Electronic Journal of Applied Statistical Analysis
EJASA, Electron. J. App. Stat. Anal.
`http://siba-ese.unisalento.it/index.php/ejasa/index`
e-ISSN: 2070-5948
DOI: 10.1285/i20705948v7n2p343

A note on ridge regression modeling techniques

By Yahya W.B., Olaifa J.B.

Published: 14 October 2014

Electronic Journal of Applied Statistical Analysis
Vol. 07, Issue 02, 2014, 343-361
DOI: 10.1285/i20705948v7n2p343

A note on ridge regression modeling techniques

Waheed Babatunde Yahya*and Julius Babatunde Olaifa

Department of Statistics, University of Ilorin, P.M.B. 1515, Ilorin, Nigeria

Published: 14 October 2014

In this study, the techniques of ridge regression model as alternative to the classical ordinary least square (OLS) method in the presence of correlated predictors were investigated. One of the basic steps for fitting efficient ridge regression models require that the predictor variables be scaled to unit lengths or to have zero means and unit standard deviations prior to parameters' estimations. This was meant to achieve stable and efficient estimates of the parameters in the presence of multicollinearity in the data. However, despite the benefits of this variable transformation on ridge estimators, many published works on ridge regression practically ignored it in their parameters' estimations. This work therefore examined the impacts of scaled collinear predictor variables on ridge regression estimators. Various results from simulation studies underscored the practical importance of scaling the predictor variables while fitting ridge regression models. A real life data set on import activities in the French economy was employed to validate the results from the simulation studies.

keywords: Ridge regression, orthogonality, shrinkage parameter, scaling, ordinary least squares, mean square error

1 Introduction

The simplest but efficient way to fit (multiple) linear regression models is through the ordinary least squares (OLS) method. This is particularly true when all the necessary assumptions underlying its application are met by the data. One of these assumptions required the predictor variables in the regression models to be purely uncorrelated, see

*Corresponding author: wb_yahya@daad-alumni.de

ISSN: 2070-5948
http://siba-ese.unisalento.it/index.php/ejasa/index

Myers (1986).
Consider a multiple linear regression model of the form

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon \tag{1}$$

where $\mathbf{Y}$ is the $n \times 1$ vector of responses, $\mathbf{X}$ is the $n \times p$ matrix of predictor variables, β is the $p \times 1$ vector of the regression coefficients while ϵ is the random noise of the model that is assumed to have Gaussian density with zero mean and a constant variance σ^2. The goal of the OLS is to minimize the error sum of squares $\epsilon'\epsilon = (\mathbf{Y} - \beta\mathbf{X})'(\mathbf{Y} - \beta\mathbf{X})$ to yield the OLS estimators

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

If some of the predictor variables in matrix $\mathbf{X}$ are correlated, the OLS estimators in (2) become less efficient and unstable, thereby rendering the resulting regression model unsuitable for meaningful inference. However, it is not uncommon in observational studies to find some of the $\mathbf{X}$ predictors to be correlated, see Yahya et al (2008). The problem of collinear predictors only becomes severe on OLS estimators when such correlations go beyond a reasonable tolerable range of values as proposed by Yahya et al (2008).
Among the earlier methods proposed in the literature to remedy the adverse effects of collinear predictors on OLS estimators is the ridge regression, see Hoerl and Kernard (1970a); Hoerl and Kernard (1970b). The ridge regression is a regression technique that allows for biased estimation of regression parameters that are quite close to the true values in the presence of correlated predictor variables in the model. All the various forms of the ridge regression techniques were meant to shrink the least square coefficients towards the origin of the parameter space and consequently reduce the mean square errors of estimates. As a result, the ridge estimators mostly yield better mean square errors than the classical OLS estimators, see Dorugade and Kashid (2010).
One of the basic procedures in ridge regression estimation, as adopted in many studies, required that the predictor variables (columns of matrix $\mathbf{X}$) be scaled to unit lengths or with zero means and unit standard deviations, seeLawless and Wang (1976); Mardikyan and Cetin (2008). This is meant to avoid over-fitting (fitting to noise in the data rather than the signal) and achieve stable estimates of the ridge regression parameters, see Cannon (2009).
Despite the benefits of scaling of the predictor matrix $\mathbf{X}$ in ridge regression estimation as demonstrated in many works, see Hoerl et al (1985); Wethril (1986); Fearn (1993); Khalaf and Shukur (2005) and Mardikyan and Cetin (2008), this desirable step was blatantly ignored in a number of studies where practical applications of ridge regression methods were presented, see Longley (1976); Myers (1986); Chatterjee and Hadi (2006). It was against this background that this present work is motivated to illustrate, with clear examples, the fundamental basis of scaling the predictor matrix in ridge regression estimation. This is aimed to guide the researchers and students alike in their future applications of the ridge regression methods.

2 Materials and Methods

2.1 *Brief Overview of Ridge Regression Method*

Consider the multiple linear regression model given by equation (3). As earlier remarked, if some pairs of predictor variables in the columns of the design matrix $\mathbf{X}$ are correlated, the OLS estimator in (2) becomes very unstable and less efficient resulting into high estimates of the mean square errors. To remedy this, Hoerl and Kernard (1970a) proposed an alternative estimator by adding a constant value k to the diagonal of $\mathbf{X}'\mathbf{X}$ matrix in the OLS estimator (2). This resulted into the ridge estimator of the form

$$\hat{\beta}^* = (\mathbf{X}'\mathbf{X} + kI)^{-1}(\mathbf{X}'\mathbf{Y}); \quad k > 0 \tag{3}$$

where k is the ridge penalty (shrinkage) parameter and I is a $p \times p$ identity matrix. The value of $k > 0$ was meant to shrink the magnitude of the estimated regression coefficients which would eventually lead to fewer effective model parameters, seeCannon (2009). It should be noted that the ridge estimator in (3) reduces to OLS estimator in (2) when $k = 0$.

