***ISSN 1684-8403***
***Journal of Statistics***
***Volume 24, 2017. pp. 20-34***

# Modified Method for Choosing Ridge Parameter

*Ayan Ullah*[1]*, Muhammad Suhail*[2] *and Maryam Ilyas*[3]

## Abstract

Multicollinearity occurs when two or more predictors are linearly related to each other. In this case, either OLS estimators do not exist or if exist the associated variances of estimated Regression co-efficients are very large, making inferences invalid. Ridge Regression is used to counter the effects of multicollinearity. This is done by introducing biasing constant $k$, called Ridge parameter in the least square objective function. Ridge parameter shrinks the estimates and their variances. Selection and choice of the unknown Ridge parameter $k$ is of prime importance in Ridge Regression analysis.

Khalaf et al. (2013) proposed some modifications of existing Ridge estimators $k_1$– $k_{16}$ by multiplying them with the factor that make use of maximum eigenvalue associated with $(X^tX)$ matrix and name resulting estimators as K1M–K16M. This study proposed some modifications of existing Ridge estimators $k_1$– $k_{16}$ by multiplying them with the factor that make use of arithmetic mean of eigenvalues associated with $(X^tX)$ matrix denoted as K1A–K16A. The comparative performance of proposed sets of estimators and Khalaf et al. (2013) was evaluated by Mean Square Error (MSE) using simulated data sets. Data sets considering different levels of collinearity (r), sample size (n), number of predictor (p), error term variances and error term distributions were generated. It was observed that proposed estimators K1A–K16A outperform K1M–K16M when error terms following normal distribution ($\sigma^2$= 0.1, 1) collinearity levels (r) are (i.e. 0.80, 0.90, 0.95) and number of predictors are (i.e. 2, 4, 6) and when error terms following non-normal distribution (F (4, 20)) collinearity levels (r) are high (i.e. 0.80, 0.90, 0.95) and number of predictors are small (i.e. 2, 4).

[1] College of Statistical and Actuarial Sciences, University of Punjab, Lahore, Pakistan.
[2] College of Statistical and Actuarial Sciences, University of Punjab, Lahore, Pakistan.
[3] College of Statistical and Actuarial Sciences, University of Punjab, Lahore, Pakistan.

**Keywords**

Multicollinearity, Ridge Regression, Mean square error

## 1. Introduction

In Regression analysis, usually we consider that the predictors are not linearly related to each other. In practice, there may be some type of relationships among the predictors. In this case, the assumption of independence of predictors is no longer valid; violation of this assumption causes the problem of multicollinearity. Regression analysis is most powerful statistical tool that helps in investigating the relationships between response variable and explanatories. Prediction and description mainly depend on the estimated Regression coefficients. Least Squares method is the mostly used method for estimating the unknown Regression coefficients. It gave good estimates only if the assumption of independence of explanatories is valid. The assumptions are that the explanatory variables are independent from each other and this is very difficult to hold in reality. When the purpose is to get more information about the outcome variable, there is need to add more predictors to Regression model. By doing so, relationships between these variables occur and the magnitude of these relationships often increases. This type of linear relationships between the predictors is called the problem of multicollinearity. Chatterjee and Hadi (2006) and Gujarati (2003) highlighted that with the existence of multicollinearity in a data set, two or more explanatories give same or approximately same information. The existence of multicollinearity among explanatories causes many problems. It affects the model's ability to estimate unknown Regression coefficients, t-test, computational accuracy, variance of LS estimated Regression coefficients, and LS estimated Regression coefficients, fitted values and predictions. Draper and Smith (1981) stated that as a result of multicollinearity, the $X^tX$ matrix is near ill conditioned (singular) that leads to large standard errors for Ordinary Least Squares (OLS) estimates.

In order to overcome the problem of multicollinearity in Multiple Linear Regression model among explanatories, Hoerl and Kennard (1970) suggested Ridge Regression (RR) method instead of OLS method in Regression analysis.

Multiple Linear Regression model can be written in matrix form as,

$$y = X\beta + \epsilon, \qquad (1.1)$$

$y$ is a vector of dependent variables with order n×1, $X$ is matrix of explanatory of order n×p, $\beta$ is a vector of unknown Regression coefficients of order p×1 and $\epsilon$ is vector of random errors of order n×1 that are distributed normally whose mean vector is zero while it's covariance matrix is $\sigma^2 \mathrm{I}_n$ ( $\mathrm{I}_n$ is identity matrix of n×n order). The OLS of the Regression coefficients $\beta$ is $\hat{\beta}_{OLS} = (X^t X)^{-1} X^t y$, and variance-covariance matrix of $\beta$ is Var ($\hat{\beta}_{OLS}$)= $\sigma^2 (X^t X)^{-1}$, both $\hat{\beta}$ and Var ($\hat{\beta}$) depend on characteristics of $X^t X$ matrix. If matrix $X^t X$ is near to singular then the variances of Ordinary Least Square (OLS) estimates becomes large. In Ridge Regression method a small positive number k (≥0) to be added to diagonal of $X^t X$ matrix to counter the effects of Multicollinearity such that the new estimates are,

$$\hat{\beta}_{RR} \ = (X^t X + k\mathrm{I}_p)^{-1} X^t y, \qquad k \geq 0 \tag{1.2}$$

For any positive value of $k$, this gave Minimum Mean Square Error (MMSE) as compared to LSE. The $k$ is known as Ridge or biasing parameter (constant) and will be finding out from data. When k=0, $\hat{\beta}_{RR}$ becomes the Ordinary Least Square estimates (OLS) and $k$ increases more bias is introduced but variance of the Regression estimator stabilizes.

