Iraqi Journal of Statistical Sciences

www.stats.mosuljournals.com

Multicollinearity in Logistic Regression Model -Subject Review-

Najlaa S. Ibrahim , Nada N. Mohammed and Shayma. W. Mahmood

Department of Statistics and Informatics, College of Computer Science and Mathematics, University of Mosul, Mosul, Iraq

Article information

Article history:
Received January 20, 2020
Accepted February 27, 2020
Available online June 1, 2020

Keywords:
logistic regression,
multicollinearity,
mean square error,
ridge estimator,
liu estimator.

Correspondence:
Najlaa S. Ibrahim
najlaa.s.a@uomosul.edu.iq

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. This research included an article review to estimate the parameters of the logistic regression model in several biased ways to reduce the problem of multicollinearity between the variables. These methods were compared through the use of the mean square error (MSE) standard. The methods presented in the research have been applied to Monte Carlo simulation data to evaluate the performance of the methods and compare them, as well as the application to real data and the simulation results and the real application that the logistic ridge estimator is the best of other method.

DOI: 10.33899/IQJOSS.2020.165448, ©Authors, 2020, College of Computer Science and Mathematics, University of Mosul.

1- Introduction

Regression is a statistical method that specializes in studying the relationship between a dependent variable and one or several other independent variables, resulting in a mathematical equation where this relationship represents the best representation. The logistic regression model is a special case of the generalized linear model which is the most common in analyzing metadata and is a logarithmic transformation of linear regression, and it has several types, but the most common is the analysis of the binary logistic regression that we will use in our research without other types of logistic regression. it is a more powerful tool because it provides a test of the significance of parameters, and it also gives the researcher an idea of how much the independent variable affects the qualitative dependent variable dual value In addition, it sees the effect of independent variables, which allows the researcher to conclude that a variable is considered stronger than the other variable in understanding the appearance of the desired result, and that the logistic regression analysis can include qualitative independent variables The effect of the interaction between the independent variables in the two-valued dependent variable [Abbas,2012]. The researcher faces many problems, most of which are the lack of analysis hypotheses when using the method of ordinary least squares, including the problem of multicollinearity that affects the results of estimates and tests, and this problem appears as a result of an association between explanatory variables that lead to giving weak estimates that cannot be relied upon as the variations of these The capabilities are amplified and unacceptable and the (OLS) method is not able to give good estimates when there is a linear relationship between the explanatory variables.

2- Logistic Regression

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 (π) probability of success, and a value (0) with a probability (1- π) 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:

$$Y = \beta_0 + \beta_1 X + \qquad (1)$$

Since (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 e represents a random error, then the model can be written as follows:

$$E(Y|X) = \beta_0 + \beta_1 X \qquad (2)$$

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

$$E(Y|X) = P_r(Y = 1) = \pi \qquad (3)$$

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 (-∞, + ∞) 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:

$$log_e\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X \qquad (4)$$

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

$$log_e\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} \qquad (5)$$

As: i = 1,2,3, ..., n $\beta_1, \beta_2, \dots \dots \dots, \beta_p$: Directed for features to be estimated. X_{ij}: are explanatory variables.

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

$$\frac{P(Y=1)}{1-P(Y=1)} = e^{\beta_0 + \sum_{j=1}^{p} \beta_j X_{ij}} \qquad (6)$$

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

$$\pi = \frac{e^{X\beta}}{1+e^{X\beta}} \qquad (7)$$

And the amount $\text{Log}_e(\pi / (1-\pi))$ 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]

$$L = \sum_{i=1}^{n} Y_i log(\pi_i) + (1 - Y_i) log(1 - \pi_i) \qquad (8)$$

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

$$0 = \sum_{i=1}^{n} X_i(Y_i - \pi_i) \tag{9}$$

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

$$\hat{\beta}_{MLE} = S^{-1}\acute{X}\widehat{W}\hat{Z} \tag{10}$$

As $S = \acute{X}\widehat{W}X$، $\widehat{W} = diag\left(\hat{\pi}_i(1-\hat{\pi}_i)\right)$ ، $\hat{Z}_i = \log(\hat{\pi}_i)$

One disadvantage of using MLE is that MSE becomes bulky when explanatory variables are Linear 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:

$$CN = \left(\frac{\lambda_{\max}}{\lambda_{\min}}\right)^{1/2} \tag{11}$$

As: $\lambda_{\max}$, $\lambda_{\min}$ 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] 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]

$$MSE(\hat{\beta}_{ML}) = \sum_{j=1}^{p} \frac{1}{\lambda_j} \tag{12}$$

As: λ_i represent the eigenvalue roots of the matrix (S).

