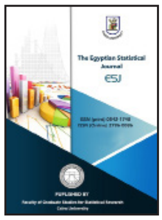
Cairo University

The Egyptian Statistical Journal

No.1 - Volume (67) , 2023

journal homepage: www.esju.journals.ekb.eg

Print ISSN 0542-1748 - Online ISSN 2786-0086



Modified Ridge Logistic Estimator Based on Singular Value Decomposition

Prof. Monira Ahmed Hussein

Prof.of Statistics at Department
of Insurance, Statistics and Mathematics
Faculty of Commerce, University of Sadat

Mostafa Kamal Abd el-Rahman

Assistant Lecturer at Department
of Insurance, Statistics and Mathematics
Faculty of Commerce, University of Sadat City

A B S T R A C T

This paper aims to introduce a modification of the ridge estimator based on the singular value decomposition (SVD) technique of the design matrix (X) to combat multicollinearity in the binary logistic model. This estimator is called a modified ridge logistic based on SVD estimator which is denoted as (MRLSVDE). We study the statistical properties of the proposed estimator in the context of the bias, variance-covariance matrix, and mean squared error. A Monte Carlo simulation study is conducted to evaluate the performance of the proposed estimator over the ridge logistic estimator (RLE) and the maximum likelihood estimator (MLE) based on the scalar mean squared error (SMSE) criterion. Moreover, an empirical application is provided to investigate the potential benefits of the proposed estimator in real-life fields. The results of the simulation study and real data application indicate that the proposed estimator outperforms the maximum likelihood and ridge logistic estimators in the scalar mean squared error sense.

Keywords:
Logistic regression model, Multicollinearity, Ridge logistic estimator, Singular value decomposition, Numerical rank.

1. **Introduction**

A binary logistic regression model is commonly used to explain the relationship between a binary response variable (dichotomous outcome variable) and one or more independent variables which are either continuous or categorical. This model is often used in applied sciences such as business, finance, biostatistics, machine learning, biology, and medical research. One of the most common and frequent methods to estimate the parameters in a logistic regression model is the maximum likelihood estimation (MLE) method. Since the likelihood equation is nonlinear in β, there is no closed-form expression for estimate β. Therefore, the ML estimator $\hat{\beta}_{ML}$ can be obtained by using numerical iterative methods such as iteratively re-weighted least squares (IRLS) by Newton–Raphson algorithm, which is an asymptotically unbiased estimate of β.

However, in the existence of the multicollinearity problem, the logistic regression model becomes unstable and the ML estimator of regression coefficients may be statistically insignificant with wrong signs and inflated variances.

To overcome the effect of multicollinearity in the logistic regression model, many biased estimators instead of the ML estimator have been proposed in the literature. The most popular estimator to handle multicollinearity is called the ridge logistic estimator (RLE) which was proposed by Schaefer *et al*. (1984). Later, Aguilera *et al.* (2006) proposed the principal component logistic estimator (PCLE), Kibria *et al.* (2012) evaluated some biasing ridge parameters (k), Mansson *et al.* (2012) introduced the Liu-estimator in logistic regression, and Nja *et al.* (2013) introduced the modified logistic ridge regression estimator (MLRE).

In addition, Inan and Erdogan (2013) proposed Liu-type estimator, Wu and Asar (2016) proposed the almost unbiased ridge logistic estimator (AURLE), Asar and Genc (2017) introduced the two-parameter ridge estimator in logistic regression, Lukman *et al.* (2020) suggested the modified ridge type logistic estimator, Abdel-Fattah (2022) developed a new class of binomial ridge-type (RT) regression estimators, Varathan (2022) proposed an improved ridge type estimator for logistic regression, and Abonazel *et al.* (2023) introduced the probit modified ridge type (PROMRT) and probit Dawoud −Kibria (PRODK) estimators for the probit regression model.

Recently, Roozbeh *et al.* (2016) proposed a modified biased estimator based on the *QR* decomposition technique to remedy the multicollinearity problem of the design matrix in linear regression models. This technique depends on decompose the design data matrix (X) into two matrices; the isometric matrix (Q) and the upper triangular matrix (R), such that $X = QR$. They suggested positive scalar values (μ_i) added to small elements of diagonal matrix R.

The main objective of this paper is to propose a modified ridge estimator based on the singular value decomposition (SVD) technique of the design matrix (X) to combat multicollinearity in the binary logistic model. In addition, the statistical properties of this estimator such as bias, variance-covariance, and matrix mean squared error (MSE) were derived. Finally, a simulation study and an empirical application were conducted to evaluate the performance of this estimator and to illustrate its potential benefits in a real-life data application.

The rest of this paper is organized as follows; In Section 2, we review the logistic regression model and maximum likelihood estimator (MLE), and some biased estimators with their asymptotic mean squared error (MSE) are presented. Section 3 introduces the proposed modified ridge logistic estimator based on the singular value decomposition technique. Then, we drive the asymptotic statistical properties of the proposed estimators in Section 4. The choice of scalar parameter for this estimator is discussed in Section 5. In Section 6, a Monte Carlo simulation study is done. In addition, an empirical application is conducted in Section 7. Finally, in Section 8, we present a summary and conclusions.

2. Logistic Regression Model

The general form of the logistic regression model can be written as follows

$$y_i = \pi_i + \varepsilon_i , \qquad i = 1, 2, ..., n \tag{1}$$

where ε_i are independent random errors such that $E(\varepsilon_i) = 0$ and $Var(\varepsilon_i) = \pi_i(1-\pi_i)$, y_i is a dichotomous random variable follows a Bernoulli distribution with parameter π_i which is the expected value of the response variable y_i and defined as

$$\pi_i = \frac{\exp(x_i'\beta)}{1+\exp(x_i'\beta)} = \frac{1}{1+\exp(-x_i'\beta)} \tag{2}$$

where x_i is the *i*th row of the design matrix X has an $n(p+1)$ dimension with p explanatory variables, and β is a $(p+1)\times 1$ vector of regression parameters.

