International Journal of Sciences: Basic and Applied Research (IJSBAR)

ISSN 2307-4531
(Print & Online)

http://gssrr.org/index.php?journal=JournalOfBasicAndApplied

Selecting the Optimal Value of Penalty Parameter K in Ridge Regression Estimators

Ahmed M. Mami[a*], Abdelbaset Abdalla[b], Eisay H. Bin Ismaeil[c]

[a,b,c]*Department of Statistics, Faculty of Science, University of Benghazi, Benghazi, Libya*

[a]*Email: ahmed.mami@uob.edu.ly*

[b]*Email: abdelbaset.abdalla@uob.edu.ly*

[c]*Email: Issawow30@gmail.com*

Abstract

Ridge regression is one of the popular parameter estimations techniques used to address the multicollinearity problem frequently arising in multiple linear regression. The ridge estimator is based on controlling the magnitude of regression coefficients. The Ridge regression constrains the sum of the absolute values of the regression coefficients to be less than some constant C which is called the penalty. The Ridge regression shrinks the ordinary least squares estimation vector of regression coefficients towards the origin, allowing a bias but providing a smaller variance. However, the choice of the optimal value of penalty parameter k in Ridge Regression estimators is critical. A Simulation study is conducted to uncover the optimal value of the penalty parameter k under different settings. This simulation study is novel in the field of Ridge Regression Estimators, and it increases the effective capabilities of using the Ridge Regression. Applications on three different real data sets are also considered to support the theoretical findings presented in the simulation study.

Key Words: Multicollinearity; Variance Inflation Factor (VIF); Shrinkage estimator; Ridge regression; Penalty parameter (k).

---

* Corresponding author.

1. Introduction

One major problem in least-squares analysis relate to failures of the basic assumptions which is the Multi-Collinearity Problem [7]. An alternative to least squares regression when some of the assumptions are not satisfied is known as the Robust regression. Robust regression refers to a general class of statistical techniques designed to reduce the sensitivity of the estimates to failures in the assumptions of the parametric model. A robust regression procedure would decrease the impact of such errors by reducing the weight given to large residuals. This can be done by minimizing the sum of absolute residuals, instead of the sum of squared residuals. See [7,3] for more details on robust statistics.

1.1 The Multicollinearity Problem

The inverse of (X^tX) may not exist, in this case, the matrix is called non-invertible or singular. One reason that this matrix might be non-invertible is that one or more of the explanatory variables are a linear combination of the other variables which is called a multicollinearity problem.

The impact of multi-collinearity on least squares is very serious if the purpose is to estimate the regression coefficients or if the purpose is to identify the important variables involved in the process. The estimates of the regression coefficients can differ greatly from the parameters they are estimating, even to the extent of having opposite sig. Moreover, the multi-collinearity allows important variables to be replaced in the model with related variables that are involved in the near singularity. Therefore, the regression analysis provides small suggestions of the relative importance of the explanatory variables

1.2 Multicollinearity Detection

Several tools have been founded to detect the presence of multicollinearity. Almost all statistical packages are not designed to inform the user automatically of the presence of near-collinearities. However, important hints are present such as unreasonable estimated values for regression coefficients, large standard errors,.etc. The solution for the multi-collinearity problem depend on the objective of the application:

- If the objective is prediction, multi-collinearity causes no serious problem within the sample explanatory variables space.
- If the objective is estimation of the regression coefficients, one of the biased regression methods may be useful.
- If the objective is to identify the important variables in the application, the regression results in the presence of severe collinearity will not be very helpful and can be misleading [11].

In this paper, our prime interest is to handle the multi-collinearity problem when estimating the coefficients for linear regression models.

1.3 Introduction to Biased Regression

If there is an exact linear relationship, it implies that such independent variables are exactly collinear. This means the correlation coefficient for these variables is equal to unity, so the variance is unacceptably large. An unbiased estimator condition can be relaxed to consider other possible estimators with better statistical properties in the presence of collinearity [9].

The biased regression term refers to a family of techniques dealing with multicollinearity that equilibrates partial regression coefficients by introducing bias. Even though, the increase in bias component there is again that more than compensates for the increase in bias which is a reduction in the variance component.

To our knowledge, the best measure of averaging nearness of an estimator to the parameter being estimated is the Mean Squared Error (MSE).

Now, let $\tilde{\theta}$ be a biased estimator with a smaller mean squared error than an unbiased estimator say, $\hat{\theta}$, the MSE of $\tilde{\theta}$ can be defined as

$$\text{MSE}(\tilde{\theta}) = E(\tilde{\theta} - \theta)^2 \qquad (1.1)$$

Recall that the variance of an estimator $\tilde{\theta}$ can also be defined as

$$\text{Var}(\tilde{\theta}) = E\left[\tilde{\theta} - E(\tilde{\theta})\right]^2 \qquad (1.2)$$

Assuming that $\tilde{\theta}$ is unbiased, then the expectation and MSE of the $\tilde{\theta}$ are $E(\tilde{\theta})=\theta$ and $\text{MSE}(\tilde{\theta}) =\sigma^2(\tilde{\theta})$, which means that $\text{Bias}(\tilde{\theta}) =0$. However, If the estimator is biased, then the MSE is given by

$\text{MSE}(\tilde{\theta}) =\sigma^2(\tilde{\theta}) + \text{Bias}^2(\tilde{\theta})$, where $\text{Bias}(\tilde{\theta}) = E(\tilde{\theta}) - \theta$.

To compensate for the high bias amount. the biased estimator obtains has variance smaller than the unbiased estimator. That represents the main strategy for the biased regression techniques [5]. As consequence might find an estimator which has an MSE that is smaller than the variance of an unbiased estimator [9]

1.4 Literature Review on Bias Regression Methods

Many biased regression methods have been proposed as solutions to the collinearity problem. Stein shrinkage (Stein(1960)), Ridge Shrinkage Regression [4], the **L**east **A**bsolute **S**hrinkage and **S**election **O**perator (**LASSO**) proposed by[9].

Although Ridge Shrinkage Regression has received the greatest acceptance, all have been used with apparent success in various problems. Ridge Shrinkage Regression and principal component regression are two used competitors for biased regression methods. For comprehensive details on the principal component regression, see [4, 7, 9 , and 10] for ridge regression.