2.2 *Assessment of Model's Performance*

The performance of the ridge regression estimators (3) can be assessed through the classical mean square error (MSE) of the estimated regression coefficients given by

$$\begin{aligned} & MSE_1 = E(\hat{\beta}^* - \beta)(\hat{\beta}^* - \beta)' \\ \Rightarrow \quad & MSE_1 = \tfrac{1}{p}\textstyle\sum_{i=1}^{p}(\hat{\beta}_i{}^* - \beta_i)^2 \end{aligned} \tag{4}$$

However, Hoerl and Kernard (1970a); Hoerl and Kernard (1970b) proposed another form of MSE to assess the performance of ridge estimators (3). This MSE is given by

$$MSE_2 = \sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i + k)^2} + k^2 \sum_{i=1}^{p} \frac{\hat{\alpha}_i{}^2}{(\lambda_i + k)^2} \tag{5}$$

Here, $\hat{\alpha} = \mathbf{P}'\hat{\beta}^*$, where $\mathbf{P}$ is the $p \times p$ matrix of eigenvectors satisfying $\mathbf{X}'\mathbf{X} = \mathbf{P} \wedge \mathbf{P}'$ with $\mathbf{PP}' = I$ and $\hat{\alpha} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_p)$. Matrix $\wedge$ is a diagonal matrix of the eigenvalues $\lambda_1, \ldots, \lambda_p$ with $\lambda_1 \geq \ldots \geq \lambda_p$.

The first component $\sigma^2 \sum_{i=1}^{p} \dfrac{\lambda_i}{(\lambda_i + k)^2}$ in (5) represents the variance of all the estimated regression coefficients while the second component represents the corresponding bias square. The whole idea is to develop a scheme that selects the shrinkage parameter k such that decrease in variance does not increase the bias of the ridge estimators. However, the mean square error for ridge estimators, MSE2 in (5) reduces to that of the OLS when the shrinkage parameter $k = 0$.

A major difference between the two mean square errors in (4) and (5) is that the MSE_1 in (4) assumes that the true parameter values in vector β are known prior to model's estimation and these are simply being compared with their corresponding estimates in parameter vector $\hat{\beta}^*$. In the contrary however, the computation of MSE_2 in (5) only requires the estimated regression parameter vector $\hat{\beta}^*$ along with the eigenvalues λ_i which are all obtained from the sample data. In a nutshell, MSE_1 is only suitable to assess the performance of the (ridge) regression estimators with simulated data in which the true parameter values of the model in vector β have been determined a priori, whereas, MSE_2 can be used to assess the performance of (ridge) estimators with both simulated and real life data sets.

2.3 *The Choice of Shrinkage Parameter* k

One of the challenges of the ridge regression in the literature is how to determine the optimal value of the shrinkage (tuning) parameter k that would yield the most efficient ridge regression models, see Hoerl et al (1975);Khalaf and Shukur (2005); Dorugade and Kashid (2010). When $k = 0$, the ridge estimator (3) reduces to the OLS estimator (2). According to Hoerl and Kernard (1970a); Hoerl and Kernard (1970b), and Faraway (2002), the reasonable values of the tuning parameter k lies within the interval $(0, 1)$ especially when each variable column in predictor matrix $\mathbf{X}$ is scaled to unit length. In the present study therefore, the best value of k within the interval $(0, 1)$ that yields the most efficient ridge parameter estimates in any given data set is determined by cross-validation search using model's assessment criteria of MSE_1 or MSE_2. By this cross-validation search criteria, the value of k_1 or k_2 ($k_1, k_2 \in k$) within the interval $(0, 1)$, that yielded the least estimated mean square error MSE_1 in (4) or MSE_2 in (5) out of a number of such estimates obtained for all the possible values of k in the interval $(0, 1)$ becomes the best tuning parameter value for the ridge regression estimators for such data. Based on these two criteria for determining the best value of k, two forms of ridge regression estimators RR_1 or RR_2, as used in this work, evolved depending on whether MSE_1 or MSE_2 has been employed to determine the best shrinkage parameter k_1 or k_2 respectively, $k_1, k_2 \in k$.

2.4 *Centering and Scaling of Correlated Predictor Variables*

As earlier remarked, one of the major steps in ridge regression estimation required that the predictor variables be centered to have zero means and scaled to unit lengths, see Mardikyan and Cetin (2008). Two types of scaling are suggested in the literature, seeHoerl et al (1975). The first one is the unit length scaling that ensures that each column in predictor matrix $\mathbf{X}$ has a zero mean and unit length. The statistic is given by

$$X^*_{ij} = \frac{x_{ij} - \tilde{x}_j}{\sqrt{\sum (x_{ij} - \tilde{x}_j)^2}} \tag{6}$$

for $i = 1, \ldots, p$ and $j = 1, \ldots, n$.
The second scaling method standardizes each column in the predictor matrix $\mathbf{X}$ to have zero mean and a unit standard deviation. Its statistic is given by

$$Z^*_{ij} = \frac{x_{ij} - \tilde{x}_j}{\sqrt{\sum \frac{(x_{ij} - \tilde{x}_j)^2}{n-1}}} \tag{7}$$

Whenever the scaling statistic (6) is used, the resulting $\mathbf{X}'\mathbf{X}$ matrix becomes the correlation matrix of the predictor variables. Nonetheless, both methods of scaling have been found to perform excellently well. However, statistic (7) as implemented in *R* statistical package *(www.cran.org)* is adopted in this study.

2.5 *Simulation Studies*

The purpose of the simulation work is to compare the relative performance of OLS and ridge estimators with respect to their MSEs when the predictor matrix $\mathbf{X}$ is $i.$ scaled or standardized and $ii.$ unscaled.
Data were simulated in line with multiple linear regression model given in (1). A $n \times p$ data matrix of $p = 6$ predictor variables $x_1, \ldots, x_6$ each of size $n = 20$ was simulated from multivariate Gaussian density with mean vector $\mu = (102, 390, 310, 260, 115, 1600)$. In order to ensure some form of dependency among the pairs of predictor variables, the correlation matrix $\rho_{ii'}$ in **?** was adapted to simulate the covariance structures $\sigma_{ii'}$ between the pairs of predictors x_i and $x_{i'}$, $i \neq i' = 1, \ldots, 6$. This is given by

$$\rho_{ii'} = \begin{bmatrix} 1.0000 & & & & & \\ 0.9916 & 1.0000 & & & & \\ 0.6206 & 06043 & 1.0000 & & & \\ 0.4647 & 0.4464 & \text{-}0.1774 & 1.0000 & & \\ 0.9792 & 0.9911 & 0.6866 & 0.3644 & 1.0000 & \\ 0.9911 & 0.9953 & 0.6683 & 0.4172 & 0.9940 & 1.0000 \\ & & & & & \end{bmatrix} \quad (8)$$

This finally yielded the following variance-covariance matrix $\sum$ as used for simulation.

$$\sum = \begin{bmatrix} 125.00 & & & & & \\ 1084.32 & 10555.00 & & & & \\ 957.00 & 8711.29 & 8000.00 & & & \\ 698.99 & 6516.83 & 5516.08 & 4250.00 & & \\ 243.00 & 2226.07 & 1946.00 & 1403.80 & 500.00 & \\ 261.62 & 2468.27 & 2189.11 & 1527.12 & 542.25 & 625.00 \end{bmatrix} \quad (9)$$

Finally, the response variable Y_i was simulated from the relationship

$$Y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \epsilon_i \quad (10)$$

where the values in the parameter vector $\beta' = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6)$ were respectively set at $(25, 10, 15, 60, 40, 35, 20)$ while $\epsilon_i \sim IIDN(0, \sigma_\epsilon^2)$ is the error term of the model with σ_ϵ^2 fixed at 25. Therefore, all the results and discussions on simulation studies are based on the regression model in (10).

