Mathematical Theory and Modeling
ISSN 2224-5804 (Paper) ISSN 2225-0522 (Online)
Vol.3, No.8, 2013

The Logistic Regression Model with a Modified Weight Function in Survival Analysis

U.P Ogoke
Dept of Mathematics and Statistics, University of Port Harcourt
uchedubem@yahoo.com
E.C Nduka
Dept of Mathematics and Statistics, University of Port Harcourt
etelnduka@yahoo.com
M.E Nja
Department of Mathematics, Federal University Lafia
mbe_nja@yahoo.com

Abstract
Most of the Ridge regression estimators can only achieve one property or the other, namely, variance reduction, bias reduction or reduced Mean Square Error. To achieve both variance and bias reduction in Logistic Ridge regression the Modified Logistic Ridge regression estimator is designed. The estimator is used to model the survival function of diabetic patients who are exposed to some specified medication. The model is formulated in such a way that the response probability is made to act as survival function. By some radical exponentiation of the weight function, the proposed estimator is found to have smaller bias than the Generalized Ordinary Logistic Ridge estimator.
Keywords: logistic ridge estimator, survival function, response probability, collinearity, means square error, bias

1.0 INTRODUCTION

There exists a large volume of literature in the solution of collinearity problem among explanatory variables in Generalized Linear Model analysis. The most celebrated Rigde regression technique came with it, the problem of bias. Many authors, including Belsley, Edwin, Welsch (1980), Madala (1992), Hawkins, Yin (2002), Carley, Kathleen, Natalia (2004), Batah (2011), Khurana, Chanbey, Chandra (2012) and Muniz, Kibira, Manson, Shukur (2012) have made significant inputs in the reduction of bias associated with the Ridge regression procedure.
Singh and Chanbey (1987), Nomura (1988) and Gruber (1998), Batah (2011), Khurana et al (2012) developed the Jackknife Ridge estimators to further reduce bias in Ridge regression for General Linear Models using canonical transformation. Even though much work has been done in Ridge regression estimation in General Linear Models, not much has been done in the case of Generalized Linear Model. In this paper, we develop the modified Logistic Ridge estimator using canonical transformation as a special case of the Generalized Linear Model estimator. Two theorems with their corresponding proofs are developed to support the theory in addition to an illustrative example. Both the theorems and the illustrative example demonstrate the fact that the proposed method is superior to its Ridge Logistic equivalent in terms of bias and variance reduction. The existence of collinearity is established among explanatory variables from the eigenvalues of the information matrix and from the collinearity matrix.

2.0 ORDINARY RIDGE REGRESSION ESTIMATOR FOR GENERAL LINEAR MODELS

The multiple linear regression model

$$Y = X\beta + e \tag{1}$$

where Y is an $(n\times 1)$ vector of observations, β is a $(p\times 1)$ vector of unknown regression coefficients, X is an $(n\times p)$ matrix of explanatory variables $X_1, X_2, ..., X_p$ and e is an $(n\times 1)$ vector of errors can be written in canonical form as

$$Y = Z\alpha + e \tag{2}$$

where $Z = XT$, T is the matrix of eigenvectors of $X^{'}X$.

$$Z'Z = T = T'X'XT = diag(\lambda_1, \lambda_2, ..., \lambda_p)$$

where λ_i is the ith eigenvalue of $X'X$.

$$\alpha = T'\beta, \qquad T'T = TT' = I_p$$

Then the OLS estimator of α is given by

$$\hat{\alpha} = (Z'Z)^{-1}Z'Y = J^{-1}Z'Y \tag{3}$$

$$\hat{\beta}_{OLS} = T\hat{\alpha}_{OLS} \tag{4}$$

$$\Rightarrow \quad \hat{\beta}_{ORE} = T\hat{\alpha}_{ORE} = T(I - kA_k^{-1})\alpha_{ORE} \tag{5}$$

where

$$A_k = diag(\lambda_1 + k, \lambda_2 + k, ..., \lambda_p + k)$$

$$k_1 = k_2 = ... = k_p = k, \qquad k \geq 0$$

K is a biasing constant. K can be generalized as $K = \left(k_1, k_2, ..., k_p\right)$ so

that $KI = diag(k_1, k_2, ..., k_p)$

to yield the Generalized Ordinary Ridge estimator,

$$\hat{\beta}_{GOR} = T\hat{\alpha}_{GOR} = T(I - KA^{-1})\alpha_{GOR} \tag{6}$$

where $A = diag(\lambda_1 + k_1, \lambda_2 + k_2, ..., \lambda_p + k_p)$.