3- **Ridge Estimator**

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] and [Kibria et al. , 2012]

$$\hat{\beta}_{LRE} = (S + kI)^{-1}\acute{X}\widehat{W}\hat{Z} \tag{13}$$

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 one of the following common formulas: [Schaefer et al., 1984] & [Smith et al., 1991]

$$k = \frac{1}{\hat{\beta}'_{ML}\hat{\beta}_{ML}}, \quad k = \frac{p}{\hat{\beta}'_{ML}\hat{\beta}_{ML}} \quad , \quad k = \frac{p+1}{\hat{\beta}'_{ML}\hat{\beta}_{ML}} \tag{14}$$

The value of the average square error of equation (13) is found according to the following formula:

$$MSE(\hat{\beta}_{LRE}) = \sum_{j=1}^{p} \frac{\lambda_j + k^2 \alpha_j^2}{(\lambda_j + k)^2} \qquad (15)$$

As: $\alpha = \gamma\hat{\beta}_{ML}$ and γ represent the eigenvalue vectors of the matrix (S).

4- **Liu Estimator**

Liu's logistic estimator was defined by the scientist (Månsson et al., 2012) as another solution to the problem of multicollinearity, and Liu's logistic estimator denoted by symbol (LLE) was defined according to the following formula:

$$\hat{\beta}_{LLE} = (S + \mathrm{I})^{-1}(S + dI)\,\hat{\beta}_{ML} \qquad (16)$$

As: (0<d <1) is the biasing parameter, regardless of the value of (d), the value of (MSE) of the Liu logistic value (LLE) is less than the value (MSE) of maximum likelihood estimate (ML). The value of d is found according to the following formula: [Månsson et al. 2012]

$$\hat{d}_{LLE} = \max(0, \frac{\sum_{j=1}^{p}((\alpha_j{}^2 - 1)/(\lambda_j + 1)^2)}{\sum_{j=1}^{p}((\lambda_j \alpha_j{}^2 + 1)/\lambda_j(\lambda_j + 1)^2)}) \qquad (17)$$

The value of the average square error of equation (16) is found according to the following formula:

$$MSE(\hat{\beta}_{LLE}) = \sum_{j=1}^{p}\left(\frac{(\lambda_j + d)^2}{\lambda_j(\lambda_j + 1)^2} + \frac{+(d-1)^2\alpha_j{}^2}{(\lambda_j + 1)^2}\right) \qquad (18)$$

5- **Liu-Type Logistic Estimator**

The Liu-Type estimator was suggested as a substitute for the ridge regression estimator in the linear regression, which was defined by the following formula:

$$\hat{\beta}_{LLTE} = \left(\acute{X}X + k\,\mathrm{I}\right)^{-1}(\acute{X}X + dI)\,\hat{\beta}_{ols} \qquad (19)$$

As: (-∞ <d <∞), (k> 0) and $\hat{\beta}$ represent the estimated value of the parameter β in the least squares method. To take into account the problem of strong linear interrelationship, a Liu-Type logistic estimator has been proposed, which can be defined according to the following formula:

$$\hat{\beta}_{LLTE} = (S + k\,\mathrm{I})^{-1}(S + dI)\,\hat{\beta}_{ML} \qquad (20)$$

And that the value of the average square error of the above equation is found according to the following formula:

$$MSE(\hat{\beta}_{LLTE}) = \sum_{j=1}^{p}\left(\frac{(\lambda_j + d)^2}{\lambda_j(\lambda_j + k)^2} + \frac{(d-k)^2\alpha_j{}^2}{(\lambda_j + k)^2}\right) \qquad (21)$$

