Available online at www.worldscientificnews.com

World Scientific News

WSN 59 (2016) 12-23 EISSN 2392-2192

Relationship between ridge regression estimator and sample size when multicollinearity present among regressors

M. C. Alibuhtto

Department of Mathematical Sciences, Faculty of Applied Sciences,
South Eastern University of Sri Lanka, Sri Lanka

E-mail address: mcabuhtto@seu.ac.lk

ABSTRACT

The problem of multicollinearity is the most common problem in multiple regression models as in such cases, the ordinary least squares (OLS) estimator is inaccurately estimated. Of many methods suggested to solve the problem of multicollinearity, ridge regression method is a one of popular method. In this paper, simulation data with different level of correlation coefficient were generated using Monte Carlo techniques in SAS. The level of multicollinearity was detected by correlation matrix, variance influence factor (VIF) and condition number. The biased parameter (k) of ridge regression has been computed by using iterative method for ordinary ridge regression in different sample sizes. According to the results of this study, it was found that biased parameter (k) of ridge regression and sample sizes are significantly negatively correlated at level of 5% significance. This study would helpful to develop biased parameter table for different level of sample sizes in present of multicollinearity.

***Keywords*:** Exponential; Multicollinearity; Variance Influence Factor; Ridge Regression; Simulation

1. INTRODUCTION

Multiple linear regressions is a widely used statistical technique that allows us to estimate models that describe the distribution of a response variable with the help of a two or more explanatory variables. The use of multiple regression mainly regards the interpretation

of the regression coefficients. In case of independent coefficients the least-squares solution gives stable estimates and useful results.

Multicollinearity is a statistical phenomenon in which there exists a perfect or exact relationship between the predictor variables. When there is a perfect or exact relationship between the predictor variables, it is difficult to come up with reliable estimates of their individual coefficients. It will result in incorrect conclusions about the relationship between outcome variable and predictor variables (Gujarati, D.N, 2004).

The presence of multicollinearity has several serious effects on the OLS estimates of regression coefficients such as high variance of coefficients may reduce the precision of estimation, it can result in coefficients appearing to have the wrong sign, the parameter estimates and their standard errors become extremely sensitive to slight changes in the data points and it tends to inflate the estimated variance of predicted values (Montgomery, 2001). Thus, multicollinearity is a serious problem in particular for predictive models, it is very important for to find a better method to deal with multicollinearity.

The method of ridge regression first introduced by Hoerl and Kennard, (1970) is nowadays established as an effective and efficient remedial method to deal with the general problems caused by multicollinearity. The main advantage of the ridge regression method is to reduce the variance term of the slope parameters. The following authors [Kibria (2003), Khalaf and Shukur (2005), Alkhamisi, Khalaf and Shukur (2006), Muniz and Kibria (2009)] were developed ridge and modified ridge parameters to solve the problem of multicollinearity. Biased estimation methods certainly compare very favorably to other methods for handling multicollinearity, such as variable elimination and variable transformation.

Furthermore, there are many biased estimators developed to solve the problems of multicollinearity. However, there is no clear evidence to use which estimates are suitable or appropriate for different size of observations. Through this research, we want to identify how the biased parameters (k) of ridge regression are associated with sample sizes of multicollinearity data.

2. METHODOLOGY

2. 1. Data

In this paper, the simulation data (Observations from 15, 20, 25, 30, 40, 50, 70, 100, 150, 200, 300, and 500) were generated using SAS software, where the correlation coefficients between the predictor variables are large ($\rho = 0.95$ *and* $\rho = 0.99$) and the number of independent variables is five. The Monte Carlo simulation procedure suggested by McDonald, G. C. and Galarneau, D. I (1975) and Gibbons (1981) was used to generate the explanatory variables:

$$X_{ij} = \left(1-\rho^2\right)^{\frac{1}{2}} Z_{ij} + \rho Z_{ip} \qquad i = 1,2,...,n \quad and \quad j = 2,..., p \qquad (1)$$

where Z_{ij} are independent standard normal distribution, ρ^2 is the correlation between any two explanatory variables and p is the number of explanatory variables.

