International Journal of Data Science and Analysis
2022; 8(6): 187-193
http://www.sciencepublishinggroup.com/j/ijdsa
doi: 10.11648/j.ijdsa.20220806.13
ISSN: 2575-1883 (Print); ISSN: 2575-1891 (Online)

Logistic Estimation Method in the Presence of Collinearity and It's Application

Runyi Emmanuel Francis[1], Maureen Tobe Nwakuya[2]

[1]Department of Statistics, Federal Polytechnic Ugep, Ugep, Nigeria

[2]Department of Mathematics and Statistics, University of Port Harcourt, Port Harcourt, Nigeria

Email address:

runyiemmanuel@gmail.com (Runyi Emmanuel Francis), maureen.nwakuya@uniport.edu.ng (Maureen Tobe Nwakuya)

To cite this article:

Runyi Emmanuel Francis, Maureen Tobe Nwakuya. Logistic Estimation Method in the Presence of Collinearity and It's Application. *International Journal of Data Science and Analysis*. Vol. 8, No. 6, 2022, pp. 187-193. doi: 10.11648/j.ijdsa.20220806.13

Received: November 18, 2022; **Accepted**: December 6, 2022; **Published**: December 15, 2022

Abstract: Regression analysis is a widely used statistical technique in investigating relationships between the response variable and outcome variable. The logistic regression examines the relationship between variables when the response variable has a dichotomous output i.e., has two possible levels and outcome variable which could be categorical or continuous. Logistic regression using maximum likelihood estimation has gained wide use in determining the parameter estimate but, in the case, where the covariates are correlated, there is an inflation in the variance, standard error of the estimator and high coefficient of determination for the regression model, leading to the problem of multicollinearity in the regression model, thereby resulting to an incorrect conclusion about the relationship among these variables, hence the traditional method of estimating the parameters fails and becomes unstable. To attempt addressing the presence of multicollinearity in the regression model, various methods have been proposed which includes Ridge estimator, Stein estimator, Bayesian estimator and Liu estimators. We therefore propose a modified estimator for estimating the parameter of the logit model in the presence of multicollinearity by modifying the existing Liu logistic estimator. The modified estimator is applied to real life data. Results showed that the Modified Liu Logistic estimator outperformed the existing estimators considered in this study, in terms of smaller variance, bias and the MSE of the estimator.

Keywords: Logistic Regression, Multicollinearity, Maximum Likelihood Estimation, Bias, Variance, Mean Square Error

1. Introduction

The regression analysis is concerned with describing the dependence and relationship between a response variable and one or more explanatory variables, with a view of estimating and or predicting the (population) mean value [20]. When the response variable is binary or dichotomous taking in two possible values, the Logistic regression model then suits the standard method of analysis.

In recent years, logistic regression has been applied extensively in numerous disciplines. There is a wide range of application of regression in areas such as engineering, the physical and chemical sciences, economics, management and the social sciences [15]. The logistic regression describes the relationship between categorical response variable and a set of explanatory variables [8]. It is one of the many cases of generalized linear models [23] characterized by three components: a random component, which identifies the probability distribution of the response variable; a systematic component, which specifies a linear function of the explanatory variables that is used as a predictor; a link function describing the functional relationship between the systematic component and the expected value of the random component [14].

In logistic regression, the response (outcome) variable is usually dichotomous, but it may be polytomous, that is, having more than two response levels. These multiple-level responses can be nominal or ordinal scaled [6] resulting to a multinomial logistic regression or ordinal logistic regression which is a simple extension of binary logistic that allows for more than two categories of the dependent or outcome variable, where nominal responses are not ordered and ordinal responses are ordered. The logistic regression uses maximum likelihood estimation to evaluate the probability of categorical membership [27].

In most cases, some explanatory variables are seen to relate with each other introducing multicollinearity into the models. Fisher, R. A. is known to be the earliest researcher on multicollinearity [3]. Multicollinearity occurs when there is interdependence, among the explanatory (independent) variables in the regression analysis that is, when a set of data can be expressed exactly or nearly as a linear combination of the other in the set of explanatory variables. When multicollinearity occurs, parameter estimates are incorrect and this renders the model unreliable. There is also variance inflation of the maximum likelihood estimates in the logistic regression which may not result to an efficient, more reliable estimate of the parameter estimate thereby affecting both prediction and inferential conclusion in the logistic regression model. When there is exact collinearity among the explanatory variables, the information matrix assumes singularity and the iterative weighted least squares method fails [16]. Severe multicollinearity can lead to instability of the regression coefficients. Multicollinearity has several manifestations, including small change in the data which can produce wide swings in the parameter estimates. Parameter coefficients can have high standard errors, high coefficient of determination (R^2) for the regression model and coefficients may have the wrong sign [5]. Consequently, the resulting model is not reliable and will result in incorrect conclusions about the relationship among the variables [18].