Now, the MSE of Ridge Regression that is introduced by Hoerl and Kennard (1970) is defined as,

$$MSE\left(\hat{\beta}(K)\right) = \sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i + k_i)^2} + \sum_{i=1}^{p} \frac{k^2{}_i \alpha^2{}_i}{(\lambda_i + k_i)^2} \tag{1.3}$$

The 1st term on right hand side of eq. (1.3) is a variance and the second term is an amount of bias

where,

$$k_i = \frac{\sigma^2}{\alpha_j^2} \tag{1.4}$$

$\sigma^2$ is the variance of the model eq. (1.3) and $\alpha_i$ is i$^{th}$ element of $\alpha$.

The unbiased estimator of $k_i$is,

$$\hat{k}_i = \frac{\hat{\sigma}^2}{\hat{\alpha}^2{}_i} \tag{1.5}$$

where,

$\hat{\sigma}^2 = \frac{(y-\hat{y})^t\,(y-\hat{y})}{(n-p-1)}$ is the residual sums of square obtained from the OLS and is an unbiased estimator of $\sigma^2$and $\hat{\alpha}^2{}_i$ is the i$^{th}$ elements of $\hat{\alpha}^2$ where $\hat{\alpha} = V^t \hat{\beta}$, $V$ is orthogonal matrix of order (p×p); the columns of $V$ are the normalized eigenvectors of correlation matrix.

Many methods for estimation of Ridge parameter $k$ have been described by many researchers such as Shehzad M. A. (2012), Vinod and Aman Ullah (1981). Some well- known existing estimators are following. These estimators make use of the canonical form of Regression model.

The canonical form of model eq. (1.3) is eq. (1.6). Consider orthogonal matrix $D$ where,
$D^T C D = \Lambda$,
where, $C = X^T X$ and $\Lambda = diag(\lambda_1, \lambda_2, \dots, \lambda_p)$ containing eigenvalues '$\lambda_i$' of matrix *C*. Model eq. (1.3) in canonical form is,

$$y = X^*\alpha + \epsilon \qquad (1.6)$$

$X^* = XD$ and $\alpha = D^T\beta$ .
The Least Square Estimators of the canonical form is,

$$\hat{\alpha} = \Delta^{-1}{X^*}^T y \qquad (1.7)$$

Ridge estimators in canonical form is

$$\hat{\alpha}(k) = ({X^*}^T X^* + KI)^{-1}{X^*}^T y \qquad (1.8)$$

K $= diag(\lambda_1, \lambda_2, \dots, \lambda_p)$. MSE of the above estimators defined as:

$$MSE(\hat{\alpha}(k)) = \sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i + k_i)^2} + \sum_{i=1}^{p} \frac{{k_i}^2 {\alpha_i}^2}{(\lambda_i + k_i)^2} \qquad (1.9)$$

First term on R.H.S eq. (1.9) is variance and second term is amount of bias.

## 2. Methodology

In this study, some successful extensions of the existing work have been proposed to deal with multicollinearity problem.

***Hoerl and Kennard estimator:*** Hoerl and Kennard (1970) explored value of $k$ which minimizing the Mean Square Error (MSE) is the following:

$$K_1 = \hat{k}_{HK} = \frac{\hat{\sigma}^2}{\hat{\alpha}^2{}_{max}}.$$

where, $\hat{\alpha}^2{}_{max}$ is the square of the maximum value of $\hat{\alpha}$.

***Kibria estimator:*** Kibria (2003) proposed the following estimators

$$K_2 = \hat{k}_{GM} = \frac{\hat{\sigma}^2}{(\prod_{j=1}^{p} \hat{\alpha}^2{}_j)^{\frac{1}{p}}}.$$

And, $K_3 = \hat{k}_{MED} = Median\{m^2{}_j\}$

where, $m_j = \sqrt{\frac{\hat{\sigma}^2}{\hat{\alpha}^2{}_j}}$

***Khalaf and Shukur estimator:*** Khalaf and Shukur (2005) proposed a new estimator as a modification of $k_{HK}$

$$K_4 = \hat{k}_{KS} = \frac{t_{max}\hat{\sigma}^2}{(n-p)\hat{\sigma}^2 + t_{max}\hat{\alpha}^2{}_{max}}\,.$$

where, $t_{max}$ is maximum eigenvalue of matrix $X^tX$.

***Alkhamisi, Khalaf and Shukur estimator:*** Alkhamisi et al. (2006) suggested that

$$K_5 = k_{S3} = \hat{k}^{KS}{}_{max} = \max(s_j);\ K_6 = \hat{k}^{KS}{}_{md} = \mathrm{med}(s_j)$$

where, $s_j = \frac{t_j\hat{\sigma}^2}{(n-p)\hat{\sigma}^2 + t_j\hat{\alpha}^2{}_j}$

***Alkhamisi and Shukur estimator:*** Alkhamisi and Shukur (2008) suggested the estimators for *k* as,

$$K_7 = k_{KM1} = \hat{k}^{KS}{}_{gm} = \left(\prod_{j=1}^{p} s_j\right)^{\frac{1}{p}}, \quad K_8 = k_{KM2} = \max\left(\frac{1}{m_j}\right),$$

$$K_9 = k_{KM4} = \left(\prod_{j=1}^{p}\frac{1}{m_j}\right)^{\frac{1}{p}}, \quad K_{10} = k_{KM5} = \left(\prod_{j=1}^{p} m_j\right)^{\frac{1}{p}},$$

$$K_{11} = k_{KM6} = \mathrm{median}\left(\frac{1}{m_j}\right), \quad K_{12} = k_{KM8} = \max\left(\frac{1}{q_j}\right),$$

$$K_{13} = k_{KM9} = \max\left(\sqrt{q_j}\right), \quad K_{14} = k_{KM10} = \left(\prod_{i=1}^{p}\frac{1}{\sqrt{q_j}}\right)^{\frac{1}{p}};$$

$$K_{15} = k_{KM11} = \left(\prod_{i=1}^{p}\sqrt{q_j}\right)^{\frac{1}{p}}, \quad K_{16} = k_{KM12} = median\left(\frac{1}{\sqrt{q_j}}\right).$$

where, $q_j = \sqrt{\frac{t_{max}\hat{\sigma}^2}{(n-p)\hat{\sigma}^2 + t_{max}\hat{\alpha}^2{}_j}}$

***Khalaf, Mansson and Shukur estimators:*** Khalaf et al. (2013) proposed modifications of all of the above estimators by multiplying them by a factor,

$$w_j = \frac{t_{max}}{\sum_{j=1}^{p}|\hat{\alpha}|_j}$$

$t_{max}$ is the maximum eigenvalue of $X^tX$ matrix and are denoted by K1M-K16M.