λ_i is the ith eigenvalue of $(X'X + kI)$.

This is now extended to model the Logistic Ridge regression estimator and its subsequent modification, the modified Logistic Ridge regression estimators as a special case a Generalized Linear model in canonical form.

3.0 THE LOGISTIC RIDGE REGRESSION ESTIMATOR

The Generalized Ridge regression estimator which we now state in canonical form is given as

$$\hat{\beta}_{GLS} = T\hat{\alpha}_{GLS} = T(I - KA^{-1})\alpha_{GLS} \tag{7}$$

where T is as earlier defined,

$K = (k_1, k_2, ..., k_p)$

$A = diag(\lambda_i + k_i)$

λ_i is the ith eigenvalue of $(X'WX + KI)$.

4.0 THE MODIFIED LOGISTIC RIDGE REGRESSION ESTIMATOR

To reduce the bias associated with the Generalized Logistic Ridge estimator and at the same time further reduce the variances of parameter estimates, the modified Logistic Ridge regression estimator is proposed in this work. It is given as follows:

$$\hat{\beta}_{MLS} = T(I - KA^{-1})\hat{\alpha}_{GLS} \tag{8}$$

where $A = diag(\lambda_i + k_i)$

λ_i is the ith eigenvalue of $(X'W^{\sqrt{1+\delta}}X + KI), \qquad 0 \leq \delta \leq 1$

In both estimators, $K = k_1, k_2, ..., k_p$ is a generalized biasing constant whose ith coordinate is obtained by Khalaf and Shukur (2005) as

$$k_i = \frac{t_{\max}\hat{\sigma}^2}{(n-p)\hat{\sigma}^2 + t_{\max}\hat{\alpha}_{\max}^{\ 2}} \tag{9}$$

where $t_{\max}$ is the maximum eigenvalue of the information matrix, $\hat{\sigma}^2$ is the Residual Mean Square estimator of the error variance σ^2.

This estimator, the Modified Logistic Ridge regression estimator is used to model the response probability associated with the survival of diabetic patients given their gender status, Fastng Sugar Blood (FSB) status and values of their Body Mass Indices (BMI) in a hypothetical illustration.

5.0 THE MODEL

The response probability μ_{ijk} is modeled as

$$\mu_{ijk} = \frac{\exp\left\{\beta_0 + \sum_{k=1}^{t} \beta_k X_{ijk}\right\}}{1+\exp\left\{\beta_0 + \sum_{k=1}^{t} \beta_k X_{ijk}\right\}} \tag{10}$$

Where μ_{ijk} is the probability that a diabetic patient of the ith sex status with a jth FBS status and kth BMI value has a survival time longer than or equal to ten years from the date of commencement of treatment. β_k is the fixed effect parameter with $k = 1,2,3,$ $X_{i,j}$ is the $(i,j)th$ element of the design matrix. This is an attempt to model a logistic regression model as a survival function $S(t)$.

The survival function, $S(t)$ measures the probability that the survival time, T is greater than or equal to some specified time t. This is achieved by defining the response variable as survival time status h such that $h = 1$ stands for survival time $\geq$10years and $h = 0$ stands for survival time < 10years. The following theorems are formulated and proven to support the analysis and conclusion in this study.

THEOREM 1

Let M be a $(p \times p)$ diagonal matrix with non-negative entries. Then the bias of the proposed Modified Logistic Ridge estimator $\hat{\alpha}_{MLS}$ is smaller than the bias of the ordinary Logistic Ridge estimator $\hat{\alpha}_{OLS}$.