6- **Tow-parameter Logistic Estimator**

The Tow-parameter estimator was suggested by [Asar and Genc, 2017] as an alternative to the ridge regression estimator in a linear regression that was defined by the formula:

$$\hat{\beta}_{LTPE} = \left(\acute{X}X + k\,\mathrm{I}\right)^{-1}(\acute{X}X + kdI)\,\hat{\beta}_{ols} \qquad (22)$$

As: (0 <d <1), (k≥0) and $\hat{\beta}$ represent the estimated value of the parameter β in the least squares method and in the ridge logistic regression model, the estimator with two parameters denoted by the symbol (LTPE) is defined according to the following formula:

$$\hat{\beta}_{LTPE} = (S + k\,\mathrm{I})^{-1}(S + kdI)\,\hat{\beta}_{ML} \qquad (23)$$

We note that $\hat{\beta}_{\mathrm{LTPE}}$combines between two different estimators, which are the liu logistic estimator (LLE) and the ridge logistic estimator (LRE), if the value of (k=1) in equation (23) we get the liu logistic estimator $\hat{\beta}_{\mathrm{LLE}}$and if the value of (k=0) in equation (23) We get the maximum likelihood estimator $\hat{\beta}_{\mathrm{ML}}$and when the value of (d = 0) in equation (23) we get the ridge logistic estimator $\hat{\beta}_{\mathrm{LRE}}$. And that the value of the average square error of equation (23) is found according to the following formula:

$$MSE(\hat{\beta}_{LTPE}) = \sum_{j=1}^{p}\left(\frac{(\lambda_j + kd)^2}{\lambda_j(\lambda_j + k)^2} + \frac{k^2(d-1)^2\alpha_j{}^2}{(\lambda_j + k)^2}\right) \qquad (24)$$

7- **The practical side:**

1- **Simulation**: For the purpose of obtaining the best capabilities, Monte Carlo simulation was used to compare the above mentioned criteria by using the standard comparison of the average squares of error. The data was generated using the MATLAB program where sample sizes were chosen (n = 50,120,200), The following formula was used to generate the explanatory variables:

$$X_{ij} = (1-\rho^2)^{1/2} w_{ij} + \rho w_{ip} \qquad i = 1,2,\dots,n \;\&\; j = 1,2,\dots,p \qquad (25)$$

As: ρ represents the value of the correlation between the explanatory variables in the studied model, and values were taken ($\rho = 0.90, 0.95, 0.99$).

n: represents the number of observation.

p: represents the number of related variables and values are taken (p = 5,10).

w_{ij}: represents random numbers that follow the standard normal distribution.

w_{ip}: represents the values of the last column of the columns of the generated variables.

The response variable for (n) of observations was found according to the formula of the logistic regression model:

$$Y \approx B\left(\frac{\exp(X\ \beta)}{1+\exp(X\ \beta)}\right) \qquad (26)$$

And $\beta_1=\beta_2=\beta_3=\ldots=\beta_p$ and the feature values were determined $\sum_{j=1}^{p}\beta_j = 1$[Kibria, 2003]. The experiment was repeated (1000) times. And the mean square error (MSE) is calculated according to the following formula:

$$MSE(\hat{\beta}_r) = \frac{1}{1000}\sum_{r=1}^{1000}(\hat{\beta}_r - \beta)^T(\hat{\beta}_r - \beta) \qquad (27)$$

As: $\hat{\beta}_r$ represents $(\hat{\beta}_{LML}, \hat{\beta}_{LRE}, \hat{\beta}_{LLE}, \hat{\beta}_{LLTE}, \hat{\beta}_{LTPE})$ Respectively

