American Journal of Applied Mathematics and Statistics, 2015, Vol. 3, No. 1, 7-11
Available online at http://pubs.sciepub.com/ajams/3/1/2

DOI:10.12691/ajams-3-1-2

On Selection of Best Sensitive Logistic Estimator in the Presence of Collinearity

C. E. ONWUKWE*, I. A. AKI

Department of Mathematics/Statistics and Computer Science University of Calabar P. M. B. 1115, Cross River State, Nigeria
*Corresponding author: mailchrisonwukwe@gmail.com

Received September 16, 2014; Revised December 16, 2014; Accepted January 08, 2015

Abstract Collinearity is a major problem in regression modeling. It affects the prediction ability of ordinary least square estimators. Collinearity is established in logistic regression models when the difference between the least and highest eigen value of the information matrix is more in relation to the least eigen value. This results in inflated variance of estimated regression parameters. Consequently, the resulting model is not reliable and will result in incorrect conclusions about the relationship among the variables. To overcome the problem of collinearity in logistic regression model a number of estimators were proposed. This article compares the performance of four estimators - ordinary logistic estimator, logistic ridge estimator, generalized logistic ridge estimator and modified logistic ridge estimator in the presence of collinearity, to ascertain which is more effective in variance reduction. To establish superiority among the above estimators, analysis is carried out on a case study in University of Calabar Teaching Hospital, Calabar Cross River State, Nigeria. Result showed that modified logistic estimator performed better than other estimator considered due to the fact that it had the smallest variance.

Keywords: *collinearity, canonical transformation, response probability, logistic ridge estimator, logit, information matrix, link function*

Cite This Article: C. E. ONWUKWE, and I. A. AKI, "On Selection of Best Sensitive Logistic Estimator in the Presence of Collinearity." *American Journal of Applied Mathematics and Statistics*, vol. 3, no. 1 (2015): 7-11. doi: 10.12691/ajams-3-1-2.

1. Introduction

Ordinary Least Squares (OLS) estimation is widely used in regression analysis. Logistic regression has proven to be one of the most versatile techniques in generalized linear models which allows for the modeling of categorical variables. Method of least squares performs well under some basic assumption such as where error are independent and following normal distribution with mean zero and having constant variance (Jadhav and Kashid, 2011). In real life situation, some variables are seen to relate with each other thereby introducing multicollinearity in models.

Presence of multicollinearity can make ordinary least square estimator to be unstable due to large variances which lead to poor prediction (Batah et al, 2008; Batah, 2011; Joshi, 2012; Nja, 2013). To overcome this problem, several measures had been presented. Remedies include ridge regression method by Hoerl and Kennard, (1970) and the iterative principal component method Marx and Smith (1990). Since multicollinearity produces large variances in ordinary least square estimation, ridge regression attempts to find parameter estimates that have smaller variance and hence smaller MSE by enlarging the small Eigen values (Nelder and Wedderburn, 1972; Hawkin and Yin, 2002; Vago and Kemeny, 2006).

2. Ordinary Ridge Regression Estimator

Consider a multiple linear regression model.

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

Where Y is (nx1) vector of observations, β is a (px1) vector of unknown regression coefficients, X is a matrix of order (nxp) of observations on p predictor (regressor) variables $x_1, x_2, \ldots x_p$ and e is an (nx1) vector of errors with $E(e) = o$ and $var(e) = \sigma^2$.

The least square estimator of β is given by $\hat{\beta} = \left(X^1 X\right)^{-1} X^1 Y$.