The maximum likelihood estimation method is the most common technique applied to estimate the logistic regression parameter vector (β). Therefore, the maximum likelihood estimator (MLE) of β is given as follows

$$\hat{\beta}_{ML} = S^{-1} X'\hat{W}\hat{z}, \tag{3}$$

where $S = X'\hat{W}X$, $\hat{W} = \text{diag}[\hat{\pi}_i(1-\hat{\pi}_i)]$ and $\hat{z}_i = \log(\hat{\pi}_i) + \frac{y_i - \hat{\pi}_i}{\hat{\pi}_i(1-\hat{\pi}_i)}$. The asymptotic variance-covariance matrix of the ML estimator $(\hat{\beta}_{ML})$ is

$$Cov\left(\hat{\beta}_{ML}\right) = S^{-1}. \tag{4}$$

Since ML estimator is an asymptotically unbiased estimate of β, the asymptotic matrix mean squared error (MMSE) of $\hat{\beta}_{ML}$ is

$$MMSE\left(\hat{\beta}_{ML}\right) = S^{-1}. \tag{5}$$

Consequently, the scalar mean squared error (SMSE) of $\hat{\beta}_{ML}$ is

$$SMSE\left(\hat{\beta}_{ML}\right) = trace\left(S^{-1}\right) = \sum_{j=1}^{p+1} \frac{1}{\lambda_j}, \tag{6}$$

where λ_j is the *j*th eigenvalues of the information matrix (S). In the presence of a multicollinearity problem, the logistic regression model becomes unstable and the estimated parameters become inaccurate. Furthermore, the mean squared error value (MSE) of the regression estimate produced by the ML estimator is inflated and leads to inefficient estimates.

As a result, there are many alternative biased estimators are introduced in the literature instead of the ML estimator to overcome the effect of multicollinearity in the logistic regression model. Schaefer *et al.* (1984) extended the ridge regression to the logistic model by adding a small positive value to the main diagonal of the information matrix S. The ridge logistic estimator (RLE) is defined as

$$\hat{\beta}_{RLE} = \left(S + kI\right)^{-1} S\,\hat{\beta}_{ML}, \qquad k > 0 \tag{7}$$

where k is the ridge parameter. A lot of literature mainly focused on different ways of estimating the ridge parameter k. The common ways to estimate the ridge parameters used for the logistic regression model are listed as follows [see, Schaefer *et al.* (1984) & Smith *et al.* (1991) & Kibria *et al.* (2012)]

$$k_1 = \frac{1}{\hat{\beta}'_{ML}\hat{\beta}_{ML}},\ k_2 = \frac{p}{\hat{\beta}'_{ML}\hat{\beta}_{ML}},\ k_3 = \frac{p+1}{\hat{\beta}'_{ML}\hat{\beta}_{ML}}. \tag{8}$$

The asymptotic variance-covariance matrix of $\hat{\beta}_{RLE}$ is defined as follows

$$Cov\left(\hat{\beta}_{RLE}\right) = \left(S + kI\right)^{-1} S \left(S + kI\right)^{-1}. \tag{9}$$

Also, the asymptotic matrix and scalar mean squared error (MSE) $\hat{\beta}_{RLE}$ are defined as

$$MMSE\left(\hat{\beta}_{RLE}\right) = \left(S + kI\right)^{-1} S \left(S + kI\right)^{-1} + \left(\left(S + kI\right)^{-1} S - I\right)\beta\beta'\left(S\left(S + kI\right)^{-1} - I\right), \tag{10}$$

and

$$SMSE\left(\hat{\beta}_{RLE}\right) = trace\left(MMSE\left(\hat{\beta}_{RLE}\right)\right) = \sum_{j=1}^{p+1} \frac{\lambda_j}{(\lambda_j + k)^2} + \sum_{j=1}^{p+1} \frac{k^2 \hat{\alpha}_j^2}{(\lambda_j + k)^2}, \tag{11}$$

where the first term in (11) is the asymptotic variance of $\hat{\beta}_{RLE}$ and the second term is its squared bias, $\hat{\alpha}^2 = \gamma\hat{\beta}_{ML}$ and γ is an orthogonal matrix whose columns are the eigenvectors corresponding to the ordered eigenvalues of the S matrix.

Mansson *et al.* (2012) generalized a Liu estimator in the linear regression for the logistic regression model. This estimator was called the logistic Liu estimator (LLE) and was defined as follows

$$\hat{\beta}_{LLE} = \left(S + I\right)^{-1} \left(S + d\,I\right)\hat{\beta}_{ML}, \tag{12}$$

where d is the shrinkage parameter, $0 < d < 1$.

Inan and Erdogan (2013) introduced the Liu-type logistic regression estimator, which is defined as

$$\hat{\beta}_{LLTE} = \left(S + kI\right)^{-1} \left(S - d\,I\right)\hat{\beta}_{ML}. \tag{13}$$

Lukman *et al.* (2020) developed the logistic version of the modified ridge-type estimator in the linear regression model which is proposed by Lukman *et al.* (2019). The logistic modified ridge-type estimator is given as

$$\hat{\beta}_{LMRT} = \left(S + k(1+d)I\right)^{-1} S\hat{\beta}_{ML}, \tag{14}$$

where $k > 0$ and $0 < d < 1$.

Recently, Roozbeh *et al.* (2016) proposed a new biased estimator based on the QR decomposition to overcome the multicollinearity in linear regression models. They used the QR decomposition technique to factorize the ill-conditional design matrix (X) into the isometric matrix Q with orthonormal columns and the upper triangular matrix R. They mentioned that, when multicollinearity occurs for the matrix (X), some diagonal entries of the matrix R become too small, and more closeness of the small entries values of the R matrix leads to more strength of the multicollinearity.

To overcome the multicollinearity problem, they added a positive scalars (μ) to the small diagonal entries of the upper triangular matrix R, and a modified version of the R matrix becomes $R_{(\mu)} = R + \text{diag}(0, ..., 0, \mu_{r+1}, ..., \mu_p)$. Consequently, the new biased estimator based on QR decomposition, which is called the QR-based least-squares estimator $(QRLSE)$ for linear regression model, is defined as

$$\hat{\beta}_{(\mu)} = \left(X'_\mu X_\mu\right)^{-1} X'Y, \tag{15}$$

where $X_\mu = QR_{(\mu)}$ is a modified design matrix obtained from $R_{(\mu)}$.

In the following section, we introduce a modified ridge estimator to overcome the multicollinearity problem in the binary logistic regression model, which is an extension to RLE proposed by Schaefer *et al.* (1984) and (QRLSE) for the linear regression model introduced by Roozbeh *et al.* (2016). While our proposed estimator is based on the singular value decomposition technique (SVD) which is applied on the design matrix (X) to overcome the multicollinearity problem for the binary logistic regression model.

3. Construction of the Proposed Estimator

In this section, we introduce the proposed estimator, and some helpful definitions and theories will be briefly presented. The proposed estimator is constructed by considering the ridge logistic estimator (RLE) introduced by Schaefer *et al.* (1984) and based on the work of Roozbeh *et al.* (2016), by using singular value decomposition (SVD) technique applied for the design matrix (X).