3 Results

The results of the simulations and applications of OLS and ridge estimators on published life data set are presented in this section.

3.1 *Simulation Results*

In order to have a quick overview of the impacts of scaling correlated predictor variables in linear regression estimation, we provide in Table 1, the OLS estimates of the regression parameters for model (10) using the simulated six predictor variables which were all scaled to have zero means and unit standard deviations according to statistic (7). Also, the OLS estimates of the model's parameters using the original (unscaled) simulated predictors were equally reported in the table. Both results were provided for ten models over ten different simulated data sets.
It is observed from the regression results in Table 1 that, except for the intercept parameter β_0, the OLS estimates of the slope parameters $\beta_1, \ldots, \beta_6$ in all the ten models are apparently stable using the raw (unscaled) values of the predictor variables given the inherent multicollinearity structure in the data. Thus, the instability of regression parameters which is one of the major consequences of multicollinearity is more pronounced on the intercept parameters than on the slope parameters in OLS estimation.
Using the raw values of the predictor variables $x_1, \ldots, x_6$, the OLS estimates of β_0 in the ten models in Table 1 were grossly unstable ranging from -805.081 to 2257.230 with a variance of 1025844. Whereas, using the standardized values of the predictor variables, the OLS estimates of β_0 in all the models were apparently stable and very close to the true value of 25. Here, the estimated $\hat{\beta}_0$ in all the ten models fall within the interval $[23.872, 25.633]$ with a relatively small variance of 0.4. These results simply confirmed that scaling of the predictor variables helps to stabilize the estimates of the intercept parameter in multiple linear regression modelling in the presence of highly correlated predictors as earlier posited by Marquardt and Snee (1975); Bradly and Srivastava (1997), even with OLS estimators.
Another important feature observed in the results in Table 1 is that, the OLS estimates of the slope parameters β_3 and β_4 of pair of orthogonal (uncorrelated) predictor variables x_3 and x_4 are apparently stable across the ten models using either the raw or standardized values of the predictors for estimation.

The collinear structure between x_3 and x_4 in all the simulated data was set at correlation value, ρ_{x_3,x_4} of -0.1774 which is not significant at 5% ($p \approx 0.4$), an indication that the predictor variables x_3 and x_4 are purely uncorrelated. In all the ten models in Table 1, the OLS estimates of the slope parameters β_3 and β_4 of the two orthogonal predictors x_3 and x_4 are very close to their true values of 60 and 40 respectively, even in the presence of other highly correlated predictor variables in the models.
Without loss of generality, it can be deduced from the results in Table 1 that in the presence of multicollinearity, scaling of predictor variables only helps to stabilize the OLS estimates of the intercept parameters while the slope parameters, especially those of the correlated predictors would be grossly unstable and inefficient for meaningful inferences.

Table 1: The OLS estimates of the regression parameters for ten simulated data sets using scaled and original (unscaled) predictor variables. It is observed that the estimates of the intercept parameter β_0 in all the models were stable and closer to their true values using the scaled (standardized) predictor variables than their estimated values using the original (unscaled) predictor variables.

Model	**Predictor Variable**	$\beta_0 = 25$	$\beta_1 = 10$	$\beta_2 = 15$	$\beta_3 = 60$	$\beta_4 = 40$	$\beta_5 = 35$	$\beta_6 = 20$
1	Unscaled	−694.696	10.531	14.749	59.953	39.982	35.527	20.452
	Scaled	25.556	15.900	−11.306	55.700	38.952	47.423	31.724
2	Unscaled	−151.070	12.254	14.602	59.950	39.964	35.852	20.017
	Scaled	24.123	28.083	−13.955	55.643	38.226	49.059	20.319
3	Unscaled	−576.140	9.675	15.049	60.020	40.011	34.422	20.421
	Scaled	25.633	6.603	19.506	62.045	40.918	22.829	29.645
4	Unscaled	−805.081	10.239	14.956	59.945	39.950	34.688	20.556
	Scaled	25.424	12.144	11.342	55.045	36.762	29.010	31.834
5	Unscaled	364.417	10.173	15.030	59.982	39.988	35.125	19.766
	Scaled	24.324	11.979	18.146	58.578	39.194	37.880	14.045
6	Unscaled	1282.708	10.083	15.147	60.018	39.993	35.170	19.158
	Scaled	24.702	10.984	31.297	61.586	39.572	39.172	−2.815
7	Unscaled	2257.230	9.084	15.705	60.104	40.055	33.532	18.567
	Scaled	23.978	−0.820	90.060	69.707	42.537	0.790	−17.409
8	Unscaled	836.714	9.638	15.085	60.038	40.039	35.222	19.465
	Scaled	24.747	6.589	22.419	63.390	42.537	39.440	8.258
9	Unscaled	34.618	8.536	15.273	60.042	40.013	34.258	20.063
	Scaled	23.872	−6.673	42.642	64.709	40.870	18.190	21.611
10	Unscaled	−713.653	10.989	14.593	59.961	40.012	35.984	20.432
	Scaled	25.013	23.776	−37.354	56.298	41.011	61.975	33.469

