Comparing Ridge Regression and Principal Components Regression by Monte Carlo Simulation Based on MSE

Kianoush Fathi Vajargah

Department of Statistics Islamic Azad University, North branch Tehran, Iran,
Corresponding addresses
k_fathi@iau-tnb.ac.ir

***Abstract*:** In multiple regression model, regression variables are usually assumed to be independent from each other. When this assumption is not established, the model would be inappropriate and therefore the results might be incorrect. So, biased regression methods are applied. Ridge regression and principal components regression are two methods of biased regression methods. In this paper, Monte Carlo simulation tests were used for estimating coefficients of ridge and principal components regression. These two methods were compared using minimum squared error (MSE).

***Keywords*:** Ridge regression, Principal components regression, Monte Carlo simulation, Minimum squared Error.

1. Introduction

In a regression model, regression variables are assumed to be independent from each other. When regression variables are not independent from each other and there is multi collinearity among them, estimating regression coefficients of minimum squared error would have large variance and they would also have averagely large length. In fact, absolute value of minimum squared estimates would be very large; thus, they are very unstable, which makes the model inappropriate and results in incorrect results. In Section 2, large variance and estimated coefficients length of minimum squares are examined. Section 3 introduces ridge regression and selection methods of ridge parameter. Section 4 introduces principal components regression and Broken Stick approach, which is one of selection methods of the number of components for remaining in regression model. Section 5 uses Monte Carlo simulation and MSE to compare ridge regression and principal components regression.

2. Regression model

2.1 Introducing regression model

Consider the following standard regression model:

$$y = X\beta + e \quad (1)$$

where y is an $n \times 1$ vector of dependent variable observations, X is an $n \times p$ matrix of observations on p regression variables $(p \leq n)$, β is a $p \times 1$ vector of unknown regression coefficients and e is an $n \times 1$ vector of random errors so that $E(e\acute{e}) = \sigma^2 I_n$ and $E(e) = 0$ in which I_n is identity matrix. In the standard model (1), assume that the variables are standardized; then, $\acute{X}X$ is correlation matrix of regression variables and $\acute{X}y$ vector is correlation coefficients between regression variables and dependent variable y. It is known that minimum squared regression coefficients are estimated as follows:

$$\hat{\beta} = (\acute{X}X)^{-1}\acute{X}y \quad (2)$$

2.2 Investigating variance of regression coefficients

In standard model (1), assume that there are only two regression variables x_1 and x_2; so, matrix $\acute{X}X$ is:

$$\acute{X}X = \begin{bmatrix} 1 & r_{12} \\ r_{21} & 1 \end{bmatrix} \quad (3)$$

where r_{12} is correlation coefficient between x_1 and x_2. Inverse of matrix $\acute{X}X$ is obtained as follows:

$$M = (\acute{X}X)^{-1} = \begin{bmatrix} \frac{1}{1-r_{12}^2} & \frac{-r_{12}}{1-r_{12}^2} \\ \frac{-r_{12}}{1-r_{12}^2} & \frac{1}{1-r_{12}^2} \end{bmatrix} \quad (4)$$

and estimating regression coefficients using Eq. (2) includes:

$$\hat{\beta}_1 = \frac{r_{1y}-r_{12}r_{2y}}{1-r_{12}^2} \quad (5)$$

$$\hat{\beta}_2 = \frac{r_{2y}-r_{12}r_{1y}}{1-r_{12}^2}$$

where r_{1y} is correlation coefficient between y and x_1 and r_{2y} is correlation coefficient between y and x_2.

Since $V(\hat{\beta}_j) = M_{jj}\sigma^2$ (M_{jj} is the j-th diagonal component of matrix M), if there is strong correlation between x_1 and x_2, then:

$$|r_{12}| \to 1 \Rightarrow V(\hat{\beta}_j) = M_{jj}\sigma^2 \to \infty \qquad j = 1,2 \quad (6)$$

$$Cov(\hat{\beta}_1, \hat{\beta}_2) = M_{12}\sigma^2 \to \pm\infty$$

Thus, the estimated variance of regression coefficients would be very large.