PROOF

$$\begin{aligned} Bias(\hat{\alpha}_{ORL}) &= E(\alpha_{ORL}) - \partial_{IWLS} \\ &= E(I - KA^{-1})\hat{\alpha}_{IWLS} - \alpha_{IWLS} \\ &= (I - KA^{-1})E(\hat{\alpha}_{IWLS}) - \alpha_{IWLS} \\ &= (I - KA^{-1})\hat{\alpha}_{IWLS} - \alpha_{IWLS} \\ &= [(I - KA^{-1}) - I]\hat{\alpha}_{IWLS} \\ &= -KA^{-1}\hat{\alpha}_{IWLS} \end{aligned} \tag{11}$$

Component-wise,

$$bias(\hat{\alpha}_{ORL}) = \frac{k_i}{(\lambda_i + k_i)}|\alpha_i| \tag{12}$$

Where λ_i is the ith eigenvalue of $\left(X'WX + KI\right)$.

Similarly,

$$bias(\hat{\alpha}_{MLR}) = \frac{k_i}{(\lambda_i^{\sqrt{1+\delta}} + k_i)}|\alpha_i| \tag{13}$$

Where $\lambda_{i\sqrt{1+\delta}}$ is the ith eigenvalue of $(X'W^{\sqrt{1+\delta}}X + KI)$, $\quad 0 \leq \delta \leq 1$

From (10) and (11), it is enough to show $bias(\hat{\alpha}_{MLR}) < bias(\alpha_{OLR})$ $\quad iff \quad$ $\lambda_{i\sqrt{1+\delta}} > \lambda_i$.

The eigenvalue of a (2×2) weighted matrix $X'WX$ are given as

$$\lambda = \frac{1}{2}[[(w_1x_{11}^2 + w_2x_{21}^2) + (w_1x_{12}^2 + w_2x_{22}^2)^2] \pm \{[(w_1x_{11}^2 + w_2x_{21}^2) + (w_1x_{12}^2 + w_2x_{22}^2)^2$$

$$-4[(w_1x_{11}^2 + w_2x_{21}^2)(w_1x_{12}^2 + w_2x_{22}^2) - (w_1x_{11}x_{12} + w_2x_{21}x_{22})(w_1x_{12}x_{11} + w_2x_{22}x_{21})]\}^{1/2}]$$

These eigenvalues can be increased by increasing the diagonal elements of $X'WX$, $i.e$ by increasing $(w_1x_{11}^2 + w_2x_{21}^2)$ and $(w_1x_{12}^2 + w_2x_{22}^2)$

Since $x_{11}, x_{12}, x_{21}, x_{22} \geq 0$, increasing $(w_1 x_{11}^2 + w_2 x_{21}^2)$ *and* $(w_1 x_{12}^2 + w_2 x_{22}^2)$ implies increasing w.

This can be generalized to any $(p \times p)$ weighted matrix. But $\lambda_{i\sqrt{1+\delta}} > \lambda_i$.

Hence $bias(\hat{\alpha}_{MLR}) < bias(\alpha_{OLR})$.

Following the lines of Dorugade and Kashid (2011) and Nja (2013), in their theorems on variance reduction for their proposed estimators, we formulate the following theorem:

THEOREM 2

Let C be a $(p \times p)$ symmetric positive definite matrix. Then the proposed Modified Logistic Ridge (MLR) estimator has smaller variance than the ordinary Logistic Ridge (OLR) estimator.

PROOF

Let $V(\hat{\alpha}_{OLR})$ be variance of the Ordinary Ridge estimator and $V(\hat{\alpha}_{MLR})$ the variance of the proposed Modified Logistic Ridge estimator.