This modification was proposed on the basis that as degree of correlation increases initial eigenvalues are larger than others. Thus, factor $w_j$ will also become larger as it is based on the maximum eigenvalue of $X^tX$ matrix. This will lead to an increase of the estimated value of the ridge parameters *k*. Hence, this modification leads to larger values of the Ridge parameter especially when the degree of correlation is high. The performance of these estimators was good for high collinearity level.

***2.1 Our Proposed estimators:*** Khalaf et al. (2013) proposed modifications by multiplying $K_1$– $K_{16}$ estimators by a factor $w_j$ which is based on maximum eigenvalue.

After exploring the trends in eigenvalues of $X^tX$, we came across a larger set of maximized eigenvalues in case of high multicollinearity (0.95, 0.99) then the others, however, as the level of multicollinearity decreases from the said level (0.95, 0.99), the difference between the maximum eigenvalue and remaining eigenvalues decreases. Hence, it was expected that some other Ridge estimators have potential to be explored in view of this phenomenon.

Thus, instead of using maximum eigenvalue of $X^tX$ in the numerator of $w_j$, Arithmetic mean of eigenvalues of the $X^tX$ matrix may be used. When degree of collinearity among regressors was not very high (0.99) then there were small difference between maximum eigenvalue and the remaining eigenvalues so in this situation arithmetic mean was expected to give good results.

Thus, a new set of Ridge estimators were developed. Which is arithmetic mean of the eigenvalues ($T_{AM}$) associated with $X^tX$ was considered for defining,

$$v_{2j} = \frac{T_{AM}}{\sum_{j=1}^{p} |\hat{\alpha}|_j}$$

Finally, the new set of estimators was defined by multiplying $K_1$– $K_{16}$ by $v_{2j}$ and the resulting estimators were denoted by K1A –K16A.

***2.2 Mean Square Error (MSE):*** The performances of Ridge Regression estimators have long being compared making use of MSE. Thus, to explore the competitive performance of the new suggested estimators and existing estimators, MSE was used. MSE is defined as,

$$MSE = \sum_{i=}^{N} \frac{(\hat{\beta} - \beta)_i{}^t (\hat{\beta} - \beta)_i}{N}$$

$\hat{\beta}$ is the estimator of $\beta$ obtained from RR or OLS and $N$ is number of replications used in Monte Carlo study.

## 3. The Monte Carlo simulation

Theoretically the proposed and the existing estimators cannot be compared, so simulation studies were designed to explore the performance of the developed and already existing Ridge estimators.

Gibbons (1981), Kibria (2003), McDonald and Galarneau (1975) and Wichern and Churchill (1978), and many other researchers used the following method to simulate or generate the predictor variables that is,

$$x_{ij} = (1-r^2)^{\frac{1}{2}} z_{ij} + r z_{ip}; \qquad i = 1, 2, 3...n; \; j = 1, 2, 3...p \tag{3.1}$$

$z_{ij}$ are standard normal random variables, '$r^2$' is level of collinearity between any two explanatories and $n$ is the number of observations. In this study, the model that is used is,

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \varepsilon_i; \quad i=1, 2, 3 \ldots n \tag{3.2}$$

$\beta_0$ is taken to be zero and $\beta_1, \beta_2, \ldots\ldots, \beta_p$ regression coefficients, are considered so that $\sum_j^p \beta_j^2 = 1$. Simulations studies are carried out using programming language R.

## 4. Factors affecting Ridge estimators

Various factors can potentially affect the performance of Ridge estimators. These include severity of multicollinearity, sample size, number of explanatory variables, error term variance (normal distribution case) and distributions of error terms. Four levels of multicollinearity between any two regressors were considered as high ($r = 0.80, 0.90$) and very high ($r = 0.95, 0.99$). The variation of sample size and number of explanatories considered as n = 30, 70, 100, 150, 200, 300 and p= 2, 4, 6, respectively. In case of normal error term distribution, the variation of error term variance was considered as $\sigma^2$= 0.1, 1. To explore the effects of error term distribution, normal and non-normal distributions were considered. For non-normal F- distribution with (4, 20) were considered.

## 5. Simulation study

In this study, a simulation study have been designed to explore the competitive performance of K1M–K16M and K1A–K16A. The comparisons of these two sets of estimators have been gauged considering different levels of sample size (n = 30, 70, 100, 150, 200, 300), number of predictors (p = 2, 4, 6), correlation levels (r = 0.80, 0.90, 0.95, 0.99) and error terms distributions (N (0, $\sigma^2$) with $\sigma^2$ = 1, 0.1 and non-normal distributions with F (4, 20)). The performances of these estimators have been evaluated making use of Mean Square Error (MSE). Our study compares K1M–K16M and K1A–K16A. Each of the two studies addresses two cases; case–1 is pertaining to the combination of the levels of sample size, number of predictors and correlation levels with normal error terms distribution ($\sigma^2$=1, 0.1). However, case-2 caters the combinations of sample size, number of predictors and correlation levels with non-normal error terms distributions (F (4, 20)). The results of the simulation studies are summarized by graphs to make comparative performance of all estimators visible in a particular scenario. The results in terms of the tables and some figures are maintained as well but are not included in this document to save space. The results of MSE for Case-1 are presented in figures 4.1 ($1-12$) and those of Case-2 are presented in figures 4.2 ($1-6$).