If there are more than two regression variables, components of main diagonal of matrix $C = (\acute{X}X)^{-1}$ are as follows:

$$M_{jj} = \frac{1}{1-R_j^2} \qquad j=1,\ldots,p$$

where R_j^2 is multiple determination coefficient, which is obtained from regression of each x_j against other regression variables. If there is strong multicollinearity between regression variables, R_j^2 would get closer to 1 and, since $V(\hat{\beta}_j) = M_{jj}\sigma^2 = \frac{\sigma^2}{1-R_j^2}$, variance of regression coefficients would be very large.

2.2 Investigating squared distance of estimated coefficients from their real value

Squared distance $\hat{\beta}$ from vector β is:

$$L^2 = (\hat{\beta} - \beta)'(\hat{\beta} - \beta) \quad (7)$$

the hope of which is:

$$E(L^2) = E\left[(\hat{\beta}-\beta)'(\hat{\beta}-\beta)\right] \quad (8)$$
$$= \sum_{j=1}^{p} E\left(\hat{\beta}_j-\beta_j\right)^2$$
$$= \sum_{j=1}^{p} V(\hat{\beta}_j) = \sigma^2 tr(\acute{X}X)^{-1}$$

Since $tr(\acute{X}X)^{-1}$ is equal to sum of its special values, then the following can be written:

$$E(L^2) = \sigma^2 \sum_{j=1}^{p} \frac{1}{\lambda_i} \quad (9)$$

where $\lambda_j > 0$ are special values of $\acute{X}X$ matrix. In case of existence of multi colinearity between regressions variables, at least one of λ_js would be large and distance of estimating minimum squares $\hat{\beta}$ from actual value of β would be large.

3. Ridge regression

3.1 Introducing ridge regression

Ridge regression is one of the methods for obtaining biased estimators of regression coefficients, which was first presented by Horl and Konard (1970). In Eq. (2), instead of using $\acute{X}X$, $\acute{X}X + kI$ in which $k>0$ is used and ridge estimation is defined as follows:

$$\hat{\beta}_k = (\acute{X}X + kI)^{-1}\acute{X}y \quad (10)$$

where I is $p \times p$ identity matrix and k is a constant larger than zero, which is called ridge parameter or bias parameter.
In Section 2, it was assumed that the variables were standardized. So, $\acute{X}X$ matrix will be correlation matrix between regression variables and its special values are obtained. Λ is matrix of special values and D is the corresponding matrix of orthogonal eigenvectors of λ_js; thus:

$$\acute{D}\acute{X}XD = \Lambda = diagonal(\lambda_1, \lambda_2, \dots, \lambda_p) \quad (11)$$

where $\boldsymbol{\Lambda_{p\times p}}$ is a diagonal matrix in which $\boldsymbol{\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0}$.
The following is defined:

$$\boldsymbol{X^* = XD} \quad , \boldsymbol{\alpha = \acute{D}\beta} \quad (12)$$

Then, regression model (1) will be as follows:

$$\boldsymbol{y = X\beta + e = XD\acute{D}\beta + e} \quad (13)$$
$$\boldsymbol{= X^*\alpha + e}$$

Minimum squared error for Eq. (13) will be obtained as follows:

$$\hat{\alpha} = (X^{*'}X^*)^{-1}X^{*'}y \quad (14)$$
$$\hat{\alpha} = \Lambda^{-1}X^{*'}y$$

Therefore, estimating ridge (0RR) is defined as follows:

$$\hat{\alpha}_k = (X^{*'}X^* + kI_p)X^{*'}y \quad (15)$$

Mean squared error for ridge regression is:

$$MSE(\hat{\alpha}_k) = \sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i+k)} + \sum_{i=1}^{p} \frac{k^2\alpha_i^2}{(\lambda_i+k)^2} \quad (16)$$

so that $\boldsymbol{\sigma^2}$ is error variance and $\boldsymbol{\alpha_1}$ is the i-th component of α.