$A_{OLR} = diag(\lambda_{i(OLR)} + K_{i(OLR)})$

and λ_i is the ith eigenvalue of $(X'WX + KI)$

$A_{MLR} = diag(\lambda_{i(MLR)} + k_{i(MLR)})$

W_{OLR} = weight matrix of the Ordinary Logistic Regression estimator

$W_{MLR} = W^{\sqrt{1+\delta}}$ =enhanced weight matrix of the proposed method where $0 \leq \delta \leq 1$

We show that

$V(\hat{\alpha}_{OLR}) - V(\alpha_{MLR}) > 0$

$$V(\hat{\alpha}_{OLR}) - V(\alpha_{MLR}) = \sigma^2 W_{OLR} A_{OLR}^{-1} W'_{OLR} - \sigma^2 W_{MLR} A_{MLR}^{-1} W'_{MLR}$$
$$= \sigma^2 [(I - KA^{-1}K) A^{-1} K (I - KA^{-1}K)']_{OLR} - \sigma^2 [(I - KA^{-1}K) A^{-1} K (I - KA^{-1}K)']_{MLR}$$
$$= \sigma^2 C$$

where

$$C = [(I - KA^{-1}K) A^{-1} K (I - KA^{-1}K)']_{OLR} - \sigma^2 [(I - KA^{-1}K) A^{-1} (I - KA^{-1}K)']_{MLR}$$

$$(I - KA^{-1}K)^2_{OLR} = diag[\frac{\lambda^2_{1(OLR)}}{(\lambda_{1(OLR)} + k_1)^2}, \frac{\lambda^2_{2(OLR)}}{(\lambda_{2(OLR)} + k_2)^2}, ..., \frac{\lambda^2_{p(OLR)}}{(\lambda_{p(OLR)} + k_p)^2}]$$

$$(A^{-1}k)_{OLR} = diag[\frac{1}{\lambda_{1(OLR)} + k_1}, \frac{1}{\lambda_{2(OLR)} + k_2}, ... \frac{1}{\lambda_{p(OLR)} + k_p}]$$

$$[(I - KA^{-1})^2 A^{-1} K]_{OLR} = diag[\frac{\lambda^2_{1(OLR)}}{(\lambda^2_{1(OLR)} + k_1)^3}, \frac{\lambda^2_{2(OLR)}}{(\lambda^2_{2(OLR)} + k_2)^3}, ..., \frac{\lambda^2_{p(OLR)}}{(\lambda^2_{p(OLR)} + k_p)^3}]$$

$$[(I - KA^{-1})^2 A^{-1} K]_{MLR} = diag[\frac{\lambda^2_{1(MLR)}}{(\lambda^2_{1(MLR)} + k_1)^3}, \frac{\lambda^2_{2(MLR)}}{(\lambda^2_{2(MLR)} + k_2)^3}, ..., \frac{\lambda^2_{p(MLR)}}{(\lambda^2_{p(MLR)} + k_p)^3}]$$

$$\therefore C = [\frac{\lambda^2_{1(OLR)}}{(\lambda_{1(OLR)} + k_1)^3} - \frac{\lambda^2_{1(MLR)}}{(\lambda^2_{1(MLR)} + k_1)^3}, \quad ..., \quad \frac{\lambda^2_{p(OLR)}}{(\lambda_{p(OLR)} + k_1)^3} - \frac{\lambda^2_{p(MLR)}}{(\lambda^2_{p(MLR)} + k_p)^3}]$$

It is left to show that $\lambda_{i(MLR)} > \lambda_{i(OLR)}$.

From theorem 1, $\lambda_{i(MLR)} > \lambda_{i(OLR)}$

Hence $\frac{\lambda^2_{i(MLR)}}{(\lambda_{i(MLR)} + k_i)^3} < \frac{\lambda^2_{i(OLR)}}{(\lambda_{i(OLR)} + k_i)^3}$

$\therefore C$ is positive definite

Thus $V(\hat{\alpha}_{MLR}) < V(\alpha_{OLR})$.

6.0 ILLUSTRATIVE EXAMPLE

The table below shows the data on diabetic patients, their sex, Fasting Blood Sugar (FBS), Body Mass Index (BMI). The study population consists of people who visited a hospital as out-patients, were placed on a particular treatment and followed up. The response survival time is dichotomous. Also two of the explanatory variables, sex and FBS status are dichotomous while BMI is continuous.