## 6. Summary and conclusions

The results of the comparative analysis of set of proposed estimators (K1A–K16A) and (K1M–K16M) indicated that distribution play vital rule. MSEs of all estimator (K1A-K16A) are minimum when error term follow normal distribution (figures (4.1 (1–4))). In case of normal distribution ($\sigma^2$= 1, p = 2) and at all levels of *r* proposed estimators, K1A–K16A out-perform K1M–K16M. MSE of estimators K2M, K3M and K10M are maximum. When p = 4, 6 and r = 0.80, 0.90, 0.95, 0.99 the proposed estimators K1A–K16A give good results. At p = 4, 6, at r = 0.95, 0.99 the estimators K1A–K16A give good results and at r = 0.80, 0.9 0, 0.95, 0.99 MSE of estimators K2M, K3M and K10M are maximum. When $\sigma^2$ = 0.1, at all levels of *r* and *p* our modified estimators K1A-K16A performs well. When p = 2 MSE of estimators K1A, K4A, K5A, K6A, K7A, K13A, when p = 4 MSE of estimators K1A, K4A, K5A, K6A, K7A, K9A, K11A, K13A and when p = 6 MSE of estimators K1A, K4A, K5A, K6A, K7A, K9A, K11A, K13A are minimum.

It was noted that MSEs of all estimators are maximum when error terms follow F-distribution (figures (4.2 (1–2)). Our Modified estimators (K1A–K16A) give smaller MSE as compared to set of estimators (K1M–K16M) when error terms follow F-distribution. When p = 2 and at all levels of *r* estimators K1A-K16A perform well. When p = 4, at r = 0.80, 0.90, 0.95, 0.99 our set of estimators give good results. At r = 0.80, 0.90, 0.95 MSE of estimators K1M, K2M, K3M and K10M are maximum.

Thus, it is concluded that our proposed estimators (K1A–K16A) out-perform K1M–K16M for collinearity levels (r = 0.80, 0.90, 0.95, 0.99), moderate number of predictors (p = 2, 4, 6) and error terms following normal distribution ($\sigma^2$ = 1, 0.1) and non- normal distribution F (4, 20)). Therefore, it is recommended to use K1A-K16A to deal the cases of high collinearity levels (r = 0.80, 0.90, 0.95, 0.99) and moderate number of predictors (p = 2, 4, 6) when error terms follow normal distribution ($\sigma^2$=1, 0.1). For non- normal distribution (F (4, 20)) to deal the cases of high collinearity levels (r = 0.80, 0.90, 0.95) and moderate number of predictors (p = 2, 4). This is due to the fact that under these conditions the proposed sets of estimators outperform the existing set of estimators (K1M–K16M).

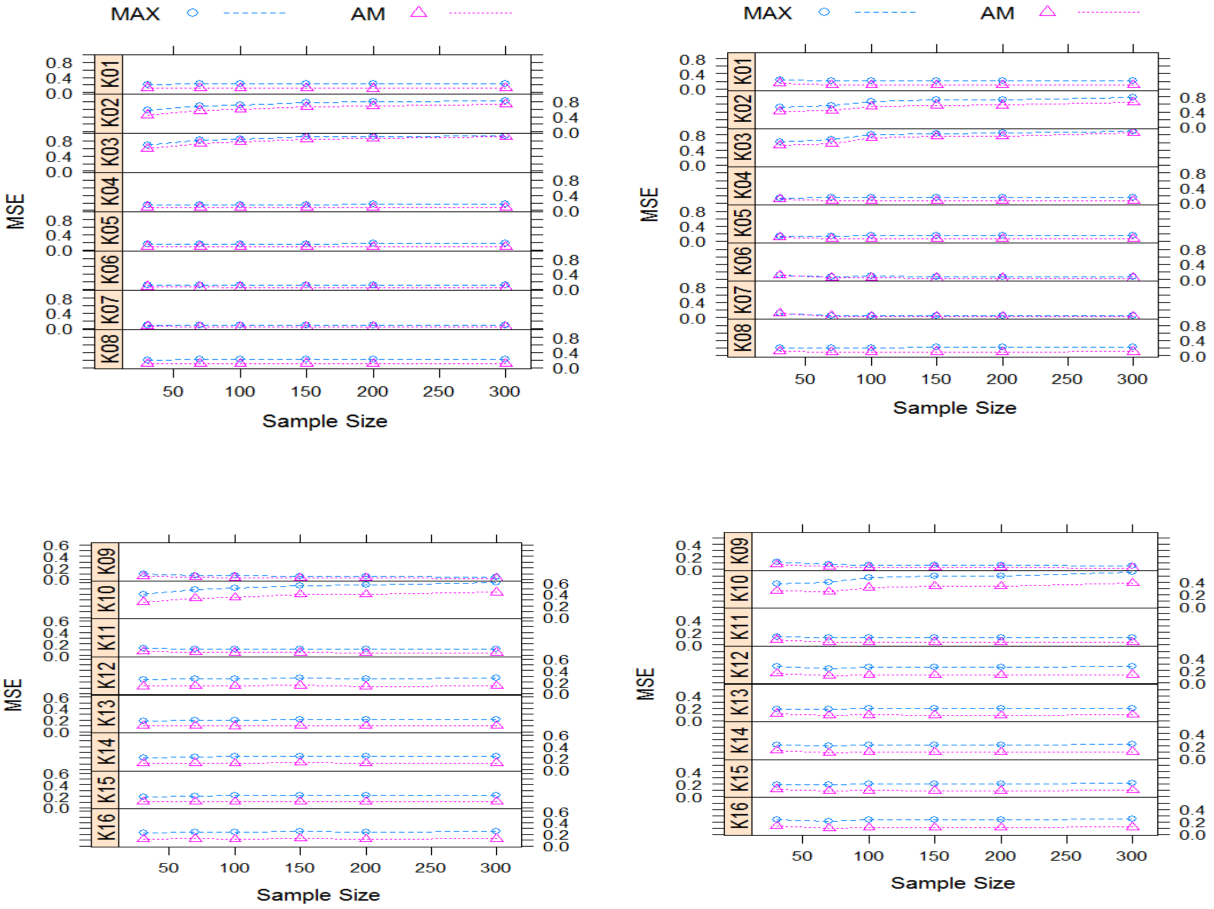

**Figure 4.1 (1):** MSE at p=2 and $\varepsilon_i$~N (0, 1). 1st column is the case of r = 0.80 and 2nd column is the case of r = 0.90.