3.2 Selection methods of ridge parameter

Below, several methods of selecting ridge parameter are introduced:

3.2.1 Horl and Konard's method:

Horl and Konard (1970) proposed estimating ridge parameter as follows:

$$\hat{k}_{HK} = \frac{\hat{\sigma}^2}{\max(\hat{\alpha}_i^2)} \quad (17)$$

3.2.2 Horl, Konard and Balvadine's method

Horl, Konard and Balvadine (1977) proposed another estimation method for ridge parameter as follows:

$$\hat{k}_{HKB} = \frac{p\hat{\sigma}^2}{\hat{\alpha}'\hat{\alpha}} = \frac{p\hat{\sigma}^2}{\sum_{i=1}^{p}\alpha_i^2} \quad (18)$$

3.2.3 Lavlez and Wang's method

Lavlez and Wang (1976) proposed the following ridge parameter estimation:

$$\hat{k}_{LW} = \frac{p\hat{\sigma}^2}{\sum_{i=1}^{p}\lambda_i\alpha_i^2} \quad (19)$$

3.2.4 Howking, Spead and Leen's method

Howking, Speed and Leen (1976) estimated ridge parameter as follows:

$$\hat{k}_{HSL} = \hat{\sigma}^2 \frac{\sum_{i=1}^{p}(\lambda_i\alpha_i)^2}{(\sum_{i=1}^{p}\lambda_i\alpha_i^2)^2} \quad (20)$$

3.2.5 Khalaf and Shoker's method

Khalaf and Shoker (2005) proposed a new method for ridge parameter selection as follows:

$$\hat{k}_{KS} = \frac{\max(\lambda_i)\hat{\sigma}^2}{(n-p-1)\hat{\sigma}^2+\max(\lambda_i)max(\alpha_i^2)} \quad (21)$$

3.2.6 Deragid and Kashier's method

Deragid and Kashier (2010) proposed another estimation method by changing ridge estimation $\hat{k}_{HKB}$ as follows:

$$\hat{k}_{DK} = \max\left(0, \frac{p\hat{\sigma}^2}{\hat{\alpha}'\hat{\alpha}} - \frac{1}{n(VIF)_{max}}\right)$$

where $\boldsymbol{VIF_j}$ is variance inflation factor of the j-th regression variable and

$$(VIF_j)_{max} = \max(VIF_j) \qquad j = 1, 2, \dots, p$$

in which components of main diagonal of matrix $(\acute{X}X)^{-1}$ are variance inflation factors.

3.2.7 Kibria's method

Kibria (2003) proposed using arithmetic mean, geometric mean and median $\frac{\hat{\sigma}^2}{\hat{\alpha}_i^2}$ as ridge parameters, which are:

$$\hat{k}_{AM} = \frac{1}{p}\sum_{1=1}^{p}\frac{\hat{\sigma}^2}{\hat{\alpha}_i^2} \quad (23)$$

$$\hat{k}_{GM} = \frac{\hat{\sigma}^2}{(\prod_{i=1}^{p}\hat{\alpha}_i^2)^{\frac{1}{p}}} \quad (24)$$

$$\hat{k}_{MED} = Medion\left\{\frac{\hat{\sigma}^2}{\hat{\alpha}_i^2}\right\} \quad , i = 1, 2, \dots, p \quad (25)$$

4. Principal components regression

In principal components regression method, instead of using regression variables, principal components are used as regression variables. Thus, the replaced regression variables are independent from each other. In principal components regression model, a subset of principal components is used instead of all components. Assume q first components are used in regression model ($q<p$); then, $\boldsymbol{\alpha}$ is estimated as follows:

$$\hat{\alpha}_q = (X_q^{*'}X_q^*)^{-1}X_q^{*'}y \quad (26)$$
$$= \Lambda_q^{-1}D_q'X'y$$

so that $\boldsymbol{X_q^* = XD_q}$ and $\boldsymbol{\Lambda_q}$ are diagonal matrix of q first special values (where $\boldsymbol{\lambda_1 \geq \lambda_2 \geq \cdots \lambda_q}$) and $\boldsymbol{D_q}$ is a matrix

with q corresponding special value vector. In Eq. (12), $\boldsymbol{\alpha}$ is defined as $\boldsymbol{\alpha} = \dot{\boldsymbol{D}}\boldsymbol{\beta}$. Then, $\boldsymbol{\beta} = \boldsymbol{D}\boldsymbol{\alpha}$ can be written and estimated value of $\boldsymbol{\beta}$ using principal component method is equal to:

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{PC}} = \boldsymbol{D}\hat{\boldsymbol{\alpha}} \qquad (27)$$

and by replacing $\hat{\boldsymbol{\alpha}}$ with $\hat{\boldsymbol{\beta}}_{\boldsymbol{PC}}$, the following is given for the reduced model:

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{PC}} = \boldsymbol{D}_{\boldsymbol{q}}\boldsymbol{\Lambda}_{\boldsymbol{q}}^{-1}\boldsymbol{D}_{\boldsymbol{q}}^{-1}\dot{\boldsymbol{X}}\boldsymbol{y} \qquad (28)$$

and minimum squared error of principal components regression are:

$$\boldsymbol{MSE}\left(\hat{\boldsymbol{\beta}}_{\boldsymbol{PC}}\right) = \boldsymbol{\sigma}^2 \sum_{i=1}^{q} \frac{1}{\lambda_i} + \sum_{i=q+1}^{p} (\boldsymbol{d}_i{}'\boldsymbol{\beta})^2 \qquad (29)$$

so that $\boldsymbol{d}_i{}'$ is the i-th vector of special values from matrix $\dot{\boldsymbol{X}}\boldsymbol{X}$. There are different ideas about selecting the number of components for presence in regression model. Kasier (1960) proposed leaving the components with special values of greater than 1 in the model. Jolaiev considered $\boldsymbol{0.7\bar{\lambda}}$ as the remaining criterion. Cumulative percent of variance was also proposed by people such as Mendel, Kerzanoski, Tang, Hang, Jackson and many others. In cumulative percent method, among total variations, the first few components that have 80 to 90 percent of changes are remained in regression model. Sometimes, considering details of the model and discretion of the researcher, the first few components that make up 70% of total changes are remained in the model. Another method was proposed by Frontier (1976) and Legendre (1983) which is called Broken Stick model and is as follows:

$\boldsymbol{b_k}$ is a criterion for selecting the number of principal components in the model and is defined as follows:

$$\boldsymbol{b_k} = \sum_{i=k}^{p} \frac{1}{i} \qquad \boldsymbol{k = 1, \dots, p} \qquad (30)$$

Values of $\boldsymbol{\lambda_i s}$ are compared with the corresponding values of $\boldsymbol{b_k}$ from large to small, respectively. For the first value of k which $b_k > \lambda_k$, the comparison is stopped and special value before k, i.e. $\lambda_{k-1}, \dots, \lambda_1$, are remained in principal components regression model. Another form of b_k was also introduced by Legendre (1998) as follows:

$$b_k = \frac{1}{p}\sum_{i=k}^{p} \frac{1}{i} \qquad (31)$$

5. Monte Carlo simulation of regression variables

McDoland and Galarnio (1975), Vichern and Churchill (1978), Gibbons (1981), Kibria (2003), Al-e Hassan and Al-e Kaseb (2010) and many other researchers have made regression variable using Monte Carlo simulation as follows:

$$x_{ij} = (1-\gamma^2)^{\frac{1}{2}} z_{ij} + \gamma z_{ip+1}, \quad i = 1, \dots, n, \quad j = 1, \dots, p \qquad (32)$$

in which z_{ij} is placebo standard random variable and γ^2 is correlation coefficient between regression variables. Large, medium and small values are considered for the correlation. Values of γ^2 are considered equal to 0.81, 0.49, 0.25, 0.16, 0.09, 0.98 and 0.94. Using Eq. (32) and the γ^2 values, regression variables are made and then the produced regression variables are standardized. n observations for the dependent variable y are obtained as follows:

$$y_i = \gamma_1 x_{i1} + \gamma_2 x_{i2} + \cdots + \gamma_p x_{ip} + e_i, \quad i = 1, \dots, n \qquad (33)$$

in which e_is are placebo normal numbers which are independent from $(0, \sigma^2)$ normal. The dependent variable is also standardized.