2. 2. Detection of Multicollinearity

2. 2. 1. Examination of Correlation Matrix

A simple method for detecting multicollinearity is to calculate the correlation coefficients between any two of the explanatory variables. A high significant value of the correlation between two variables may indicate that the variables are collinear. This method is easy, but it cannot produce a clear estimate of the degree of multicollinearity (El-Dereny and Rashwan, 2011). The correlation coefficients are greater than 0.80 or above is an indication of multicollinearity (Kennedy P, 2003).

2. 2. 2. Variance Inflation Factor (VIF)

The VIF quantifies the severity of multicollinearity in an ordinary least squares regression analysis. Let R_j^2 denote the coefficient of determination when X_j is regressed on all other predictor variables X_1, X_2,..., $X_{j-1, \ldots,} X_p$ in the model. The VIF is given by:

$$VIF = \frac{1}{1-R_j^2} \qquad j = 1,2,3...p-1 \qquad (2)$$

The VIF provides an index that measures how much the variance of an estimated regression coefficient is increased because of the multicollinearity. As per practical experience, if any of the VIF values exceeds 10, it is an indication that the associated regression coefficients are poorly estimated because of multicollinearity (Montgomery, 2001).

2. 2. 3. Condition Number

The eigenvalues of the correlation matrix can also be used to measure the presence of multicollinearity. If multicollinearity is present in the predictor variables, one or more of the eigenvalues will be small (near to zero).

Let $\lambda_1, \lambda_2, ..., \lambda_p$ be the eigenvalues of correlation matrix. The condition number of correlation matrix is defined as:

$$K = \frac{\lambda_{\max}}{\lambda_{\min}} \text{ and } K_j = \frac{\lambda_{\max}}{\lambda_j} \qquad j = 1,2,..., p \qquad (3)$$

where $\lambda_{\max}$ is the largest eigenvalue.

$\lambda_{\min}$ is the smallest eigenvalue

λ_j is the eigenvalue of j^{th} independent variable

If the condition number is less than 100, there is no serious problem with multicollinearity and if a condition number is between 100 and 1000 implies a moderate to strong multicollinearity. Also, if the condition number exceeds 1000, severe multicollinearity is indicated (Montgomery, 2001).

2. 3. Ridge Regression

The ridge regression estimator is much more stable than the OLS estimator in the presence of multicollinearity. The ridge estimator restricts the length of the coefficients estimator in order to reduce the effects of multicollinearity (Hocking et al., 1976). In the presence of multicollinearity, Hoerl and Kennard (1970) introduced the ridge estimator as an alternative to the OLS estimator when the independent assumption is no longer valid.

$$\hat{\beta}_R = \left(X'X + kI\right)^{-1} X'Y \qquad (4)$$

where: I denotes an identity matrix and k is known as ridge parameter.

2. 3. 1. Determining the ridge parameter (k)

The most important case in ridge regression is determining the ridge parameter k. Researchers have been suggested different methods for determining k. Firstly, Hoerl and Kennard (1970) suggested Ridge Trace for determining k. Ridge trace is an easily applicable method. It is obtained by plotting $\hat{\beta}_R$ versus k values which are usually taken in the interval of [0, 1]. Hoerl et al. (1975) suggested another method for determining ridge parameter. According to the method k can be taken as;

$$k = \frac{p\hat{\sigma}^2}{\hat{\beta}'\hat{\beta}} \qquad (5)$$

where, p is the number of independent variables. $\hat{\sigma}^2$ and $\hat{\beta}$ are the estimations which are obtained from OLS estimation. Based on this estimation, Hoerl and Kennard suggested an iterative method for determining the ridge parameter. According to this method, iteration starts from a point (k_0) as;