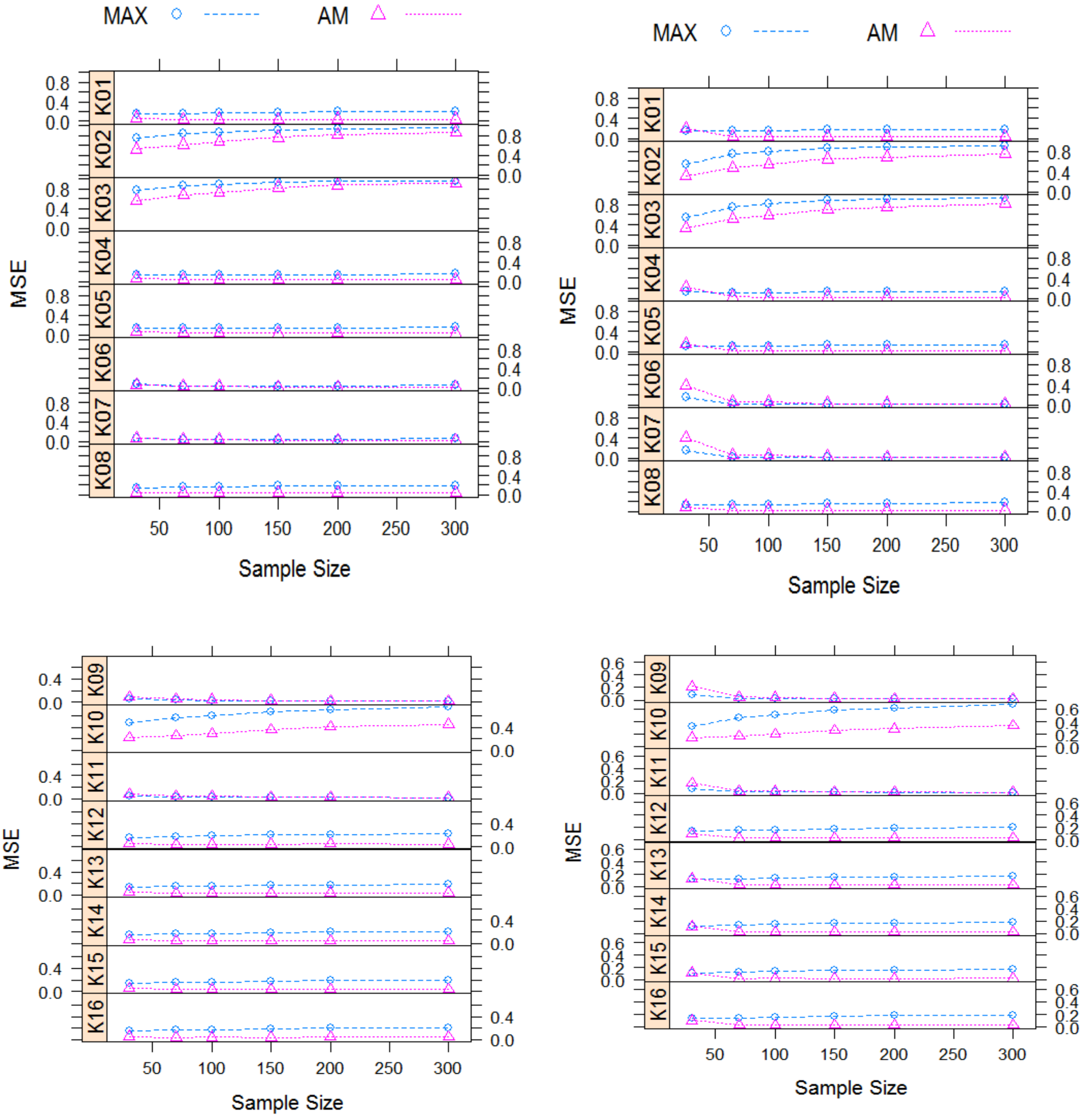

**Figure 4.1 (2):** MSE at p = 4 and $\varepsilon_i \sim N$ (0, 1). 1st column is the case of r = 0.80 and 2nd column is the case of r = 0.90.