2. The Ridge Shrinkage Regression

The problem of multicollinearity arises when independent variables are very correlated. One of the helpful tools that can be considered to deal with this problem is the Ridge Shrinkage Regression. If there is a multicollinearity problem, the estimates obtained through ordinary least squares will be unbiased, but their variances become much larger so those estimates may be far from the real value. To motivate the theoretical development of the Ridge Shrinkage Regression estimator, take a closer look at the mean squared error of the least squares estimator of $\boldsymbol{\beta}$

$$\text{MSE}(\widehat{\boldsymbol{\beta}}) = \text{E}||\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}||^2 \qquad (2.1)$$

Now, we can rewrite the MSE as the sum of the bias square and the variance, and can be written in the following form:

$$E\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right\|^2 = \sum_j E(b_j - \beta_j)^2 = \sum_j \{E(b_j) - \beta_j\}^2 + \sum_j Var(b_j) \quad (2.2)$$

According to the Gauss-Markov theorem, the least squares approach achieves the smallest variance among all unbiased linear estimates. Although, the minimum MSE is not necessarily guaranteed. To explain how ridge regression works, let $\widehat{\boldsymbol{\beta}}^{LS}$ denote the ordinary least squares estimator of $\boldsymbol{\beta}$ and consider the following linear regression model below,

$$y = \text{X}\beta + \varepsilon$$

the estimator $\hat{\beta}^{LS} = (\text{X}^t\text{X})^{-1}\text{X}^t y$ is unbiased estimator of β in addition ,

$\text{E}(\hat{\beta}^{LS}) = \beta$ and $Cov(\hat{\beta}^{LS}) = \sigma^2.(\text{X}^t\text{X})^{-1}$

We have

$\text{MSE}(\hat{\beta}^{LS}) = \text{E}\left\|\hat{\beta}^{LS}\right\|^2 - \|\beta\|^2.$

$$= \text{tr}\{\sigma^2(\text{X}^t\text{X})^{-1}\} = \sigma^2.tr\{(\text{X}^t\text{X})^{-1}\} \qquad (2.3)$$

Therefore, by rearrange (2.3), we get

$$E\left(\left\|\hat{\beta}^{LS}\right\|^2\right) = \|\beta\|^2 + \sigma^2.tr\{(\text{X}^t\text{X})^{-1}\} \qquad (2.4)$$

Because of the ill-conditioned in X^tX, the resultant least square estimate of $\hat{\beta}^{LS}$would be large in length$\left\|\hat{\beta}^{LS}\right\|$ and related to large standard errors. As well, this large variation would lead to the poor model prediction. The Ridge Shrinkage Regression is a constrained type of least squares. However, It solves the estimation problem by producing a biased estimator, with small variances [10].

2.1 Derivation of Ridge Shrinkage Estimator

Let $\hat{\beta}$ be the least squares estimator, the least squares criterion can be rewritten as its minimum, reached at$\|\hat{\beta}^{LS}\|$. The quadratic form in b:

$$Q(b) = \|y - X\hat{\beta}^{LS} + X\hat{\beta}^{LS} - Xb\|^2$$

$$= (y - X\hat{\beta}^{LS})^t (y - X\hat{\beta}^{LS}) + (b - \hat{\beta}^{LS})^t X^t X (b - \hat{\beta}^{LS})$$

$$= Q_{min} + \phi(b) \qquad (2.5)$$

Contours for each constant of the quadratic form $\phi(b)$ are hyperellipsoids centered at the ordinary LSE $\hat{\beta}^{LS}$.It is reasonable to expect from (7) that, if one moves away from Q_{min}, the movement is in a direction that shortens the length of $\hat{\beta}$.

In Ridge Shrinkage Regression, the optimization problem can be defined as:

$$\text{minimizing } \|\beta\|^2 \text{subject to } (\beta - \hat{\beta}^{LS})^t X^t X (\beta - \hat{\beta}^{LS}) = \phi_0 \qquad (2.6)$$

for some constant ϕ_0.The imposed constrain guarantees a reasonably small residual sum of squares Q(β) when compared to its minimum Q_{min}. Figure (1) displays the contours of residual sum of squares together with theL_2 ridge shrinkage constraint in the two-dimensional case [8].

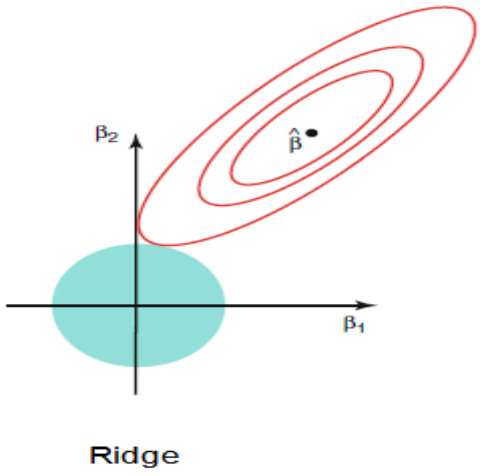

Figure1: Contours of the Sum of Squares of the Residual and the Constraint Functions in Ridge Shrinkage Regression.

In the view of Lagrangian problem, it is equivalent to minimizing

$$Q^*(\beta) = \|\beta\|^2 + (1/k)\{(\beta - \hat{\beta}^{LS})^t X^t X (\beta - \hat{\beta}^{LS}) - \phi_0\} \qquad (2.7)$$

where k is the deflection factor chosen to satisfy the constraint.

Therefore, differentiate $Q^*(\beta)$with respect to β

$$\frac{\partial Q^*(\beta)}{\partial \beta} = 2\beta + (1/k)\{2(X^tX)\beta - 2(X^tX)\hat{\beta}^{LS}\} = 0 \qquad (2.8)$$

that yields the Ridge Shrinkage estimator as follows

$$\hat{\beta}^R = \{X^tX + k\mathrm{I}\}^{-1}X^ty \qquad (2.9)$$

An alternative way is to state the Ridge Shrinkage problem in the constrained least squares form by

minimizing$\|y - X\beta\|^2$, subjectto$\|\beta\|^2 \leq s$,

for some constant value of *s*.

Hence, the Lagrangian problem becomes simply minimizing that

$$\|y - X\beta\|^2 + \lambda . \|\beta\|^2$$

which produces the same estimator given in (2.9). The penalty parameter $\lambda \geq 0$ controls the amount of shrinkage in $||\beta||^2$. As the value gets larger, the greater amount of shrinkage. For this reason, the Ridge Shrinkage estimator is often called the shrinkage estimator. There is a one-to-one correspondence among four values, s, k, and ϕ_0 [1]. It is extremely important to note that the formal Ridge Shrinkage solution is not invariant under the scaling of the explanatory variables. Therefore, standardization of both the explanatory variables and the response is essential.

