Estimating Shrinkage Parameter of Generalized Liu Estimator in Logistic Regression Model

Najlaa Saad Alsharabi

Department of Statistics and Informatics, University of Mosul, Mosul, Iraq
E-mail: najlaa.s.a@uomosul.edu.iq, mob: 07703880964

Abstract:

The logistic regression model is one of the modern statistical methods developed to predict the set of quantitative variables (nominal or monotonous), and it is considered as an alternative test for the simple and multiple linear regression equation as well as it is subject to the model concepts in terms of the possibility of testing the effect of the overall pattern of the group of independent variables on the dependent variable and in terms of its use for concepts of standard matching criteria, and in some cases there is a correlation between the explanatory variables which leads to contrast variation and this problem is called the problem of Multicollinearity. In this study a generalized Liu estimator was introduced to combat the multicollinearity in the logistic regression model. The generalized Liu coefficient (shrinkage coefficient) was estimated in different ways and a comparison of these methods was performed using the mean square error criterion by applying to Monte Carlo simulation data and compared of road performance. Simulation results demonstrate that shrinkage parameter selection based on the work by Kibria (2003) ie (D5) is more efficient than methods.

Key words: logistic regression, multicollinearity, mean square error, ridge estimator, liu estimator
1- Introduction

The logistic regression model is an important statistical model in analyzing binary data (0 or 1) as the primary goal of most studies is to analyze and evaluate relationships between a set of variables to obtain a formula by which we describe the model and uses the logistic regression model to describe the relationship between the response variable of the discontinuous type and the explanatory variables, prediction, estimation and control of the values of the dependent variable according to the changes in the values of the variable with interpretation (Farhood, 2014).

One of the characteristics of the binary response logistic regression is that the dependent variable \((Y)\) of the response variable follows the Bernoulli distribution taking the value \((1)\) with a probability of \((\pi)\) probability of success, and a value \((0)\) with a probability \((1-\pi)\) of failure probability (Qasim, 2011). As we work in linear regression whose independent and dependent variables take continuous values, the model that links the variables is as follows:

\[
Y = \beta_0 + \beta_1X + \varepsilon \tag{1}
\]

Whereas \((Y)\): represents a continuous observational variable and assuming that the average values of \((Y)\) observation or actual at a given value of the variable \(x\) which is \(E(Y)\) and that the variable \(\varepsilon\) represents a random error, then the model can be written as follows:

\[
E(Y|X) = \beta_0 + \beta_1X \tag{2}
\]

In regression (the other end), it is known that models have values \((-\infty, + \infty)\), but when the variable \((Y)\) is:

\[
E(Y|X) = P(Y = 1) = \pi \tag{3}
\]

Thus, the value of the right side is confined between the two numbers \((0, 1)\), and thus the model is not applicable from the regression point of view, and one of the methods of solving this problem is to enter an appropriate mathematical transformation on the dependent variable \((Y)\). Since \((0 \leq \pi \leq 1)\), then the ratio \((\pi / (1-\pi))\) is a positive amount confined between \((0, \infty)\) i.e. \((0 \leq \pi / (1-\pi) \leq \infty)\) and taking the natural logarithm For the base \((e)\) of the amount \((\pi / (1-\pi))\) the value field becomes between \((-\infty, + \infty)\) and is \((-\infty \leq \log_e (\pi / (1-\pi)) \leq \infty)\). Therefore, the regression model can be written in the case of one explanatory variable as follows:

\[
\log_e \left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1X \tag{4}
\]

But if we have more than one explanatory variable, then the model is formulated as follows:

\[
\log_e \left(\frac{\pi}{1-\pi}\right) = \beta_0 + \sum_{j=1}^{p} \beta_jX_j \tag{5}
\]
As: \( i = 1,2,3, \ldots, n \). \( \beta_1, \beta_2, \ldots, \beta_p \): Directed for features to be estimated. \( X_{ij} \): are explanatory variables.

As for \( \frac{\pi}{(1-\pi)} \) odds of success rate or preference ratio for the desired event and its mathematical formula are as follows:

\[
\frac{P(Y=1)}{1-P(Y=1)} = e^{\beta_0 + \sum_{j=1}^{p} \beta_j X_{ij}}
\]

(6)

The probability formula for the logistic regression model is written as follows:

\[
\pi = \frac{e^{X^\beta}}{1+e^{X^\beta}}
\]

(7)

And the amount \( \log(\pi / (1-\pi)) \) is called the logs odds of success logarithm.