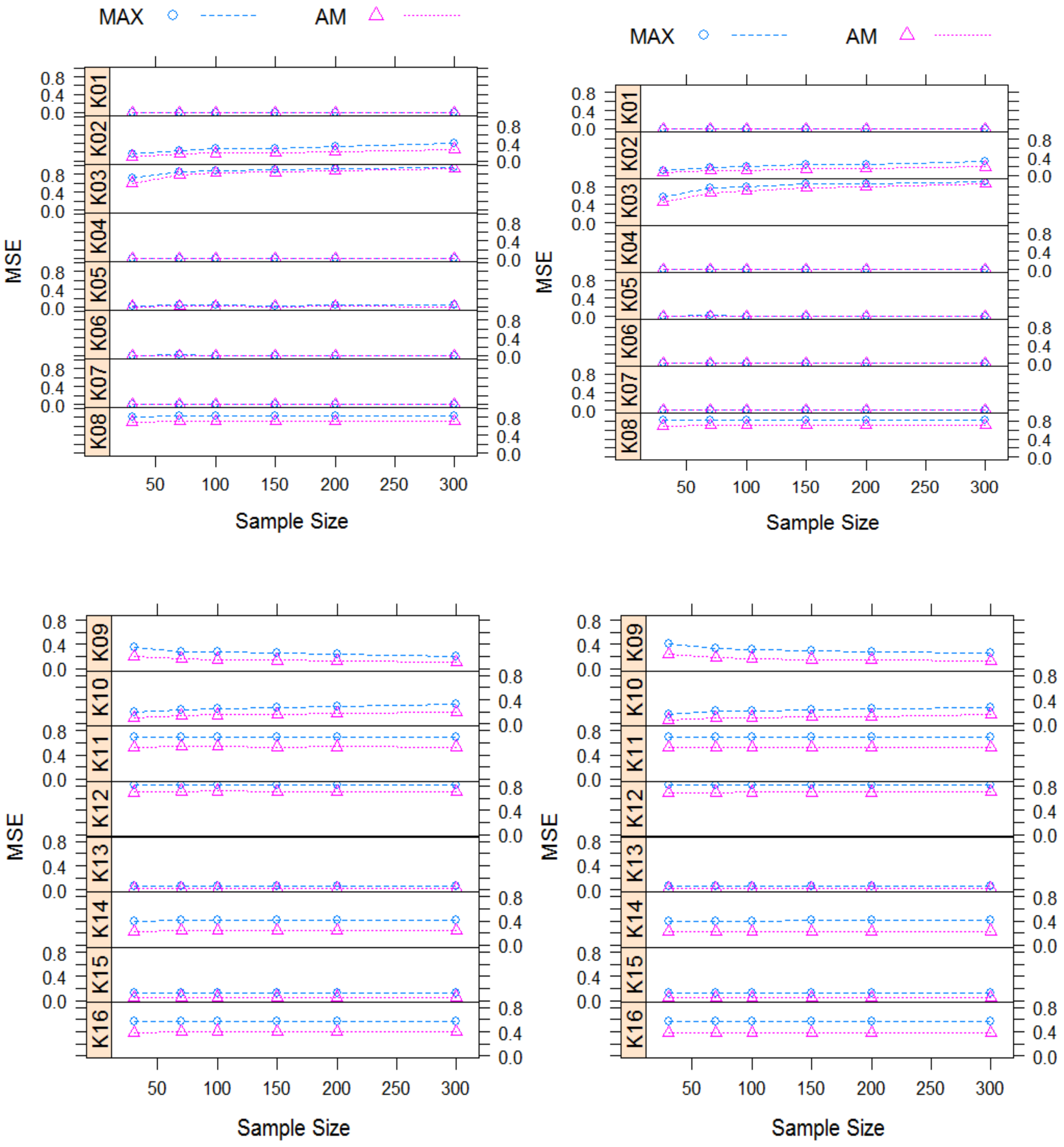

**Figure 4.1 (3):** MSE at p = 2 and $\varepsilon_i$~N (0, 0.1). 1st column is the case of r = 0.80 and 2nd column is the case of r = 0.90.