It is very important to note that the formal Ridge Shrinkage solution is not invariant under the scaling of the explanatory variables. Therefore, standardization of both the explanatory variables and the response is essential, that is:

$$x'_{ij} = \frac{x_{ij} - \bar{x}_j}{s_{xj}} \ \ and \ \ y'_i = \frac{y_i - \bar{y}}{s_y}$$

before using the Ridge Shrinkage estimator in (2.9). It is helpful to adopt the following standardized variables notation, the matrices X^tX and X^ty become as follows:

$X^tX = R_{XX}$ and $X^ty = r_{XY}$

Note that $\mathbf{R_{XX}}$ denotes the correlation matrix among $\boldsymbol{X_j}$'s, and $\mathbf{r_{XY}}$denotesthe correlation vector between *Y* and all $\boldsymbol{X_j}$'s. Now, the Ridge Shrinkage estimator can be written as :

$$\hat{\beta}^R = \{R_{XX} + k\mathrm{I}\}^{-1}r_{XY} \qquad (2.10)$$

If the explanatory variables are orthogonal ($\mathbf{X^tX = I}$), then the Ridge Shrinkage estimates are just a scaled version of least squares estimates (it is equivalent to , $\widehat{\boldsymbol{\beta}}^{\boldsymbol{R}} = \frac{\mathbf{1}}{\mathbf{1+k}}.\widehat{\boldsymbol{\beta}}^{\boldsymbol{LS}}$for some shrinkage constant ($0 \leq \frac{\mathbf{1}}{\mathbf{1+k}} \leq \mathbf{1}$).

In addition, the intercept **value** $\boldsymbol{\beta_0}$ goes to**0** when working with standardized data. Having obtained a Ridge Shrinkage estimator $\widehat{\boldsymbol{\beta}}^{\boldsymbol{R}}$, transformation step of its components is necessary in order to get the fitted linear regression equation between the original Y and $\boldsymbol{X_j}$ values. It is suitable to express in matrix form the normalization and its inverse transformation involved. Let $\mathbf{X_0}$ be the original design matrix. Its centered version is given by :

$$\mathrm{X}_C = (\mathrm{I} - \mathrm{j}_n\mathrm{j}_n^t/n)\mathrm{X}_0$$

and its normalized version is

$\mathrm{X} = \mathrm{X}_C\mathrm{L}^{-1/2}$

where$\mathbf{j_n}$be the n-dimensional vector with all elements are ones and L be a diagonal matrix with diagonal elements from the matrix $\mathbf{X_c^tX_C}$, i.e.,

$\mathrm{L} = \mathrm{diag}\,(\mathrm{X}_c^t\mathrm{X}_C)$

Likewise, the original response vector y_0 can be normalized as

$$\mathrm{y} = \frac{(\mathrm{I} - \mathrm{j}_n\mathrm{j}_n^t/n)\mathrm{y}_0}{s_y}$$

wheres_y is the sample standard deviation of y_0 [2] .

It is straightforward to use the Ridge Shrinkage estimator $\hat{\beta}^R$ in (2.9) to predict with a new data matrix X_{new} (which is m×p on the original data scale), .

The predicted vector $\widehat{\boldsymbol{y}}_{new}$is then given as :

$$\hat{y}_{new} = s_y\,.\{(\mathrm{X}_{new} - \mathrm{j}_m\mathrm{j}_n^t\mathrm{X}/n)\mathrm{L}^{-1/2}\hat{\beta}^R + \mathrm{j}_m\mathrm{j}_n^t y/n\} \qquad (2.11)$$

Thus, the computation of the expectation and variance of $\hat{\beta}^R$ can be obtained using the following relation

$$\hat{\beta}^R = \mathrm{Z}\,\hat{\beta}^{LS} \qquad (2.12)$$

where

$\mathrm{Z} = \{\mathrm{I} + k(\mathrm{X^tX})^{-1}\}^{-1}$

It follows that

$$E(\hat{\beta}^R) = Z\,\beta \qquad (2.13)$$

$$Cov(\hat{\beta}^R) = \sigma^2.Z(X^tX)^{-1}Z^t \qquad (2.14)$$

Finally, comparison can be achieved between $\hat{\beta}^R$ with $\hat{\beta}^{LS}$ to see which estimator having a smaller MSE for certain values of k.

Let the ascending order sequence of the eigen values of Z matrix as follows:

$$\lambda_{max} = \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m = \lambda_{min} > 0$$

From standard least square estimation, it is well known that

$MSE(\hat{\beta}^{LS}) = \sigma^2.\sum_j 1/\lambda_j$

For the Ridge Shrinkage estimator, the components of the Mean squared errors can be found from (2.13) and (2.14). The first component is the sum of their squared biases is

$$\sum_j \{E(\hat{\beta}_j^R) - \beta_j\}^2 = \{E(\hat{\beta}^R) - \beta\}^t\{E(\hat{\beta}^R) - \beta\}$$

$$\sum_j \{E(\hat{\beta}_j^R) - \beta_j\}^2 = \beta^t(I - Z)^t(I - Z)\beta$$

$$= k^2\beta^t(X^tX + kI)^{-2} \qquad (2.15)$$

and the second component is sum of their variances is

$tr\{Cov(\hat{\beta}^R)\} = \sigma^2.tr\{(X^tX)^{-1}z^tz\}$

$$= \sigma^2 \sum_j \left\{\frac{1}{\lambda_j}.\frac{\lambda_j^2}{(\lambda_j+k)^2}\right\}$$

$$= \sigma^2 \sum_j \left\{\frac{\lambda_j}{(\lambda_j+k)^2}\right\} \qquad (2.16)$$

Therefore, the MSE for the Ridge Shrinkage estimator is as follows

$$MSE(\hat{\beta}^R, k) = \sigma^2 \sum_j \left\{\frac{\lambda_j}{(\lambda_j+k)^2}\right\} + k^2\beta^t (X^tX) + kI)^{-2}\beta$$

$$= \gamma_1(k) + \gamma_2(k) \qquad (2.17)$$

It is worth noting that the first quantity$\gamma_1(k)$ is a monotonic decreasing function of k while the second quantity

$\gamma_2(k)$ is monotonically increasing. The constant k reflects the amount of bias increased and the variance decreased. However, when k = 0, it turns into the usual Least Squares Estimate [4] had showed that there always exists a k > 0 such that

$\text{MSE}(\hat{\beta}^R, k) < \text{MSE}(\hat{\beta}^R, 0) = \text{MSE}(\hat{\beta}^{LS})$

Finally, the Ridge Shrinkage estimator can be superior in comparing with the Least Squares Estimator in terms of providing a smaller MSE. However, in practice the right choice of k is yet to be determined and hence there is no guarantee that a smaller MSE always be achieved by the Ridge Shrinkage Regression.

2.2 Optimal Value of Shrinkage Coefficient k

Choosing the optimal value of k is the most important issue in Ridge Shrinkage Regression. Many strategies have been developed. One choice is using a trial-and-error strategy. Several choices of k are tried until the desired amount of smoothness is achieved.