The presence of multicollinearity in the logistic regression may indicate that some explanatory variables are linear combination of the other variables. This does not improve explanatory power of a model and could be dropped from the model. Since the problem of collinearity was first revealed, researchers have tried to develop statistical remedies to combat the problem introduced into the model as a result of the presence of multicollinearity. To address the multicollinearity in logistic regression, several approaches have been proposed. For linear regression with continuous dependent variables, there are multiple options, including shrinkage methods such as ridge regression [11], Liu-type regression [11], Least Absolute Shrinkage and Selection Operator [22] and Elastic Net [28]; Dimension reduction methods such as Principal Component Analysis [13]. Pure Bayesian regression [28, 9] and Bridge regression which is ridge or lasso regression with Bayesian prior [4]. For dichotomous and polytomous outcomes, the properties of logistic regression make it less flexible in addressing multicollinearity. The commonly used statistical methods for overcoming multicollinearity for a logistic regression are Ridge Logistic Estimator (RLE) [21], Liu Logistic Estimator (LLE) [10, 23, 12], Partial Least Squares Multinomial Logistic Regression (PLSMLR) [26], Principal Component Logistic Estimator (PCLE) [1], Modified Logistic Ridge Estimator (MLRE) [17].

2. Materials and Methods of Research

2.1. Binary Logistic Model

The binary logistic regression model:

$$y_i = \pi_i + \varepsilon_i \tag{1}$$

where $\varepsilon_i (i = 1,2, \dots, n)$ are disturbance assumed to be distributed with mean 0 and variance $\omega_i = \pi_i(1 - \pi_i)$. π_i is the expectation of y_i when i^{th} value of the dependent variable is distributed as Bernoulli $Be(\pi_i)$ such that

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

where β is a $(p+1) \times 1$ vector of coefficients, x_i is the row of X which is an $n \times (p+1)$ matrix.

Let the relationship between the dependent variable y and the independent variable $x_1, x_2, \dots, x_p$ be as follows [8, 2].

$$y_i = \frac{exp^{\beta_0 + \sum_{j=1}^{p} \beta_j + x_{ij}}}{1 + exp^{\beta_0 + \sum_{j=1}^{p} \beta_j + x_{ij}}} + \varepsilon_i \tag{3}$$

Where, $i = 1,2, \dots, n$; $\text{n} = \text{sample size}$;

$\text{p} =$ number of the explanatory (independent) variables;

$x_{ij} =$ the measurment of the j^{th} explanatory variable for the i^{th} observation ($\text{i} = 1, 2, \dots, \text{n}; \text{j} = 1,2, \dots, \text{p}$);

$$\beta_j = j^{th}\ regression\ parameters;$$

$$\varepsilon_i = disturbance\ (random\ error)\ term\ for\ the\ i^{th}\ observation.$$

$$\begin{cases} 1\ i^{th}\ observation\ under\ consideration \\ \qquad 0\ Otherwise \end{cases}$$

The fitted binary logistic model is therefore as follows:

$$Log\left(\frac{\hat{\pi}_1(x)}{\hat{\pi}_0(x)}\right) = \beta_0 + \sum_{j=1}^{p} \beta_j + x_{ij} \tag{4}$$

where β_0 = the constant;

β_j = the regression coefficient parameter.

2.2. Maximum Likelihood Estimator

The most common method of estimating the parameters of the model in a logistic regression is to apply the maximum likelihood method. The resulting log-likelihood equation of model (2) above is given by

$$L(\beta) = \sum_{i=1}^{N} y_i Log(\pi_i) + \sum_{i=1}^{N} (1 - y_i) Log(1 - \pi_i) \tag{5}$$

where π_i is the i^{th} element of the vector π_i, $i = 1,2, \dots, n$. Solving the above equation (5) by taking the first derivative and equals the expression to zero, the maximum likelihood estimation are obtained.