The linear model can be written in canonical form as

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

where Z = XT, T is the matrix of eigen vectors of X'X

$$Z'Z = T = T'X'XT = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_p)$$

where λ_i is the i^{th} eigen value of X'X

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

The OLS estimator of α is given by

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

where

$$(\mathrm{Z'Z}) = \mathrm{J},\ \beta_{OLS} = T\alpha_{OLS} \tag{4}$$

$$\mathrm{A_K} = \mathrm{diag}\left(\lambda_1, +\mathrm{K}, \lambda_2 + \mathrm{K}, \ldots, \lambda_\mathrm{p} + \mathrm{k}\right) \tag{5}$$

where

$$\mathrm{K_1} = \mathrm{K_2} = \ldots = \mathrm{K_p}, \mathrm{K} \geq 0$$

K is a biasing constant.

K can be generalized as k = (K_1, K_2,…k_p) so that

$$\mathrm{KI} = \mathrm{diag}\left(\mathrm{k_1},\ \mathrm{K_2}, \ldots \mathrm{k_p}\right)$$

The generalize ordinary Ridge estimator is obtained as

$$\beta_{\mathrm{GOR}} = \mathrm{T}\alpha_{\mathrm{GOR}} = \mathrm{T}\left(1 - \mathrm{KA}^{-1}\right)\alpha_{\mathrm{GOR}} \tag{6}$$

where

$$\mathrm{A} = \mathrm{diag}\left(\lambda_1 + \mathrm{K_1},\ \lambda_2 + \mathrm{K_2}, \ldots,\ \lambda_\mathrm{p} + \mathrm{K_p}\right)$$

λ_i is the i^{th} eigen value of (X'X + KI)

This procedure is extended to model logistic ridge estimator and its subsequent modification, the modified logistic ridge regression estimator.

3. Ordinary Logistic Regression Estimator

The ordinary logistic estimator uses the iterative weighted least squares method. The ordinary logistic estimate of β is given by

$$\hat{\beta} = \left(X^1 WX\right)^{-1} X^1 WZ \tag{7}$$

where,

W is a diagonal matrix of weights

Z is a column matrix of adjusted dependent variables.

4. Logistic Ridge Regression Estimator

The generalized ridge regression can be expressed in canonical form as

$$\beta_{\mathrm{GLS}} = \mathrm{T}\alpha_{\mathrm{GLS}} = \mathrm{T}\left(1 - \mathrm{KA}^{-1}\right)\alpha_{\mathrm{GLS}} \tag{8}$$

$$\mathrm{K} = \left(\mathrm{K_1},\ \mathrm{K_2}, \ldots, \mathrm{K_P}\right), \mathrm{A} = \mathrm{diag}\left(\lambda_1 + \mathrm{K}\right)$$

λ_i is the i^{th} eigen value of (X'WX + KI)

The logistic Ridge regression estimator of β is given by

$$\beta = \left(\mathrm{X'WX} + \mathrm{KI}\right)^{-1} \mathrm{X'WZ} \tag{9}$$

5. Modified Logistic Ridge Regression Estimator

Modified logistic ridge regression estimator was proposed by Nja et al (2013). This is given in canonical form as follows

$$\beta_{MLS} = \mathrm{T}\left(1 - \mathrm{KA}^{-1}\right)\alpha_{GLS} \tag{10}$$

Where

$$\mathrm{A} = \mathrm{diag}\left(\lambda_\mathrm{i} + \mathrm{K_i}\right)$$

λ_i is the i^{th} eigen value of ($X'W_{\sqrt{1+\gamma}}X + KI$) $0 \leq \gamma \leq 1$.

The modified logistic ridge estimator of β is given by

$$\beta = \left(X'W_{\sqrt{1+\gamma}}X + KI\right)^{-1} X'W_{\sqrt{1+\gamma}}Z_{\sqrt{1+\gamma}} \tag{11}$$

6. Methodology

If the probability of an event taking place is P, then the odd of that event is given by:

$$\mathrm{Odd} = \frac{p}{1-p}$$

That is, odd is the probability of an event taking place divided by the probability of the event not taking place. The log of the odds is known as logit given as

$$\mathrm{logit}\ (\mathrm{P}) = \log\left(\frac{p}{1-p}\right)$$

Logistic regression like other regression has a dependent variable and independent variable(s). In logistic regression the dependent variable is a logit which is the natural log of the odds,

$$\mathrm{Log}\ (\mathrm{odds}) = \mathrm{logit}\ (\mathrm{P}) = \log\left(\frac{p}{1-p}\right)$$