Parameter estimators of the response probability are obtained using our proposed Modified Logistic Ridge regression estimators aided by MATLAB software.

The table is as follows:

Table 1: Diabetic Data

Sex	FSB	BMI	Surv. Time ≥ 10	Surv.T<10	Total
Male	< 6.5	10.1	7	4	11
Male	≥ 6.5	15.3	4	8	12
Female	< 6.5	16.6	5	4	9
Female	≥ 6.5	36	6	8	14

The following solutions were obtained for the first and second iterations.

Table 2: Parameter estimates and their corr. Variance

First Iteration

	Parameters Estimates	**Variance of Estimators**
Ordinary Ridge	$\beta_0 = 0.0645$	0.5701
	$\beta_1 = -0.6123$	1.8097
	$\beta_2 = -1.4455$	1.3725
	$\beta_3 = 0.0468$	0.0070
Proposed	$\beta_0 = 0.0618$	0.3784
	$\beta_1 = -0.6215$	1.2528
	$\beta_2 = -1.4549$	0.8995
	$\beta_3 = 0.0474$	0.0046

Second Iteration

	Parameters Estimates	**Variance of Estimators**
Ordinary Ridge	$\beta_0 = 0.0409$	0.5886
	$\beta_1 = -0.6444$	1.8707
	$\beta_2 = -1.4932$	1.4815
	$\beta_3 = 0.0502$	0.0072
Proposed	$\beta_0 = 0.0345$	0.3872
	$\beta_1 = 0.5461$	1.4966
	$\beta_2 = 1.3795$	1.1094
	$\beta_3 = 0.0445$	0.0052

Table 3: Bias of the estimators at second iteration

Ordinary	λ_1	λ_2	λ_3	λ_4 6.06×10^{-4}
	6.12×10^{-1}	1.3964	4.840×10^{-2}	
Proposed	2.39×10^{-4}	6.11×10^{-1}	1.3897	4.836×10^{-2}

RESPONSE PROBABILITIES

Looking at the table of the illustrative example (Table 1), the response probabilities are shown according to the subpopulations as follows:

1. The probability that a male with FBS < 6.5 whose BMI is 10.1 having been subjected to some medication will survived for 10 or more years is 0.7496.
2. Probability that a male with FBS $\geq$ 6.5, BMI 15.3 will survive 10 or more years is 0.4721
3. Probability that a female with FBS < 6.5, BMI 16.6 will survive 10 or more years is 0.5565
4. Probability that a female with FBS $\geq$ 6.5, BMI 36 will survive 10 or more years is 0.4236.

7.0 DISCUSSION

For the same value of the biasing constant k, the Modified Logistic Ridge estimator reduces the bias of the Ordinary Logistic Ridge estimator. This is seen from the result of the illustrative example as shown in table 3.

The bias of the Ordinary Logistic Ridge estimator component-wise for $(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$ are $6.06\times10^{-4}, 6.12\times10^{-1}, 1.3964$ *and* 4.840×10^{-2} while those of the Modified Logistic Ridge estimator (proposed) are 2.39×10^{-4}, 6.11×10^{-1}, 1.3897 *and* 4.863×10^{-2}. The variances of parameter estimates show significant difference between the two methods. The proposed method has smaller variances of parameter estimates as shown in table 2. The variances of estimates for the Ordinary Logistic Ridge estimator are 0.5886, 1.8707, 1.4815, and 0.0072 while those of the proposed method are 0.3872, 1.4966, 1.1094 and 0.0052. Two theorems, one on the comparison of bias and the other on the comparison of variances of parameter estimates between the two methods are formulated and proven in this paper to provide a theoretical basis for effective assessment of the methods.