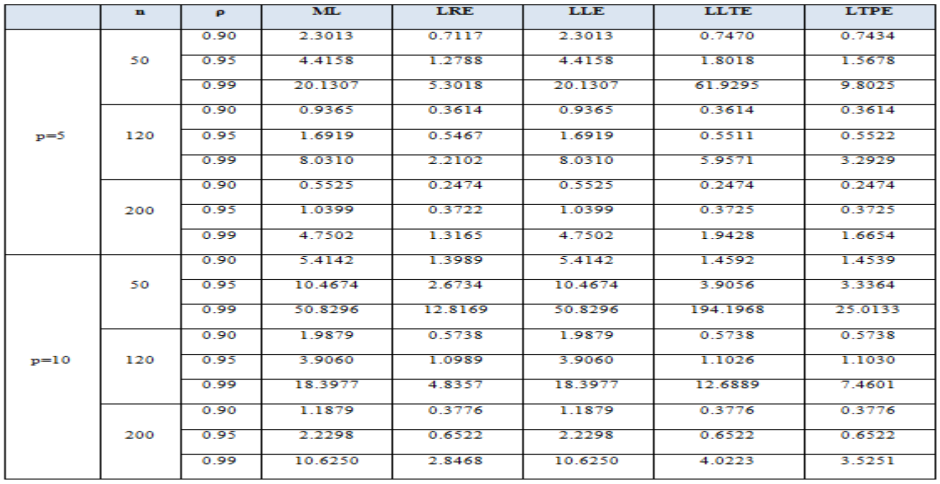
We conclude from the results of Table (1) the following three points:

1- 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 (LRE) is better than the rest of the estimators.

2- 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 (LRE) is better than the rest of the estimators.

3- As the sample size increases, the value of MSE decreases when taking different values for each correlation coefficient and the number of explanatory variables.

Table (1): shows MSE values for different values of ρ, p, n for data generated for each of the capabilities ML, LRE, LLE, LLTE, LTPE.



	n	ρ	ML	LRE	LLE	LLTE	LTPE
p=5	50	0.90	2.3013	0.7117	2.3013	0.7470	0.7434
		0.95	4.4158	1.2788	4.4158	1.8018	1.5678
		0.99	20.1307	5.3018	20.1307	61.9295	9.8025
	120	0.90	0.9365	0.3614	0.9365	0.3614	0.3614
		0.95	1.6919	0.5467	1.6919	0.5511	0.5522
		0.99	8.0310	2.2102	8.0310	5.9571	3.2929
	200	0.90	0.5525	0.2474	0.5525	0.2474	0.2474
		0.95	1.0399	0.3722	1.0399	0.3725	0.3725
		0.99	4.7502	1.3165	4.7502	1.9428	1.6654
p=10	50	0.90	5.4142	1.3989	5.4142	1.4592	1.4539
		0.95	10.4674	2.6734	10.4674	3.9056	3.3364
		0.99	50.8296	12.8169	50.8296	194.1968	25.0133
	120	0.90	1.9879	0.5738	1.9879	0.5738	0.5738
		0.95	3.9060	1.0989	3.9060	1.1026	1.1030
		0.99	18.3977	4.8357	18.3977	12.6889	7.4601
	200	0.90	1.1879	0.3776	1.1879	0.3776	0.3776
		0.95	2.2298	0.6522	2.2298	0.6522	0.6522
		0.99	10.6250	2.8468	10.6250	4.0223	3.5251

2- **Real data**: Data were taken that dealt with anemia on two levels, namely acute anemia that was symbolized (0), and chronic anemia, which was symbolized (1). The explanatory variables are the gender represented by the variable (X_1), the age represented by the variable (X_2), the hemoglobin ratio (hp) represented by the variable (X_3), the ferritin ratio in the blood represented by the variable (X_4), the retic count(They are immature red blood cells) ratio represented by the variable (X_5), the MCV ratio represented by the variable (X_6), iron deficiency in the blood represented by the variable (X_7), the rate of transferrin in the blood represented by the variable (X_8), the cause of poverty is hemorrhage represented by the variable (X_9) anemia, chronic diseases represented by variable (X_{10}), and anemia is a decrease in blood cells Red represented by the variable (X_{11}). After conducting the initial data analysis in the Minitab program, he found the following:

1- The number of people with severe anemia is (67) patients with a percentage of 47.9%, while those with chronic anemia are (73) patients with a rate of 52.1% as shown in Table (2).

Table (2): Shows the number of patients with anemia.

Types of anemia	Number of people with types of the disease	The proportion of injured
Severe anemia	67	47.9
Chronic anemia	73	52.1
Total	140	100.0

2- As for the number of males and females in the sample, they were as in Table (3) as follows:

Table (3): Shows the number of males and females in the sample.

