaluation of some ridge type estimators. *Journal of the American Statistical Association*, *70*, 407-416.

Murthy, K. P. N. (2003). *An introduction to Monte Carlo simulation in statistical physics*. India: Indira Gandhi Center for Atomic Research.

Pliskin, L. J. (1987). A ridge-type estimator and good prior means. *Communications in Statistics*, *16*(*12*), 3427-3429.

Swindel, B. F. (1976). Good ridge estimators based on prior information. *Communications in Statistics*, *A5*(*11*), 1065-1075.