Theorem 1. (Watkins, 2002): *Let* $A \in \mathbb{R}^{n\times m}$ *be a nonzero matrix with rank* r*, it is possible to factorize* A *as a product*

$$A = U \sum V^T,$$

where $U \in \mathbb{R}^{n\times n}$ *and* $V \in \mathbb{R}^{m\times m}$ *are orthogonal matrices* such as $U'U = V'V = I$, *and* $\sum \in \mathbb{R}^{n\times m}$ *is a rectangular diagonal matrix* its elements called the singular values, which are listed in descending order such as $\sigma_1 = \sigma_{\max} \geq \sigma_2 \geq ... \geq \sigma_r = \sigma_{\min} \geq 0$.

By using **Theorem 1**, the $n\times p$ design matrix X can be factorized into the product of three matrices as follows

$$X_{n\times p} = \mathcal{U}_{n\times n} \mathcal{D}_{n\times p} \mathcal{V}'_{p\times p}, \tag{16}$$

where $\mathcal{U}$ and $\mathcal{V}$ are orthogonal matrices, and $\mathcal{D}$ is diagonal matrix with elements called the uniquely singular values (δ's), which are listed in descending order as $\delta_1 = \delta_{\max} \geq \delta_2 \geq ... \geq \delta_p = \delta_{\min} \geq 0$, p is the number of explanatory variables which refers to the exact rank of the full column rank matrix X.

In the existence of multicollinearity, the design matrix (X) becomes an ill-conditioned matrix, and the diagonal matrix $\mathcal{D}$ becomes having r large singular values, while the others are relatively small which perhaps close to zero [see, e.g. Kibria *et al.* (2012) & Roozbeh *et al.* (2016)]. To identify which singular values of $\mathcal{D}$ are small, we need to determine a cutoff value or a positive constant (ω) that separates the large and small singular values [see, e.g. Watkins (2002), pp. 269-272], such as

$$\delta_1 \geq \delta_2 \geq \ldots \geq \delta_r >> \omega \geq \delta_{r+1} \geq \ldots \geq \delta_p, \tag{17}$$

where r is the numerical rank of ill-conditioned matrix X, which is defined as the number of singular values of X that are substantially larger than ω. Cattell (1966) introduced the scree plot that draws the singular values in a coordinate system and then r is chosen as the "large gap" or "elbow" of the graph.

To reduce the effect of the multicollinearity, we can keep the large singular values $(\delta_1 \geq \delta_2 \geq ... \geq \delta_r)$ as they are because they are large enough. On the other hand, it is reasonable to increase the small singular values $(\delta_{r+1} \geq ... \geq \delta_p)$ of the diagonal matrix $\mathcal{D}$ by some positive scalar values (τ_i) as follows

$$\delta_i \leftarrow \delta_i + \tau_i, \qquad i = r+1, \ldots, p, \tag{18}$$

where τ_i are positive scalar parameters will be discussed in following section. Therefore, we get a modified version of the ill-conditioned design matrix (X) such as

$$X_\tau = \mathcal{U}\mathcal{D}_\tau \mathcal{V}', \tag{19}$$

where $\mathcal{D}_\tau = \mathcal{D} + \text{diag}(0, ..., 0, \tau_{r+1}, ..., \tau_p)$.

Consequently, the modified information matrix becomes as follows

$$S_{\tau} = X_{\tau}' \hat{W} X_{\tau}, \tag{20}$$

where $\hat{W} = \text{diag}[\hat{\pi}_i(1-\hat{\pi}_i)]$ which is a weight matrix estimated based on MLE. Hence, considering the ridge logistic estimator of Schaefer *et al.* (1984) and the modified information matrix (S_{τ}), the proposed estimator can be obtained. The proposed estimator called a modified ridge logistic estimator based on singular value decomposition and denoted (MRLSVDE) can be defined as follows

$$\hat{\beta}_{SVD}^{MRL} = \left(S_{\tau} + kI\right)^{-1} S\,\hat{\beta}_{ML}, \qquad k > 0 \tag{21}$$

where k is the ridge parameter, τ_i are positive scalar values.

Lemma 1. *The proposed estimator is considered a general case, such as*

i. $\lim\limits_{\tau_i \to 0} \hat{\beta}_{SVD}^{MRL} = \lim\limits_{\tau_i \to 0} \left((S_{\tau} + kI)^{-1} S\,\hat{\beta}_{ML}\right) = (S + kI)^{-1} S\,\hat{\beta}_{ML} = \hat{\beta}_{RLE}$

ii. $\lim\limits_{k,\tau_i \to 0} \hat{\beta}_{SVD}^{MRL} = \lim\limits_{k,\tau_i \to 0} \left((S_{\tau} + kI)^{-1} S\,\hat{\beta}_{ML}\right) = (S)^{-1} S\,\hat{\beta}_{ML} = \hat{\beta}_{ML}$

Since the proposed estimator is a modification of the ridge logistic estimator, we hope that this estimator reduces the mean squared error (MSE) of the logistic regression estimation and makes our estimator more efficient than other biased estimators.

4. Asymptotic properties of the proposed estimator

In this section, we consider some statistical properties of the proposed modified ridge logistic estimator $(\hat{\beta}_{SVD}^{MRL})$. The bias, variance-covariance matrix, and mean squares error (MSE) are derived.

The asymptotic bias of $\hat{\beta}_{SVD}^{ML}(\tau)$ can be obtained as follows

$$\begin{aligned} Bias\left(\hat{\beta}_{SVD}^{MRL}\right) &= E\left(\hat{\beta}_{SVD}^{MRL}\right) - \beta \\ &= E\left(\left(S_{\tau} + kI\right)^{-1} S\,\hat{\beta}_{ML}\right) - \beta \\ &= \left(S_{\tau} + kI\right)^{-1} S\; E\left(\hat{\beta}_{ML}\right) - \beta \end{aligned}$$

Since $\hat{\beta}_{ML}$ estimator is an asymptotically unbiased estimate of β. Therefore, the asymptotic bias of the proposed $\hat{\beta}_{SVD}^{MRL}$ estimator can be obtained as

$$Bias\left(\hat{\beta}_{SVD}^{MRL}\right) = \left(\left(S_{\tau} + kI\right)^{-1} S - I\right)\beta. \tag{22}$$

The asymptotic variance-covariance matrix of $\hat{\beta}_{SVD}^{MRL}$ can be derived as follows

$$Cov\left(\hat{\beta}_{SVD}^{MRL}\right)=Cov\left(\left(S_{\tau}+kI\right)^{-1} S\,\hat{\beta}_{ML}\right)$$

$$=\left(S_{\tau}+kI\right)^{-1} S\;Cov\left(\hat{\beta}_{ML}\right)S\left(S_{\tau}+kI\right)^{-1}$$

Since $Cov\left(\hat{\beta}_{ML}\right)=S^{-1}$, then

$$Cov\left(\hat{\beta}_{SVD}^{MRL}\right)=\left(S_{\tau}+kI\right)^{-1} S\;\left(S_{\tau}+kI\right)^{-1}.$$

