Journal of Modern Applied Statistical Methods

Volume 11 | Issue 2 Article 15

11-1-2012

A Proposed Ridge Parameter to Improve the Least Square Estimator

Ghadban Khalaf
King Khalid University, Saudi Arabia

DOI: 10.22237/jmasm/1351743240

A Proposed Ridge Parameter to Improve the Least Square Estimator

Cover Page Footnote
This research was financed by King Khalid University, Abha, Saudi Arabia, Project No. 182.

Journal of Modern Applied Statistical Methods
November 2012, Vol. 11, No. 2, 443-449

1538 – 9472/12/$95.00

A Proposed Ridge Parameter to Improve the Least Squares Estimator

Ghadban Khalaf
King Khalid University,
Saudi Arabia

Ridge regression, a form of biased linear estimation, is a more appropriate technique than ordinary least squares (OLS) estimation in the case of highly intercorrelated explanatory variables in the linear regression model $\vec{Y} = X\vec{\beta} + \vec{u}$. Two proposed ridge regression parameters from the mean square error (MSE) perspective are evaluated. A simulation study was conducted to demonstrate the performance of the proposed estimators compared to the OLS, HK and HKB estimators. Results show that the suggested estimators outperform the OLS and the other estimators regarding the ridge parameters in all situations examined.

Key words: Multicollinearity, ridge regression, Monte Carlo simulation.

Introduction

Consider the standard model for multiple linear regression

$$\vec{Y} = \beta_0 1 + X\vec{\beta} + \vec{u}, \qquad (1)$$

where $\vec{Y}$ is a (n × 1) column vector of observations on the dependent variable, β_0 is a scalar intercept, 1 is a (n × 1) vector with all components equal to unity, X is a (n × p) fixed matrix of observations on the explanatory variables and is of full rank p, $\vec{\beta}$ is a (p × 1) unknown column vector of regression coefficients and $\vec{u}$ is a (n × 1) vector of random errors, $E(\vec{u}) = 0$, $E(uu') = \sigma^2 I_n$, where I_n denotes the (n × n) identity matrix and the prime denotes the transpose of a matrix.

The *OLS* estimator, $\hat{\vec{\beta}}$, of the parameters is given by

$$\hat{\vec{\beta}} = (X'X)^{-1} X'\vec{Y} \qquad (2)$$

Ghadban Khalaf is an Associate Professor of Statistics in the Department of Mathematics. He is a member of the Faculty of Science. Email him at: albadran50@yahoo.com.

where $\hat{\vec{\beta}}$ is an unbiased estimator of $\vec{\beta}$. Multiple linear regression is very sensitive to predictors that are in a configuration of near collinearity. When this is the case, the model parameters become unstable (large variances) and cannot be interpreted. From a mathematical standpoint, near-collinearity makes the $X'X$ matrix ill-conditioned (with X the data matrix), that is, the value of its determinant is nearly zero, thus, attempts to calculate the inverse of the matrix result in numerical snags with uncertain final values.

Exact collinearity occurs when at least one of the predictors is a linear combination of other predictors. Therefore, X is not a full rank matrix, the determinant of X is exactly zero, and inverting $X'X$ is not simply difficult, it does not exist.

When multicollinearity occurs, the least squares estimates remain unbiased and efficient. The problem is that the estimated standard error of the coefficient β_i (for example, S_{bi}) tends to be inflated. This standard error has a tendency to be larger than it would be in the absence of multicollinearity because the estimates are very sensitive to any changes in the sample observations or in the model specification. In other words, including or excluding a particular variable or certain observations may greatly

change the estimated partial coefficient. If S_{bi} is larger than it should be, then the *t*-value for testing the significance of β_i is smaller than it should be. Thus, it becomes more likely to conclude that a variable X_i is not important in a relationship when, in fact, it is important.

Several criteria have been put forth to detect multicollinearity problems. Draper and Smith (1998) suggested the following:

(1) Check if any regression coefficients have the wrong sign, based on prior knowledge.