$$\frac{\delta l}{\delta \beta_0} = \sum_{i=1}^{N} (y_i - \pi_i) = 0$$

$$\frac{\delta l}{\delta \beta_j} = \sum_{i=1}^{N} (y_i - \pi_i) x_i = 0$$

Since the above equations are nonlinear in β, the iterative weighted least squares (IWLS) algorithm is applied. Therefore, maximum likelihood estimator (MLE) of β can be obtained using the iterative weighted least squares (IWLS) which is given as follows:

$$\hat{\beta}_{MLE} = (X'\widehat{W}X)^{-1}X'\widehat{W}\hat{Z}. \tag{6}$$

where $\widehat{W} = diag[\pi_i(1-\pi_i)]$ and $\hat{Z}$ is a column vector with elements $z_i = Log(\pi_i) + \frac{(y_i-\pi_i)}{\pi_i(1-\pi_i)}$ which is an unbiased estimator of β. The variance-covariance matrix of $\hat{\beta}_{MLE}$ is given by

$$Var(\hat{\beta}_{MLE}) = Cov(\hat{\beta}_{MLE}) = \left(X'\widehat{W}X\right)^{-1} \tag{7}$$

$$\{X'diag[\pi_i(1-\pi_i]X\}^{-1}$$

The MSE of the asymptotically unbiased $\hat{\beta}_{MLE}$ is

$$MSE = E\left(\hat{\beta}_{MLE} - \beta\right)'\left(\hat{\beta}_{MLE} - \beta\right) \tag{8}$$

$$tr\left[Var\left(\hat{\beta}_{MLE}\right)\right] = \sum_{j=1}^{p} \frac{1}{\lambda_i} \tag{9}$$

For a long time, the MLE has being treated as the best estimator. However, the variance of the MLE becomes inflated in the presence of multicollinearity. To address the effect of the presence of multicollinearity, several biased estimators have been developed.

2.3. Ridge Logistic Estimator

Horel and Kennard [7] proposed the Ridge estimator in other to control the inflation and general variability associated with the Maximum Likelihood Estimator. Scheafer *et al.,* extended the estimator to the Logit model [21]. The idea behind the ridge regression is that by adding a positive constant k > 0 to the diagonal of the information matrix $(X'\widehat{W}X)$, by so doing, one can obtain a smaller condition number and the variance is decreased [25].

Ridge Estimator is given as follows:

$$\hat{\beta}_{RE} = (X'\widehat{W}X + kI)^{-1}X'\widehat{W}Z \tag{10}$$

where $k = \frac{1}{\hat{\beta}'_{MLE}\hat{\beta}_{MLE}}$ the biasing constant

Let $C_k = X'\widehat{W}X + kI$, $C = X'\widehat{W}X$ and $I = pxp\ Identity\ Matrix.$

The bias, variance and mean squared error expression for the Ridge Logistic Estimator are given as:

$$\text{Bias } (\hat{\beta}_{RE}) = -k\mathrm{C}_k^{-1}\hat{\beta}_{MLE} \tag{11}$$

$$\text{Var } (\hat{\beta}_{RE}) = Q\mathrm{C}_k^{-1}\, C\, \mathrm{C}_k^{-1}Q' \tag{12}$$

$$\text{MSE } (\hat{\beta}_{RE}) = \sum_{j=1}^{J} \frac{\sigma^2\lambda_j}{(\lambda_j+k)^2} + \sum_{j=1}^{J} \frac{k^2\alpha_j^2}{(\lambda_j+k)^2} \tag{13}$$

where Q is the matrix whose columns are eigen vectors of $X'\widehat{W}X$.

where $\alpha_j^2 = Q'\hat{\beta}_{MLE}$; λ_j is the j^{th} eigen value of $X'\widehat{W}X$

2.4. Liu logistic Estimator

To address the effect of the presence of multicollinearity, Liu, K introduced the Liu estimator by combining the Stein estimator $\hat{\beta}_{SE} = C\hat{\beta}_{MLE}$, where $0 < c < 1$, and the biasing parameter obtained as $c = \frac{\hat{\beta}'_{MLE}\hat{\beta}_{MLE}}{\hat{\beta}'_{MLE}\hat{\beta}_{MLE} + tr(X'\widehat{W}X)^{-1}}$ with the Ridge estimator $\hat{\beta}_{RE} = (X'\widehat{W}X + kI)^{-1}X'\widehat{W}Z$, where the ridge biasing parameter $k = \frac{1}{\hat{\beta}'_{MLE}\hat{\beta}_{MLE}}$ to form the Liu estimator [10, 24] given as

$$\hat{\beta}_{LE} = \left(X'\widehat{W}X + I\right)^{-1}(X'\widehat{W}X + dI)\,\hat{\beta}_{MLE} \tag{14}$$

where the Liu biasing parameter is obtained as suggested by Liu, K $d = \frac{(\gamma_j^2-1)}{\frac{1}{q_j}-\gamma_j^2}$.