Consequently, the asymptotic matrix mean squared error (MMSE) of proposed estimator ($\hat{\beta}_{SVD}^{MRL}$) can be obtained as

$$MMSE\left(\hat{\beta}_{SVD}^{MRL}\right)=Cov\left(\hat{\beta}_{SVD}^{MRL}\right)+Bias\left(\hat{\beta}_{SVD}^{MRL}\right)\left(Bias\left(\hat{\beta}_{SVD}^{MRL}\right)\right)'$$

$$=\left(S_{\tau}+kI\right)^{-1} S\;\left(S_{\tau}+kI\right)^{-1}+\eta\,\eta',\qquad (23)$$

where $\eta=\left(\left(S_{\tau}+kI\right)^{-1} S-I\right)\beta$. While, the asymptotic scalar mean squared error (SMSE) of $\hat{\beta}_{SVD}^{MRL}$ can be presented as

$$SMSE\left(\hat{\beta}_{SVD}^{MRL}\right)=trace\left(MMSE\left(\hat{\beta}_{SVD}^{MRL}\right)\right)$$

$$=trace\left(Cov\left(\hat{\beta}_{SVD}^{MRL}\right)\right)+\left(Bias\left(\hat{\beta}_{SVD}^{MRL}\right)\right)'\;Bias\left(\hat{\beta}_{SVD}^{MRL}\right),$$

where $\lambda_j(\tau)$ is the eigenvalues of S_{τ} matrix.

5. The Choice of the Scalar Parameter (τ)

It is reasonable to conclude that the performance of the proposed estimator is affected by the values of scalar parameters as well as the ridge parameter (k). Whereas, Roozbeh *et al.* (2016) pointed out that there is no closed-form expression for the scalar parameter in their QRLSE for the linear regression model, and they conducted some simulation and graphical results to find the best scalar parameter. So, we believe that the construction of the best formula for the scalar parameters perhaps depends on the following points

1. The range between any small singular value and the next one of matrix $\mathcal{D}$.
2. The ridge parameter (k), which is somewhat similar to the scalar parameters (τ_i). Since ridge parameter (k) is added to all diagonal elements of the matrix S while the scalar parameters (τ_i) are added to the last small singular values only of the matrix $\mathcal{D}$.
3. The number of explanatory variables (*p*) in the logistic model.
4. The eigenvalues of S matrix.

In this context, by means of Monte Carlo simulations based on the mean squared error (MSE), we suggest that the scalar parameter may be defined as follows

$$\tau_i = \left(\frac{2(p+1)}{\lambda_{\min}} \right) \times \max\left(k, \frac{\delta_{r+i-1,r+i-1} - \delta_{r+i,r+i}}{\delta_{r+i,r+i}} \right), \quad i = 1, \ldots, (p-r) \qquad (24)$$

where $\lambda_{\min}$ is the minimum eigenvalue of S matrix, p is the number of explanatory variables, k is ridge parameter and δ's are singular values of the matrix $\mathcal{D}$.

The proposed estimator ($\hat{\beta}_{SVD}^{MRL}$) with the suggested scalar parameter formula is evaluated by means of Monte Carlo simulations in the scalar mean squared error (SMSE) criterion.

In literature, multicollinearity is a concern when the condition index or condition number is greater than 15. Since the condition number is the square of the largest condition index, we suggest using the square of the condition index of the S matrix to determine the small singular values (δ's) in the $\mathcal{D}$ matrix which are increasing by scalar parameters (τ_i 's). Such that the small singular values for which square of whose condition index is greater than 15.

6. Simulation Study

In this section, we conduct a Monte-Carlo simulation to compare the performance of the proposed MRLSVDE estimator with the well-known MLE and RLE estimators based on the mean squared error (MSE) criteria. The explanatory variables are generated at different strengths of the correlation using the simulation equations presented in Kibria (2003) and Lukman *et al.* (2019) as follows

$$x_{ij} = \left(1-\gamma^2\right)^{1/2} w_{ij} + \gamma\, w_{ip}, \quad i = 1, 2, \ldots, n; \; j = 1, 2, \ldots, p, \qquad (25)$$

where w_{ij} are pseudo-random numbers from the standard normal distribution and γ is the degree of correlation between the explanatory variables. We consider five different levels of correlations such as γ = 0.85, 0.90, 0.95, 0.99, and 0.999. Then, the observations of the response variable are generated from the *Be*(π_i) distribution where

$$\pi_i = \frac{\exp\left(x_i'\beta\right)}{1+\exp\left(x_i'\beta\right)} = \frac{1}{1+\exp\left(-\left(\beta_1 x_{i1} + \ldots + \beta_p x_{ip}\right)\right)}, \qquad (26)$$

where β is the true parameter vector which is restricted in many simulation studies to be chosen as the normalized eigenvector corresponding to the largest eigenvalue of $X'X$ so that $\sum_{j=1}^{p} \beta_j^2 = 1$. To illustrate the effect of the sample size n, four various samples 75, 100, 150, and 200 are taken. Further, we consider different numbers of explanatory variables (p)

corresponding to 4, 6, and 8. In this study, we choose different values of ridge parameter ($k1$, $k2$, and $k3$) defined in Eq. [8], and its corresponding scalar parameters ($\tau1$, $\tau2$, and $\tau3$) respectively, defined in Eq. [24] for the proposed estimator (MRLSVDE).

All computations were conducted using the R packages version 4.3.0, and the used packages and required libraries in the source file were “dplyr”, “blorr”, “MASS”, “Metrix”, “glmnet”, “pracma” and “car”. Then the simulation experiment is replicated 1000 times, and the estimated MSE values of the estimators are obtained according to the following formula

$$MSE\left(\hat{\beta}\right)=\frac{1}{1000}\sum_{j=r}^{1000}(\hat{\beta}_r-\beta)'(\hat{\beta}_r-\beta), \qquad (27)$$

where $\hat{\beta}_r$ are the estimated parameters in the rth replication. The estimated mean squared errors (MSEs) of the MLE, RLE with $k1$, $k2$, & $k3$, and its corresponding proposed MRLSVDE estimator with $\tau1$, $\tau2$, & $\tau3$, are summarized for different values of ρ, n and p in Tables 6.1-6.3.