(2) Check if predictors anticipated to be important based on prior knowledge have regression coefficients with small *t*-statistics.

(3) Check if deletion of a row or a column of the *X* matrix produces a large change in the fitted model.

(4) Check the correlations between all pairs of predictor variables to determine if any are unexpectedly high.

(5) Examine the variance inflation factor (VIF). The VIF of X_i is given by:

$$VIF_i = \frac{1}{1-R_i^2}, \tag{3}$$

where R_i^2 is the squared multiple correlation coefficient resulting from the regression of X_i against all other explanatory variables.

If X_i has a strong linear relation with other explanatory variables, then R_i^2 will be close to one and VIF values will tend to be very high. However, in the absence of any linear relation among explanatory variables, R_i^2 will be zero and the VIF will equal one. It is known that a VIF value greater than one indicates deviation from orthogonality and has tendencies to col linearity. Leclerc and Pireaux (1995) suggested that a VIF value exceeding 300 may indicate the presence of multicollinearity. Conversely, examining a pairwise correlation matrix of explanatory variables might be insufficient to identify collinearity problems because near linear dependencies may exist among more complex combinations of regressors, that is, pairwise independence does not imply independence. Because VIF is a function of the multiple correlation coefficient among the explanatory variables, it is a much more informative tool for detecting multicollinearity than the simple pairwise correlations.

Many procedures have been suggested in an attempt to overcome the effects of multicollinearity in regression analysis. Horel and Kennard (1970) proposed a class of biased estimator called ridge regression estimators as an alternative to the OLS estimator in the presence of collinearity. Freund and Wilson (1998) summarize these into three classes: variable selection, variable redefinition and biased estimation, such as ridge regression. Ridge regression is a variant of ordinary multiple linear regression whose goal is to circumvent the problem of predictors collinearity. Ridge regression gives up the OLS estimator as a method for estimating the parameters of the model and focuses instead on the $X'X$ matrix; this matrix will be artificially modified in order to make its determinant appreciably different from zero. The idea is to add a small positive quantity, for example k, to each of the diagonal elements of the matrix $X'X$ to reduce linear dependencies observed among its columns. A solution vector is thus obtained by the expression

$$\vec{\hat{\beta}}^{*} = (X'X + k\, I_p)^{-1} X'\vec{Y}, \tag{4}$$

where the ridge parameter $k > 0$ represents the degree of shrinkage. By adding the term kI_p, I_p is an identity matrix of the same order as *X'X*, the ridge-regression model reduces multicollinearity and prevents the matrix *X'X*

from being singular even if X itself is not of full rank.

Note that if $k = 0$, the ridge-regression coefficients, defined by (4), are equal to those from the traditional multiple-regression model given by (2). This makes the new model parameters somewhat biased, that is, $E(\vec{\hat{\beta}}^{*}) \neq \vec{\beta},$ (whereas the parameters as calculated by the OLS method are unbiased estimators of the true parameters). However, the variances of the new parameters are smaller than that of the OLS parameters and, in fact, so much smaller than their MSE may also be smaller than that of the parameters of the least squares model. This is an illustration of the fact that a biased estimator may outperform an unbiased estimator provided its variance is small enough.

Perhaps the best way for choosing the ridge regression parameter (k) would be to minimize the expected squared difference between the estimate and the parameter being estimated, that is, the MSE. This would reveal the ideal balance between increase in bias and reduction in variance of the estimator, where

$$MSE = Variance + (Bias)^2. \tag{5}$$

Therefore, it is helpful to allow a small bias in order to achieve the main criterion of keeping the MSE small: this is precisely what ridge regression seeks to accomplish.