	Male and female number	Male and female ratio
Male	83	59.3
Female	57	40.7
Total	140	100.0

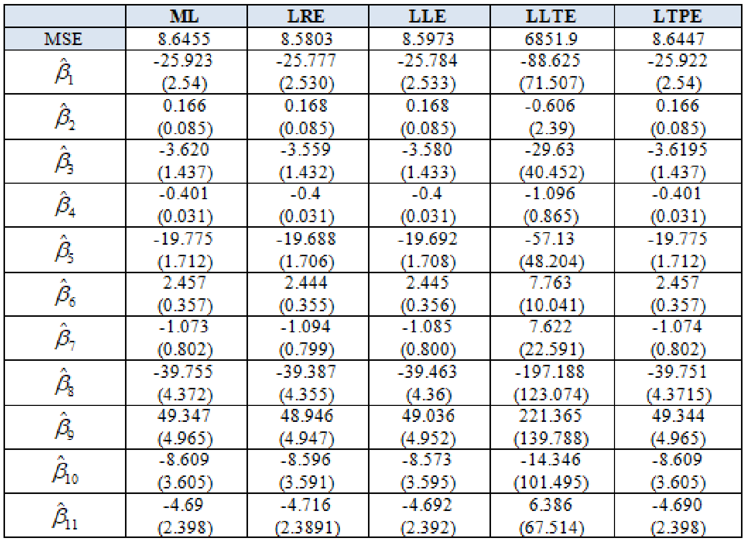
To test the existence of the problem of linear relationship between the data, the eigenvalue roots of the matrix (S) were found and the values of the roots were as shown in Table (4), as we note that the value of CN = 726.9358 is greater than (30) and this is evidence of the existence of a problem of multicollinearity between the explanatory variables.

Table (4): shows the values of the eigenvalue roots of the matrix (S).

λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8	λ_9	λ_{10}	λ_{11}
143077	33144.1	1762.75	31.27	13.65	7.12	2.35	1.8	0.93	0.27	0.38

The following table shows the estimated binary logistic regression parameters, standard error, and MSE values for each of the ML, LRE, LLE, LLTE, and LTPE estimators. We note that the best estimation is LRE having the lowest value for MSE.

Table (5): shows the estimated parameters, standard error, and MSE values for ML, LRE, LLE, LLTE, LTPE.



	ML	LRE	LLE	LLTE	LTPE
MSE	8.6455	8.5803	8.5973	6851.9	8.6447
$\hat{\beta}_1$	-25.923 (2.54)	-25.777 (2.530)	-25.784 (2.533)	-88.625 (71.507)	-25.922 (2.54)
$\hat{\beta}_2$	0.166 (0.085)	0.168 (0.085)	0.168 (0.085)	-0.606 (2.39)	0.166 (0.085)
$\hat{\beta}_3$	-3.620 (1.437)	-3.559 (1.432)	-3.580 (1.433)	-29.63 (40.452)	-3.6195 (1.437)
$\hat{\beta}_4$	-0.401 (0.031)	-0.4 (0.031)	-0.4 (0.031)	-1.096 (0.865)	-0.401 (0.031)
$\hat{\beta}_5$	-19.775 (1.712)	-19.688 (1.706)	-19.692 (1.708)	-57.13 (48.204)	-19.775 (1.712)
$\hat{\beta}_6$	2.457 (0.357)	2.444 (0.355)	2.445 (0.356)	7.763 (10.041)	2.457 (0.357)
$\hat{\beta}_7$	-1.073 (0.802)	-1.094 (0.799)	-1.085 (0.800)	7.622 (22.591)	-1.074 (0.802)
$\hat{\beta}_8$	-39.755 (4.372)	-39.387 (4.355)	-39.463 (4.36)	-197.188 (123.074)	-39.751 (4.3715)
$\hat{\beta}_9$	49.347 (4.965)	48.946 (4.947)	49.036 (4.952)	221.365 (139.788)	49.344 (4.965)
$\hat{\beta}_{10}$	-8.609 (3.605)	-8.596 (3.591)	-8.573 (3.595)	-14.346 (101.495)	-8.609 (3.605)
$\hat{\beta}_{11}$	-4.69 (2.398)	-4.716 (2.3891)	-4.692 (2.392)	6.386 (67.514)	-4.690 (2.398)

8- Conclusions:

1- The simulation results showed that the best way to address the problem of multicollinearity is the ridge logistic regression method.