where $\gamma_j^2 = \sum_{i=1}^{p} \upsilon_j(\hat{\beta}_{MLE})$ and q_j is the j^{th} eigenvalue of the weighted information matrix $\left(X'\widehat{W}X\right); j = 1,2,\ldots,p$; υ_j is the j^{th} eigenvector corresponding to the j^{th} eigenvalue of $\left(X'\widehat{W}X\right)$.

where $0 < d < 1$ and $\hat{\beta}_{MLE} = (X'WX)^{-1}X'WZ$.

$$\text{Let C} = X'\widehat{W}X \text{ then } \hat{\beta}_{LE} = (C+I)^{-1}(C+dI)\,\hat{\beta}_{MLE} \tag{15}$$

The bias, variance and mean squared error of the Liu logistic estimator are given by

$$\text{Bias } (\hat{\beta}_{LE}) = (d-1)\Lambda_1^{-1}\hat{\beta}_{MLE} \tag{16}$$

$$\text{Var } (\hat{\beta}_{LE}) = Q\Lambda_1^{-1}\Lambda_d\Lambda^{-1}\Lambda_d\Lambda_1^{-1}Q' \tag{17}$$

where Q is the matrix whose columns are Eigen vectors of $X'\widehat{W}X$, Λ is the diagonal matrix containing the Eigen values of $X'\widehat{W}X$, Λ_1 is $(C+I)$ and Λ_d is $(C+dI)$.

$$\text{MSE } (\hat{\beta}_{LE}) = \sum_{j=1}^{J} \frac{(\lambda_j+d)^2}{\lambda_j(\lambda_j+d)^2} + (d-1)^2 \sum_{j=1}^{J} \frac{\alpha_j^2}{(\lambda_j+1)^2} \tag{18}$$

where $\alpha_j^2 = Q'\hat{\beta}_{MLE}$; λ_j is the j^{th} eigen value of $X'\widehat{W}X$.

$d = \frac{(\gamma_j^2-1)}{\frac{1}{q_j}-\gamma_j^2}$. Where $\gamma_j^2 = \sum_{i=1}^{p} \upsilon_j(\hat{\beta}_{MLE})$ and q_j is the j^{th} eigenvalue of the weighted information matrix $\left(X'\widehat{W}X\right); j = 1,2,\ldots,p$; υ_j is the j^{th} eigenvector corresponding to the j^{th} eigenvalue of $\left(X'\widehat{W}X\right)$.

2.5. Modified Logistic Ridge Estimator

To address the presence of multicollinearity Ogoke *et al.,* presented a modification on the Ridge Logistic estimator [17] developed by Scheafer *et al.,* [21] by exponentiating the response probability, which enhanced the weighted matrix thereby reducing the variances of the parameter estimates in the logistic regression. Their modified estimator is given by

$$\beta_{MLRE} = (X'\widehat{W}^{\sqrt{1+\delta}}X + kI)^{-1}X'\widehat{W}^{\sqrt{1+\delta}}\hat{Z}^{\sqrt{1+\delta}} \tag{19}$$

where

$$\widehat{W}^{\sqrt{1+\delta}} = diag[\pi_i^{\sqrt{1+\delta}}\left(1-\pi_i^{\sqrt{1+\delta}}\right)]\ 0 \le \delta \le 1.$$

$$\hat{Z}^{\sqrt{1+\delta}} = z_i = \eta + \frac{(y_i - \pi_i^{\sqrt{1+\delta}})}{\pi_i^{\sqrt{1+\delta}}(1-\pi_i^{\sqrt{1+\delta}})}$$

where

$$\eta = x_{ij}\hat{\beta}_0 + x_{ij}\hat{\beta}_1 + \cdots + x_{ij}\hat{\beta}_{p-1}.$$

