DIGITALCOMMONS @WAYNESTATE

Journal of Modern Applied Statistical Methods

Volume 11 | Issue 1 Article 12

5-1-2012

Improved Estimator in the Presence of Multicollinearity

Ghadban Khalaf
King Khalid University, Saudi Arabia

Journal of Modern Applied Statistical Methods
May 2012, Vol. 11, No. 1, 152-157

Improved Estimator in the Presence of Multicollinearity

Ghadban Khalaf
King Khalid University,
Saudi Arabia

The performances of two biased estimators for the general linear regression model under conditions of collinearity are examined and a new proposed ridge parameter is introduced. Using Mean Square Error (MSE) and Monte Carlo simulation, the resulting estimator's performance is evaluated and compared with the Ordinary Least Square (OLS) estimator and the Hoerl and Kennard (1970a) estimator. Results of the simulation study indicate that, with respect to MSE criteria, in all cases investigated the proposed estimator outperforms both the OLS and the Hoerl and Kennard estimators.

Key words: Regression, multicollinearity, ridge regression, Monte Carlo simulation.

Introduction

Multiple regression fits a model to predict a dependent variable (*Y*) from two or more independent variables (*X*):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + e.$$

If the model fits the data well, the overall R^2 value will be high and the corresponding *P* value will be low. In addition to the overall *P* value, multiple regression also reports an individual *P* value for each independent variable; a low *P* value indicates that a particular independent variable significantly improves the fit of the model.

If the overall *P* value is very low, but all the individual *P* values are high, this indicates that a model fits the data well, even though none of the *X* variables has a statistically significant impact on predicting *Y*. This occurs when two or more variables are highly correlated. If both variables are removed from the model, the fit would be much worse; thus, the overall model fits the data but neither *X* variable makes a significant contribution when it is added to the model last. When this occurs, the *X* variables are collinear and the results show multicollinearity, meaning the variables are related.

If the goal is simply to predict *Y* from a set of *X* variables, then multicollinearity is not problematic. The predictions will be accurate and the overall R^2 quantifies how well the model predicts the *Y* values. However, if the goal is to understand how the various *X* variables impact *Y*, then multicollinearity poses a big problem. These problems are summarized as:

(1) The individual *P* values can be misleading, that is, a *P* value can be high, even though the variable is important.

(2) The confidence intervals on the regression coefficients will be very wide – and may include zero – which indicates that a researcher cannot be confident whether an increase in *X* values is associated with an increase or decrease in *Y* values. In addition, wide confidence intervals can change the coefficients and/or their signs.

For these reasons, multicollinearity must be examined and removed. Different methods exist that can be used to reduce or to eliminate the impact of multicollinearity, examples include:

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

(1) Removing a variable: If one of the variables does not seem logically essential to the model, then removing it may be helpful.

(2) Combining the variables; for example, if height and weight are collinear independent variables, then it would be logical to remove height and weight from the model and instead use a variable such as surface area (calculated from height and weight).

(3) Increasing sample size: Another way to reduce the impact of collinearity is to increase sample size, this results in narrower confidence intervals, despite multicollinearity, with more data.

(4) Using a standard technique called ridge regression: Ridge regression was originally developed to overcome multicollinearity.

Consider the standard linear regression model:

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

formulated to result in a $X'X$ in correlation form and where $X'Y$ is the vector of correlation coefficients of the dependent variable with each explanatory variable. Also assume that X is $n \times p$ of full rank $p < n$, $E(e) = 0$ and $E(ee') = \sigma^2 I_n$. The p- vector of the OLS estimator ($\hat{\vec{\beta}}$), is then given by the solution of:

$$X'X\,\hat{\vec{\beta}} = X'\vec{Y} \qquad (2)$$

so that,

$$\hat{\vec{\beta}} = (X'X)^{-1}\,X'\vec{Y}\,. \qquad (3)$$

Clearly, $\hat{\vec{\beta}}$ is an unbiased estimator of $\vec{\beta}$. There are many reasons why a data analyst is often not satisfied with OLS estimates. One of the reasons is prediction accuracy: OLS estimates often have low bias but large variance. Thus, prediction accuracy can occasionally be improved by shrinking some coefficients to zero. In doing this, a little bias is sacrificed to reduce the variance of the predicted values and hence may improve overall prediction accuracy.