2- The higher the correlation coefficient value, the greater the MSE value.

3- The more the number of explanatory variables (p) increases, the value of (MSE) increases, and that this increase affects the amount of estimators, however the estimated performance (LRE) is better than the rest of the estimators.

4- As the sample size increases, the value of (MSE) decreases when taking different values for each correlation coefficient and the number of explanatory variables.

5- The results of the application on the data showed the fact that the ridge logistic regression method is the best method presented in the search because it has the lowest value for the average squares of the error, and that the value of the standard error for the estimated parameters was almost close to all methods.

Reference:

1- Abbas, Ali Khudair (2012), "Using the Logistic Regression Model in Predicting Functions of Qualitative Economic Variables", Kirkuk Journal of Administrative and Economic Sciences, Volume 2, No. 2, pp. 234-253.

2- Farhood, Suhaila 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.

3- 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.

4- Asar, Y., Genc, A. (2015)." A New Two-Parameter Ridge Estimator in Binary Logistic Regression",Communications in Statistics - Simulation and Computation, ISSN: 0361-0918 (Print) 1532-4141

5- 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 $^{(73-84)}$.

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

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

8- Inan, D., and Erdogan, B. E. (2013). Liu-type logistic estimator. Comm. Statist. Sim. Comp., 42(7), 1578-1586.

9- Kibria, B. M. G. (2003). Performance of some new ridge regression estimators. Communications in Statistics—Theory and Methods 32:419–435.

10- Kibria, B. M. G., Mansson, K. and Shukur, G. (2012). Performance of some logistic ridge regression estimators. Comp. Econ., 40(4), 401-414.

11- Mansson, K., Kibria, B. and Shukur, G. (2012). On Liu estimators for the logit regression model. Econ. Model., 29(4), 1483-1488.

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

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

14- Smith, K. R., Slattery, M. L., French, T. K. (1991). Collinear nutrients and the risk of colon cancer. Journal of Clinical Epidemiolgy 44:715–723.

15- Siray, G.U., Toker, S., Ka¸cıranlar, S. (2015). On the Restricted Liu Estimator in the Logistic Regression Model. Comm. Statist. Sim. Comp. 44:217-232.

تعدد العلاقة الخطية بنموذج الانحدار اللوجستي – مراجعة مقال–

نجلاء سعيد إبراهيم ، ندى نزار محمد & شيماء وليد محمود

قسم الاحصاء والمعلوماتية/ كلية علوم الحاسوب والرياضيات/ جامعة الموصل/ الموصل/ العراق

الخلاصة: يعتبر نموذج الانحدار اللوجستي من الطرق الاحصائية الحديثة الموضوعة للتنبؤ بمجموعة المتغيرات الكمية (اسمية او رتيبة)، ويعد كاختبار بديل عن معادلة الانحدار الخطي البسيط والمتعدد وكذلك فهو يخضع الى مفاهيم النموذج من حيث امكانية اختبار اثر النسق الكلي لمجموعة المتغيرات المستقلة على المتغير التابع ومن حيث توظيفها لمفاهيم معايير المطابقة النمذجية، وفي بعض الحالات يكون هناك ارتباط بين المتغيرات التوضيحية مما يؤدي تضخم التباين وهذه المشكلة تدعى بمشكلة تعدد العلاقة الخطية. تناول هذا البحث دراسة مراجعة مقال لتقدير معلمات نموذج الانحدار اللوجستي بعدة طرق متحيزة لتقليل من مشكلة تعدد العلاقة الخطية بين المتغيرات وتمت المقارنة بين هذه الطرق من خلال استخدام معيار متوسط مربع الخطأ (MSE). ولقد تم تطبيق الطرق المقدمة في البحث على بيانات محاكاة مونت كارلو لتقييم اداء الطرق والمقارنة بينهم وكذلك التطبيق على بيانات حقيقة وكانت نتائج المحاكاة والتطبيق الحقيقي بان مقدر انحدار الحرف اللوجستي هو افضل الطرق الاخرى.

الكلمات المفتاحية: الانحدار اللوجستي – التعدد الخطي – متوسط مربع الاخطاء– مقدر الحرف –مقدر ليو.