Several methods for estimating k have been proposed, for example see: Hoerl and Kennard (1970), Hoerl, et al. (1975), McDonald and Galarneau (1975), Lawless and Wang (1976), Hocking, et al. (1976), Wichern and Churchill (1978), Nordberg (1982), Saleh and Kibria (1993), Singh and Tracy (1999), Wencheko (2000), Kibria (2003), Khalaf and Shukur (2005), Alkhamisi, et al. (2006), Alkhamisi & Shukur (2007), Khalaf (2011) and Khalaf, et al., (2012).

The Main Result

Identifying the optimal method for choosing k is beyond the goal of this study; Hoerl and Kennard (1970) showed that the optimal values for k_i will be

$$\hat{k}_i = \frac{\hat{\sigma}^2}{\hat{\beta}_i^2}, \quad i = 1, 2, ..., p. \tag{6}$$

The acronym HK is used for this estimator. Hoerl and Kennard (1970) stated that "based on experience the best method for achieving a better estimator $\vec{\hat{\beta}}^{*}$ is to use $\hat{k}_i = k$ for all i." Thus, the $\hat{k}_i -$ values of (6) can be combined to obtain a single value of k. Thereby it is not advisable to use an ordinary average because a large k and too much bias would result. Hoerl, et al. (1975) proposed a more reasonable averaging, namely the harmonic mean given by

$$\hat{k}_{HKB} = \frac{p\hat{\sigma}^2}{\hat{\beta}'\hat{\beta}}, \tag{7}$$

where p denotes the number of parameters and $\hat{\sigma}^2$ is given by

$$\hat{\sigma}^2 = \frac{RSS}{n-p}, \tag{8}$$

where RSS denotes the residual sum of squares and the acronym HKB is used for estimator (7). The original definition of k provided by Horel and Kennard (1970) and Hoerl, et al. (1975) is used throughout this article to suggest the proposed estimators as modifications of their estimators. It is known that the denominator $(n-p+2)$ yields an estimator of σ^2 with a lower MSE than the unbiased estimator given by (8) (Rao, 1973). Thus, the use of $\hat{\sigma}^2$ is suggested and is defined by

$$\hat{\sigma}^{*2} = \frac{RSS}{n-p+2}, \tag{9}$$

to estimate $\hat{\sigma}^2$ in both (6) and (7). This leads to the following new estimators

$$\hat{k}_1^* = \frac{\hat{\sigma}^{*2}}{\hat{\beta}_i^2}, \quad i = 1, 2, ..., p \tag{10}$$

and

$$\hat{k}_2^* = \frac{p\hat{\sigma}^{*2}}{\hat{\beta}'\hat{\beta}}. \tag{11}$$

This investigation shows that both $\hat{k}_1^*$ and $\hat{k}_2^*$ in (10) and (11) perform very well relative to the OLS estimator from the MSE point of view.

Methodology

The Simulation

A simulation study was conducted to evaluate the performance of the proposed estimators and to illustrate their superiority. The simulation study concerns a regression model, without the intercept term, with p = 6. The simulation procedure suggested by McDonald and Galarneau (1975), Gibbons (1981) and Kibria (2003) was used to generate the explanatory variables:

$$X_{ij} = (1-\rho^2)^{\frac{1}{2}} z_{ij} + \rho z_{ip}, \quad i = 1, 2, ..., n, \quad j = 1, 2, ..., p, \tag{12}$$

where z_{ij}'s are independent standard normal distribution, ρ^2 is the correlation between any two explanatory variables and p is the number of explanatory variables. The value of ρ^2 is taken as 0.9, 0.99, 0.999 and 0.9999, respectively. The resulting condition numbers (CN) of the generated X equal: 87.36, 368.62, 867.05 and 4250.64, respectively. The n observations for the dependent variable $\vec{Y}$ are determined by:

$$Y_i = \beta_0 + \beta_i X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + u_i, \quad i = 1, 2, ..., n \tag{13}$$

where u_i are independent normal $(0, \sigma^2)$ pseudo-numbers and β_0 is assumed to be identically zero. In this study n is 10, 100 and 1,000 in order to cover both small and large sample sizes. The parameter values were chosen so that $\beta'\beta = 1$, which is a common restriction in simulation studies (Muniz & Kibria, 2009). For given values of p, n and ρ^2, the experiment was repeated 10,000 times by generating 10,000 samples. For each replicate, the values of k for different proposed estimators and the corresponding ridge estimators were calculated using equation (4) where k is given by (6), (7), (10) and (11).