Table 6.1. The estimated MSE values of the estimators for different ρ when $p = 4$

	ρ	MLE	RLE			MRLSVDE		
			with $k1$	with $k2$	with $k3$	with $\tau1$	with $\tau2$	with $\tau3$
$n=75$	0.85	2.1118	1.5025	0.7760	0.6650	0.7763	0.4713	0.4194
	0.90	4.0616	2.7703	1.3571	1.1463	0.7508	0.4820	0.4355
	0.95	6.1141	3.9052	1.7583	1.4522	0.9387	0.5329	0.4655
	0.99	35.7219	20.941	8.8256	7.2124	0.1279	0.1142	0.1107
	0.999	365.107	212.748	90.1339	73.8182	0.1144	0.1139	0.1137
$n=100$	0.85	1.5715	1.1460	0.6030	0.5162	0.6765	0.3812	0.3320
	0.90	2.7489	2.0194	1.1139	0.9666	0.5823	0.4987	0.4535
	0.95	4.3455	2.7889	1.2631	1.0444	0.4389	0.2336	0.2028
	0.99	26.1301	15.5392	6.7102	5.5122	0.0684	0.0637	0.0624
	0.999	269.2637	153.954	65.3283	53.6403	0.0555	0.0550	0.0549
$n=150$	0.85	0.9394	0.7079	0.3878	0.3337	0.4883	0.2655	0.2306
	0.90	1.1707	0.8584	0.4499	0.3828	0.5534	0.2961	0.2550
	0.95	2.4960	1.7324	0.8442	0.7060	0.4021	0.1784	0.1493
	0.99	12.8202	7.9133	3.5472	2.9354	0.0752	0.0526	0.0482
	0.999	128.9405	76.673	32.8409	26.8846	0.0410	0.0405	0.0404
$n=200$	0.85	0.7196	0.5879	0.3844	0.3476	0.4325	0.2512	0.2290
	0.90	0.9191	0.7001	0.3891	0.3346	0.4781	0.2682	0.2323
	0.95	1.9356	1.3566	0.6737	0.5661	0.2636	0.1107	0.092
	0.99	8.5420	5.3628	2.4377	2.0215	0.0971	0.0414	0.0365
	0.999	100.8600	58.9659	25.7827	21.3102	0.0243	0.0240	0.0239

Table 6.2. The estimated MSE values of the estimators for different ρ when $p = 6$

	ρ	MLE	RLE			MRLSVDE		
			with $k1$	with $k2$	with $k3$	with $\tau 1$	with $\tau 2$	with $\tau 3$
$n = 75$	0.85	4.9615	3.6534	1.3387	1.1589	2.4379	0.9837	0.8628
	0.90	8.6089	5.8759	1.9445	1.6696	4.5585	1.9558	1.7210
	0.95	12.0540	8.2471	2.8217	2.4340	1.5918	0.7133	0.6350
	0.99	37.0153	24.0343	7.6824	6.5945	0.1566	0.1319	0.1293
	0.999	943.0959	615.829	203.5419	175.7291	0.2571	0.2552	0.2548
$n = 100$	0.85	2.3096	1.7251	0.7025	0.6192	1.1090	0.5475	0.4909
	0.90	5.3987	3.9050	1.4324	1.2410	1.3670	0.5074	0.4470
	0.95	10.8337	7.4726	2.7037	2.3507	0.7678	0.3393	0.3071
	0.99	67.1738	43.0526	14.7624	12.8511	0.1319	0.1234	0.1219
	0.999	512.5147	334.493	115.0579	99.8651	0.0992	0.0985	0.0983
$n = 150$	0.85	1.5525	1.2007	0.5199	0.4611	0.7713	0.3848	0.3466
	0.90	2.3951	1.7907	0.7261	0.6386	0.9336	0.4367	0.3912
	0.95	4.5332	3.2746	1.2669	1.1086	0.6497	0.2133	0.1897
	0.99	28.2067	18.7235	6.7686	5.9115	0.1442	0.059	0.0546
	0.999	316.2853	208.066	73.7943	64.2931	0.0396	0.0388	0.0386
$n = 200$	0.85	1.1460	0.9190	0.4305	0.3844	0.6283	0.3247	0.2932
	0.90	1.7732	1.3679	0.5954	0.5276	0.8064	0.3954	0.3552
	0.95	3.7565	2.7232	1.0874	0.9566	0.4425	0.1375	0.1228
	0.99	23.2268	15.4990	5.6722	4.9613	0.3081	0.0539	0.0479
	0.999	219.9348	142.716	51.2684	44.872	0.0298	0.0296	0.0295

Table 6.3. The estimated MSE values of the estimators for different ρ when $p = 8$

	ρ	MLE	RLE			MRLSVDE		
			with $k1$	with $k2$	with $k3$	with $\tau 1$	with $\tau 2$	with $\tau 3$
$n = 75$	0.85	7.9062	5.5000	1.5429	1.3742	3.0143	1.2532	1.1427
	0.90	16.5145	11.5143	3.0522	2.6993	5.1181	1.8840	1.7050
	0.95	34.0302	21.3293	5.2128	4.6121	14.833	4.6984	4.1819
	0.99	115.5238	78.0677	21.9662	19.6202	0.6249	0.4404	0.4264
	0.999	1155.39	813.130	232.096	206.7431	0.2832	0.2817	0.2815
$n = 100$	0.85	4.6352	3.5247	1.1732	1.0563	2.0966	0.9254	0.8495
	0.90	8.9893	6.4909	1.9451	1.7368	2.6833	1.1481	1.0526
	0.95	25.4968	17.3581	4.8819	4.3585	2.4150	1.2443	1.1610
	0.99	94.7457	65.2526	19.2361	17.2334	0.4094	0.3273	0.3228
	0.999	1161.465	768.186	225.2209	202.4131	0.2305	0.2292	0.2290
$n = 150$	0.85	2.9898	2.3397	0.8410	0.7600	1.3419	0.5909	0.5417
	0.90	3.8156	2.9500	1.0378	0.9356	1.6025	0.6622	0.6048
	0.95	10.7659	7.7908	2.5542	2.3001	0.8507	0.2939	0.2713
	0.99	45.4548	31.3182	9.8785	8.9038	0.1393	0.0580	0.0560
	0.999	435.3315	303.825	95.9183	86.3507	0.0779	0.0764	0.0762
$n = 200$	0.85	1.6816	1.3751	0.5766	0.5299	0.8992	0.4401	0.4096
	0.90	2.9559	2.2929	0.8470	0.7677	1.2463	0.5536	0.5086
	0.95	6.6218	4.9375	1.6881	1.5230	0.6921	0.1115	0.1016
	0.99	28.1343	20.2552	6.6225	5.9681	0.1943	0.0430	0.0407
	0.999	288.9609	200.662	63.5103	57.2252	0.0434	0.0382	0.0378

Source: by the researcher from R outputs.

Tables 6.1-6.3. show that the proposed MRLSVDE estimator outperforms the classical MLE and RLE in the sense of minimum mean squared error criterion for all scenarios. The superiority of the MRLSVDE estimator appears clearly in the presence of high correlation degrees among the explanatory variables.