Another choice of choosing the value of k is selected by some measure of prediction error such as the cross-validation technique. The optimal value of k is the one that maximizes the prediction accuracy. However, the formal two choices are not particle and can lead to overfitting of the underlying model [8].

Since the estimation of the regression model parameters is the main goal in this paper, it is really to claim that the explanatory variables can be treated as a random sample. Then, the usual regression assumptions must be met. for example, there are correlated explanatory variables, whether the resulting biases are likely to be large enough to matter.

Also, the Ridge Shrinkage Regression introduces some additional problems. Starting with the estimation of the regression model coefficients:

Firstly, the fitted values are biased by design.

Secondly, if hypothesis tests are undertaken and the usual regression output is used, then the reported p-values are no longer accurate.

Thirdly, if standard confidence intervals are constructed, then they do not have their usual coverage.

Lastly, the used regression estimates are necessary to be balanced by some efficient Shrinkage Coefficient k, unfortunately, it is unknown.

The main purpose of the Ridge Shrinkage Regression is to address the instability of estimated regression coefficients when explanatory variables are highly correlated. Moreover, the Ridge Shrinkage Regression also provides another way to adjust the smoothness of the fitted values.

As it has been mentioned in the problem statement of this paper, finding the optimal value of shrinkage

coefficient k is one of our main goals.

3. The Simulation Study

Reference [6] introduced the Ridge regression estimator as an alternative to the ordinary least squares estimator to overcome the potential effects of multicollinearity. In this section, a simulation study is performed to select the ridge parameter (k) when there is multicollinearity between the columns of the design matrix in the form of mean square errors (MSE). Several factors that can affect the properties of this technique have been varied.

3.1 Description of The Experiment

In this simulation study, the main goal is to determine the best value of k that will make the ridge regression estimator work optimally. In order to better quantify some of the commonly mentioned advantages of the Ridge Regression Estimator, we conduct the following simulation study. The simulation setting includes the use of different correlation options $\rho = (0.10,0.25,0.25,0.50, and\ 0.90)$,several choices of the number of independent variables (p = 2, 5, 10 and 20), several taken into account Sample sizes (n = 25, 50, and 100).

The model was of the form Y= $X\beta + \varepsilon$ with the normal distribution as the marginal distributional of errors were used.

3.2 The Numerical Summary

The numerical results include the means and standard deviations of the resulting k value and their corresponding counterparts MSE values and standard deviations acquired by utilizing the **Ridge** Regression estimators. The graphical summary of the results for case P = 2, 5, 10, and 20 in numerous choices of correlation and sample sizes are shown in four panels in Figure 1.

After taking a closer look at various panels in Figure 1, we found that the 95% confidence interval of the optimal value of k resulting from our simulation, according to the number of explanatory variables (P), and the can be summarized as follows:

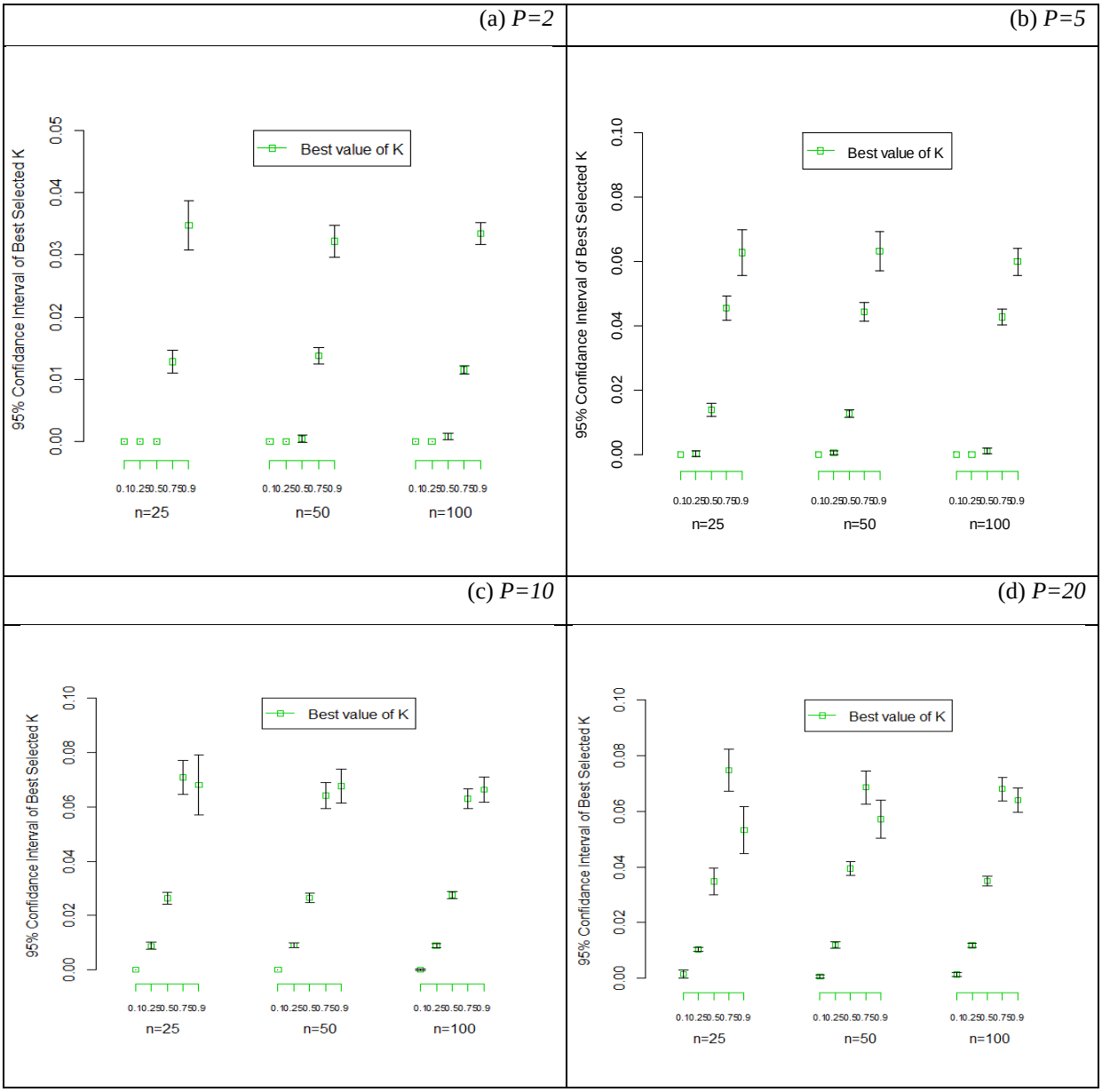

Figure 1: The 95% confidence intervals of the Selected k value using Ridge Regression estimator for various choices of correlation and sample sizes for (a)p=2, (b)p=5, (c)p=10, and (d)p=20.