To investigate whether the ridge estimator is better than the OLS estimator, the MSE was calculated using the equation

$$MSE(\vec{\beta}^*) = \frac{1}{10000}\sum_{r=1}^{10000}(\beta^* - \beta)'(\beta^* - \beta). \tag{14}$$

Results

Ridge estimators are constructed with the aim of having smaller MSE than the MSE for the least squares. Improvement, if any, can therefore be studied by looking at the amounts of these MSE's. The detailed results of the simulations are shown in Tables 1 – 3. The results concerning the MSE's and the comparisons of ridge estimators with least squares is then dealt with. To summarize these findings:

(1) Regardless of the condition of $X'X$, the values of MSE of the estimators relative to the OLS estimator are small and therefore the improvement of the ridge estimators over the OLS estimator is remarkable. This may indicate that the influence of multicollinearity upon the MSE criterion is relatively weak. Consequently, the two proposed estimators, given by $\hat{k}_1^*$ and $\hat{k}_2^*$, are far more effective than HK and HKB in improving the OLS estimator.

(2) Regardless of sample size, the differences of the values of each type of the suggested

estimators are trivial. The $\hat{k}_2^*$ estimator, defined by (11), performed very well; it appears to outperform $\hat{k}_1^*$, and it is also considerably better than HK and HKB.

In summary, the proposed estimators can greatly improve the OLS estimator, as well the HK and HKB estimators, under the MSE criterion. The proposed estimators appear to offer an opportunity for a large reduction in MSE when the degree of multicollinearity as measured by the CN is high.

Table 1: The MSE of the Suggested Estimators, HK, HKB and the OLS Estimator (n = 20)

ρ^2	0.9	0.99	0.999	0.9999
CN	***87.36***	***368.62***	***867.05***	***4250.64***
OLS	0.190	0.284	0.817	4.213
$\hat{k}_1^*$	0.125	0.156	0.360	1.578
$\hat{k}_2^*$	0.141	0.153	0.240	0.259
HK	0.197	0.207	0.280	0.626
HKB	0.180	0.264	0.688	2.363

Table 2: The MSE of the Suggested Estimators, HK, HKB and the OLS Estimator (n = 100)

ρ^2	0.9	0.99	0.999	0.9999
CN	***87.36***	***368.62***	***867.05***	***4250.64***
OLS	0.40	0.058	0.169	0.940
$\hat{k}_1^*$	0.034	0.046	0.086	0.360
$\hat{k}_2^*$	0.032	0.036	0.070	0.224
HK	0.045	0.045	0.083	0.250
HKB	0.039	0.056	0.154	0.631

Table 3: The MSE of the Suggested Estimators, HK, HKB and the OLS Estimator (n = 1,000)

ρ^2	0.9	0.99	0.999	0.9999
CN	***87.36***	***368.62***	***867.05***	***4250.64***
OLS	0.030	0.045	0.130	0.658
$\hat{k}_1^*$	0.026	0.036	0.073	0.229
$\hat{k}_2^*$	0.023	0.028	0.058	0.156
HK	0.027	0.031	0.065	0.183
HKB	0.029	0.044	0.108	0.449

Conclusion

Ridge regression is more than a last resort attempt to salvage least square linear regression in the case of near or full collinearity of predictors. It is to be considered a major linear regression technique that proves its usefulness when collinearity is problematic. From the MSE point of view, it is not surprising that the use of traditional multiple linear regression suffers from multicollinearity problems and clearly shows that ridge regression performs best when the input data are multicollineared.