$$\begin{aligned} k_0 &= \frac{p\hat{\sigma}^2}{\hat{\beta}'\hat{\beta}} \\ k_1 &= \frac{p\hat{\sigma}^2}{\hat{\beta}_R'(k_0)\hat{\beta}_R(k_0)} \\ k_2 &= \frac{p\hat{\sigma}^2}{\hat{\beta}_R'(k_1)\hat{\beta}_R(k_1)} \qquad (6) \\ &\vdots \\ k_{j+1} &= \frac{p\hat{\sigma}^2}{\hat{\beta}_R'(k_j)\hat{\beta}_R(k_j)} \end{aligned}$$

The changes in the values of i k , is used to terminate the iteration procedure. Such that; the iteration will stop, if $\frac{k_{j+1}-k_j}{k_j} < 20T^{-1.3}$

Where,

$$T = {Tr(X'X)^{-1}}/{p} \qquad (7)$$

3. RESULTS AND DISCUSSIONS

3. 1. Detection of Multicollinearity

The correlation matrix based on a set of simulated data for all conditions such as two different correlation coefficients ($\rho = 0.95$ and $\rho = 0.99$) and different sample sizes (n = 15, 20, 25, 30, 40, 50, 70, 100, 150, 200, 300 and 500) are shown the correlation between independent variables are highly correlated. This implies that the multicollinearity exits.

This results further confirmed by VIF and eigenvalues structure and the results are given in table 1 & 2.

Table 1. VIF values of independent variables.

X	VIF ($\rho = 0.95$)					
	Sample size (n)					
	15	***20***	***25***	***30***	***40***	***50***
X_1	30.155	15.218	13.386	13.221	11.620	10.097
X_2	27.043	22.891	14.651	15.256	13.854	14.479
X_3	28.563	18.978	14.699	11.866	10.058	10.326
X_4	29.097	13.172	7.964	6.807	7.291	8.492
X_5	15.383	11.812	10.395	8.790	10.699	11.660
	70	***100***	***150***	***200***	***300***	***500***
X_1	8.079	9.166	9.265	9.021	8.076	9.281
X_2	12.294	13.447	12.478	10.690	9.615	10.803
X_3	9.623	10.109	9.274	8.795	8.059	9.760
X_4	8.281	9.532	10.643	9.723	8.488	9.192
X_5	10.416	10.572	11.810	10.838	9.689	10.545

X	VIF ($\rho = 0.99$)					
	Sample size (n)					
	15	***20***	***25***	***30***	***40***	***50***
X_1	40.188	35.148	32.490	28.733	30.361	30.271
X_2	47.432	49.966	42.053	30.827	43.347	38.655
X_3	83.763	75.105	76.259	57.286	67.192	61.664
X_4	77.196	77.002	70.723	55.976	56.075	52.512
X_5	71.454	60.250	47.888	42.145	52.416	43.085
	70	***100***	***150***	***200***	***300***	***500***
X_1	35.077	39.751	48.503	53.561	49.763	48.331
X_2	43.702	40.473	38.092	43.528	44.146	46.446
X_3	64.880	51.106	55.108	61.376	57.436	51.777
X_4	52.757	45.418	45.446	46.440	44.662	45.152
X_5	45.966	39.442	38.959	42.755	43.000	43.874

Table 1 shows that almost all the VIF of most of the variables is greater than 10 in two different correlation coefficients which implies that the variables are themselves correlated.

Table 2. Results of Eigen analysis.