Table 7

	P = 2	P = 5	P = 10	P = 20
$\rho = 0.1$	$0.0000 \leq k \leq 0.0000$	$0.0000 \leq k \leq 0.0000$	$0.0000 \leq k \leq 0.0001$	$\mathbf{0.0006 \leq k \leq 0.0016}$
$\rho = 0.25$	$0.0000 \leq k \leq 0.0000$	$0.0000 \leq k \leq 0.0004$	$0.0088 \leq k \leq 0.0090$	$\mathbf{0.0104 \leq k \leq 0.0120}$
$\rho = 0.50$	$0.0000 \leq k \leq 0.0008$	$0.0012 \leq k \leq 0.0140$	$0.0264 \leq k \leq 0.0275$	$\mathbf{0.0348 \leq k \leq 0.0394}$
$\rho = 0.75$	$0.0115 \leq k \leq 0.0138$	$0.0428 \leq k \leq 0.0456$	$0.0630 \leq k \leq 0.0708$	$\mathbf{0.0680 \leq k \leq 0.0748}$
$\rho = 0.90$	$0.0322 \leq k \leq 0.0348$	$0.0600 \leq k \leq 0.0632$	$0.0663 \leq k \leq 0.0680$	$\mathbf{0.0532 \leq k \leq 0.0640}$

Important remarks:

In the case of number of explanatory variables P = 2

we have noticed the following remarks:

1- At n = 25, the value of k at = 0.10, 0.25, and 0.50 was zero, which means that the OLS estimator can also be used. In other words, there is no serious multicollinearity other than it appears to be increasing.

2- When n = 50, the value of k starts to appear initially when = 0.50, and its value increases. However, the first two values of k were set to zero when the value of = 0.10 and 0.25.

3- When n = 100, the value of k starts to appear when = 0.50 and its value continues to increase. However, the first two values of k were set to zero when the value of = 0.10 and 0.25.

4- From Table 3.9 we found that the value of k is limited to ($0.0004 \leq k \leq 0.0348$), since the lowest value at n = 100 at = 0.50 and the highest value at n = 25 at = 0.90 . was measured.

In the cases of number of explanatory variables P = 5, 10, and 20

we noticed the following remarks:

if p = 5, we notice that the initial value of k is increasing in all values of (n), so we find that in certain cases it is exactly zero, namely:

when n = 25 at $\rho = 0.10$,

when n = 50 at $\rho = 0.10$.

when n = 100 at $\rho = 0.25$, and $\rho = 0.10$

The value of k lies between ($0.0004 \leq k \leq 0.0632$), with the lowest value recorded at n = 25 at = 0.25 and the highest value recorded at n = 50 at = 0.90.

When p = 10, we notice that the initial value of k has evolved with the increasing value of (n) increases, except when $\rho = 0.75$ at n = 25, where it reached the largest value and then decreased when $\rho = 0.90$.

Likewise, the value of k was equal to zero in two cases, namely when n = 25 at $\rho = 0.10$, when n = 50 when $\rho = 0.10$.

The value of k ranges between ($0.0088 \leq k \leq 0.0708$), with the lowest value recorded at n = 25 at = 0.25 and the highest value recorded at n = 25 at = 0.75.

When p = 20, we notice that the initial value of k was scaled with increasing value of (n) increases, except in three cases when ρ = 0.75 with (n = 25, n = 50, n=100) where it reached the largest value and then decreased when ρ = 0.90.

Likewise, the value of k ranges between (0.0104≤k ≤0.0748), where the lowest value being recorded at n = 25 at ρ = 0.25, and the largest value was recorded when n = 25 at ρ = 0.75.

Broadly speaking, we notice that the optimal value of k it ranges between a high and a low value (0.0004 ≤ k ≤ 0.0748) regardless the selected number of explanatory variables (p = 2, p = 5, p = 10 and p = 20), or the different values of (ρ = 0.10, ρ = 0.25, ρ = 50, ρ = 75 and ρ = 0.90). The major effect on choosing the optimal k values was due to the correlation coefficient (ρ) first and then the number of explanatory variables (p).

4. Applications on Real Data

In this section, we looked at three different data sets. There are two main steps in dealing with these data sets: (1) recognizing the presence of multicollinearity; (2) Implementing solutions to get more consistent results.

4.1 The Hald Data

The Hald data are used by Hoerl, Kennard, and Baldwin (1975). This data is simulated data utilized by the authors to investigate the multicollinearity problem. The first column is the answer on the logarithmic scale, the remaining columns are the predictors. There are a total of 13 observations, and it can be found in R in the Ridge package under the name (Hald data).

The variance inflation factor (VIF) is a measure of multi-collinearity. It is the reciprocal of 1- R^2_x, where R^2_x is the R^2 obtained when this variable is regressed on the remaining independent variables. A VIF of 10 or more for large data sets indicates a multi-collinearity problem since the Rx2 with the remaining X's is 90 percent. For small datasets, even VIF's of 5 or more can signify multi-collinearity.

Least Squares Multi-collinearity detecting:

Table 1: Values of VIF and Rx2 of the independent variables in the Hald Data.

Independent Variable	VIF	Rx2 vs Other X's	Tolerance
$\boldsymbol{X_1}$	38.4962	0.9741	0.0259
$\boldsymbol{X_2}$	254.423	0.9961	0.0039
$\boldsymbol{X_3}$	46.8684	0.9786	0.0214
$\boldsymbol{X_4}$	282.513	0.9965	0.0035

Since all VIF's are greater than 10, the multicollinearity problem exists.

Table 2 summarizes the numerical values of the inference results obtained using the OLS and Ridge, including the estimated parameters, the standard error of the estimated parameter, the *P-values* of the test statistic of the estimated parameter, and the value of the VIF. Other important criteria such as MSE, Adj, AIC, and BIC are also taken into account for comparison purposes.

Table 2: The Results of fitting the OLS, and Ridge estimators for the Hald Data.