Although an increase in the degree of correlation among explanatory variables has a negative effect on the MLE and RLE estimators, the MRLSVDE estimator works very well with high correlation degrees. Also, with more explanatory variables included in the model, the values of MSEs for all estimators increase, while the performance of the MRLSVDE estimator remains the best estimator. On the other hand, increasing the sample size has a positive effect on all estimators and the values of MSEs decrease for any correlation degree or each number of explanatory variables.

Finally, it is important to note that with each improvement in the performance of the ridge estimator, the proposed MRLSVDE estimator also improves. Where the RLE estimator performs better with ridge parameter k_3 than k_1 and k_2 for all scenarios, therefore MRLSVDE estimator performs better with the corresponding scalar parameter τ_3 than τ_1 and τ_2.

7. Empirical Application

In this section, we conduct an empirical application in order to illustrate the performance of the proposed MRLSVD estimator with the existing MLE and RLE estimators. In addition, the potential benefits of this estimator in real-life fields are shown. Also, the results and conclusions are discussed.

The real data used in this paper is secondary data which was taken from the http://paulblanche.com/files/DataFramingham.html website. This data was prepared by the famous "Framingham Heart Study”, which initially planned as a 20 years team study of residents aged 30-59 in Framingham town, Massachusetts, in 1948. The used data in our paper is a sample consists of 150 persons only. The data set contains variables that are indicated to be associated with the heart disease. We fit a logistics regression model where the heart disease is a response variable y and defined as:

$$y_i = \begin{cases} 1 \text{ if heart disease occured,} \\ 0 \quad \text{otherwise.} \end{cases}$$

We fit a logistic regression model where the response variable is explained by some explanatory variables defined as

X_1: (AGE) Age of the person in years,

X_2: (FRW) Framingham relative weight,

X_3: (SBP) Systolic blood pressure at baseline mmHg,

X_4: (DBP) Diastolic blood pressure at baseline mmHg,

X_5: (CHOL) Cholesterol at baseline mg/100ml.

The condition number can be used to detect the multicollinearity among explanatory variables. According to the literature, if the condition index or condition number is 15, multicollinearity is a concern; if it is greater than 30, multicollinearity is a serious concern.

Table 7.1 gives the eigenvalues, condition indices (CI), condition number (κ) of the S matrix, eigenvalues and singular values of $X'X$ matrix, where

$$CI = \sqrt{\frac{\lambda_{\max}}{\lambda_j}} \text{ and } \kappa = \frac{\lambda_{\max}}{\lambda_{\min}}, \qquad j = 1,\ 2, ..., p$$

where $\lambda_{\max}$ and $\lambda_{\min}$ are the maximum and minimum eigenvalues of the information matrix (S) respectively [see, e.g. Weissfeld and Sereika (1991), Lukman *et al.* (2020) & Awwad (2022)].

Table 7.1. Condition indices & number of $X'\widehat{W}X$, and singular values of $X'X$ matrix

	$X'\widehat{W}X$			$X'X$	
λ	**Eigenvalue**	**Condition index**	**Condition index square**	**Eigenvalue**	**Singular values of $\mathcal{D}$**
1	1724915.213	1	1	13839189.89	3720.106
2	29236.97	7.680999	58.99775	240769.109	490.6823
3	3411.1505	22.4871	505.6697	28531.224	168.9119
4	605.2539	53.38449	2849.904	4852.193	69.65768
5	547.3829	56.13559	3151.204	4323.587	65.75399
Condition Number	**3151.204**				

Source: by the researcher through R outputs.

Table 7.1. Shows a high value of the condition number $(\kappa) > 30$. Hence, this indicates that a serious multicollinearity problem exists in this data set.

We determine the small singular values in the $\mathcal{D}$ matrix which are increasing by scalar parameters $(\tau_i\text{'s})$ using the square of condition index of the S matrix. Table 7.1. indicates that, there is three of squared condition indices are greater than 15. Consequently, we can conclude that there are three small singular values in the $\mathcal{D}$ matrix which can be increased by the positive scalars (τ_i) values to obtain the adjusted $\mathcal{D}_\tau$ and then a modified X_τ matrix.

In the literature, there are many rules on how to choose the ridge parameter (k) for the ridge logistic estimator. In this paper, we take different values of k, such as k_1, k_2, and k_3, which are defined in Eq. [8]. In addition, we consider the best value of k (k_{opt}) that produces the minimum SMSE value for the ridge logistic estimator. By plotting the SMSE of the ridge logistic estimator with lots of ridge parameters k. With every value of k, we consider a corresponding scalar parameter (τ_i) for the proposed MRLSVDE estimator, which is defined in Eq. [24]. In the same way, the optimal scalar parameter (τ_{opt}) is considered.