Many attempts have been made to improve the OLS estimator procedure. Hoerl and Kennard (1970a) suggested a new technique called ridge regression to improve OLS estimates. The ridge regression estimators $\hat{\vec{\beta}}^*$, for a fixed $k > 0$, satisfy,

$$(X'X + kI_p)\hat{\vec{\beta}}^* = X'\vec{Y}\,, \qquad (4)$$

so that,

$$\hat{\vec{\beta}}^* = (X'X + kI_p)^{-1}\,X'\vec{Y}\,, \qquad (5)$$

as an alternative to the OLS estimator for use in the presence of multicollinearity, where I denotes an identity matrix, and k is a positive number known as ridge parameter, which must be estimated from the real data. The ridge regression MSE is given by:

$$MSE(\hat{\vec{\beta}}^*) = \sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i + k_i)^2} + \sum_{i=1}^{p} \frac{k_i^2 \beta_i^2}{(\lambda_i + k_i)^2},$$
$$= Variance + (Bias)^2, \qquad (6)$$

where σ^2 represents the error variance of the model given by (1). When the reduction in variance exceeds the square of the bias, ridge estimates are preferred.

When using ridge estimates, the choice of k-values in (5) is crucially important and several methods have been proposed for this purpose (see Hoerl & Kennard, 1970a; Saleh & Kibria, 1993; Singh & Tracy, 1999; Khalaf & Shuker, 2005; Alkhamisi & Shukur, 2008; Khalaf, 2011; Khalaf, 2011).

The Proposed Estimator

The general form of ridge regression suggested by Hoerl and Kennard (1970a) reduces $X'X$ to a diagonal matrix by applying an orthogonal transformation Q, thus,

$$Q(X'X)Q' = \Lambda\,,$$

where Q is a $p \times p$ orthogonal matrix, Λ is a diagonal matrix whose diagonal elements $\lambda_1, \lambda_2, \cdots, \lambda_p$ are the eigenvalues of $X'X$. If $X^* = XQ'$ and $\alpha = Q\beta$, then model (1) may be rewritten as,

$$\vec{Y} = X^* \alpha + \vec{e},$$

where

$$(X^*)'(X^*) = \Lambda.$$

The general ridge estimation procedure is defined as,

$$\hat{\alpha}^* = \left((X^*)'(X^*) + K\right)^{-1} (X^*)'\vec{Y}, \quad (7)$$

where K is a diagonal matrix with non-negative diagonal elements $k_1, k_2, \ldots, k_p, k_i > 0.$ It follows from Hoerl and Kennard (1970a) that the value of k_i which minimizes the MSE of $\hat{\alpha}^*$ given by:

$$MSE(\vec{\hat{\alpha}}^*) = \sigma^2 \sum_{i=}^{p} \frac{\lambda_i}{(\lambda_i + k_i)^2} + \sum_{i=1}^{p} \frac{k_i^2 \alpha_i^2}{(\lambda_i + k_i)^2},$$

is:

$$k_i = \frac{\sigma^2}{\alpha_i^2}, \quad (8)$$

where σ^2 represents the error variance of model (1) and α_i is the i^{th} element of $\vec{\alpha}$. Equation (8) gives a value of k_i that is fully dependent on the unknown σ^2 and α_i, and therefore must be estimated from observed data. Hoerl and Kennard (1970a) recommended replacing σ^2 and α_i by their corresponding unbiased estimators, that is:

$$\hat{k}_i = \frac{\hat{\sigma}^2}{\hat{\alpha}_i^2}, \quad (9)$$

where $\hat{\sigma}^2 = \frac{\sum_i e_i^2}{n-p}$ is the residual MSE, which is an unbiased estimator of σ^2, and $\hat{\alpha}_i$ is the element of $\vec{\alpha}$ which is an unbiased estimator of $\vec{\alpha}$. Hoerl and Kennard found that the best method for achieving a better estimator $\vec{\hat{\alpha}}^*$ is to use $k_i = k$ for all *i*, and they suggested *k* to be $\hat{k}_{HK}$ where:

$$\hat{k}_{HK} = \frac{\hat{\sigma}^2}{\max(\hat{\alpha}_i^2)}. \quad (10)$$

They showed that the estimator $\hat{k}_{HK}$ (HK) is sufficient to give ridge estimators with smaller MSEs than an OLS estimator. This article proposes a modification of the Hoerl and Kennared (1970a) estimator shown in (10) to obtain a new estimator, given by:

$$\begin{aligned} \hat{k}_{GK} &= \frac{\hat{\sigma}^2}{\max(\hat{\alpha}_i^2)} + \frac{1}{\frac{1}{2}(\lambda_{\max} + \lambda_{\min})}, \\ &= \frac{\hat{\sigma}^2}{\max(\hat{\alpha}_i^2)} + \frac{2}{(\lambda_{\max} + \lambda_{\min})}, \end{aligned} \quad (11)$$

where $\lambda_{\max}, \lambda_{\min}$, are the largest and smallest eigenvalues of the matrix $X'X$, respectively. This estimator will be denoted by GK. Because $\frac{2}{\lambda_{\max} + \lambda_{\min}} > 0,$ then GK is greater than HK.

Monte Carlo Simulation

Monte Carlo simulation was used to investigate the properties of the considered estimators. It is convenient to make the comparison among the OLS estimator, HK estimator given by (10) and the new proposed GK estimator given by (11). These choices were made for many reasons: First is that; interest herein lies in studying the properties of the proposed GK estimator as an alternative to the OLS estimator in the presence of multicollinearity. Second, GK is a modified version of HK, so it is necessary to make a comparison between them. Finally, the HK estimator was the first ridge estimator proposed among all other estimators, therefore, most studies comparing ridge estimators consider the HK estimator.

A comparison was made based on MSE criterion. Following McDonald and Galarneau (1975), Wichern and Churchill (1978), Gibbons (1981) and Kibria (2003), the explanatory variables were generated using the device:

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

where z_{ij} are independent standard normal pseudo-random numbers, ρ is specified so that the correlation between any two explanatory variables is given by ρ^2 and p is the number of explanatory variables. Once more, the variables are standardized so that $X'X$ and $X'Y$ are in correlation forms. Four sets of correlations were considered corresponding to ρ = 0.7, 0.8, 0.9 and 0.99. Using the condition number, $CN = \frac{\lambda_{\max}}{\lambda_{\min}}$, it can be shown that these values of ρ will include a wide range of low, moderate and high correlations between variables. The n observations for the dependent variable Y are determined by:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_p X_{ip} + e_i, \quad i = 1, 2, ..., n$$

where e_i are independent normal $\left(0,\ \sigma^2\right)$ pseudo-numbers and β_0 is considered identically zero without loss of generality. Three different sample sizes, n = 20, 30, 50 were used with 10, 15 and 20 explanatory variables, respectively. These choices of p were chosen to study the behavior of the estimators for small, moderate and large number of explanatory variables.