2.6. Modified Liu Logistic Estimator

The Modified Liu Logistic Regression estimator is a combination of [10] and [17] Modified logistic ridge estimator is stated as follows:

$$\hat{\beta}_{MLLE} = (X'\hat{W}^{\sqrt{1+\delta}}X + I)^{-1}(X'\hat{W}^{\sqrt{1+\delta}}X + dI)\tilde{\beta} \quad (20)$$

where
I= Identity Matrix and

$$\tilde{\beta} = (X'\hat{W}^{\sqrt{1+\delta}}X + I)^{-1}X'\hat{W}^{\sqrt{1+\delta}}\hat{Z}^{\sqrt{1+\delta}}$$

$$\hat{W}^{\sqrt{1+\delta}} = diag[\pi_i^{\sqrt{1+\delta}}\left(1 - \pi_i^{\sqrt{1+\delta}}\right)]$$

$$\hat{Z}^{\sqrt{1+\delta}} = z_i = Log\left(\pi_i^{\sqrt{1+\delta}}\right) + \frac{(y_i - \pi_i^{\sqrt{1+\delta}})}{\pi_i^{\sqrt{1+\delta}}(1-\pi_i^{\sqrt{1+\delta}})}$$

where
$\pi_i = i^{th}$ Response probability and $0 \le d \le 1$ and $0 \le \delta \le 1$

The Modified Liu logistic estimator for the logit model is a biased estimator, and a direct modification of the one proposed by Liu, K. for the linear regression model [10]. The parameter δ may take values between zero and one and when δ is equal to 0, we have $\hat{\beta}_{MLLE} = \hat{\beta}_{LE}$. When δ is less than or equal to one, we have then $\hat{\beta}_{MLLE} < \hat{\beta}_{LE}$. Since $\hat{\beta}_{LE}$ addresses the problem of multicollinearity, $\hat{\beta}_{MLLE}$ is assumed to perform better than $\hat{\beta}_{LE}$ in such situation. The presented estimator is compared with the existing Liu, Ridge and Modified Ridge Logistic estimators in terms of smaller bias, variance and mean squared error. The bias, variance and mean squared error of the presented estimator are given by

$$\text{Bias}\ (\hat{\beta}_{MLLE}) = (d-1)\,\tilde{\Lambda}_1^{-1}\tilde{\beta} \quad (21)$$

$$\text{Var}\ (\hat{\beta}_{MLLE}) = Q\tilde{\Lambda}_1^{-1}\tilde{\Lambda}_d\tilde{\Lambda}^{-1}\tilde{\Lambda}_d\tilde{\Lambda}_1^{-1}Q' \quad (22)$$

where Q is the matrix whose columns are eigenvectors of $X'\hat{W}^{\sqrt{1+\delta}}X$, $\tilde{\Lambda}$ is the diagonal matrix containing the eigen values of $X'\hat{W}^{\sqrt{1+\delta}}X$, $\tilde{\Lambda}_1 = (\tilde{C} + I)$, $\tilde{\Lambda}_d = (\tilde{C} + dI)$ and $\tilde{C} = X'\hat{W}^{\sqrt{1+\delta}}X$.

$$\text{MSE}\ (\hat{\beta}_{LE}) = \sum_{j=1}^{J}\frac{(\lambda_j^{\sqrt{1+\delta}}+d)^2}{\lambda_j^{\sqrt{1+\delta}}(\lambda_j^{\sqrt{1+\delta}}+d)^2} + (d-1)^2\sum_{j=1}^{J}\frac{\alpha_j^2}{(\lambda_j^{\sqrt{1+\delta}}+1)^2} \quad (23)$$

where $\alpha_j^2 = Q'\hat{\beta}_{MLE}$; λ_j is the j^th eigen value of $X'\hat{W}^{\sqrt{1+\delta}}X$.

$$d_M = \frac{(\gamma_j^2 - 1)}{\frac{1}{q_j} - \gamma_j^2}$$

where $\gamma_j^2 = \sum_{i=1}^{p} \upsilon_j\tilde{\beta}$ and q_j is the j^{th} eigenvalue of the weighted information matrix $\left(X'\hat{W}^{\sqrt{1+\delta}}X\right)$; $j = 1,2,\ldots,p$; υ_j is the j^{th} eigenvector corresponding to the j^{th} eigenvalue of $\left(X'\hat{W}^{\sqrt{1+\delta}}X\right)$ for the modified biasing parameter.

3. Application, Analysis, Results and Discussion

3.1. Application and Analysis

We present the results of the correlation matrix. Also, the results of the maximum likelihood estimator of the logistic regression, Liu logistic estimator, ridge logistic estimator, modified ridge logistic estimator and the modified logistic estimator when compared in terms of bias, variance and mean squared error criterion using a life dataset.

A secondary dataset on hypotensive patients was used for illustration. The variables considered are Hypertension, Age (AGE), Body Mass Index (BMI), Marital Status (MS), Diabetic (D), and Smoking Habit (SH). We take response variable (Categorical) Y= Hypertension with predictor variables X1=AGE, X2=BMI, X3=MS, X4=D, X5=SH.

