REVSTAT – Statistical Journal
Volume 18, Number 3, July 2020, 325–336

EFFICIENCY OF THE PRINCIPAL COMPONENT LIU-TYPE ESTIMATOR IN LOGISTIC REGRESSION

Authors: Jibo Wu
– School of Mathematics and Finance, Chongqing University of Arts and Sciences, Chongqing, China
linfen52@126.com

Yasin Asar
– Department of Mathematics–Computer Sciences, Necmettin Erbakan University, Konya, 42090, Turkey
yasar@erbakan.edu.tr , yasinasar@hotmail.com

Received: September 2017 Revised: February 2018 Accepted: February 2018

Abstract:

- In this paper we propose a principal component Liu-type logistic estimator by combining the principal component logistic regression estimator and Liu-type logistic estimator to overcome the multicollinearity problem. The superiority of the new estimator over some related estimators are studied under the asymptotic mean squared error matrix. A Monte Carlo simulation experiment is designed to compare the performances of the estimators using mean squared error criterion. Finally, a conclusion section is presented.

Key-Words:

- *Liu-type estimator; logistic regression; mean squared error matrix; maximum likelihood estimator; multicollinearity.*

AMS Subject Classification:

- 62J07, 62J12.

1. INTRODUCTION

Consider the following binary logistic regression model

$$\pi_i = \frac{\exp(x_i'\beta)}{1+\exp(x_i'\beta)}\,, \qquad i=1,...,n\,, \tag{1.1}$$

where $x_i' = (1\ x_{i1} \cdots x_{iq})$ denotes the i-th row of X which is an $n \times p$ $(p=q+1)$ data matrix with q known covariate vectors, y_i shows the response variable which takes on the value either 0 or 1 with $y_i \sim \text{Bernoulli}(\pi_i)$, y_i's are supposed to be independent of one another and $\beta' = (\beta_0\ \beta_1 \cdots \beta_q)$ stands for a $p \times 1$ vector of parameters.

Usually the maximum likelihood (ML) method is used to estimate β. The corresponding log-likelihood equation of model (1.1) is given by

$$L = \sum_{i=1}^{n} y_i \log(\pi_i) + (1-y_i)\log(1-\pi_i)\,, \tag{1.2}$$

where π_i is the i-th element of the vector π, $i=1,2,...,n$.

ML estimator can be obtained by maximizing the log-likelihood equation given in (1.2). Since Equation (1.2) is non-linear in β, one should use an iterative algorithm called iteratively re-weighted least squares algorithm (IRLS) as follows (Saleh and Kibria, [18]):

$$\hat\beta^{t+1} = \hat\beta^{t} + \left(X'V^tX\right)^{-1} X'V^t\left(y-\hat\pi^t\right), \tag{1.3}$$

where π^t is the estimated values of π using $\hat\beta_t$ and $V^t = \text{diag}\big(\hat\pi_i^t(1-\hat\pi_i^t)\big)$ such that $\hat\pi_i^t$ is the i-th element of $\hat\pi_t$. After some algebra, Equation (1.3) can be written as follows:

$$\hat\beta_{\text{ML}} = \left(X'VX\right)^{-1} X'Vz\,, \tag{1.4}$$

where $z' = (z_1 \cdots z_n)$ with $\eta_i = x_i'\beta$ and $z_i = \eta_i + (y_i - \pi_i)\,(\partial\eta_i/\partial\pi_i)$.

In linear regression analysis, multicollinearity has been regarded as a problem in the estimation. In dealing with this problem, many ways have been introduced to deal with this problem. One approach is to study the biased estimators such as ridge estimator (Hoerl and Kennard, [7]), Liu estimator (Liu, [14]), Liu-type estimator (Huang *et al.*, [8]), modified Liu-type estimator (Alheety and Kibria, [2]) and improved ridge estimators (Yüzbaşı *et al.*, [21]). Alternatively, many authors such as Xu and Yang ([20]) and Li and Yang ([13]), have studied the estimation of linear models with additional restrictions.

As in linear regression, estimation in logistic regression is also sensitive to multicollinearity. When there is multicollinearity, columns of the matrix $X'VX$ become close to be dependent. It implies that some of the eigenvalues of $X'VX$ become close to zero. Thus, mean squared error value of MLE is inflated so that one cannot obtain stable estimations. Thus many authors have studied how to reduce the multicollinearity, such as Lesaffre and Max ([12]) discussed the multicollinearity in logistic regression, Schaefer *et al.* ([19]) proposed the ridge logistic (RL) estimator, Aguilera *et al.* ([1]) proposed the principal component logistic regression (PCLR) estimator, Månsson *et al.* ([15]) introduced the Liu logistic (LL) estimator, by combining the principal component logistic regression estimator and ridge logistic estimator to deal with multicollinearity. Moreover, Inan and Erdoğan ([9]) proposed Liu-type logistic estimator (LTL) and Asar ([3]) studied some properties of LTL.