One of the alternative efficient techniques to estimate multiple linear regression models in the presence of multicollinearity is offered by the ridge regression estimators, see EI-Dereny and Rashwan (2011) and Hoerl et al (1975), the results of which are presented in Table 2 for ridge regression type RR_1 and RR_2 as earlier described in Section 2.2 for various simulated data according to the simulation scheme in Section 4.1. For each simulated data set, the regression results of the OLS estimators are also presented in the table. However, regression results of OLS, RR_1 and RR_2 estimators for ten different simulated data sets are presented in Table 2 due to space. The three regression types were fitted using the scaled predictor variables according to (7).
For each simulated data, the best shrinkage parameter values k_1 and k_2 for ridge regression estimators RR_1 and RR_2 were determined by cross-validation from 1000 possible values of the shrinkage parameter k within the interval $(0, 1)$ using the respective mean square errors MSE_1 and MSE_2 as described in Section 2.2. The regression coefficients of the OLS estimators in Table 2 were obtained at value of $k = 0$ in all cases. For more understanding of how the best shrinkage parameter values k_1 and k_2 were determined for the two ridge regression estimators RR_1 and RR_2, the plot of the all the ridge regression parameter estimates at various possible values of the shrinkage parameter k within the interval $(0, 1)$ for RR_1 and RR_2 models are obtained as shown in Fig 1 and Fig 2. The values of MSE_1 and MSE_2 yielded by OLS estimator in each data are equally reported in Table 2. The plots of the graphical display of how the optimal shrinkage parameter estimates, k_1 and k_2, of the two ridge estimators (RR_1 and RR_2) were determined are presented by Fig 1 and Fig 2. In both graphs, the best ridge regression models that yielded the least MSE_1 and MSE_2 values were obtained at $k(k_1) = 0.048$ and $k(k_2) = 0.051$ for RR_1 and RR_2 estimators respectively over 1000 cross-validation search for the best value of k.

Various results in Table 2 indicated that, with scaled predictor variables, the two ridge estimators RR_1 and RR_2 are more efficient than the OLS estimators. In all the results, the estimated MSE_1 and MSE_2 values of the two ridge estimators are relatively smaller than that of the OLS estimators.
It is very instructive to remark that, the better performance of the ridge estimator as demonstrated by RR_1 and RR_2 estimators over OLS depends largely on the degree of multicollinearity in the data. If not all the predictor variables in a multiple linear regression model are correlated, the OLS estimator might still be efficient by some chance factor. In the present study, two (x_3 and x_4) of the six simulated predictor variables are purely uncorrelated ($p \approx 0.4$) while the remaining four predictors are significantly correlated ($p < 0.05$). For data sets with this kind of multicollinearity structure, about three out of ten OLS models fitted to such data would still be efficient despite the presence of multicollinearity as shown by the results in Table 3. Out of between 500 and 10000 data sets simulated, the OLS and the ridge (RR_2) estimators have average relative efficiencies of about 28% and 72% respectively. The relative efficiency (RE) of an estimator, in this context, is determined by the proportion of times (expressed in percentages in parenthesis) the estimator yielded the best models (i.e. the least mean square error, MSE_2) out of the total number of the fitted models. It is observed (results not shown) that the

Table 2: The regression results of the ordinary least squares (OLS) estimator, ridge regression estimator 1 (RR_1) and estimator 2 (RR_2) for ten simulated data sets. The estimated regression parameters of the models, their mean square errors (MSE_1 and MSE_2) and the values of the shrinkage parameter k (k_1 for RR_1 and k_2 for RR_2) that yielded the best ridge regression models (the least MSE_1 or MSE_2) in each data are presented. The value of $k = 0$ for all the OLS estimators. The mean square errors of the OLS estimators for MSE_1 and MSE_2 are also reported.

Data	Model Type	k	$\beta_0(25)$	$\beta_1(10)$	$\beta_2(15)$	$\beta_3(60)$	$\beta_4(40)$	$\beta_5(35)$	$\beta_6(20)$	MSE$_1$	MSE$_2$
1	OLS	0.000	24.484	8.262	7.537	49.853	45.090	49.398	20.146	56.455	5321.1
	RR1	0.099	24.484	13.147	11.356	46.767	41.770	44.153	23.010	42.075	–
	RR2	0.058	24.484	11.424	9.968	47.895	42.975	45.958	22.022	–	1539.9
2	OLS	0.000	23.442	13.443	16.640	70.332	38.999	32.895	7.658	40.212	5321.1
	RR1	0.113	23.442	17.932	16.818	61.488	37.683	31.033	14.898	16.858	–
	RR2	0.042	23.442	15.444	16.536	66.394	38.555	31.997	11.005	–	1797.3
3	OLS	0.000	25.688	5.162	9.454	66.436	44.408	40.878	11.400	32.002	5321.1
	RR1	0.132	25.688	11.852	12.292	57.995	40.962	36.276	18.248	2.982	–
	RR2	0.036	25.688	7.462	10.267	63.553	43.355	39.199	13.880	–	1928.2
4	OLS	0.000	25.177	13.274	27.155	56.448	31.386	34.185	15.089	38.586	5321.1
	RR1	0.011	25.177	13.716	26.974	55.630	31.426	33.931	15.849	38.313	–
	RR2	0.133	25.177	17.227	26.082	49.397	31.319	32.121	21.246	–	1036.9
5	OLS	0.000	23.289	10.734	21.506	67.273	32.465	32.465	14.879	26.694	5321.1
	RR1	0.065	23.289	13.431	21.040	62.090	32.261	31.608	19.089	18.256	–
	RR2	0.061	23.289	13.291	21.051	62.357	32.281	31.661	18.883	–	1497.8
6	OLS	0.000	27.080	6.246	24.816	69.697	37.208	28.276	15.171	40.733	5321.1
	RR1	0.100	27.080	10.452	24.035	61.787	36.369	27.507	21.155	22.860	–
	RR2	0.048	27.080	8.521	24.258	65.353	36.864	27.792	18.577	–	1497.8
7	OLS	0.000	24.732	9.292	10.529	67.848	40.544	33.555	17.005	13.357	5321.1
	RR1	0.074	24.732	12.716	11.818	62.599	38.842	31.937	19.535	5.096	–
	RR2	0.045	24.732	11.532	11.305	64.422	39.479	32.459	13.357	–	1674.7
8	OLS	0.000	23.907	17.691	2.230	43.535	51.477	32.204	33.957	118.41	5321.1
	RR1	0.154	23.907	21.353	9.568	42.339	44.846	30.346	32.413	95.820	–
	RR2	0.078	23.907	19.858	6.617	43.048	47.535	30.985	32.936	–	1632.4
9	OLS	0.000	26.726	12.411	29.279	61.321	36.201	19.412	21.457	67.710	5321.1
	RR1	0.042	26.726	13.616	28.591	58.274	36.009	20.014	23.519	65.235	–
	RR2	0.085	26.726	14.636	28.173	55.815	35.714	20.579	25.045	–	1207.3
10	OLS	0.000	24.412	7.790	9.900	58.522	39.750	42.391	21.916	13.111	5321.1
	RR1	0.039	24.412	9.867	11.049	56.454	38.666	40.645	23.554	10.689	–
	RR2	0.060	24.412	10.818	11.615	55.495	38.143	39.877	24.267	–	1426.1