For this analysis, we first observed the correlation matrix to detect the presence of multicollinearity and afterwards advanced to use the Maximum Likelihood Estimator (MLE) estimation method to obtain the parameters of the logistic model and then used the Liu logistic estimator, Ridge logistic estimator, Modified Ridge logistic estimator and the Modified Liu logistic estimator to address the problem of multicollinearity as well to test its competency.

3.2. Results

Table 1 shows the classification table of the variables of the model for the prediction of hypertension case, in order to describe the nature of the variables under study. The following is an analytical presentation of these measures for each variable of the model. The result revealed that age, BMI and marital status are significant to the cause of hypertension while diabetic and smoking habit are not significant to the cause of hypertension.

Table 1. *Classification table for the prediction of hypertension.*

Predictor (s)	Hypertensive (n=21)	Not hypertensive (=129)	P-value
Age			
Young (18-40 Years)	2 (9.5)	109 (84.5)	<0.0001
Old (Above 40 Years)	19 (90.5)	20 (15.5)	
BMI			
Not obese (<30)	1 (48)	117 (90.7)	<0.0001

Predictor (s)	Hypertensive (n=21)	Not hypertensive (=129)	P-value
Obese (≥30)	20 (90.5)	12 (9.3)	
Marital Status			
Married	3 (14.3)	116 (89.9)	<0.0001
Single	18 (85.7)	13 (10.1)	
Diabetic			
Non diabetic	1 (4.8)	113 (87.6)	0.063
Diabetic	20 (95.2)	16 (12.4)	
Smoking Habit			
Non-smokers	5 (23.8)	112 (86.8)	0.559
Smokers	16 (76.2)	17 (13.2)	

Percentages in bracket

Table 2. *Correlation matrix.*

Variable	X_1 (Age)	X_2 (BMI)	X_3 (MS)	X_4 (D)	X_5 (SH)
X_1 (Age)	1				
X_2 (BMI)	0.804	1			
X_3 (MS)	0.824	0.860	1		
X_4 (D)	0.841	0.889	0.908	1	
X_5 (SH)	0.886	0.884	0.802	0.832	1

Table 3. *β Parameter estimates.*

Estimator (s)	β_0	β_1	β_2	β_3	β_4	β_5
MLE	-5.589	1.374	4.399	1.008	2.159	-2.943
Liu Logistic	-24.723	-1.644	-4.569	-0.515	2.391	0.276
Ridge Logistic	-22.521	-1.022	-2.596	-0.055	4.39	2.412
Modified Ridge Logistic	-18.716	-2.114	-4.143	-0.502	3.513	0.278
Modified Liu Logistic	-18.998	-2.65	-4.332	-0.504	2.331	0.274

Table 4. *Variance of β parameter estimates.*

Estimator (s)	β_0	β_1	β_1	β_3	β_4	β_5
MLE	2.741	1.080	2.833	1.729	2.663	4.461
Liu Logistic	1.833	0.442	0.565	0.976	1.933	2.225
Ridge Logistic	1.921	0.661	0.721	0.792	2.015	2.699
Modified Ridge Logistic	1.746	0.301	0.391	0.925	0.972	0.765
Modified Liu Logistic	1.741	0.299	0.379	0.775	0.921	0.642

The variance of the proposed MLLE has the least variance

Table 5. *Standard errors of β parameter estimates.*

Estimator (s)	β_0	β_1	β_1	β_3	β_4	β_5
MLE	1.513	1.715	1.698	1.700	2.024	1.665
Liu Logistic	0.784	0.993	0.422	1.246	1.935	1.225
Ridge Logistic	0.943	1.959	0.951	1.551	2.016	2.002
Modified Ridge Logistic	0.789	0.952	0.492	1.121	1.446	2.055
Modified Liu Logistic	0.688	0.933	0.396	1.119	1.366	1.116

The standard errors of the proposed MLLE has the least standard error values

Table 6. *Bias of β parameter estimates.*

Estimator (s)	β_0	β_1	β_1	β_3	β_4	β_5
Liu Logistic	2.239	-0.614	-2.299	0.436	-1.062	1.770
Ridge Logistic	2.920	-0.421	-2.143	-0.309	-1.093	2.966
Modified Ridge Logistic	1.535	-1.135	-1.535	-0.437	-1.892	0.499
Modified Liu Logistic	1.523	-1.153	-1.545	-0.574	-1.937	0.429

The bias of the proposed MLLE has the least bias values

***Table 7.** Computed MSE values of the estimators.*

	MLE	LLE	LRE	MLRE	MLLE
MSE (β)	17.8522	8.2420	9.9959	7.9959	**7.665389**

The mean square error of the proposed MLLE has the least mean square error

3.3. Discussion of the Findings

To establish the existence of the presence of multicollinearity, Table 2 shows the correlation matrix obtained which showed that all the bivariate correlations are greater than 0.88 which means that there exists the presence of multicollinearity [19].