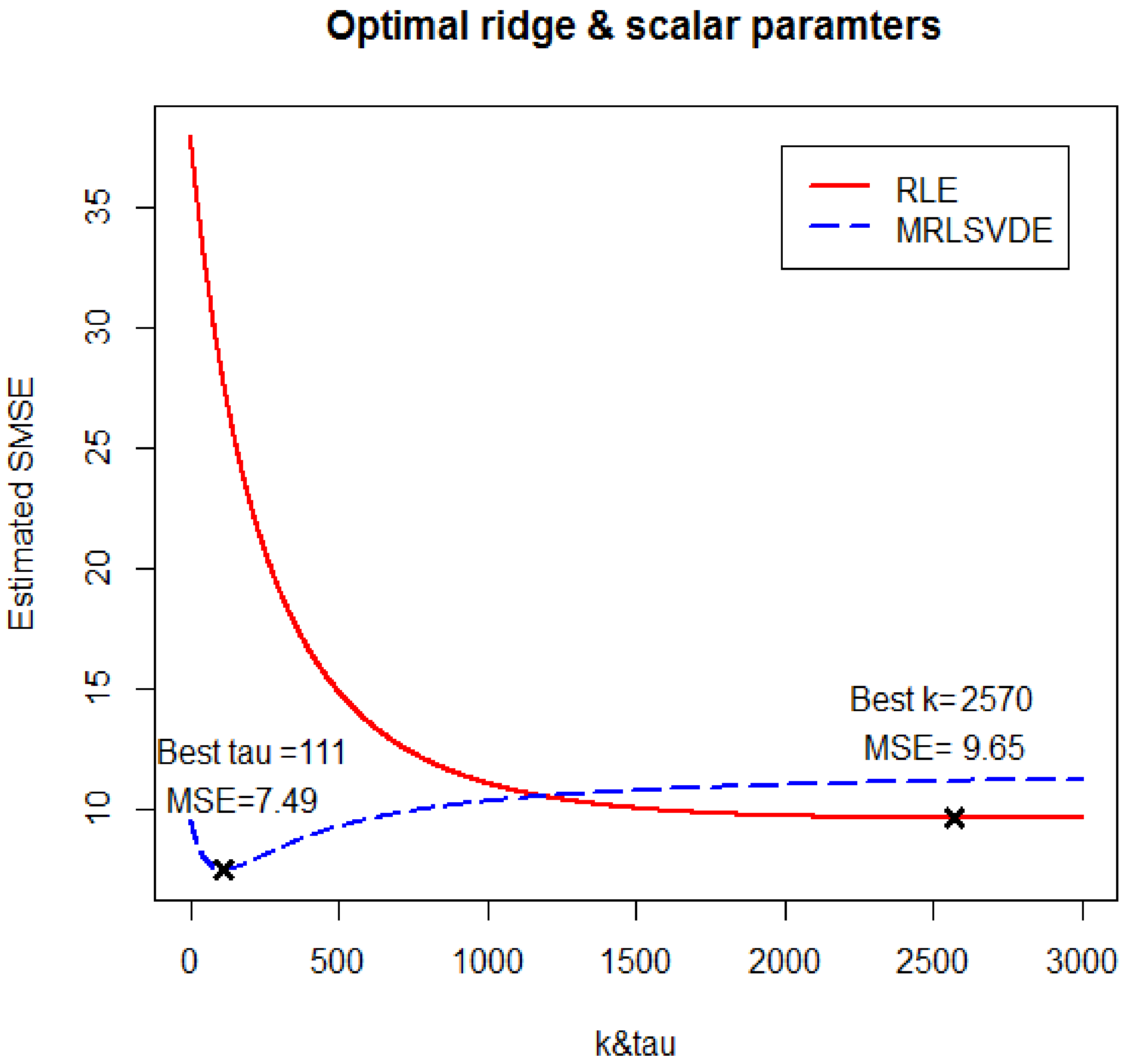


Fig. 7.1. The best value of k (k_{opt}) for the RLE and best τ (τ_{opt}) for the MRLSVDE.

In Figure 1, we found the best values of k $(k_{opt} = 2570)$ and τ $(\tau_{opt} = 111)$ by plotting the scalar mean squared errors (SMSEs) for the RLE and MRLSVDE estimators with different values of k and τ respectively, from zero to 3000.

Table 7.2 gives the regression coefficients, standard errors, and the MSE values of MLE, RLE, and the proposed MRLSVDE for the considered values of ridge parameter (k) and corresponding scalar parameters (τ_i) for the MRLSVDE estimator.

Table 7.2. The coefficients, standard errors and SMSE values (in 10^{-4}) of estimators.

		$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	SMSE
with $\boldsymbol{k_1}$**= 825.9584 and corresponding** $\boldsymbol{\tau_i}$**= 18.10707**							
MLE	*Estimate*	-225.651	-119.655	-58.5847	227.8569	-22.0159	38.0701
	Std. Error	405.8625	177.8478	208.1953	373.0396	42.9335	
RLE	*Estimate*	-102.41	-104.113	-17.3716	84.9596	-27.2454	11.8439
	Std. Error	166.8251	123.7467	111.3083	154.4226	38.5235	
MRLSVDE	*Estimate*	-87.0192	-102.144	-27.8154	96.0183	-29.6284	9.1945
	Std. Error	133.8219	104.6177	92.3867	123.9459	35.8049	
with $\boldsymbol{k_2}$**= 4129.792 and corresponding** $\boldsymbol{\tau_i}$**= 90.5353**							
MLE	*Estimate*	-225.651	-119.655	-58.5847	227.8569	-22.0159	38.0701
	Std. Error	405.8625	177.8478	208.1953	373.0396	42.9335	
RLE	*Estimate*	-41.276	-69.5567	-14.2775	13.7718	-31.0967	9.8266
	Std. Error	50.6966	67.9458	55.56	49.4351	34.1553	
MRLSVDE	*Estimate*	-32.7652	-67.7184	-27.1831	41.8286	-37.0564	8.1724
	Std. Error	32.9381	43.4181	36.324	32.4686	25.6649	
with $\boldsymbol{k_3}$**= 4955.75 and corresponding** $\boldsymbol{\tau_i}$**= 108.6424**							
MLE	*Estimate*	-225.651	-119.655	-58.5847	227.8569	-22.0159	38.0701
	Std. Error	405.8625	177.8478	208.1953	373.0396	42.9335	
RLE	*Estimate*	-37.0992	-65.0015	-15.4367	9.0482	-31.5383	9.9564
	Std. Error	43.3376	61.5476	50.7384	42.9432	33.3315	
MRLSVDE	*Estimate*	-29.3434	-63.625	-27.3463	37.8697	-38.0518	8.3501
	Std. Error	27.1862	37.7272	31.8787	27.269	23.973	
with $\boldsymbol{k_{opt}}$**= 2570 and corresponding** $\boldsymbol{\tau_i}$**= 56.3408**							
MLE	*Estimate*	-225.651	-119.655	-58.5847	227.8569	-22.0159	38.0701
	Std. Error	405.8625	177.8478	208.1953	373.0396	42.9335	
RLE	*Estimate*	-54.6805	-81.4251	-12.1947	29.183	-29.9665	9.6498
	Std. Error	75.1159	85.2459	69.6362	71.2248	35.8794	
MRLSVDE	*Estimate*	-44.035	-78.6605	-26.5002	53.7733	-34.575	7.8432
	Std. Error	52.6652	60.1795	49.7803	50.2707	29.5739	
with $\boldsymbol{k_{opt}}$**= 2570 and corresponding** $\boldsymbol{\tau_{opt}}$**= 111**							
MLE	*Estimate*	-225.651	-119.655	-58.5847	227.8569	-22.0159	38.0701
	Std. Error	405.8625	177.8478	208.1953	373.0396	42.9335	
RLE	*Estimate*	-54.6805	-81.4251	-12.1947	29.183	-29.9665	9.6498
	Std. Error	75.1159	85.2459	69.6362	71.2248	35.8794	
MRLSVDE	*Estimate*	-35.3947	-73.2228	-31.1846	57.4988	-37.7964	7.4898
	Std. Error	36.2705	44.3761	36.9389	35.3311	24.9955	

Source: by the researcher through R outputs.

According to Table 7.2, it is observed that the MRLSVD estimator has less SMSE than the MLE and RLE estimators. Therefore, the results reveal that the proposed estimator works well and outperforms the MLE and LRE in the SMSE sense. In addition, the parameters of the proposed estimator have fewer standard errors than all parameters of MLE and some parameters of RLE. One can note that, if the positive scalar (τ_i) values equal zero, we obtain the ML estimator.