1.pdf

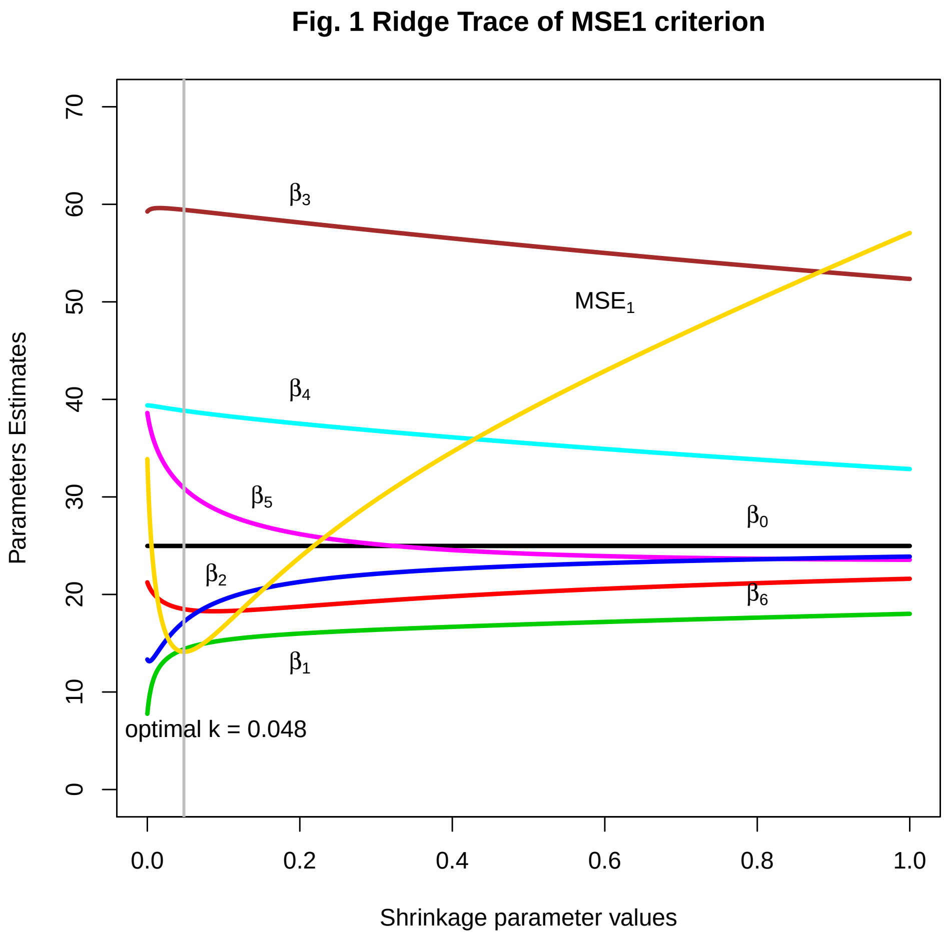

Figure 1: **The plots of the ridge regression parameter estimates at various possible values of the shrinkage parameter k within the interval $(0, 1)$ for RR_1 model. The best ridge regression model that yielded the least MSE_1 value was obtained at $k(k_1) = 0.048$ over 1000 cross-validation search.**

2.pdf

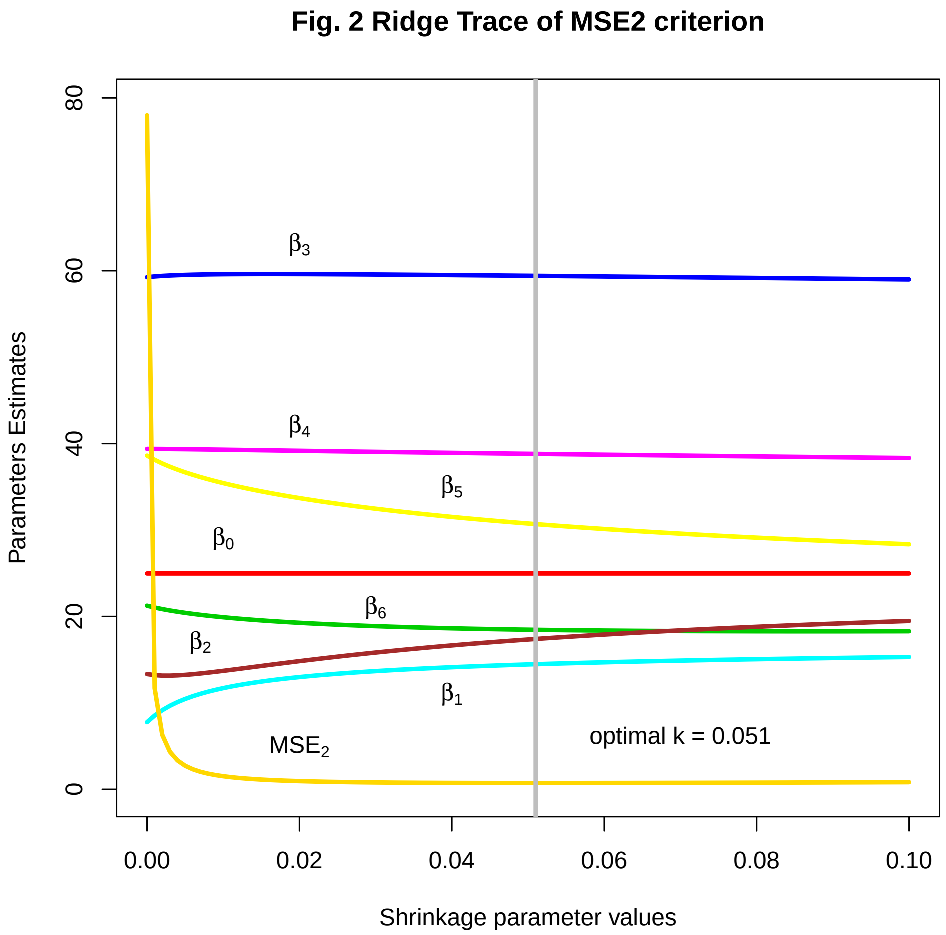

Figure 2: **The plots of the ridge regression parameter estimates at various possible values of the shrinkage parameter k within the interval $(0, 1)$ for RR_2 model. The best ridge regression model that yielded the least MSE_2 value was obtained at $k(k_2) = 0.051$ over 1000 cross-validation search.**

Table 3: The table shows the number (percentage) of best regression models (models with the least MSE_2 values) yielded by the ordinary least squares (OLS) and ridge regression (RR_2) estimators out of a number of fitted models (from 500 to 10,000 as indicated in the first column). At each iteration (number of models fitted), the relative efficiency (RE) of each estimator is determined by the proportion of times (expressed in % in the parenthesis) the estimator yielded the least mean square error (MSE_2) out of the total number of fitted models.