In this project, n was equal to 50 observations and p was equal to 20. Using MATLAB software, values of x_{ij} and y_{ij} were estimated. Different methods of ridge regression were applied using Eqs. (17) to (25) for estimating value of ridge parameter. MSE values of different ridge regression methods were obtained by Eq. (16). The number of principal components for remaining in principal components regression model was obtained using Broken Stick method and MSE value of principal components regression was obtained by Eq. (29). In the following tables, results of the simulation are presented. In each table, the methods are sorted in terms of MSE value from small to large. Values of k and number of components for remaining in regression model are also given in the tables. Moreover, these tables show ratio of MSE of ridge regression methods to MSE of Broken Stick method.

Table 1.

$\gamma^2 = 0.16$				$\gamma^2 = 0.09$			
$\frac{MSE(\hat{\alpha}_k)}{MSE(\hat{\beta}_{PC})}$	MSE	k	Method	$\frac{MSE(\hat{\alpha}_k)}{MSE(\hat{\beta}_{PC})}$	MSE	k	Method
0.780851	0.38261	0.00176000	AM	0.746567	0.359291	0.00210000	GM
0.780853	0.382618	0.00106300	MED	0.746588	0.359301	0.00140000	ME
0.780894	0.382638	0.00134000	GM	0.7446588	0.359301	0.00140000	HKB
0.780918	0.38265	0.00110000	HKB	0.746648	0.35933	0.00038066	DK
0.780967	0.382674	0.00068350	DK	0.746648	0.35933	0.00038660	HK
0.780988	0.382684	0.00049513	HK	0.746661	0.359336	0.00038064	KS
0.780988	0.382684	0.00049513	KS	0.746694	0.359352	0.00010313	HSL
0.781024	0.382702	0.00001989	LW	0.746694	0.359352	0.00001858	LW
0.791228	0.387702	0.00860000	HSL	0.741335	0.359784	0.01055000	AM
1	0.49	Based on the 16 compone	PC	1	0.481254	Based on the 18 compone	PC

Table 2.

$\gamma^2 = 0.49$				$\gamma^2 = 0.25$			
$\frac{MSE(\hat{\alpha}_k)}{MSE(\hat{\beta}_{PC})}$	MSE	k	Method	$\frac{MSE(\hat{\alpha}_k)}{MSE(\hat{\beta}_{PC})}$	MSE	K	Method

0.951562	0.451950	0.0002450	HK	0.936204	0.441420	0.0084500	GM
0.941565	0.451951	0.0004607	GM	0.936267	0.44145	0.0008490	MED
0.941569	0.0004358	0.4519530	HKB	0.936282	0.441457	0.0003658	DK
0.941571	0.4519540	0.0002456	KS	0.936314	0.441472	0.0005185	AM
0.941571	0.4519540	0.0002456	DK	0.936373	0.441500	0.0003700	HK
0.941575	0.451956	0.0004798	MED	0.936373	0.441500	0.0003700	KS
0.941581	0.451959	0.0004887	AM	0.936373	0.441500	0.0003700	HKB
0.941592	0.451964	0.0001220	HSL	0.936729	0.441668	0.0001155	HSL
0.941598	0.451967	0.00006013	LW	0.944284	0.445230	0.0000123	LW
1	0.48	Based on the 16 compone	PC	1	0.471	Based on the 17 compone	PC

Table 3.