	OLS				RIDGE			
	$\hat{\beta}$	Stander ϵ	P-Value	VIF	$\hat{\beta}$	Stander ϵ	P-Value	VIF
Intercept	62.405	2129.7	0.5762		82.675	308.64	0.0406	
X_1	1.5511	14.308	0.0582	38.496	1.3152	3.9697	0.0000***	3.163
X_2	0.5102	36.784	0.4761	254.423	0.3061	5.3166	0.0127	5.675
X_3	0.1019	15.787	0.8897	46.868	-0.1290	3.9468	0.0486**	3.127
X_4	-0.1441	38.761	0.8348	282.512	-0.3429	5.4433	0.0053*	5.948
MSE		3309.505			392.5526			
R^2		0.98240			0.9719			
Adj R^2		0.97650			0.9625			
AIC		24.94429			23.27944			
BIC		60.54843			58.36274			

Based on the results in Table 2, we noted the following remarks:

1. Using the OLS estimator, all variables have VIF values greater than 30. This is a strong indication of the presence of multicollinearity, even though the model explains greater than 98.24%.
2. When the ridge regression estimator is used with k = 0.01, all variables have VIF values less than 10. This is strong evidence that the multicollinearity has been removed, even though the model explained = 97.19% less.
3. All estimation parameters of the Ridge regression estimator are statistically significant at 0.05, while most of the estimated parameters of the OLS estimator are statistically insignificant at 0.05. Therefore, the Ridge Regression Estimator provided a valid model for the Hald data.
4. Most importantly, the Ridge regression estimator has a value of MSERidge = 392.5526, which is much smaller than the value of MSEOLS = 3309.505 obtained using the OLS estimator
5. The AIC and BIC criteria also support the superiority of the Ridge Regression estimator over the traditional OLS estimator. Since the value of AICRidge = 23.27944 and the value of BICRidge= 58.36274, whereas the value of AICOLS = 24.94429 and the value of BICOLS= 60.54843.

Choosing Optimal Ridge k Coefficient

One of the main practical problems in using ridge regression is choosing an appropriate value for k. Hoerl and Kennard (1970), the inventors of ridge regression, suggested using a graph they called the ridge trace. Figure 3 shows the comb regression coefficients as a function of k. Looking at the ridge curve, the analyst selects a value for k for which the regression coefficients have become stable. Often the regression coefficients vary widely for small values of k and then become stable. Choose the smallest possible value of k (which introduces the smallest bias) after which the regression coefficients appear to remain constant. Note that increasing k eventually drives the regression coefficients to zero. The Hald data was found to be k = 0.012, which is consistent with the result obtained using other criteria (minimum CV at 0.01000 and HKB (1975) at 0.01162).

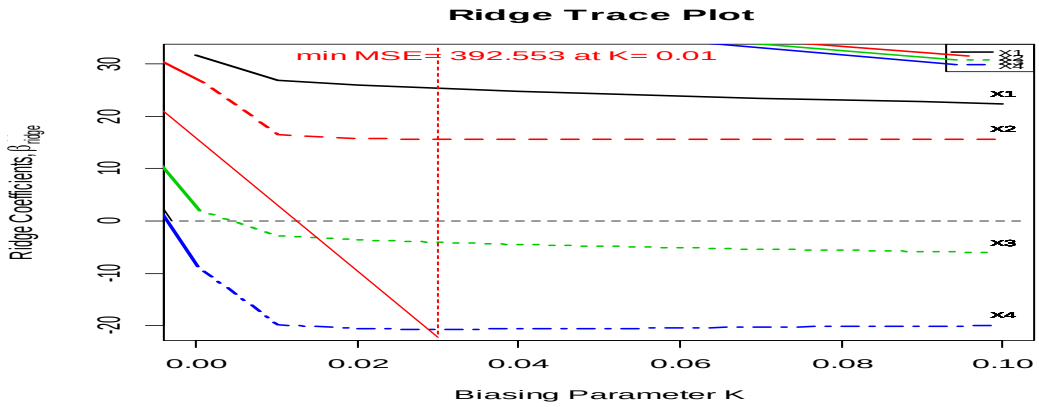

Figure 3: The Trace plot of the optimal Ridge k coefficient using Hald data.

According to the simulation results, one can determine the optimal range of the ridge coefficient *k* by using the confidence interval provided, for a certain number of independent variables (p), and correlation coefficient (ρ). In Hald data, it was found that p=4 ≈ 5 and for ρ, it can be computed by using the greatest eigenvalue of the spectrum analysis of the correlation matrix of the independent variables divided by (p), and it was found to be = 0.5589 ≈ 0.5. Therefore, the appropriate optimal value of *k* is within the range of ($0.0012 \leq k \leq 0.0140$). That agrees with the Ridge Coefficient *k* obtained from the numerical analysis of the Hald data.

4.2 Consumption Car (CC) Data

The consumption vehicle data is very typical data showing the presence of multicollinearity. The goal of consumption vehicle data is to predict the consumption of cars (consumption) based on various characteristics (price, displacement, horsepower, and weight). A data frame with 27 observations on 5 variables, which can be found in R in *the Ridge package* under the name Cars93 data).

Least Squares Multi-collinearity detecting:

Table 3: Values of VIF and Rx2 of the independent variables in the CC Data.

Independent Variable	VIF	Rx2 vs Other X's	Tolerance
PRIX	19.7919	0.9495	0.0505
CYLINDREE	12.8689	0.9223	0.0777
PUISSANCE	14.8923	0.9329	0.0671
POIDS	10.2259	0.9022	0.0978

Since all VIF's are greater than 10, multi-collinearity problem exists.

We utilize the two different estimators to handle the regression curve estimation (OLS and Ridge) for the CC data. Table 4. consists of the numerical values of the inferential results.

Table 4: The Results of fitting the OLS, and Ridge estimators for the CC Data

	OLS				RIDGE			
	$\hat{\beta}$	Stander ϵ	P-Value	VIF	$\hat{\beta}$	Stander ϵ	P-Value	VIF
Intercept	1.8380	80212.0	0.3553		1.96802	62397.4	0.2109	
PRIX	0.0000	2.8333	0.4497	19.791	0.00004	2.20346	0.2921	2.886
CYLINDREE	0.0012	2.2846	0.1013	12.868	0.00109	1.90114	0.0756*	2.913
PUISSANCE	-0.0037	2.4577	0.8014	14.892	-0.00137	1.96898	0.9083	2.805
POIDS	0.0037	2.0365	0.0077***	10.225	0.00356	1.67593	0.0024***	2.591
MSE	23.43428				9.13219			
R^2	0.92950				0.9217			
Adj R^2	0.92030				0.9114			
AIC	-20.69476				-21.55163			
BIC	73.47615				72.01848			

By carefully examining the results in Table 4, we noted the following remarks

1- Using the OLS estimator, all variables have VIF values greater than 10. This is a strong indication of the presence of multicollinearity presence, even though the model explained more than 92.95%.

2. When using the Ridge Regression estimator with k = 0.06, all variables have VIF values less than 10. This is strong evidence that the Multi-collinearity has been removed, even though the model explained less with = 92.17%.