X	Sample size (n) for $\rho = 0.95$							
	15		***20***		***25***		***30***	
	λ_j	K_j	λ_j	K_j	λ_j	K_j	λ_j	K_j
X_1	4.819	1.00	4.768	1.00	4.694	1.00	4.663	1.00
X_2	0.100	48.06	0.103	46.45	0.126	37.33	0.144	32.37
X_3	0.033	144.25	0.053	90.14	0.081	57.94	0.079	59.01
X_4	0.029	167.72	0.042	113.52	0.054	86.34	0.067	69.14
X_5	0.018	266.67	0.034	140.25	0.045	104.17	0.047	99.30

	40		***50***		***70***		***100***	
X	λ_j	K_j	λ_j	K_j	λ_j	K_j	λ_j	K_j
X_1	4.662	1.00	4.675	1.00	4.646	1.00	4.682	1.00
X_2	0.138	33.89	0.118	39.49	0.114	40.87	0.099	46.86
X_3	0.079	59.38	0.090	52.11	0.103	44.92	0.086	54.36
X_4	0.068	68.47	0.067	68.06	0.079	58.70	0.075	62.21
X_5	0.054	86.80	0.048	96.71	0.058	80.75	0.056	82.91
	150		***200***		***300***		***500***	
X	λ_j	K_j	λ_j	K_j	λ_j	K_j	λ_j	K_j
X_1	4.690	1.00	4.666	1.00	4.625	1.000	4.671	1.00
X_2	0.095	49.42	0.0981	47.55	0.114	40.62	0.092	50.95
X_3	0.086	54.33	0.0888	52.56	0.096	48.41	0.089	52.63
X_4	0.069	67.79	0.0784	59.51	0.091	50.88	0.083	56.56
X_5	0.060	78.21	0.0683	68.34	0.074	62.28	0.066	70.44

X	Sample size (n) for $\rho = 0.99$							
	15		***20***		***25***		***30***	
	λ_j	K_j	λ_j	K_j	λ_j	K_j	λ_j	K_j
X_1	4.939	1.00	4.934	1.00	4.925	1.00	4.913	1.00
X_2	0.026	193.28	0.027	185.09	0.030	162.01	0.034	144.93
X_3	0.016	305.84	0.020	250.66	0.023	216.71	0.026	190.35
X_4	0.012	396.48	0.012	405.54	0.014	356.29	0.015	337.64
X_5	0.007	683.92	0.008	625.32	0.008	590.77	0.013	382.18
	40		***50***		***70***		***100***	
X	λ_j	K_j	λ_j	K_j	λ_j	K_j	λ_j	K_j
X_1	4.927	1.00	4.921	1.00	4.929	1.00	4.924	1.00

X_2	0.029	167.41	0.030	164.56	0.025	196.79	0.023	218.45
X_3	0.019	262.43	0.021	231.41	0.019	260.09	0.021	235.40
X_4	0.014	364.82	0.015	329.68	0.015	325.66	0.017	284.14
X_5	0.011	433.08	0.013	386.45	0.012	411.14	0.015	325.55
	150		***200***		***300***		***500***	
X	λ_j	K_j	λ_j	K_j	λ_j	K_j	λ_j	K_j
X_1	4.926	1.00	4.932	1.00	4.931	1.00	4.931	1.00
X_2	0.024	206.92	0.022	225.74	0.022	226.76	0.020	243.06
X_3	0.021	236.89	0.020	251.29	0.019	256.53	0.018	270.20
X_4	0.015	323.13	0.014	354.56	0.015	329.06	0.016	315.77
X_5	0.014	345.85	0.013	388.85	0.014	365.09	0.015	331.61

From the Table 2, the corresponding condition indices are significantly large in two different set of data. This indicates that there is multicollinearity between independent variables. According to the above results, there is multicollinearity exist in the generated independent variables. The estimates of OLS and ridge regression are given in Table 3.

Table 3. The estimates of OLS and Ridge regression ($\rho = 0.95$).