The parameter values were chosen so that $\sum_{j=1}^{p} \beta_j^2 = 1,$ which is a common restriction in simulation studies (see Muniz & Kibria, 2009). For given values of p, n and ρ, the experiment was repeated 5,000 times by generating 5,000 samples. For each replicate r (r = 1, 2, …, 5,000) the values of k different proposed estimators and the corresponding ridge estimators were calculated using:

$$\hat{\alpha}^* = (\Lambda + \hat{k}I)^{-1} X'^* \vec{Y}, \tag{13}$$

where $\hat{k} = HK, GK$. The MSE for the estimators were calculated as follows:

$$MSE(\vec{\hat{\alpha}}^*) = \frac{1}{5000} \sum_{r=1}^{5000} (\vec{\alpha}_r^* - \vec{\alpha})'(\vec{\alpha}_r^* - \vec{\alpha}). \tag{14}$$

Results

The results of the simulations that compared the MSE to the other estimators are summarized in Tables 1-3. To compare the performances of the considered estimators, the MSE was calculated for each. The estimator that resulted in the minimum MSE was considered to be the best. The statistics package Minitab 14 was used for all calculations.

Tables 1-3 show that both HK and GK are better than the OLS estimator, and the GK estimator performs better than the HK estimator. This also reveals that for low correlation, r = 0.7, the performance of the GK estimator is slightly better than the HK estimator. Moreover, it was observed that, for given n and p, the MSE for all estimators increased as the correlation among the explanatory variables increased. Conversely, as the sample size and the number of explanatory variables increase, the MSE of all estimators decreased.

Table 1: Estimated MSE and the values of CN with $p = 10$ and $n = 20$.

	ρ	0.99	0.9	0.8	0.7
	CN	980.58	90.17	44.24	18.93
Estimators	OLS	0.03877	0.02844	0.02147	0.01679
	HK	0.03453	0.02613	0.02015	0.01611
	GK	0.02820	0.02241	0.01826	0.01393

Table 2: Estimated MSE and the values of CN with $p = 15$ and $n = 30$.

	ρ	0.99	0.9	0.8	0.7
	CN	2834.43	255.92	89.68	64.19
Estimators	OLS	0.01886	0.01299	0.00933	0.00738
	HK	0.01822	0.01275	0.00922	0.00733
	GK	0.01641	0.01112	0.00875	0.00701

Table 3: Estimated MSE and the values of CN with $p = 20$ and $n = 50$.

	ρ	0.99	0.9	0.8	0.7
	CN	25266.30	2235.51	945.37	399.63
Estimators	OLS	0.01454	0.00922	0.00665	0.00505
	HK	0.01416	0.00910	0.00660	0.00503
	GK	0.00694	0.00436	0.00320	0.00488

Conclusion

Several procedures for constructing ridge estimators have been proposed in the literature. These procedures aim at a rule for selecting the constant *k* in equation (13). The best method for estimating *k* remains an unsolved problem and no constant value of *k* is certain to yield an estimator that is uniformly better in terms of MSE than the OLS estimators in all cases.

This study investigated the properties of a newly proposed method for estimating the ridge parameter (*k*) in the presence of multicollinearity. The investigation used Monte Carlo experiments, where levels of correlation, numbers of explanatory variables and sample sizes were varied. Each combination was replicated 5,000 times. The evaluation of the new estimator was accomplished by comparing the MSE of this estimator with the OLS estimator and the Hoerl and Kennard (1970a) estimator. Results show that the proposed estimator uniformly dominates the other estimators.

References

Alkhamisi, M., & Shukur, G. (2008). Developing ridge parameters for SUR model. *Communications in Statistics – Theory and Methods*, *37*, 544-564.

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

Hoerl, A. E., & Kennard, R.W. (1970a). Ridge regression: biased estimation for non - orthogonal problems. *Technometrics*, *12*, 55-67.

Khalaf, G. (2011). Suggested ridge regression estimators under multicollinearity. To appear in *Journal of Natural and Applied Sciences*, *University of Aden*, *15*(*2*).

Khalaf, G. (2011). Suggested values of the ridge parameters. *International Journal of Statistics and Analysis*, *1*(*2*), 109-118.

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

Kibria, B. M. G. (2003). Performance of some new ridge regression estimators. *Communications in Statistics – Theory and Methods*, *32*, 419-435.

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.

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*, 147-157.

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