Logistic regression does not require many assumptions. It only requires that there is no correlation between the explanatory variables and that the volume of observations is large in each group that is assumed to be greater than five times the number of parameters used in the final model (Demosthenes, 2006).

The estimation of the parameters of the logistic regression model is carried out using the Maximum Likelihood Method (ML), which is one of the most famous estimation methods in statistics. Assuming that the observations are independent, the logarithmic likelihood function is defined by the following formula: (Hosmer and Lemeshow, 2000)

\[
L = \sum_{i=1}^{n} Y_i \log(\pi_i) + (1-Y_i) \log(1-\pi_i)
\]

(8)

By maximizing the likelihood function (L) and taking the derivative with respect to the parameters (\( \beta \)) and equating the result of the equation with zero, the possibility function is given as:

\[
0 = \sum_{i=1}^{n} X_i (Y_i - \pi_i)
\]

(9)

Since equation (9) is a nonlinear parameter, some special methods should be used to obtain the appropriate solutions. Therefore, Iteratively Re-Weighted Least Squares (IRLS) can be applied to obtain appropriate solutions. The maximum likelihood estimator (MLE) of the parameters (\( \beta \)) can be found using the IRLS algorithm as follows:

\[
\hat{\beta}_{MLE} = S^{-1}X^\hat{W}\hat{Z}
\]

(10)

As \( S = \hat{X}\hat{W}X, \quad \hat{W} = \text{diag}(\hat{\pi}_i(1-\hat{\pi}_i)) \quad , \quad \hat{Z}_i = \log(\hat{\pi}_i) \)
One disadvantage of using MLE is that MSE becomes bulky when explanatory variables are linearly dependent, which is called the problem of multicollinearity. A condition number (CN) has been developed to test the existence of the problem of multicollinearity between the variables known as the following formula:

$$CN = \left( \frac{\lambda_{\text{max}}}{\lambda_{\text{min}}} \right)^{1/2}$$

(11)

As: $\lambda_{\text{max}}$, $\lambda_{\text{min}}$ They represent the largest and smallest eigenvalue roots of the matrix (S), if the value of CN <10 this means there is no problem of multicollinearity between the explanatory variables and if it is 10< CN <30 then there is a problem of moderate multicollinearity between the explanatory variables and if the value CN> 30 This means that there is a strong multicollinearity problem between the explanatory variables (Inan and Erdogan, 2013; Algamal, 2018) Also when the eigenvalue root values of the matrix (S) are close to zero, this indicates that there is a problem of multicollinearity between the variables and this will lead to an increase in the value of (MSE).

The value of the mean square error of equation (10) is found according to the following formula: [Siray et al. 2015]

$$MSE(\hat{\beta}_{ML}) = \sum_{j=1}^{p} \frac{1}{\lambda_j}$$

(12)

As: $\lambda_j$ represent the eigenvalue roots of the matrix (S).

When there is multicollinearity, the maximum likelihood estimator method (ML) suffer from inflation in the variations of the estimated parameters and the occurrence of instability, and this inflation is represented by the diagonal elements of the matrix (S). To solve this problem, (Schaefer et al., 1984) suggested a logistic ridge estimator (LRE) that was first introduced by 1970 (Horal & Kennard), and used it to estimate the parameters for the Multiple Linear Regression Model. This method is summarized by adding a small positive constant quantity (k) whose value falls between zero and one (0≤ k ≤1) to the diagonal elements of the information matrix (S) to obtain more accurate estimator, and this method works to decouple the links between the explanatory variables and the logistic character estimator is defined according to the formula next: (Månsson and Shukur, 2011; Alanaza and Algamal, 2018)

$$\hat{\beta}_{LRE} = (S + kI)^{-1} \hat{X} \hat{W} \hat{Z}$$

(13)

The estimator (ML) can be considered a special case of equation (13) when the value of (k = 0).The value of k in logistic regression models is found according to the formulas

$$k = \frac{1}{\hat{\beta}'_{ML} \hat{\beta}_{ML}}$$

(Schaefer et al., 1984).
2. Generalized Liu Estimator (GL)

The researcher Liu proposed in 1993 a new estimator to address the problem of multicollinearity. It combined the features Stein estimator in 1956 and Ordinary Ridge Regression estimator (ORR). where it is estimated:

\[ \hat{\alpha}_{RR} = \left( X''WX' + KI \right)^{-1} X''WZ \]  