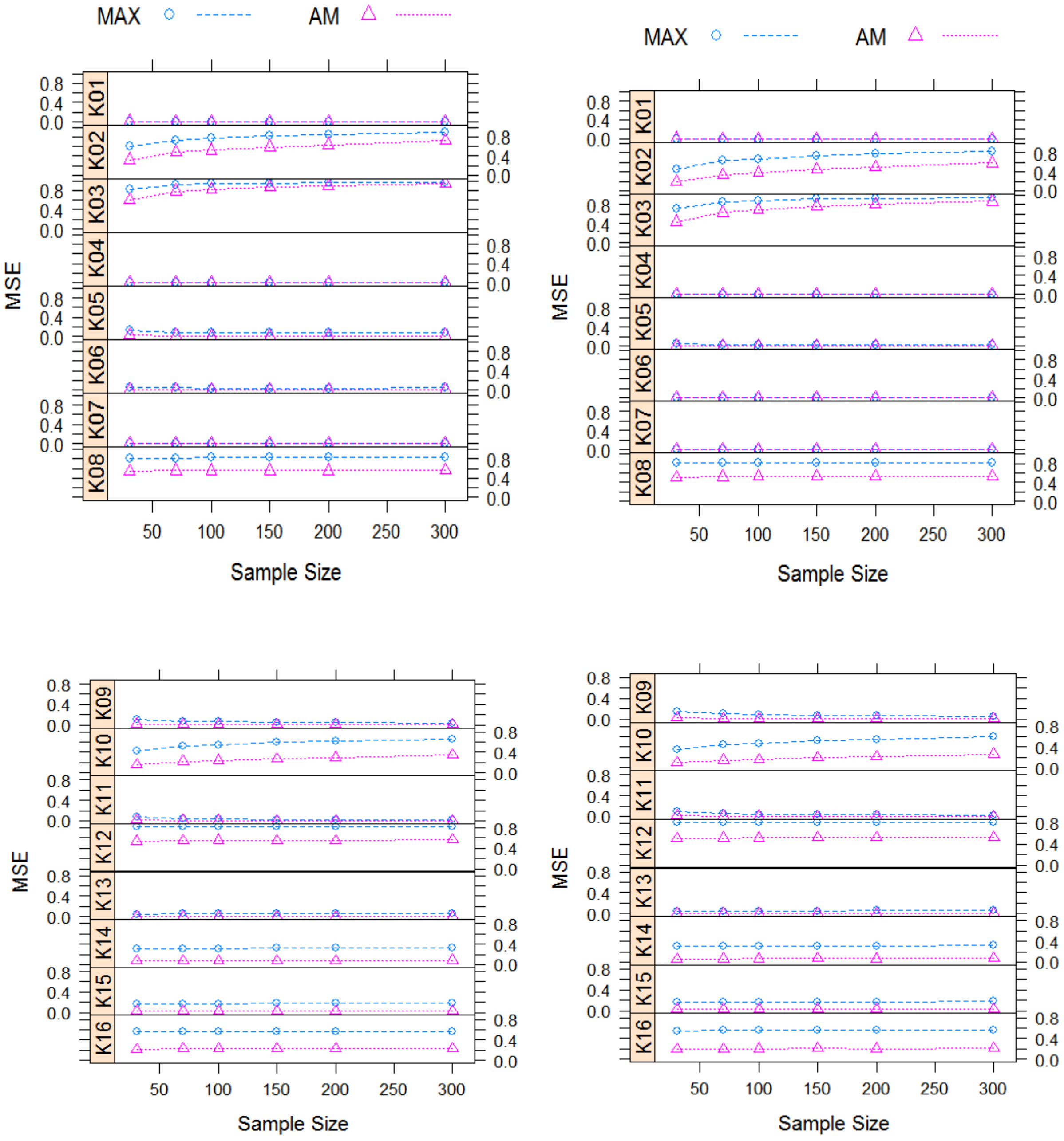

**Figure 4.1 (4):** MSE at p = 4 and $\varepsilon_i$~N (0, 0.1). 1st column is the case of r = 0.80 and 2nd column is the case of r = 0.90.

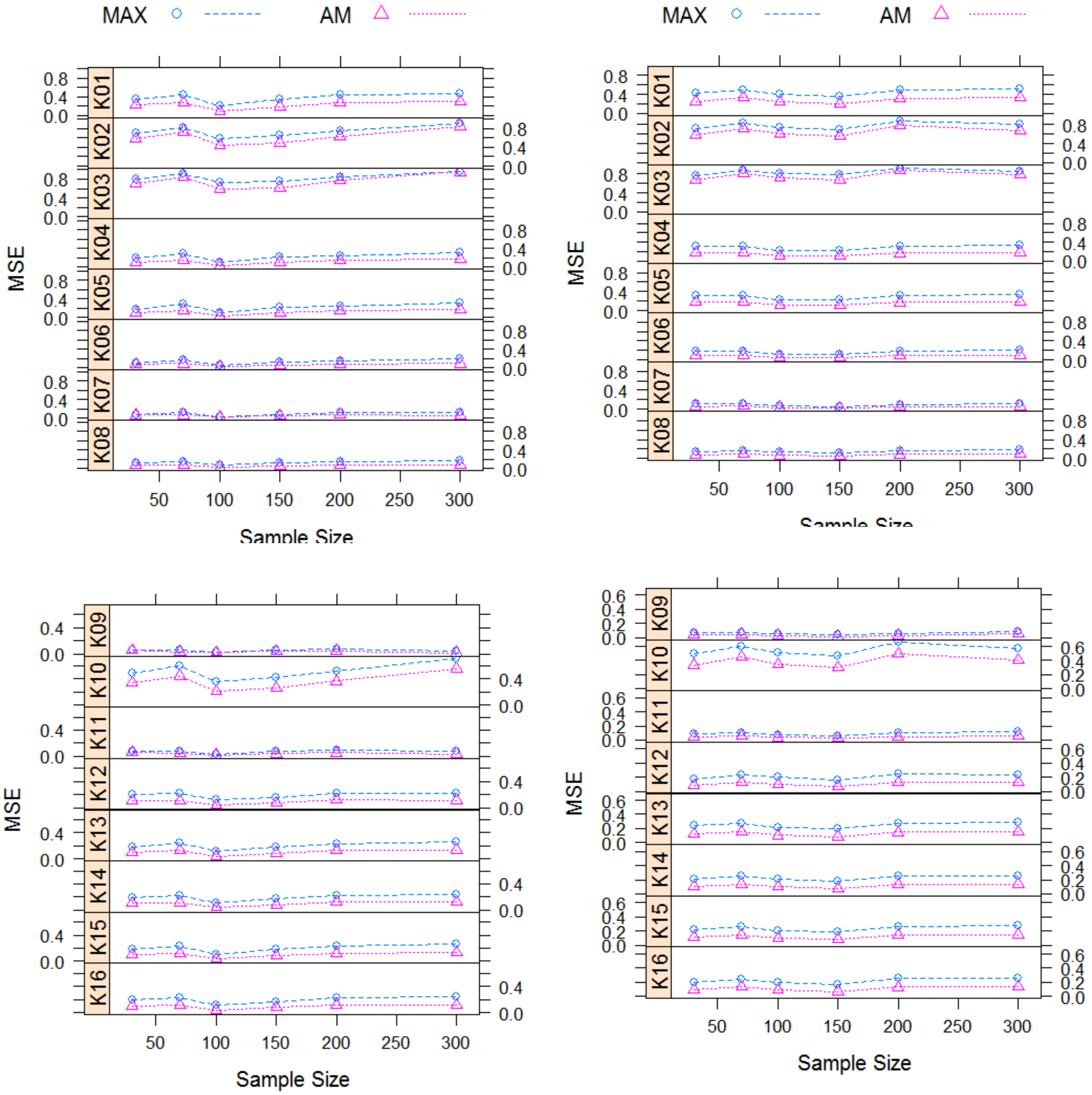

**Figure 4.2 (1):** MSE at p = 2 and $\varepsilon_i$~F (4, 20). 1st column is the case of r = 0.80 and 2nd column is the case of r = 0.90.