$\gamma^2 = 0.94$				$\gamma^2 = 0.81$			
$\frac{MSE(\hat{\alpha}_k)}{MSE(\hat{\beta}_{PC})}$	MSE	k	Method	$\frac{MSE(\hat{\alpha}_k)}{MSE(\hat{\beta}_{PC})}$	MSE	k	Method
0.749423	0.361297	0.0001770	GM	0.706268	0.333500	0.0000041	LW
0.749601	0.361383	0.0001784	MED	0.706692	0.33370	0.0001980	KS
0.749601	0.361383	0.0001766	HKB	0.706692	0.33370	0.0001971	HK
0.749601	0.361383	0.00017709	HK	0.706692	0.33370	0.0001971	DK
0.749601	0.361383	0.00017709	DK	0.706713	0.33371	0.0001525	HSL
0.749734	0.361447	0.0000016	KS	0.706722	0.333714	0.0002302	HKB
0.749734	0.361447	0.0001775	AM	0.857687	0.405	0.01070	GM
0.749759	0.361459	0.000003616	LW	0.945284	0.46363	0.0109603	MED
0.750773	0.3661948	0.0006257	HSL	0.945637	0.44653	0.0108074	AM
1	0.4821	Based on the 7 compone	PC	1	0.472	Based on the 14 compone	PC

Table 4.

$\gamma^2 = .9801$			
$\frac{MSE(\hat{\alpha}_k)}{MSE(\hat{\beta}_{PC})}$	MSE	K	Method
0.840976	0.41006	0.0001548	KS
0.841051	0.410096	0.0001632	MED
0.841084	0.4101126	0.0001648	GM
0.841124	0.410132	0.0001633	AM
0.841218	0.4101779	0.00015483	HKB
0.841675	0.4104009	0.00016108	HSL
0.845785	0.412405	0.00016763	DK
0.845786	0.412405	0.00016763	HK
0.855185	0.41699	0.000047	LW
1	0.4876	Based on the 14 compone	PC

6. Conclusion

6.1 Comparing different selection methods of ridge parameter

In the above tables, at all levels of correlation, $\hat{k}_{LW}$ value was smaller than other selection methods of ridge parameter and MSE value was larger than them. At all correlation levels of Kibria mean methods, two or three means had very close MSE to each other and were slightly different from other ks. In high and low correlation values, two methods of KS and HK had very close MSE value to each other. In general, MSE values of different methods of selecting k in ridge regression were close to each other.

6.2 Comparing ridge regression and principal components

According to the above tables, with increasing value of correlation coefficient, a less number of principal components were selected for remaining in principal components regression model. At all correlation levels, MSE of principal components regression was larger than that of ridge regression. In fact, based on MSE, ridge regression worked better than principal components regression. As can be seen in the tables, with increasing correlation value, MSE value of ridge regression and principal components regression got closer to each other.

References

[1] Batach, SH.M., Ramanathan, T.V., Gore, the efficiency of modified jackknife and ridge type regression estimators: *a comparsion, surveys in mathematics and its application*, 3, pp.111-122, (2008).

[2] Cadima, J., Jolliffe, I.T., variable selection and the interpretation of principal subspace, journal of agricultural, *biological and environmental statistics*,6 , pp.62-79, (2001).

[3] Chen, P., a confidence interval for the number of principal components, *journal of statistical planning and inference*,16, pp.2630-2639, (2006).

[4] Dorugade, A.V., Kashid, D.N., alternative method for choosing ridge parameter for regression, *applied mathematical science*, vol4, no9, pp.447-456, (2010).

[5] Gibbons, D.G., a simulation study of some ridge estimators, *journal of American statistical association*, 76, pp.131-139, (1981).

[6] Hoerl, E., Kennard, R.W., ridge regression: application of nonorthogonal problems, *technometrics*, vol12, no1, pp.69-82, (1970).

[7] Hoerl, E., Kennard, R.W., 1970, ridge regression: biased estimation for nonothogonal problems, *technometrics*, vol12, pp.55-67, (1970).

[8] Fathi Vajargah B and Fathi Vajargah K., Parallel Monte Carlo computation for solving SLAE with minimum communication, *Applied Mathematics and Computation* (2006), 1-9.

[9] Orsythe, S. and Liebler, Matrix Inversion by a Monte Carlo method .*Math. Tables other Aids Compul.*, 1950, 4:127-129.

[10] D.C. and Peck E. A., *Introduction to linear regression analysis*, 1991.