Logistic regression is a modeling strategy that relates the logit to a set of explanatory variable with a linear model (Bender and Groven, 1997; Hosmer and Lemeshow, 2008; Lamote 2012). That is,

$$In\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \mathrm{X_1}$$

where:

β_0 = the constant

β_1 = the regression coefficient

X = the predictor variable

So that

$$\left(\frac{p}{1-p}\right) = e^{\beta_0 + \beta_1 X}\ (\mathrm{var}\ (\beta) = \sigma^2 \left(\mathrm{X'WX}\right)^{-1})$$

$$\sigma^2 = \frac{\sum_i^{\mu} (1-\mu_i)^2 + (0-\mu_i)^2}{N-1},\ P = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$

7. The Model

We are modeling the probability that a person selected from a subpopulation has respiratory infection given by,

$$P = \frac{\exp\left(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3\right)}{1 + \exp\left(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3\right)}$$

where,

β_0 = constant

β_1 = sex

β_2 = location
β_3 = % of exposure

The estimation of βs are as follows:

i. Ordinary logistic estimator

$$\beta = (X'WX)^{-1}X'WZ$$

W is a diagonal matrix of weights given by

$$W = m_i \mu_i (1-\mu_i), i = 1,2,3,4$$

where:
m_i is the sub population total
μ_i is the response probability
and Z is a column matrix of adjusted dependent variate given by

$$Z_i = \eta_i + \left(\frac{y_i}{m_i} - \mu_i\right)\frac{1}{\mu_i(1-\mu_i)}, i = 1,2,3,4$$

where:
η_i is the link function
y_i is number of favourable outcome
ii. Logistic Ridge Estimator

$$\beta = (X'WX+KI)^{-1}X'WZ$$

Computation for Z and W are the same as those of ordinary logistic estimator.

KI is diagonal matrix of Tikhonov constants (small positive biasing constants).
where:

$$K_1 = K_2 = K_3 = K_4$$

iii. Generalized Logistic Ridge Estimator

$$\beta = (X'WX+K1)^{-1}X'WZ$$

The computation for Z, W and K are the same as those of logistic ridge except that;

$$K_1 \neq K_2 \neq K_3 \neq K_4$$

iv. Modified Logistic Ridge Estimator

$$\beta = \left(X'W_{\sqrt{1+\gamma}}X + K1\right)^{-1} X'W_{\sqrt{1+\gamma}}Z_{\sqrt{1+\gamma}}$$

where,

$$W_{\sqrt{1+\gamma}} = m_i \mu^{\sqrt{1+\gamma}}\left(1-\mu^{\sqrt{1+\gamma}}\right)$$

$$Z_{\sqrt{1+\gamma}} = \eta_i + \left(\frac{y_i}{m_i} - \mu^{\sqrt{1+\gamma}}\right)\frac{1}{\mu^{\sqrt{1+\gamma}}\left(1-\mu^{\sqrt{1+\gamma}}\right)}$$

The variance of the parameter is given by Var(β) = $\sigma^2(X'WX)^{-1}$
where:

$$\sigma^2 = \frac{\sum_i^N e_i^2}{N-1}$$

where e_i is the error.

The estimation of parameters and calculation of variances were done with MATLAB iteratively.

8. Data Collection

The data for this research were obtained from the University of Calabar Teaching Hospital, Calabar, in Cross River State of Nigeria. This was facilitated by a well structured questionnaire that was administered to patients attending the family medicine clinic of the hospital within a period of two weeks. A total of 180 questionnaires were issued out and 169 were properly filled and returned which is presented in Table 1. Data are obtained on location of patients' residents, sex and levels of exposure. The explanatory variables are sex, location and percentage level of exposure of which the first two are dichotomous and the third is continuous. The response variable is dichotomous.