Where \( X' = XP \) and \( P \) is represents a perpendicular matrix, whose columns represent distinctive vectors corresponding to the characteristic roots of the matrix of information \( (X'WX) \) and \( P'P = PP' = I \). This model called the Canonical Linear Model or Uncorrelated components model, and the estimation MLE of the \( (\alpha) \) is given:

\[ \hat{\alpha}_{MLE} = \left( \chi''W\chi' \right)^{-1} \chi''WZ \]  

It has advantages and an advantage. It is advantageous in the practical application but it is a complex function of \((K)\). (Algamala, 2018)

Akdeniz & Kaciranlar proposed in 1995 a new estimator named (GL). It is the general state of estimator (LE) there is a special advantage to estimating (LE) overcomes the estimator (ORR) where (LE) is a linear function with a bias parameter \( (d) \). So it is easier to calculate then the character parameter \( k \) for estimator (ORR). The character is also estimated as a decreasing function in \( k \) while Liu is estimating an increasing function in \( (d) \). The general Liu is indicated by:

\[ \hat{\alpha}_{GL} = \left( \Lambda + I \right)^{-1} \left( X''WZ + D\hat{\alpha}_{MLE} \right) \]

\[ \hat{\alpha}_{GL} = \left( \Lambda + I \right)^{-1} \left( \Lambda\hat{\alpha}_{MLE} + D\hat{\alpha}_{MLE} \right) \]

\[ \hat{\alpha}_{GL} = \left( \Lambda + I \right)^{-1} \left( \Lambda + D \right)\hat{\alpha}_{MLE} \]

Which can be written as follows:

\[ \hat{\alpha}_{GL} = \left( I - \left( \Lambda + I \right)^{-1} \left( \Lambda - D \right) \right)^{-1} \]  

\( D = \text{diag} \left( d_i \right), 0 < d_i < 1 \) represents a diagonal matrix with bias parameters \( (d_i) \) and \( \Lambda = X''WX'' \) (Algamala and Asar, 2018).

The forecast for the estimate \( \hat{\alpha}_{GL} \) as follows:
The estimator of (GL) is biased for parameter ($\alpha$) and the biased estimated is:

$$\text{Bias}(\hat{\alpha}_{GL}) = E(\hat{\alpha}_{GL} - \alpha) = -\left(\Lambda + I\right)^{-1}(I - D)\alpha$$  \hspace{1cm} (18)

The variance matrix of the (GL) is estimated as follows:

$$\text{Var}(\hat{\alpha}_{GL}) = (I - (\Lambda + I)^{-1}(I - D))\text{Var}(\hat{\alpha}_{MLE})(I - (\Lambda + I)^{-1}(I - D))'$$

$$= \hat{\sigma}^2(I - M)\Lambda^{-1}(I - M)'$$  \hspace{1cm} (19)

where:

$$M = (\Lambda + I)^{-1}(I - D)$$

The matrix of average error squares to (GL) estimator are as follows:

$$\text{MSE}(\hat{\alpha}_{GL}) = \text{Var}(\hat{\alpha}_{GL}) + (\text{Bias}(\hat{\alpha}_{GL}))^2$$

$$= \hat{\sigma}^2(I - M)\Lambda^{-1}(I - M)' + M\alpha\alpha'M'$$  \hspace{1cm} (20)

3. Estimating the shrinkage parameter

In order to estimate the optimal value of ($D$) in Eq.(16), several methods will be proposed. The idea behind these proposed estimators are obtained from the work of Hoerl and Kennard (1970), Kibria (2003) and Khalaf and Shukur (2005). Where several different methods of estimating the shrinkage parameter for linear ridge regression have been proposed. The first estimator which is based on the work by Hoerl and Kennard (1970) is the following:

$$D_1 = \frac{\hat{\alpha}_j^2 - 1}{1 + \hat{\alpha}_j^2}$$  \hspace{1cm} (21)

$$D_2 = \max \left(0, \frac{\hat{\alpha}_{\max}^2 - 1}{1 + \hat{\alpha}_{\max}^2} \right)$$  \hspace{1cm} (22)
Where we define $\hat{\alpha}_{\text{max}}^2$ and $\hat{\lambda}_{\text{max}}$ to be the maximum element of $\hat{\alpha}_j^2$ and $\hat{\lambda}_j$, respectively. Furthermore, the following estimators, which are based on the ideas in Akdeniz and Kaciranlar (1995), are proposed:

$$D_4 = \frac{\hat{\lambda}_j \left( \hat{\alpha}_j^2 - \hat{\sigma}^2 \right)}{\left( \hat{\lambda}_j \hat{\alpha}_j^2 + \hat{\sigma}^2 \right)}$$  \hspace{1cm} (23)$$