We can note that, with an improvement in the ridge estimator performance, the proposed MRLSVDE estimator also improves. For instance, at the optimal ridge parameter (k_{opt}), the RLE estimator gives its minimum SMSE over the other ridge parameters ($k1$, $k2$, & $k3$), and the MRLSVDE estimator performs better than the corresponding $\tau 1$, $\tau 2$, and $\tau 3$. Also, with the optimal ridge parameter (k_{opt}) and optimal scalar parameter (τ_{opt}), the RLE and MRLSVDE estimators have their minimum SMSE, while the performance of the MRLSVDE estimator remains the best estimator.

Finally, the results of the empirical study are similar to the simulation results and they verify the theoretical findings.

8. Conclusion

This paper proposed a new estimator to combat the multicollinearity in the binary logistic model which is called a modified ridge logistic based on SVD estimator denoted as (MRLSVDE). This proposed estimator based on the singular value decomposition (SVD) technique for the design matrix (*X*) in logistic regression. It is summarized by adding the positive scalars (τ_i) to the last $(p-r)$ singular values only, which are too small of the diagonal matrix $\mathcal{D}$. Also, we derived some statistical properties of this estimator such as bias, variance-covariance matrix, and scalar mean squared error (SMSE).

The results of the simulation study and the real data application reveal that the proposed estimator outperforms the MLE and RLE estimators in the SMSE criterion. Furthermore, choosing the ridge parameter (k) and then the corresponding scalar parameters (τ_i) affects the performance of the MRLSVDE estimator.

Moreover, the benefits of using the MRLSVDE estimator increase in the presence of high multicollinearity among the explanatory variables. In contrast, the results show that the ML estimator provides the least performance as expected when multicollinearity exists. So, the ML estimator should not be used in the presence of a multicollinearity problem since the parameter becomes unstable and it has a large SMSE. This problem is especially severe when the correlation between explanatory variables is high and the data size is small.

There may be a potential limitation in this paper. There is no closed-form expression for the scalar parameter (τ_i), such as Roozbeh *et al.* (2016) pointed out in their study about QRLSE estimator for the linear regression model. This is due to the scalar parameter (τ_i) being built in the $\mathcal{D}$ matrix which is also built into the X matrix. Therefore, it is difficult to differentiate SMSE with respect to τ_i in order to obtain the optimal one.

So, following Roozbeh *et al.* (2016) we carried out some numerical comparisons and graphical results by estimating the proposed MRLSVD estimator with many τ_i (say from 1 to 10000) and finding the optimal scalar parameter (τ_{opt}) which is corresponding to the minimum SMSE value. Hence, we found that the suggested scalar parameter formula in Eq. [24] for our proposed estimator and the optimal scalar parameter (τ_{opt}) which is considered in the empirical application, yield the minimum MSE values and gain some benefits.

REFERENCES

1) Abonazel, M., Dawoud, I., Awwad, F. and Tag-Eldin, E. (2023). New estimators for the probit regression model with multicollinearity. Scientific African, Vol, 19.

2) Abdel-Fattah, M. (2022). On a new class of binomial ridge-type regression estimators. Communications in Statistics-Simulation and Computation, Vol. 51 (6), pp. 3272-3290.

3) Aguilera, A., Escabias, M. and Valderrama, M. (2006). Using principal components for estimating logistic regression with high-dimensional multicollinear data. Computational Statistics & Data Analysis, Vol. 50, pp. 1905-1924.

4) Asar, Y. and Genç, A. (2016). New Shrinkage Parameters for the Liu-type Logistic Estimators. J. Communications in Statistics, Vol. 45, pp. 1094–1103.

5) Asar, Y. and Genç, A. (2017). Two-parameter ridge estimator in the binary logistic regression. Communications in Statistics-Simulation and Computation, Vol. 46(6), pp. 7088-7099.

6) Cattell, R. (1966). The scree test for the number of factors. Multivar Behav Res Vol. 1(2), pp. 245–276.

7) Inan, D. and Erdogan, B. (2013). Liu-type logistic estimator. Comm. Stat. and Simulation and Computation, Vol 42, pp. 1578-1586.

8) Jadhav, N. (2020). On linearized ridge logistic estimator in the presence of multicollinearity. Computational Statistics, Vol. 35(2), pp. 667-687.

9) Kibria, B. (2003). Performance of some new ridge regression estimators. Commun. Statist. Theor. Meth., Vol. 32, pp. 419-435.

10) Kibria, B., Månsson, K. and Shukur, G. (2012). Performance of some logistic ridge regression estimators. Comput Econ, Vol. 40, pp. 401-414.

11) Liu, X. (2010). A Class of generalized shrunken least squares estimators in linear model. MSc. thesis. University of Manitoba, Canada.

12) Lukman, A., Emmanuel, A., Clement, O., and Ayinde, K. (2020). A modified ridge-type logistic estimator. Iranian Journal of Science and Technology, Transactions A: Science, Vol. 44, pp. 437-443.

13) Lukman, A., Ayinde, K., Binuomote, S. and Onate, A. (2019) Modified ridge-type estimator to combat multicollinearity: application to chemical data. Journal of Chemometrics. Vol. 33.

14) Mansson, K., Kibria, B. and Shukur, G. (2012). On Liu estimators for the logit regression model. Economic Model, Vol. 29, pp. 1483–1488.

15) Mansson, K. and Shukur, G. (2011). On ridge parameters in logistic regression. Comm Statist Theory Methods, Vol. 40, pp. 3366–3381.

16) Midi, H., Sarkar, S. and Rana, S. (2010). Collinearity diagnostics of binary logistic regression model. Journal of interdisciplinary mathematics, Vol. 13(3), pp.253-267.

17) Nja, M., Ogoke, U. and Nduka, E. (2013). The logistic regression model with a modified weight function. Journal of Statistical and Econometric Method, Vol. 2, pp.161‑171.

18) Roozbeh, M., Kafakia, S. and Arashi, M. (2016). A class of biased estimators based on QR decomposition. Linear Algebra and its Applications, ELSEVIER, Vol. 508, pp. 190–205.

19) Schaefer, R., Roi, L. and Wolfe, R. (1984). A ridge logistic estimator. Comm. Stat. Theory Methods, Vol. 13, pp. 99–113.

20) Smith, K., Slattery, M. and French, T. (1991). Collinear nutrients and the risk of colon cancer. Journal of Clinical Epidemiolgy, Vol.44, pp.715-723.

21) Varathan, N. (2022). An improved ridge type estimator for logistic regression. Statistics in Transition New Series, Vol.23, pp.113-126.

22) Watkins, D. (2002). Fundamentals of matrix computations. 2nd ed., John Wiley, New York.

23) Wu, J. and Asar, Y., (2016). On almost unbiased ridge logistic estimator for the logistic regression model. Hacettepe Journal of Mathematics and Statistics, Vol. 45(3), pp. 989‑998.