Table 1. Data on Respiratory Infection

Location	Gender	% level of exposure	Disease	No Disease	Total
Rural	Female	20	31	17	48
Rural	Male	26	20	9	29
Urban	Female	39	32	28	60
Urban	Male	42	15	17	32

9. Result of Analysis

Table 2. Parameter estimates

1st iteration				
Estimators	β_0	β_1	β_2	β_3
Ordinary logistic:	-2.4499	-3.3470	-0.7292	0.1519
Logistic ridge:	-0.0601	-1.2495	-0.2128	0.0355
Generalized logistic ridge:	-0.1551	-1.3357	-0.2332	0.0402
Modified logistic ridge:	0.1091	-1.3275	-0.2343	0.0403
2nd iteration				
Estimators	β_0	β_1	β_2	β_3
Ordinary logistic:	-2.4420	-3.3576	-0.7150	0.1521
Logistic ridge:	-0.0330	-1.2453	-0.2079	0.0350
Generalized logistic ridge:	-0.1283	-1.3311	-0.2278	0.0396
Modified logistic ridge:	-0.1034	-1.3056	-0.2262	0.0400
3rd iteration				
Estimators	β_0	β_1	β_2	β_3
Ordinary logistic:	-2.4449	-3.3600	-0.7155	0.1523
Logistic ridge:	-0.0316	-1.2442	-0.2074	0.0349
Generalized logistic ridge:	-0.1280	-1.3313	-0.2278	0.0396
Modified logistic ridge:	0.0994	-1.3120	-0.2280	0.0402

Table 3. Variances of the different estimators

Estimators	$Var(\beta_0)$	$Var(\beta_1)$	$Var(\beta_2)$	$Var(\beta_3)$
Ordinary logistic:	5.0261	3.8151	0.2438	0.0118
Logistic ridge:	1.0639	0.8090	0.0720	0.0025
Generalized logistic ridge:	1.1313	0.8644	0.0748	0.0026
Modified logistic ridge:	1.0629	0.8084	0.0719	0.0025

10. Fitted Model

The probabilities that a person selected from a sub group has respiratory infection as given by the different estimators are as follows:

1. Ordinary logistic estimator

$$P = \frac{\exp\begin{pmatrix} -2.4449 - 3.3600X_1 \\ -0.7155X_2 + 0.1523X_3 \end{pmatrix}}{1 + \exp\begin{pmatrix} -2.4449 - 3.3600X_1 \\ -0.7155X_2 + 0.1523X_3 \end{pmatrix}}$$

2. Logistic ridge estimator

$$P = \frac{\exp\begin{pmatrix} -0.0316 - 1.2442X_1 \\ -0.2074X_2 + 0.0349X_3 \end{pmatrix}}{1 + \exp\begin{pmatrix} -0.0316 - 1.2442X_1 \\ -0.2074X_2 + 0.0349X_3 \end{pmatrix}}$$

3. Generalized logistic ridge estimator

$$P = \frac{\exp\begin{pmatrix} -0.1280 - 1.3313X_1 \\ -0.2278X_2 + 0.0396X_3 \end{pmatrix}}{1 + \exp\begin{pmatrix} -0.1280 - 1.3313X_1 \\ -0.2278X_2 + 0.0396X_3 \end{pmatrix}}$$

4. Modified logistic ridge estimator

$$P = \frac{\exp\begin{pmatrix} 0.0994 - 1.3120X_1 \\ -0.2280X_2 + 0.0402X_3 \end{pmatrix}}{1 + \exp\begin{pmatrix} 0.0994 - 1.3120X_1 \\ -0.2280X_2 + 0.0402X_3 \end{pmatrix}}$$

From Table 3 (variances of the different estimators) we can see that modified logistic ridge estimator has the least variances of the parameters and hence we take the model obtained using modified logistic estimator.