Two methods for specifying k were proposed herein and were evaluated in terms of MSE via simulation techniques. Comparisons were made with other ridge-type estimators evaluated elsewhere. The simulation study showed that the OLS estimator is dominated by these estimators in all cases investigated and that the improvement of the suggested estimators is substantial from the MSE point of view. Finally, although there are many strategies for choosing an optimal value for k, there is no consensus regarding the best or most general way to choose k. In other words, the best method for estimating k is an unsolved problem and there is no rule for choosing k evaluated to date that assures the corresponding ridge estimator is uniformly better (in terms of MSE) than the OLS estimator.

Acknowledgement

This research was financed by King Khalid University, Abha, Saudi Arabia, Project No. 182.

References

Alkhamisi, M. A., & Shukur, G. (2007). A Monte Carlo study of recent ridge parameters. *Communications in Statistics - Simulation and Computation*, *36*(*3*), 535-547.

Alkhamisi, M. A., Khalaf, G., & Shukur, G. (2006). Some modification for choosing ridge parameter. *Communications in Statistics - Theory and Methods*, *35*, 2005-2020.

Draper, N. R., & Smith, H. (1998). *Applied regression analysis*, *3rd Ed*. New York, NY: John Wiley and Sons.

Freund, R. J., & Wilson, W. J. (1998). *Regression analysis: Statistical modeling of a response variable*, *1st Ed.* San Diego, CA: Academic Press.

Gibbons, D. G. (1981). A simulation study of some ridge estimators. *Journal of the American Statistical Association*, *76*(*373*), 131-139.

Hocking, R. R., Speed, F. M., & Lynn, M. J. (1976). A class of biased estimators in linear regression. *Technometrics*, *18*(*4*), 425-437.

Horel, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*, 55-67.

Hoerl, A. E., Kennard, R. W., & Baldwin, K. F. (1975). Ridge regression: Some simulation. *Communications in Statistics*, *4*, 105-123.

Khalaf, G. (2011). Ridge regression: An evaluation to some new modifications. *International Journal of Statistics and Analysis*, *1*(*4*), 325-342.

Khalaf, G., Månsson, K., Shukur, G., & Sjölander, P. (2012). A Tobit ridge regression estimator. To appear in *Communications in Statistics - Theory and Methods*.

Khalaf, G., & Shukur, G. (2005). Choosing ridge parameter for regression problems. *Communications in Statistics - Theory and Methods*, *34*, 1177-1182.

Kibria, B. M. (2003). Performance of some new ridge regression estimators. *Communications in Statistics - Simulation and Computation*, *32*(*2*), 419-435.

Lawless, J. F., & Wang, P. (1976). A simulation study of ridge and other regression estimators. *Communications in Statistics*, *A5*, 307-323.

Leclerc, G., & Pireaux, J. J. (1995). The use of least squares for XPS peak parameters estimation, part 3: Multicollinearity, ill-conditioning and constraint-induced bias. *Journal of Electron Spectroscopy Related Phenomena*, *71*, 179-190.

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

Muniz, G., & Kibria, B .M. G. (2009). On some ridge regression estimators: An empirical comparison. *Communications in Statistics - Simulation and Computation*, *38*, 621-630.

Nordberg, L. (1982). A Procedure for determination of a good ridge parameter in linear regression. *Communications in Statistics*, *A11*, 285-309.

Rao, C. R. (1973). *Linear statistical inference and its applications*. New York, NY: John Wiley and Sons.

Saleh, A. K., & Kibria, B. M. (1993). Performances of some new preliminary test ridge regression estimators and their properties. *Communications in Statistics - Theory and Methods*, *22*, 2747-2764.

Singh, S., & Tracy, D. S. (1999). Ridge regression using scrambled responses. *Metrika*, *b*, 147-157.

Wencheko, E. (2000). Estimation of the signal-to-noise in the linear regression model. *Statistical Papers*, *41*, 327-343.

Wichern, D., & Churchill, G. (1978). A comparison of ridge estimators. *Technometrics*, *20*, 301-311.