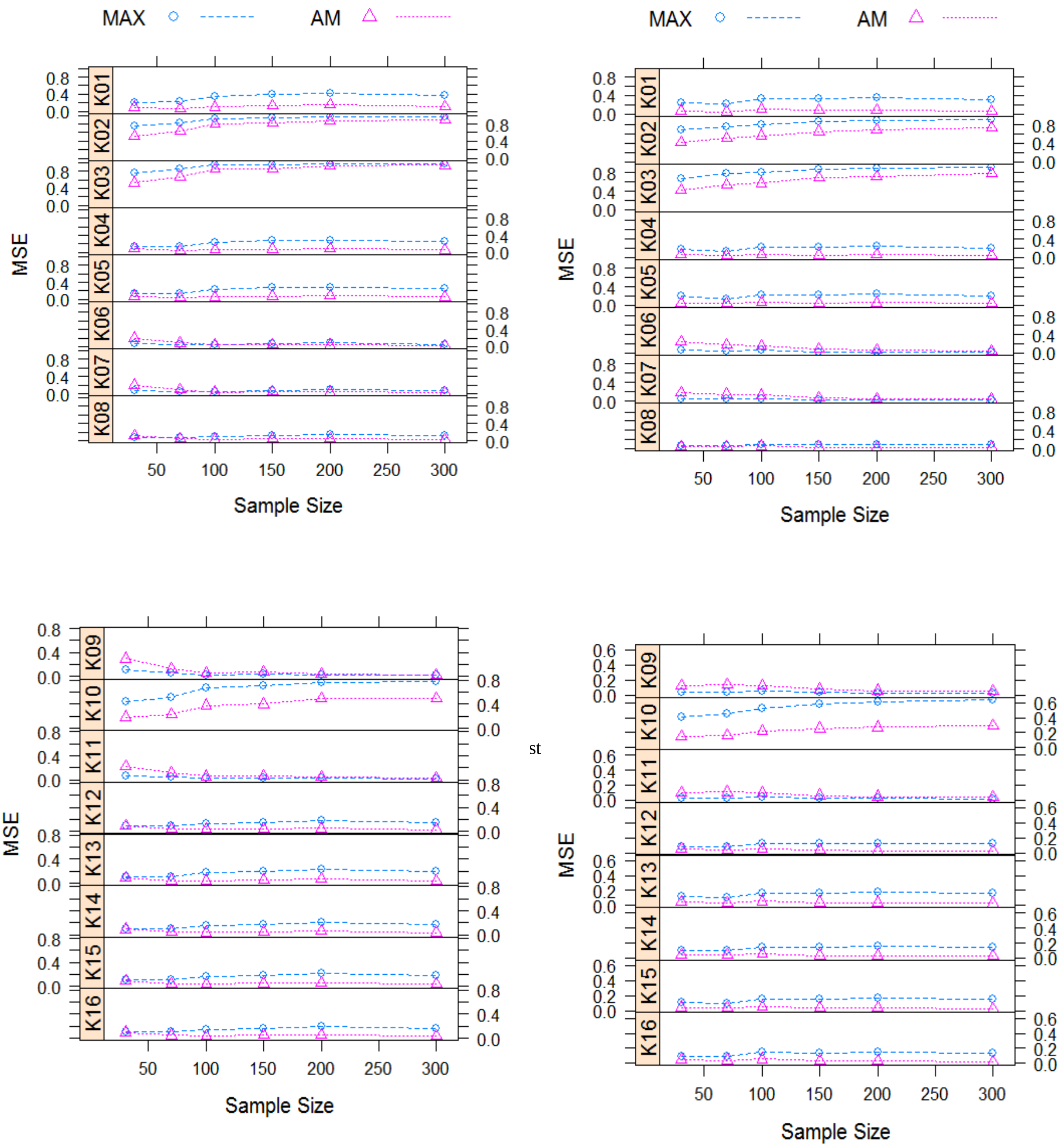

**Figure 4.2 (2):** MSE at p = 4 and $\varepsilon_i$~F (4, 20). 1st column is the case of r = 0.80 and 2nd column is the case of r = 0.90.

## 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. Alkhamisi, M. A. and Shukur, G. (2007). A Monte Carlo study of recent Ridge parameters. *Communications in Statistics—Simulation and Computation*, **36(3)**, 535-547.
3. Chatterjee, S. and Hadi, A. S. (2006). *Analysis of collinear data: Regression analysis by example* (4th ed.). John Wiley, New Jersy.
4. Draper, N. R. and Smith, H. (1981). An introduction to nonlinear estimation. *Applied Regression Analysis* (3rd ed.), 505-565.
5. Gibbons, D. G. (1981). A simulation study of some Ridge estimators. *Journal of the American Statistical Association*, **76(373)**, 131-139.
6. Gujarati, D. N. and Porter, D. C. (2003). *Basic Econometrics* (4th ed.). McGraw-Hill, New York.
7. Hoerl, A. E. and Kennard, R. W. (1976). Ridge Regression iterative estimation of the biasing parameter. *Communications in Statistics-Theory and Methods*, **5(1)**, 77-88.
8. Khalaf, G., Mansoon, K. and Shukur, G. (2013). Modified Ridge Regression estimators. *Communications in Statistics-Theory and Methods*, **42(8)**, 1476-1487.
9. Khalaf, G. and Shukur, G. (2005). Choosing Ridge parameter for Regression problems. *Communications in Statistics—Theory and Methods*, **34**, 1177–1182.
10. Kibria, B. G. (2003). Performance of some new Ridge Regression estimators. *Communications in Statistics-Simulation and Computation*, **32(2)**, 419-435.
11. McDonald, G. C. and Galarneau, D. I. (1975). A Monte Carlo evaluation of some ridge-type estimators. *Journal of the American Statistical Association*, **70(350)**, 407-416.
12. Shehzad M. A. (2012). Penalization and data reduction of auxiliary variables in survey sampling. PhD Thesis.
13. Vinod H. V. and Aman Ullah (1981). *Recent advances in Regression methods, Statistics: Textbooks and Monographs*, Marcel Dekker Inc. 41, New York.
14. Wichern, D. W. and Churchill, G. A. (1978). A comparison of Ridge estimators. *Technometrics*, **20(3)**, 301-311.