The model given by modified logistic ridge estimator can be explained as follows:

1. The probability that a female living in a rural area with 20% level of exposure is 0.7116
2. The probability that a male living in a rural area with 26% level of exposure is 0.7144.
3. The probability that a female living in an urban centre with 39% level of exposure is 0.5879.
4. The probability that a male living in an urban centre with 42% level of exposure is 0.5616.

11. Discussion of Findings

Result presented in Table 2 show significant difference in the parameter estimates by the different estimators. It is observed that the estimates obtained by using ordinary logistic estimator is significantly different from those of the ridge estimators. In Table 3 it is seen that there is significant difference in the variances of the parameter estimates from the different estimators. Looking closely at the result, modified logistic ridge estimator is more sensitive and performs better than the other estimator due to its ability to reduce the variance associated with multicollinearity. The probability shows that males living in rural area with an exposure level of 39% have a higher probability of having respiratory infection.

12. Conclusion

Base on the findings of this study, it can be concluded that modified logistic ridge estimator is more superior to other estimators (ordinary logistic, logistic ridge and generalized logistic ridge) on the basis of variances of the parameter estimates. Also persons living in rural areas are seen to be more prone to having respiratory infection.

Acknowledgement

I acknowledge the efforts of Dr. M. E. Nja who has contributed immensely to the success of this work by putting me through the computational procedures involve. I also appreciate the effort of Mr. Kayode which has led to the actualization of the goal of this work.

References

[1] Batah, F. S. M. Ramanathan, T. V., Gore, S. D. (2008). The Efficiency of modified Jackknife and Ridge Type Regression Estimators: A comparison Surveys in Mathematics and its application 3 111-122.

[2] Batah, F. S. (2011). A new Estimator by generalized Modified Jackknife Regression. Estimator: Journal of Basarah Researches (Sciences), 37 (4) 138-149.

[3] Bender, R. and Grooven, U. (1997). Ordinal Logistic Regression in Medical Research. Journal of the Royal College of Physician of London. Sept/Oct 1997: v 31 (5): 546-551.

[4] Hawkins, D. M. Yin, X. (2002). A faster algorithm for ridge regression. Computational statistics and data analysis. 40, 253-262.

[5] 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.

[6] Hosmer, D. w. and Lemeshow, S. (2008). Applied Logistic Regression 2nd Edition. Wiley.

[7] Joshi, H. (2012). Multicollinearity Diagnosis in Statistical Modeling and remedies to deal with it using bars. Cytel Statistical Software Services PVT Ltd. Pone India.

[8] Judhav, N. H. & Kashid, D. N. (2011). A jackknife Ridge M. Estimator for Regression models with multicollinearity and outliers. Journal of statistical theory and practice. 5: 4, 659-673.

[9] Lamote, W. W. (2012). Multiple Logistic Regression. Boston. Boston University Press.

[10] Marx, B. D. and Smith, E. P. (1990). Principal component estimation for generalized regression. Biometrika. 77 (1): 23-31 (1990).

[11] Nelder, J., Wedderburn, R. W M. (1972). Generalized Linear Models. Journal of the Royal Statistical society, A 135, 370-384.

[12] Nja, M. E. (2013). A new Estimation procedure for Generalized Linear Regression Designs with near Dependencies. Accepted for publication. Journal of Statistical; Econometric Methods.

[13] Nja, M. E., Ogoke, U. P. & Nduka, E. C. (2013). The logistic Regression model with a modified weight function. Journal of statistical and econometric Method, Vol. 2 No. 4 2013. 161-171.

[14] Vago, E. & Kemeny, S. (2006). Logistic Ridge Regression for clinical Data Analysis (A case study). Applied Ecology and Environmental Research 4 (2) 171-179.