Number of fitted Models (Iteration)	**Number (%) of best models yielded by OLS and Ridge estimators using MSE_2 criteria**	
	OLS	Ridge (RR_2)
500	143(28.6%)	357(71.4%)
1000	254(25.4%)	746(74.6%)
1500	418(27.9%)	1082(72.1%)
2000	537(26.9%)	1463(73.1%)
2500	685(27.4%)	1815(72.6%)
3000	873(29.1%)	2127(70.9%)
3500	976(27.9%)	2524(72.1%)
4000	1140(28.5%)	2860(71.5%)
4500	1243(27.6%)	3257(72.4%)
5000	1384(27.7%)	3616(72.3%)
10000	2762(27.6%)	7238(72.4%)
Average % of RE	27.69%	72.31%

fewer the number of collinear predictors (i.e. the lower the degree of multicollinearity) in the model, the higher the RE score of the OLS estimator and vice-versa. However, the results of the ridge estimators become that of the OLS when all the predictor variables in the model are purely uncorrelated. Without loss of generality, the OLS estimator appears optimistic in the presence of some levels of multicollinearity in the model, the chance that this estimator would yield efficient results on data with correlated predictors is relatively small, less than 30% in this case. Therefore, it is strongly recommended that whenever multicollinearity is suspected in a data set, an alternative robust estimator like the ridge should be employed to guarantee the reliability of the results obtained for meaningful inference.

The MSE_2 in (5) was chosen to assess the performance of the regression models in Table 3 mainly for practical purposes, since it is the most appropriate model's assessment criterion for real life data among the two MSEs (MSE_1 and MSE_2) employed here. This consequently informed the choice of the ridge regression estimator, RR_2 against which the results of the OLS estimator were compared as shown in Table 3.

To assess the stability of the OLS and the two ridge regression estimators (RR_1 and

Table 4: Table of some summary statistics (mean, median and variance) of the estimated regression parameters by ordinary least squares (OLS) and the two ridge regression estimators (RR_1 and RR_2) over 500 fitted models (iterations). The three models were estimated using the scaled predictor variables.

Estimator	Parameter	$\beta_0 = 25$	$\beta_1 = 10$	$\beta_2 = 15$	$\beta_3 = 60$	$\beta_4 = 40$	$\beta_5 = 35$	$\beta_6 = 20$
OLS	Mean	25.054	10.164	15.281	60.463	39.408	35.096	19.574
	Median	25.017	10.145	14.846	60.354	39.128	34.660	19.634
	Variance	1.137	35.920	44.551	52.150	37.102	61.003	52.913
RR_1	Mean	25.054	12.993	16.171	56.639	38.154	33.441	22.526
	Median	25.017	12.631	15.672	57.346	38.066	33.651	21.844
	Variance	1.137	19.227	28.100	21.915	25.259	38.747	24.416
RR_2	Mean	25.054	12.839	16.254	56.753	38.114	33.666	22.283
	Median	25.017	13.156	16.363	56.350	37.744	33.025	22.710
	Variance	1.137	35.915	33.333	50.537	30.139	46.126	42.381

RR_2) using the scaled predictor variables, we present in Table 4, some summary statistics (mean, median and variance) of the three regression estimators over 500 fitted models (iterations). The results in Table 4 showed that, although, the median and the mean estimated by the three regression estimators are very close, an indication that they are all consistent. A closer look at the estimated variances showed that the two ridge estimators, RR_1 and RR_2 are quite more stable than the OLS estimators. The variances of the OLS estimators are relatively larger than those provided by RR_1 and RR_2 estimators across the six estimated slope parameters in the models.

From all the results in Table 1 through Table 4 as discussed so far, the positive impact of scaled predictor variables at improving the ridge regression estimators has been largely demonstrated. While OLS estimator may seem promising in few instances despite the presence of multicollinearity, the ridge regression estimator with a higher relative efficiency of about 70% still remains a good alternative to OLS to model data with correlated predictor variables.

3.2 *Results From Real Life Data*

The impact of scaled and unscaled correlated predictor variables on the performance of OLS and ridge regression estimators are presented using a real life data set on the French economy.

The French data analysed here is an historical data set on import activities in French economy. The data were first analysed by Malinvaud (1968) and later by Chatterjee and Hadi (2006) among others. The response variable is imports (IMPORT) which was regressed on domestic production (DOPROD), stock formation (STOCK) and domestic consumption (CONSUM), all measured in billions of French francs for 18 years beginning

Table 5: The correlation matrix showing the extent of linear relationship between the predictor variables. The estimated p-values of the correlation tests are reported in parentheses. Only the amount spent on domestic production (DOPROD) and domestic consumptions (CONSUM) are highly and significantly correlated ($p < 0.001$), an indication that multicollinearity exist in the data.

	DOPROD	STOCK	CONSUM
DOPROD	1		
STOCK	$-0.106(p = 0.771)$	1	
CONSUM	$0.997(p < 0.001)$	$-0.101(p = 0.782)$	1

from 1949 to 1966. This resulted into the multiple linear regression model of the form

$$IMPORT = \beta_0 + \beta_1^* DOPROD + \beta_2^* STOCK + \beta_3^* CONSUM + \epsilon \tag{11}$$

Due to the violation of the basic assumption of constancy of error term across all the 18 sample units in the data, as established by Chatterjee and Hadi (2006), data set for 11 years beginning from 1949 to 1959, as used in that work were equally employed here for easy comparison of results.
In order to examine the existence of linear relationship among the three predictor variables, the correlation tests were performed using their sample pair-wise correlation coefficients, as presented in the correlation matrix in Table 5 with their respective p-values. The results in the table showed that French's domestic productions (DOPROD) and domestic consumptions (CONSUM) are the only pair of predictor variables that are significantly correlated ($corr. = 0.997, p < 0.001$), indicating the existence of multicollinearity in the data.