Table 3 indicates the β parameter estimates by the traditional maximum likelihood estimator which suffer from multicollinearity as shown in the correlation matrix. For addressing multicollinearity, the Liu Logistic estimator, Ridge Logistic estimator and Modified Ridge estimator were applied. For the same value of the biasing constant d in the Liu logistic estimator, the Modified Liu logistic estimator reduces the bias, variance and MSE as likened with the parent Liu logistic estimator. The results are showed in tables 4, 5 and 6. The bias of the ordinary Liu Logistic estimator are 2.2391, -0.6139, -2.2985, -0.4360, -1.0617, 1.7697, bias values for Ridge Logistic estimator are 2.9201, -0.4211, -2.1432, -0.3039, -1.0928, 2.9664, bias values for Modified Ridge Logistic estimator are 1.5347, -1.1347, -1.5345, 0.4371, -1.8919, 0.4993 while bias values of the Modified Liu Logistic estimator are 1.5227, -1.1527, -1.5449, -0.5737,-1.9373, 0.4294. The modified Liu logistic estimator has the smallest variance of the parameter estimates as showed in table 4 above. The variance of estimates of the ordinary Liu logistic are 1.8331, 0.4418, 0.5646, 0.9758, 1.9327, 2.6423, for the Ridge Logistic are 1.9210, 0.6613, 0.7214, 0.7921, 2.0150, 2.6991, for Modified Ridge Logistic are 1,7459, 0.3011, 0.3912, 0.9247, 0.9721, 0.7649, while those of the presented estimator are 1.7410, 0.2992, 0.3789, 0.7749, 0.9214, 0.6419. The MSE of the parameter estimates using the MLE is 17.852, for the Liu Logistic estimator is 8.242, for Logistic Ridge estimator is 9.996, for Modified Logistic Ridge estimator is 7.996 and for the presented estimator Modified Liu Logistic estimator has the least MSE of 7.665.

According to the results, the modified Liu logistic estimator MLLE has an improved performance in comparison to the Liu Logistic estimator, Ridge Logistic estimator and the Modified Ridge Logistic estimator. This is based on bias, variance and MSE values.

4. Summary and Conclusion

4.1. Summary

This study modified the existing Liu logistic estimator by exponentiating the response probability to obtain a new estimator called the Modified Liu Logistic Estimator (MLLE). The Modified Liu Logistic Estimator (MLLE) is examined and compared against the Liu Logistic, Ridge Logistic and Modified Ridge Logistic estimators in terms of the bias, variance and MSE criterion. Expressions for these criterions were stated in each case.

To demonstrate the efficiency of the modified Liu Logistic estimator, a real life dataset of contributing factors responsible for a person being prone to hypertension survey was used. Results were obtained and generated using both R statistical software and MathLab. The results showed that the modified Liu Logistic estimator have smaller values for the bias, variance and mean squared error as compared to the values obtained from the traditional MLE, Liu logistic estimator, Ridge Logistic estimator and the Modified Ridge Logistic estimator indicating its superiority.

4.2. Conclusion

This study presented an estimator, named the Modified Liu Logistic estimator which is an extension of the Liu Logistic estimator. The modified Liu logistic estimator outperformed the Liu Logistic estimator, Ridge Logistic estimator and the Modified Ridge Logistic estimator in terms of smaller bias, variance, and mean squared error (MSE) values using a life dataset. The findings showed that the presented estimator can be used in place of Liu Logistic estimator, Ridge Logistic estimator and the Modified Ridge Logistic estimator to address multicollinearity issues arising from real life data situations.

References

[1] Aguilera, A. M., M. Escabias and M. J Valderrama (2006). Using Principal Component for estimating Logistic Regression with high-dimensional multicollinear data. Comput. Stat. Data Anal., 50 (8): 1905-1924.

[2] Agresti, A. (2007). An introduction to categorical data analysis, 2nd Edition. New York: Wiley.