Akdeniz et al. proposed method in 1999 (Alheety and Kibria, 2009) as following:

$$D_5 = \left( 1 - \frac{\hat{\sigma}^2 \left( \hat{\lambda}_j + 1 \right)^2}{\hat{\lambda}_j \hat{\alpha}_j^2 + \hat{\sigma}^2} \right)$$  \hspace{1cm} (24)$$

The following estimators, which are based on the ideas in Kibria (2003), are proposed:

$$D_6 = \text{Max} \left\{ 0, \text{Median} \left( \frac{\hat{\alpha}_j^2 - 1}{\frac{1}{\hat{\lambda}_j} + \hat{\alpha}_j^2} \right) \right\}$$  \hspace{1cm} (25)$$

Using the average value and median is very common when estimating the shrinkage parameter for the ridge regression. Finally, the following estimators are proposed:

$$D_6 = \text{Max} \left\{ 0, \text{Max} \left( \frac{\hat{\alpha}_j^2 - 1}{\frac{1}{\hat{\lambda}_j} + \hat{\alpha}_j^2} \right) \right\}$$  \hspace{1cm} (26)$$

For these estimators other quintiles than the median is used which was successfully applied by Khalaf and Shukur (2005).

4. Monte Carlo simulation study

In this section, a comprehensive simulation study was conducted to evaluate the performance of the Estimating the shrinkage parameter ($D$) of Liu estimator. The explanatory variables $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{in})$ have been generated from the following formula:

$$x_{ij} = (1 - \rho^2)^{1/2} w_{ij} + \rho w_{ip} \quad \text{for} \quad i = 1, 2, \ldots, n \text{ and } j = 1, 2, \ldots, p$$  \hspace{1cm} (27)$$

where $\rho$ represents the correlation between the explanatory variables, $p$ represents the number of explanatory variables, and $w_{ij}$ are independent standard normal pseudo-random numbers and $w_{ip}$ : represents the values of the last column of the variables generated. The
response variable for \((n)\) of observations was found according to the formula of the logistic regression model:

\[
Y \approx B \left( \frac{\exp(X\beta)}{1 + \exp(X\beta)} \right)
\]

and \(\beta = \beta_1 + \beta_2 + \beta_3 + \ldots + \beta_p\) with \(\sum_{j=1}^{p} \beta_j = 1\) and \(\beta_1 = \beta_2 = \beta_3 = \ldots = \beta_p\) (Kibria, 2003; Månsson and Shukur, 2011). Because the sample size has direct impact on the prediction accuracy three representative values of the sample size are considered: 50, 75 and 150. In addition, the number of the explanatory variables are considered as \(p = 5\) and \(p = 8\). Further, because we are interested in the effect of multicollinearity, in which the degrees of correlation considered more important, three values of the pairwise correlation are considered with \(\rho = (0.90, 0.95, 0.99)\). The experiment was repeated (1000) times. And the mean square error (MSE) is calculated according to the following formula:

\[
MSE(\hat{\beta}_r) = \frac{1}{1000} \sum_{r=1}^{1000} (\hat{\beta}_r - \beta)^T (\hat{\beta}_r - \beta)
\]

where \(\hat{\beta}_r\) is the obtained liu estimator with different shrinkage parameter \(D_1, D_2, D_3, D_4, D_5\) and \(D_6\). We conclude from the results of Table (1) The lowest value for MSE when \(n = 150, \ p = 5\) and \(\rho = 0.90\), the MSE of the \(D_5\) was about 0.7747. As the correlation coefficient value increases, the MSE value increases when taking all the probabilities of the number of explanatory variables \((p)\) and the sample size \((n)\). In addition, the estimated performance \((D_5)\) is better than the rest of the estimators. The more the number of explanatory variables \((p)\) increases, the value of \((MSE)\) increases, and this increase affects the quantity of estimators. However, the estimated performance \((D_5)\) is better than the rest of the estimators. As the sample size increases, the value of MSE decreases when taking different values for each correlation coefficient and the number of explanatory variables. The best performance is performance shrinkage parameter \(D_5\) for liu estimator. The performance of the parameter \((D_5)\) of Liu estimation was the worst for having the highest values of the MSE.