3- Although not all of the estimated parameters of the ridge regression estimator are statistically significant at 0.05, whereas most of the estimated parameters of the OLS estimator are not statistically significant at 0.05. Therefore, the ridge regression estimator has provided a better model for the CC data.

4- Most importantly, the Ridge Regression estimator has a value of MSERidge = 9.132191, which is much smaller than the value of MSEOLS = 23.43428 that obtained through using the OLS estimator

5- The AIC and BIC criteria provide additional support for the superiority of the Ridge Regression estimator over the traditional OLS estimator. Since the value of AICRidge = -21.55163 and the value of BICRidge= 72.01848, whereas the value of AICOLS = -20.69476 and the value of BICOLS= 73.47615.

Choosing Optimal Ridge k Coefficient

Figure 4 shows the ridge regression coefficients as a function of *k*. Looking at the ridge trace, the analyst picks a value for *k* for which the regression coefficients have become stable. For the CC data, it was found that *k*= 0.06, which agrees with the result obtained using other criteria (Minimum CV at 0.06000, and Minimum GCV at 0.08000).

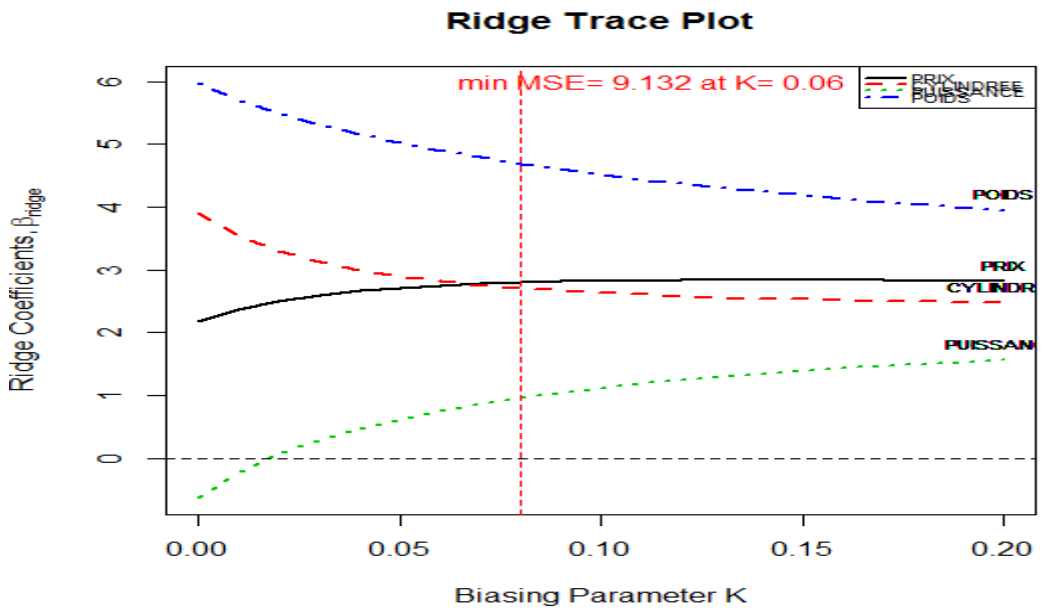

Figure 4: The Trace plot of the optimal Ridge coefficient k using CC data.

According to the simulation results, one may determine the optimal range of the Ridge Coefficient *k* by using the confidence interval provided, for a specific number of independent variables (p), and correlation coefficient (ρ). In CC data, it was found that p=4 ≈ 5. For and ρ, it can be computed using the greatest eigenvalue of the spectrum analysis of the correlation matrix of the independent variables divided by (p), and it was found to be = 0.93277 ≈ 0.9. Therefore, the appropriate optimal value of *k* is within the range of (0.0600 ≤ k ≤ 0.0632). That agrees with the Ridge *k* Coefficient obtained from the numerical analysis of the CC data.

4.4 Housing Values in Suburbs of Boston Data

The Boston data are used by Harrison and Rubinfeld (1978). These data are concerned with the housing values problem. The Boston data frame has 506 rows and 14 columns. The last column is the response on the log scale, the remaining columns are the predictors. There are 506 observations in total, and it can be found in R in the MASS package under the name (Boston data).

Least Squares Multi-collinearity detecting:

Table 5: Values of VIF and Rx2 of the independent variables in the Boston Data.

Independent Variable	VIF	Rx2 vs Other X's	Tolerance
Crime	1.7922	0.4420	0.5580
Zn	2.2988	0.5649	0.4351
Indus	3.9916	0.7495	0.2505
Chas	1.0740	0.0689	0.9311
Nox	4.3937	0.7724	0.2276
Rm	1.9337	0.4829	0.5171
Age	3.1008	0.6775	0.3225
Dis	3.9559	0.7472	0.2528
Rad	7.4845	0.8664	0.1336
Tax	9.0086	0.8889	0.1111
Ptratio	1.7991	0.4442	0.5558
Black	1.3485	0.2584	0.7416
Lstat	2.9415	0.6600	0.3400

Since most of the VIF values are less than 5, we can identify the source of the multi-collinearity problem. The variables tax and rad have VIF values that are more than 5 with a coefficient of determination ≥ 86%. We utilize the two different estimators to handle the regression curve estimation (OLS, and Ridge) for the CC data. Table 6 consists of the numerical values of the inferential result.

Table 6: The Results of fitting the OLS, and Ridge estimators for the Boston Data. The Significant codes are 0 '***' 0.001 '**' 0.01 '*' 0.05.

	OLS				RIDGE			
	$\hat{\beta}$	Stander ϵ	P-Value	VIF	$\hat{\beta}$	Stander ϵ	P-Value	VIF
Intercept	36.46	6162.9	*		34.695	5467.9	***	
Crim	-0.108	6.3462	**	1.792	-0.1035	6.20	***	1.707
Zn	0.046	7.1874	***	2.298	0.0434	6.93	***	2.138
Indus	0.020	9.4710	0.73	3.991	0.0052	8.90	***	3.517
Chas	2.68	4.9127	**	1.074	2.7463	4.85	***	1.046
Nox	-17.76	9.9366	***	4.393	-16.62	9.41	***	3.938
Rm	3.80	6.5921	***	1.933	3.86	6.38	***	1.811
Age	0.0007	8.3476	***	3.101	-0.00034	8.01	***	2.854
Dis	-1.47	9.4286	***	3.956	-1.413	8.96	***	3.576
rad	0.306	12.969	***	7.485	0.27	11.43	***	5.813
Tax	-0.012	14.228	**	9.009	-0.0106	12.44	***	6.882
Ptratio	-0.95	6.3584	***	1.799	-0.93	6.18	***	1.697
Black	0.009	5.5049	***	1.349	0.0093	5.42	***	1.307
Lstat	-0.525	8.1303	***	2.942	-0.516	7.82	***	2.721
MSE	1014.01				1004.27			
R^2	0.74060				0.7287			
Adj R^2	0.73				0.72			
AIC	1587.64				1587.3**2**			
BIC	4793.21				4791.12			

Having looked carefully at Table 6., we noted the following remarks

1- Using the OLS estimator, two variables have VIF values greater than 5. This is a strong indication that multicollinearity is still present, even though the model is more declared as 74.06%.