The parameter estimates for the Ordinary Logistic Ridge estimator are $\hat{\beta}_0 = 0.0409, \beta_1 = -0.6444, \hat{\beta}_2 = -1.4932,$ *and* $\beta_3 = 0.0502$ while those of the proposed method are $\hat{\beta}_0 = 0.0345, \beta_1 = 0.5461, \hat{\beta}_2 = 1.3795,$ *and* $\beta_3 = 0.0445$. The response probabilities show that people with lower Fasting Blood Sugar have a higher probability of surviving with diabetic medication for 10 or more years irrespective of their sexes. This is also true for people with lower Body Mass Indices.

The Kaplan Meir survival function estimator lacks the ability to model survival probability as a function of both categorical and continuous explanatory variables. By a careful design of the model, the response probability has been made in this work to behave like the survival function in that it is made to measure the probability of survival time. This response probability can be used to obtain the product binomial probability.

8.0 CONCLUSION

The proposed Modified Logistic Ridge regression model is superior to the Generalized Ordinary Logistic Ridge regression model in terms of bias and variance of parameter estimates. The proposed estimator is one estimator that satisfies both properties simultaneously. By a careful design of the model, the response probability is made to behave like the survival function in that it is formulated to measure the probability of survival time. It is demonstrated that people with lower Fasting Blood Sugar have a higher probability of surviving with diabetic medication for 10 or more years irrespective of their sexes. The same goes for people with lower Body Mass Indices.

REFERENCES

1. Batah, F.S. (2011). A new Estimator By Generalized Modified Jackknife Regression Estimator: Journal of Basarah Researches (Sciences), 37(4) 138-149
2. Belsley, D.A: Edwin, K: Welsch, R.E. (1980). Regression Diagnostics: Identifying influential data and sources of collinearity. Wiley, New York
3. Carley, Kathleen M.,Natalia Y.K. (2004). A network of Optimization Approach for Improving Organizational Design: Carnegie Mellon University, School of Computer Science. Technical report. CMU-ISRI-04-102.
4. Hawkin, D.M: Yin, X (2002). A faster algorithm for ridge regression. Computational statistic and data analysis, 40, 253-262
5. Khurana, M: Chaubey, Y.P: Chandra, S. (2012). Jackknifing the Ridge Regression Estimator: A Revisit: Technical Report 12(1) 1-19
6. Madala, G.S. (1992). Introduction to Econometrics. Macmillan, New York.
7. Singh, B: Chaubey, Y.P., Divivedi, T.D. (1986). An almost Unbiased r idge estimator. Sankya, B 48 342-360.
8. Nomura M. (1988). On the Almost Unbiased Ridge Regression Estimation, Communication Statistics-Simulation, vol. 17(3), pp 729-743.
9. Gruber, M.H.J. (1998). Improving Efficiency by Shrinkage: The James-Stein and Ridge Regression Estimators, New York: Marcel Dekker.
10. Nja, M.E. (2013). A new Estimation Procedure for Generalized Linear Regression Designs with Near Dependencies. Accepted for Publication. Journal of Statictical Econometric Methods
11. Dorugade, A.V., Kashid, D.N. (2011). Parameter Estimation Method in Ridge Regression. Shivaji University, India.

This academic article was published by The International Institute for Science, Technology and Education (IISTE). The IISTE is a pioneer in the Open Access Publishing service based in the U.S. and Europe. The aim of the institute is Accelerating Global Knowledge Sharing.

More information about the publisher can be found in the IISTE's homepage: http://www.iiste.org

CALL FOR PAPERS

The IISTE is currently hosting more than 30 peer-reviewed academic journals and collaborating with academic institutions around the world. There's no deadline for submission. **Prospective authors of IISTE journals can find the submission instruction on the following page:** http://www.iiste.org/Journals/

The IISTE editorial team promises to the review and publish all the qualified submissions in a **fast** manner. All the journals articles are available online to the readers all over the world without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself. Printed version of the journals is also available upon request of readers and authors.

IISTE Knowledge Sharing Partners

EBSCO, Index Copernicus, Ulrich's Periodicals Directory, JournalTOCS, PKP Open Archives Harvester, Bielefeld Academic Search Engine, Elektronische Zeitschriftenbibliothek EZB, Open J-Gate, OCLC WorldCat, Universe Digtial Library , NewJour, Google Scholar