Table 1: Average MSE values for different values of \(\rho, \ n\) and \(p\).

| \(n\) | \(\rho\) | \(D_1\) | \(D_2\) | \(D_3\) | \(D_4\) | \(D_5\) | \(D_6\) |
|-----|-----|-----|-----|-----|-----|-----|-----|
| 50  | 0.90 | 2.8794 | 1.7449 | 3.6189 | 3.2373 | 1.2587 | 1.6581 |
| 0.95 | 5.3993 | 3.3228 | 5.5453 | 5.7579 | 1.5086 | 2.8787 |
| 0.99 | 31.6187 | 17.5280 | 39.7088 | 26.5289 | 3.3218 | 11.8269 |
| 75  | 0.90 | 2.3207 | 1.2486 | 1.8853 | 2.7669 | 1.0973 | 1.2277 |
| 0.95 | 4.2766 | 2.0539 | 3.5611 | 5.1903 | 1.3191 | 1.9051 |
| 0.99 | 21.3497 | 11.0621 | 24.1213 | 20.3833 | 2.4319 | 7.9190 |
| 150 | 0.90 | 1.4638 | 0.7814 | 1.1045 | 1.9021 | 0.7747* | 0.7808 |
| 0.95 | 2.8528 | 1.1400 | 2.0094 | 3.9486 | 1.0342 | 1.1257 |
| 0.99 | 14.4391 | 4.8206 | 11.4733 | 18.2836 | 1.7055 | 3.8633 |
**5- Conclusion**

In this paper, a compare of different shrinkage parameter selection of the liu regression model. Simulation results demonstrate that shrinkage parameter selection based on the work by Kibria (2003) ie $(D_4, D_5, D_6, D_7)$ methods when $\rho \geq 0.90$. As the sample size increases, the value of (MSE) decreases when taking different values for each correlation coefficient and number of explanatory variables.

**References**

1- Akdeniz, F., and Kacıranlar, S.(1995)," On the almost unbiased generalized Liu estimator and unbiased estimation of the bias and MSE". Common Statist, Theory, Meth. 24:1789–1797.

2- Alanaza, Mazin M. and Algamal, Zakariya Y. Proposed methods in estimating the ridge regression parameter in Poisson regression model. EJASA, 2018. Vol. 11, Issue 02, 506-515.

3- Alheety, M.I., and Kibria, B. M. G.(2009)," On the Liu and almost unbiased Liu estimators in the presence of multicollinearity with heteroscedastic or correlated errors". Surveys in Mathematics and its Applications. 4, 155-167.

4- Algamal, Zakariya Y. (2018).’ biased estimators in poisson regression model in the presence of multicollinearity:a subject review’. QJAE. Vol. 20, Issue 01, 2312-9883.

5- Algamal, Zakariya Y. and Asar, Yasin,(2018)," Liu-type estimator for the gamma regression model". Communications in Statistics - Theory and Methods Taylor & Francis Group, LLC..

6- Demosthenes B. Panagiotakos (2006) ," A comparison between Logistic Regression and Linear Discriminant Analysis for the Prediction of Categorical Health Outcomes", International Journal of Statistical Sciences, Number 5, pp (73-83).

7- Farhood, Suhaile Hammoud Abdullah (2014), “Using Logistic Regression to Study the Factors Affecting Stock Performance (An Applied Study on the Kuwait Stock Exchange), The Public Authority for Applied Education and Training, The State of Kuwait, Statistics Department, Al-Azhar Magazine, No. 16, Pg. 47- 68.

8- Hosmer, D. D. and Lemeshow, S. (2000). Applied Logistic Regression: John Wiley and Sons.

9- Hoerl, A. E., and R. W. Kennard,.(1970)," Ridge regression: Biased estimation for nonorthogonal problems", Technometrics. 12 (1):55–67.

10- Inan, D., and Erdogan, B. E. (2013). Liu-type logistic estimator. Comm. Statist. Sim. Comp., 42(7), 1578-1586.
11- Kibria, B. M. G, (2003), "Performance of some new ridge regression estimators"). Communications in Statistics—Theory and Methods. 32:419–435.

12- Khalaf, G., and Shukur, G. (2005), "Choosing ridge parameters for regression problems". Communications in Statistics—Theory and Methods. 34:1177–1182.

13- Mansson, K. and Shukur, G. (2011). On ridge parameters in logistic regression. Comm. Statist. Theo. Meth., 40(18), 3366-3381.

14- Qasem, Bahaa Abdul-Razzaq (2011), “Analysis of the effect of some variables on the incidence of periodontal disease using the logistic regression model, Journal of Statistical Sciences, University of Basra, No. 27, pp. 139-164.

15- Schaefer, R. L., Roi, L. D. and Wolfe, R. A. (1984). A ridge logistic estimator. Comm. Statist. Theo. Meth., 13(1), 99-113.