2- If ridge regression estimator is used with k = 0.01, again 2 variables have VIF values greater than 5. This is a strong evidence that the Multi-collinearity has not been removed, even though the model explained less with R^2=72.87%.

3- All estimated parameters of Ridge Regression estimator are statistically significant at 0.001, whereas 6 of estimated parameters of OLS estimator are not statistically significant at 0.01. Therefore, the Ridge

Regression estimator has provided a valid model for the Boston data.

4- Most important that Ridge Regression estimator has a value of MSE_{Ridge} = 1004.27, which is much smaller the value of MSE_{OLS} = 1014.01 that obtained through using the OLS estimator.

5- The AIC and BIC criteria provide additional support for the superiority of the Ridge Regression estimator over the traditional OLS estimator. Since the value of AIC_{Ridge} = 1587.319 and the value of BIC_{Ridge}= 4791.121, whereas the value of AIC_{OLS} = 1587.64282 and the value of BIC_{OLS}= 4793.21539.

Choosing Optimal Ridge k Coefficient

Figure 5 shows the ridge regression coefficients as a function of *k*. When viewing the ridge trace, the analyst picks a value for *k* for which the regression coefficients have become stable. For the Boston data, it was found that *k*= 0.01, it agree with the result obtained using other criteria (HKB (1975) at 0.01071, Minimum CV and Minimum GCV at 0.01000).

According to the simulation results, one may determine the optimal range of the Ridge *k* Coefficient through using confidence interval provided, for a specific number of independent variables (p), and correlation coefficient (ρ). In Boston data, it was found that p=13 ≈ 10. For and ρ, it can be computed using the greatest eigen value of the spectrum analysis of the correlation matrix of the independent variables divided by (p), and it was found to be = 0.38232 ≈ 0.25. Therefore, the appropriate optimal value of *k* is within the range of (0088 ≤ k ≤ 0.0090). That is very close to the Ridge *k* Coefficient obtained from the numerical analysis of the Boston data.

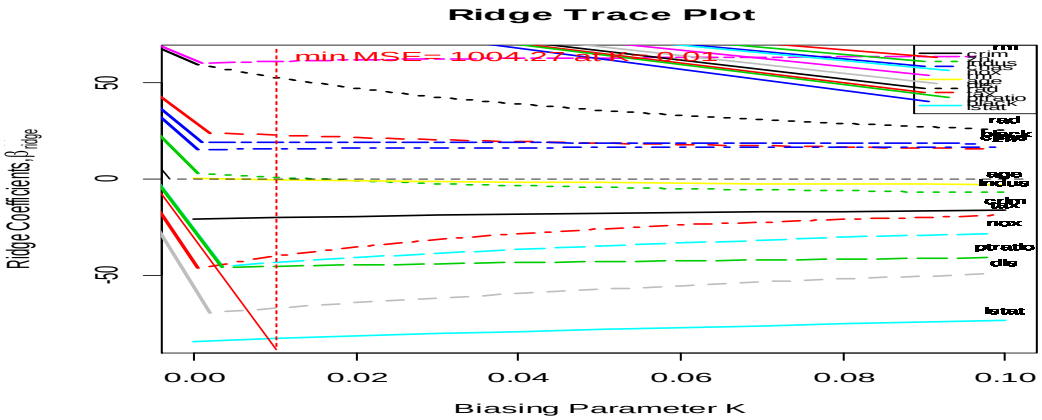

Figure 5: The Trace plot of the optimal Ridge coefficient k using Boston data.

5. The Discussion and Conclusion

The crucial task in Ridge Regression is the choice of the Penalty parameter (k) as it is usually a judgmental one. As for the simulation study, the main goal was to determine the range in which the value of k ranges. Our strategy was to rely on the value of correlation as well as the number of explanatory variables used to determine the interval of the optimal value of k. As the number of the explanatory variables (P) increases and the

correlation among them gets bigger, the value of the Penalty parameter (k) begins to appear (i.e., the number of the explanatory variables and the correlation coefficient have a fundamental effect on the value of the Penalty parameter (k)). The optimal interval of the Penalty parameter (k) has paved the way for research avenues in determining the optimal values of the Penalty parameter (k). The sample size has no significant effect in determining the Penalty parameter (k).

Moreover, we have considered three different datasets, namely the Hald Data, the Consumption Car (CC)Data, the Housing Values in Suburbs of Boston Data. The numerical results have indicated unequivocally that the Ridge estimator has the advantage over the OLS estimator. The AIC and BIC criteria have added more support to the MSE values in making the Ridge estimator more preferred. From this discussion, we can see that the use of Ridge Regression is perfectly practical in circumstances in which it is believed that large beta-values are unlikely from a practical point of view. However, it must be understood that the choice of Penalty parameter (k) is essentially equivalent to an expression of how big one believes those betas to be.

References

[1] Bates, D.M. and Watts, D.G (1988). Nonlinear Regression Analysis and Its Applications. John Wiley and Sons.

[2] Development, R. Core Team, R: A Language and environment for Statistical computing, R Foundation for statistical computing Vienna, ISBN 3-900051-07-0, 2015.

[3] Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. Robust Statistics, The Approach Based on Influence Functions.Wiley, New York, 1986.

[4] Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: biased rstimation for nonorthogonal problems. Technometrics, 12, pp. 55{67}.

[5] Hoerl, A. E. et al, (1975) Ridge Regression: some simulations, Comm Stat Theory Method 4:105.

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

[7] Huber P. J. Robust Statistics. Wiley, New York, 1981.

[8] Kotz, S. and Nadarajah, S. (2004). Multivariate t Distributions and Their Applications. Cambridge University Press.

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

[10] Smith, G., & Campbell, F. (1980). A critique of some ridge regression methods. Journal of the American Statistical Association, 75(369), 74-81.

[11] Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society, Series B, 58, pp. 267{288}.

[12] Weisberg, S. (2005). Applied Linear Regression. 3rd edition. Wiley and Sons, Inc.

[13] Wood, S. N. (2006). Generalized Additive Models: An Introduction with R. Chap-man and Hall.