In Table 6, we present the results of the OLS and ridge regression (RR_2) of model (11) fitted to the data, as equally reported in Chatterjee and Hadi (2006). The two regression models were fitted using the raw (unscaled) values of the three predictor variables DOPROD, STOCK and CONSUM. From the results in Table 6, the OLS representation of the fitted remodel is

$$\widehat{IMPORT} = -10.13 + 0.0514^* DOPROD + 0.5869^* STOCK + 0.2868^* CONSUM \tag{12}$$

while its ridge regression representation is

$$\widehat{IMPORT} = -8.5537 + 0.0635^* DOPROD + 0.5859^* STOCK + 0.1156^* CONSUM \tag{13}$$

The simple interpretation of the estimated intercept parameters in the OLS and ridge regression equations (12) and (13) is that the expected amount of imports (IMPORT) into the French economy from 1949 to 1959 are about -10 and -9 billions of francs respectively given that the values of the domestic production (DOPROD), stock formation (STOCK) and domestic consumption (CONSUM) are all zero. These two results

Table 6: The results of the OLS and ridge regression models on the French economy data from 1949 to 1959 using the raw(unscaled) values of the predictor variables as reported by Chatterjee and Hadi (2006).

Estimator	Estimated Models' parameters			
	Intercept	DOPROD	STOCK	CONSUM
OLS	−10.1300	−0.0514	0.5869	0.2868
Ridge (RR_2)	−8.5537	0.0635	0.5859	0.1156

Table 7: The results of the OLS and ridge regression models on the French economy data from 1949 to 1959, see Chatterjee and Hadi (2006), using the scaled values of the predictor variables. The optimal shrinkage parameter of the ridge regression was determined to be 0.07. The mean square error of the OLS and the best ridge regression models are provided in the table which shows better performance of the ridge estimators (with smaller MSE_2) over the OLS estimators.

Estimator	Shrinkage parameter k	Estimated Model's parameters				MSE_2
		Intercept	DOPROD	STOCK	CONSUM	
OLS	0.000	21.891	−1.542	0.968	5.919	62.422
Ridge	0.07	21.891	−0.830	0.974	5.205	18.211

are unrealistic, because the amount of imports (in billions of French francs) into the French economy, quantified in monetary terms, cannot be negative as portrayed by the two results. The main cause of the unrealistic and unstable results of both the OLS and ridge estimators in Table 6 is the presence of multicollinearity in the data as shown in Table 5.

To demonstrate the impact of scaling, the values of the three predictor variables in the French data were scaled to zero means and unit standard deviations using statistic (7). The OLS and the ridge regression (RR_2) models were fitted to the transformed data, the results of which are presented in Table 7. The optimal shrinkage parameter k_2 of the ridge regression estimator RR_2 was determined to be 0.07 through 1000 cross-validation search for the best shrinkage parameter value within the interval $(0, 1)$ according to the procedures detailed in Section 2.2. This value of $k_2 = 0.07$ is the optimal value of the shrinkage parameter k of the ridge regression estimator RR_2 for the data which yielded the least mean square error (MSE_2) estimate of 18.211.

The plot of the various estimates of the ridge regression parameters against the values of the shrinkage parameter k at which they were obtained is presented in Fig 3.The graph showed the optimal shrinkage parameter value of 0.07 at which stable estimates of the ridge regression parameters were obtained.

Based on the results in Table 7, the most efficient regression model for the French data is the fitted ridge regression model

$$\widehat{IMPORT} = 21.891 + 0.830^{*}DOPROD + 0.974^{*}STOCK + 5.205^{*}CONSUM \quad (14)$$

3.pdf

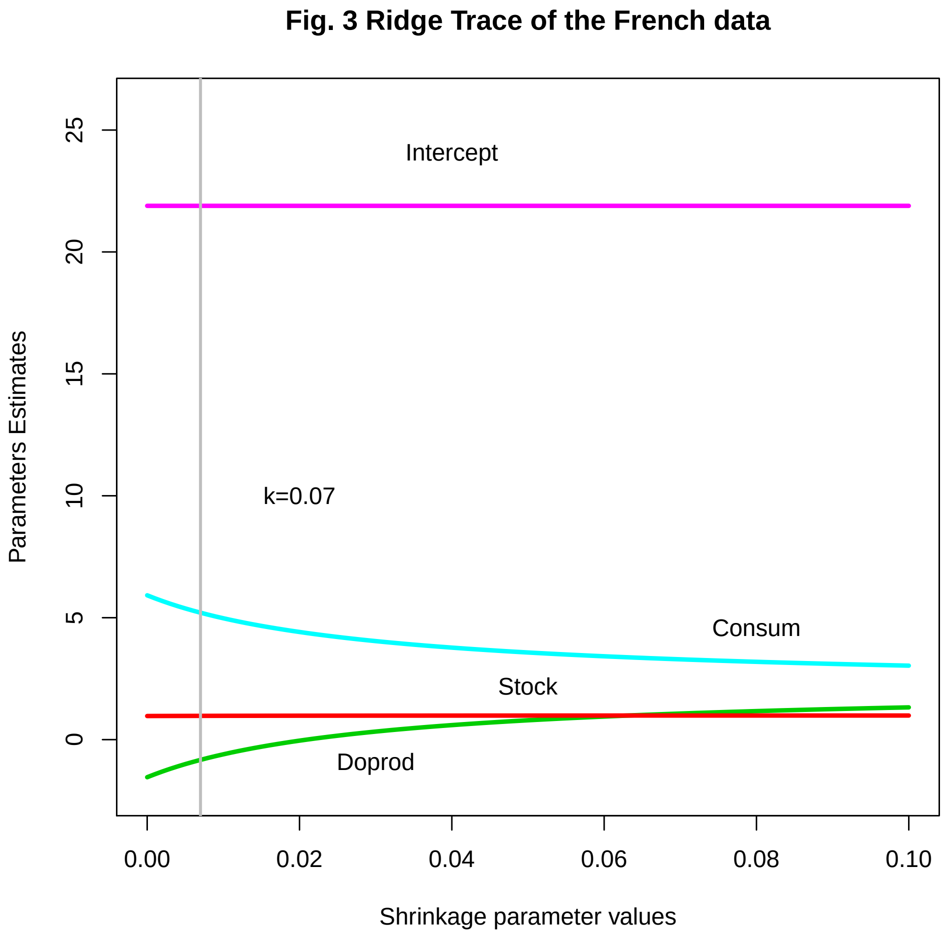

Figure 3: **The plot of various regression parameters against the shrinkage parameter values. The best shrinkage parameter value k_2 that yielded the least mean square error (MSE_2) as determined by cross-validation is 0.07 for the data.**

From these regression results, it is quite obvious that scaling of the predictor variables has greatly assisted to stabilize the estimates of both the OLS and ridge regression parameters, most especially the intercepts. In both models, the value of the intercept parameter was estimated to be 21.891 which reasonably translates to the expected amount of imports (IMPORT) into the French economy (in billions of francs) between 1949 to

1959. However, the results showed that the ridge estimator with estimated mean square error (MSE_2) of 18.211 is still better than the OLS estimator with relatively higher MSE_2 value of 62.422.