[3] Fisher, R. A. (1935). "The fiducial argument in statistical inference". Annals of Eugenics. 8 (3): 391–398.

[4] Frank, I. E and J. H Friedman, 1993. A Statistical view of some Chenometrics regression tools. Technometrics, 35: 109-135.

[5] Grenne, W (2000) Econometric Analysis, 5th Ed. New Jersey: Prentice Hall.

[6] Hansman, J. and McFadden, D. (1984). Specification tests for the multinomial logit model, Econometrica 52 (5): 1219-1240.

[7] Hoerl, A. E. and Kennard, R. W. (1970). Ridge Regression Biased Estimation for non-Orthogonal Problems. Communication is statistics: Theory and Methods (4) 105-123.

[8] Hosmer, D. W. and Lemeshow, S. (2000). Applied Logistic Regression, 2nd edn. New York: Wiley.

[9] Leamer, E. E. (1978), Specification searches: Ad Hoc Inference with Non-experimental Data, New York: John Wiley.

[10] Liu, K (1993) A New class of biased estimate in linear regression. Communications in Statistics. Theory and Methods, (22): 393-402.

[11] Liu, K (2003). Using Liu-Type estimator to combat collinearity. Communications in Statistics-Theory and Methods, 32 (5): 1009-1020.

[12] Mansson, K., Kibria, B. M. G and Shukur, G. (2012) On Liu Estimators for the logit Regression Model. The Royal Institute of Technology Center of Excellence for Science and Innovation studies (CESIS), Sweden. 259-272.

[13] Massy, W. F. (1965). Principal component Regression in Explanatory Statistical Research. Journal of American Statistical Association, 60 (309): 234-256.

[14] Maura E Stokes, Charles, S. Davis (2000). Categorical Data Analysis using the SAS System. Second Edition. Cary; NC; SAS Institute Inc.

[15] Montgomery, D. C., Peck, E. A., and Vining, G. G. (2001). Introduction to Linear Regression Analysis (3rd ed.): John Wiley and Sons, Inc.

[16] Nja, M. E., Ogoke, U. P and Nduka, E. C. (2013) The Logistic Regression Model with a Modified Weight Function. Journal of Statistical and Econometrics Method, 2 (3); 161-171.

[17] Ogoke U. P., Nduka E. C., and Nja M. E. (2013). A New Logistic Ridge Regression Estimator Using Exponentiated Response Function. Journal of Statistical and Econometric Method. Vol. 2 (4); p. 161-171.

[18] Onwukwe, C. E., and I. A. Aki. (2015) "On Selection of Best Sensitive Logistic Estimator in the Presence of Collinearity." American Journal of Applied Mathematics and Statistics 3.1 (2015): 7-11.

[19] Rasha A. F., and Samah M. A., (2017). Evaluating the performance of the Liu Logistic regression estimator. Research Journal of Mathematics and Statistics, 9 (2): 11-19.

[20] Ryan, T. P. (1997). Modern Regression Methods. John Wiley & Sons, Inc., New York.

[21] Schaefer, R. L, Roi, L. D, and Wolfe, R. A (1984). A Ridge Logistic Estimator. Communications in statistics. Theory and Methods, 13: 99-133.

[22] Tibshirani, R. (1996). Regression Shrinkage and Selection via the LASSO. Journal of Royal Statistical Society. 58: 267-288.

[23] Traissac P., Martin-Prevel Y., Delpeuch F., & Maire B. (1999). Logistic regression vs other generalized linear models to estmate prevalence rate ratios. Journal of the Royal Statistical Society 47 (6), 593-604.

[24] Urgan, N. N and Tez, M (2008) Liu Estimator in Logistic Regression when the data are collinear. International Conference. "Continuous optimization and knowledge based technologies" 323-327.

[25] Vago, E. and S. Kememy, (2006). Logistic Ridge regression for clinical data analysis. Applied Ecol. Environ. Res., 4: 171-179.

[26] Wang H. W. (2006) PLS Logistic regression on computational data and its Application. The Journal of Quantitative & Technical Econometrics 9: 156-161.

[27] Xinfeng, C. (2015) On the Almost Unbiased Ridge and Liu Estimator in the logistic regression model. International Conference on Social Science, Education, Management and Sports Education, Atlentis Press, Amsterdam, (17) 1663-1665.

[28] Zou, H. and Hastie, T. (2005). "Regularization and Variable Selection via the elastic net." Journal of Royal Statistical Society, 67: 301-320.