X	Sample size (n)							
	15		***20***		***25***		***30***	
	$\hat{\beta}_{OLS}$	$\hat{\beta}_R$	$\hat{\beta}_{OLS}$	$\hat{\beta}_R$	$\hat{\beta}_{OLS}$	$\hat{\beta}_R$	$\hat{\beta}_{OLS}$	$\hat{\beta}_R$
X1	-0.390	0.095	0.044	0.138	0.132	0.162	0.107	0.148
X2	-0.064	0.159	-0.020	0.151	0.117	0.170	0.080	0.155
X3	0.319	0.227	0.333	0.225	0.297	0.225	0.361	0.260
X4	0.543	0.189	0.223	0.183	0.118	0.155	0.117	0.142
X5	0.581	0.281	0.409	0.265	0.316	0.245	0.317	0.255
X	***40***		***50***		***70***		***100***	
	$\hat{\beta}_{OLS}$	$\hat{\beta}_R$	$\hat{\beta}_{OLS}$	$\hat{\beta}_R$	$\hat{\beta}_{OLS}$	$\hat{\beta}_R$	$\hat{\beta}_{OLS}$	$\hat{\beta}_R$
X1	0.299	0.239	0.196	0.192	0.187	0.189	0.196	0.196

X2	0.085	0.154	0.115	0.163	0.170	0.181	0.235	0.220
X3	0.232	0.214	0.293	0.245	0.264	0.238	0.296	0.269
X4	0.043	0.098	0.128	0.150	0.156	0.164	0.175	0.176
X5	0.322	0.263	0.244	0.213	0.198	0.193	0.068	0.102
X	***150***		***200***		***300***		***500***	
	$\hat{\beta}_{OLS}$	$\hat{\beta}_R$	$\hat{\beta}_{OLS}$	$\hat{\beta}_R$	$\hat{\beta}_{OLS}$	$\hat{\beta}_R$	$\hat{\beta}_{OLS}$	$\hat{\beta}_R$
X1	0.251	0.241	0.213	0.209	0.207	0.205	0.188	0.189
X2	0.202	0.199	0.177	0.180	0.139	0.145	0.182	0.183
X3	0.207	0.204	0.212	0.209	0.261	0.254	0.229	0.227
X4	0.214	0.208	0.179	0.181	0.148	0.152	0.182	0.183
X5	0.106	0.124	0.193	0.192	0.221	0.217	0.198	0.197

Table 4. The estimates of OLS and Ridge regression ($\rho = 0.99$)

X	Sample size (n)							
	15		***20***		***25***		***30***	
	$\hat{\beta}_{OLS}$	$\hat{\beta}_R$	$\hat{\beta}_{OLS}$	$\hat{\beta}_R$	$\hat{\beta}_{OLS}$	$\hat{\beta}_R$	$\hat{\beta}_{OLS}$	$\hat{\beta}_R$
X1	0.299	0.216	0.165	0.174	0.137	0.165	0.118	0.161
X2	0.252	0.189	0.405	0.238	0.382	0.237	0.207	0.195
X3	-0.312	0.108	-0.109	0.145	-0.106	0.148	0.045	0.158
X4	0.455	0.231	0.321	0.214	0.368	0.223	0.339	0.240
X5	0.297	0.236	0.205	0.207	0.208	0.206	0.282	0.229
X	***40***		***50***		***70***		***100***	
	$\hat{\beta}_{OLS}$	$\hat{\beta}_R$	$\hat{\beta}_{OLS}$	$\hat{\beta}_R$	$\hat{\beta}_{OLS}$	$\hat{\beta}_R$	$\hat{\beta}_{OLS}$	$\hat{\beta}_R$
X1	0.086	0.146	0.013	0.076	0.085	0.118	0.210	0.205
X2	0.239	0.197	0.239	0.221	0.222	0.222	0.267	0.250
X3	0.047	0.131	0.141	0.174	0.317	0.268	0.173	0.180
X4	0.314	0.267	0.191	0.202	0.098	0.139	0.143	0.157