4 Conclusion

Ridge regression estimator, has been established to be a credible alternative to the classical OLS estimators when some of the predictor variables are correlated. It is a biased but efficient regression technique in the presence of multicollinearity in multiple linear regression models, seeMuniz and Kibria (2009) and Kibria (2003).
The basic procedures for fitting ridge regression model to data with inherent collinear structure are examined in this study. The widely adopted ridge regression technique, as proposed by Hoerl and Kernard (1970a), required that the optimal value of the shrinkage parameter k be nonnegative ($k > 0$), and indeed, that the value of k should fall within the interval $(0, 1)$. This has resulted into the development of various forms of ridge estimator, k based on this earlier proposition as reported in Lawless and Wang (1976); Lin and Kmenta (1982); Hoerl et al (1985) and in few other works. However, it has been clearly demonstrated in this work that, following this traditional ridge regression techniques, efficient ridge regression models might not be achieved using the raw (unscaled) values of the predictor variables for estimation in the presence of multicollinearity. It is therefore necessary and desirable to scale the predictor variables in ridge regression modelling when the presence of multicollinearity is suspected.
Another important result from this study is that, the OLS estimators might sometimes yield good regression results like the ridge estimator in the presence of multicollinearity if the predictor variables are scaled. This is evident from the Monte-Carlo results in Table 3 in which the average relative efficiency of OLS estimator was about 30% despite the inherent collinear structure in the data. However, this optimistic behaviour of the OLS depends largely on the degree of multicollinearity in the data. The OLS estimator would have appreciable relative efficiency while modelling data with fewer numbers of correlated predictors than data with much number of collinear predictors.
Finally, as reported by Yahya et al (2008) and several others, inter-dependencies among the pairs of explanatory variables in general regression estimation is inevitable in many practical real life situations. When multicollinearity is suspected in a data set, thorough examination is needed to determine the severity of such collinear structure. This will inform the proper choice of suitable estimation techniques to model the data. However, whenever the ridge regression modelling technique as proposed by Hoerl and Kernard (1970a) is adopted, it is desirable for the investigators to work with the standardize values of the predictor variables as implemented in many works, see Marquardt and Snee (1975); Fearn (1993) and Bradly and Srivastava (1997). By this, efficient estimates of the ridge regression models can be guaranteed in the presence of collinear predictors in the data.

References

Bradly, R. A., Srivastava, S. S. (1997). Correlation in polynomial regression. URL: http://stat.fsu.edu/techreports/M409.pdf

Cannon, A. J. (2009). Negative ridge regression parameters for improving the covariance structure of multivariate linear downscaling models. Int. J. Climatol., 29, 761-769.

Chatterjee, S., Hadi, A. S. (2006). Regression Analysis by Example. John Wiley & Sons, Inc., Hoboken, New Jersey.

Dorugade, A. V., Kashid, D. N. (2010). Alternative method for choosing ridge parameter for regression. Applied Mathematical Science, 4(9), 447-456.

El-Dereny, M. and Rashwan, N. I. (2011). Solving Multicollinearity Problem Using Ridge Regression Models. Int. J. Contemp. Math. Sciences, 6(12), 585-600.

Faraway, J. J. (2002). Practical regression and ANOVA using R. http://cran.r-project.org/doc/contrib/Faraway-PRA.pdf

Fearn, T. (1993). A misuse of ridge regression in the calibration of a near infrared reflectance instrument. Applied Statistics, 32, 73-79.

Hoerl, A. E., Kennard, R. W., Hoerl, R. W.(1985). Practical use of ridge regression: A challenge met. Applied Statistics, 34(2), 114-120.

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

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

Hoerl, A. E., Kennard, R. W. (1970b). Ridge regression: Applications to nonorthogonal problems. Technometrics, 12, 69-82, 1970b.

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

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

Lawless, J. F., Wang, P. A. (1976). Simulation study of ridge and other regression estimators. Communications in Statistics -Theory and Methods, 14, 1589-1604.

Longley, J. W.(1976). An appraisal of least-squares programs from the point of view of the user. Journal of the American Statistical Association, 62, 819-841.

Lin, L., Kmenta, J. (1982). Ridge Regression under Alternative Loss Criteria. The Review of Economics and Statistics, 64(3), 488-494.

Malinvaud, E. (1968). Statistical Methods of Econometrics, Rand-McNally, Chicago.

Mardikyan, S., Cetin, E.(2008). Efficient Choice of Biasing Constant for Ridge Regression. Int. J. Contemp. Math. Sciences, 3, 527-547.

Marquardt, D. W., Snee, R. D. (1975). Ridge regression in practice. The American Statistician, 29(1), 3-20.

Muniz, G., Kibria, B. M.(2009). On Some Ridge Regression estimator: An empirical comparison. Communication in Statistics-Simulation and Computation, 38, 62-630.

Myers, R. H.(1986). Classical and Modern Regression with Applications. PWS-KENT Publishing Company, Massachusetts.

Sparks, R. (2004). SUR Models Applied To an Environmental Situation with Missing Data and Censored Values. Journal of Applied Mathematics and Decision Sciences, 8(1), 15-32, 2004.

Wethril, H. (1986). Evaluation of ordinary Ridge Regression. Bulletin of Mathematical Statistics, 18, 1-35, 1986.

Yahya, W.B., Adebayo, S.B., Jolayemi, E.T., Oyejola, B.A., Sanni, O.O.M. (2008). Effects of non-orthogonality on the efficiency of seemingly unrelated regression (SUR) models. InterStat Journals, 1-29.