X5	0.309	0.247	0.409	0.319	0.270	0.245	0.203	0.202
X	***150***		***200***		***300***		***500***	
	$\hat{\beta}_{OLS}$	$\hat{\beta}_R$	$\hat{\beta}_{OLS}$	$\hat{\beta}_R$	$\hat{\beta}_{OLS}$	$\hat{\beta}_R$	$\hat{\beta}_{OLS}$	$\hat{\beta}_R$
X1	0.148	0.158	0.174	0.178	0.171	0.173	0.248	0.245
X2	0.212	0.209	0.214	0.212	0.257	0.251	0.208	0.206
X3	0.170	0.176	0.192	0.193	0.159	0.165	0.139	0.144
X4	0.222	0.216	0.215	0.212	0.217	0.215	0.172	0.174
X5	0.243	0.235	0.200	0.200	0.190	0.191	0.228	0.226

Tables 3 & 4 show the parameter estimates of both OLS and ridge regression.

3. 2. Biased parameter versus sample sizes

The biased ridge parameter (k) computed using iterative method for different sample sizes and it is given in Table 5.

Table 5. Results of k and sample sizes.

Sample Size (n)	Ridge biased estimator (K)	
	$\rho = 0.95$	$\rho = 0.99$
15	0.152600	0.081575
20	0.129838	0.067876
25	0.111436	0.047451
30	0.098192	0.035876
40	0.065808	0.015832
50	0.063703	0.012950
70	0.046216	0.009749
100	0.030769	0.006170
150	0.017389	0.003962
200	0.015873	0.003019
300	0.010364	0.001972
500	0.005449	0.001223

Table 5 shows that the biased ridge parameter significantly decreased by its sample sizes in two different correlation.

Table 6. Correlation coefficient of k and sample size

Correlation Coefficient	K for $\rho = 0.95$	K for $\rho = 0.99$
Sample size (n)	-0.729 (0.007)	-0.584 (0.046)

Table 6 shows that the correlation between biased parameter and sample sizes are negatively correlated and it is significant at 5% level of significance.

4. CONCLUSIONS

This paper is an attempt to explore the relationship between biasing parameter of biased estimators and sample size when the multicollinearity present among the regressors. This study has been done using Monte Carlo simulation data, where levels of correlation and the sample sizes have been varied. According to the results of this study the multicollinearity was detected using calculating the variance inflation factor (VIF) and Eigen value analysis. Also, it was found that biased parameter (k) of ridge regression and sample sizes are significantly negatively correlated at level of 5% significance.

References

[1] Alkhamisi, M., Khalaf, G. and Shukur, G. (2006). Some modifications for choosing ridge parameters. *Communications in Statistics - Theory and Methods*, 35(11), 2005-2020.

[2] El-Dereny, M., Rashwan, N. I. (2011). Solving multicollinearity problem using ridge regression models, *Int. J. Contemp. Math. Sci.* 6, No. 9-12, 585-600.

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

[4] Gujarati, D. N. (2004). Basic Econometrics, 4th edition, Tata McGraw-Hill, New Delhi.

[5] Hocking, R. R., Speed, F. M. and Lynn, M. J. (1976). A class of biased estimators in linear regression. *Technometrics*, 18, 425-438.

[6] Hoerl, A. E., Kennard, R. W. and Baldwin, K. F. (1975). Ridge regression: some simulation. *Communications in Statistics*, 4, 105-123.

[7] Hoerl, A. E. and Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Non-orthogonal Problems Regression Analysis and Biased Estimation. *Technometrics*, pp. 55-67.

[8] Kennedy, P. (2003). A Guide to Econometrics, 5th edition, The MIT Press, Cambridge.

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

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

[11] McDonald, G. and Galarneau,D., (1975). A Monte Carlo evaluation of some ridge type estimators. *Journal of the American Statistics Association*, 70 (1975), 407-416.

[12] Montgomery, D. C., Peck, E. A., Vining, G. G. (2001). Introduction to linear regression analysis, 3rd edition, Wiley, New York.

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

(Received 24 October 2016; accepted 07 November 2016)