In this study, by combining the principal component logistic regression estimator and the Liu-type logistic estimator, the principal component Liu-type logistic estimator is introduced as an alternative to the PCLR, ML and LTL to deal with the multicollinearity.

The rest of the paper is organized as follows. In Section 2, the new estimator is proposed. Some properties of the new estimator are presented in Section 3. A Monte Carlo simulation is given in Section 4 and some concluding remarks are given in Section 5.

2. THE NEW ESTIMATOR

The logistic regression model is expressed by Aguilera *et al.* ([1]) in matrix form in terms of the logit transformation as $L = X\beta = XTT'\beta = Z\alpha$ where $T = [t_1, ..., t_p]$ shows an orthogonal matrix with $Z'VZ = T'X'VXT = \Lambda$ and $\Lambda = \mathrm{diag}(\lambda_1, ..., \lambda_p)$, $\lambda_1 \geq \cdots \geq \lambda_p$, is the ordered eigenvalues of $X'VX$. Then T and Λ may be written as $T = (T_r\ T_{p-r})$ and $\begin{bmatrix} \Lambda_r & O \\ O & \Lambda_{p-r} \end{bmatrix}$ where $Z_r'VZ_r = T_r'X'VXT_r = \Lambda_r$ and $Z_{p-r}'VZ_{p-r} = T_{p-r}'X'VXT_{p-r} = \Lambda_{p-r}$. The Z matrix and the α vector can be partitioned as $Z = (Z_r\ Z_{p-r})$ and $\alpha = (\alpha_r'\ \alpha_{p-r}')'$. The handling of multicollinearity by means of PCLR corresponds to the transition from the model $L = X\beta = XT_rT_r'\beta + XT_{p-r}T_{p-r}'\beta = Z_r\alpha_r + Z_{p-r}\alpha_{p-r}$ to the reduced model $L = Z_r\alpha_r$. Then by Equation (1.1) and PCLR method we get the PCLR estimator.

Inan and Erdoğan ([9]) proposed Liu-type logistic estimator (LTL) as

$$\hat{\beta}(k,d) = \left(X'VX + kI\right)^{-1}\left(X'Vz - d\hat{\beta}_{\mathrm{ML}}\right), \tag{2.1}$$

where $-\infty < d < \infty$ and $k > 0$ are biasing parameters.

The principal component logistic regression estimator (Aguilera *et al.*, [1]) is defined as

$$\hat{\beta}_r = T_r\left(T_r'X'VXT_r\right)^{-1}T_r'X'Vz\,. \tag{2.2}$$

We can write (2.2) as follows:

$$\hat{\beta}_r = T_r\left(T_r'X'VXT_r\right)^{-1}T_r'X'Vz = T_rT_r'\hat{\beta}_{\mathrm{ML}}\,. \tag{2.3}$$

Then we can introduce a new estimator by replacing $\hat{\beta}^*(k,d)$ with $\hat{\beta}_{\mathrm{ML}}$ in (2.3), and we get

$$\begin{aligned} \hat{\beta}_r(k,d) &= T_rT_r'\hat{\beta}(k,d) \\ &= T_r\left(T_r'X'VXT_r + kI_r\right)^{-1}\left(T_r'X'VXT_r - dI_r\right)\left(T_r'X'VXT_r\right)^{-1}T_r'X'Vz\,, \end{aligned} \tag{2.4}$$

where $-\infty < d < \infty$ and $k > 0$ are biasing parameters. We call this estimator as the principal component Liu-type logistic regression (PCLTL) estimator.

Remark 2.1. It is obvious that

$$\hat{\beta}_r(k,d) = T_r\big(T_r'X'VXT_r + k\,I_r\big)^{-1}\big(T_r'X'VXT_r - dI_r\big)\,T_r'\hat{\beta}_r\,.$$

Thus we can see the PCLTL estimator as a linear combination of the PCLR estimator.

Remark 2.2. It is easy to obtain the followings:

(**a**) $\hat{\beta}_r(0,0) = \hat{\beta}_r = T_r\big(T_r'X'VXT_r\big)^{-1}T_r'X'Vz$, PCLR estimator;

(**b**) $\hat{\beta}_p(0,0) = \hat{\beta}_{\mathrm{ML}} = \big(X'VX\big)^{-1}X'Vz$, ML estimator;

(**c**) $\hat{\beta}_p(k,d) = \hat{\beta}(k,d) = \big(X'VX + kI\big)^{-1}\big(X'Vz - d\hat{\beta}_{\mathrm{ML}}\big)$, LTL estimator.

Thus, the new estimator in (2.4) includes the PCLR, ML and LTL estimators as its special cases.

In the next section, we will study the properties of the new estimator.

3. THE PROPERTIES OF NEW ESTIMATOR

For the sake of convenience, we present some lemmas which are needed in the following discussions.

Lemma 3.1 (Farebrother, [6]; Rao and Tountenburg, [17])**.** *Suppose that M is a positive definite matrix, namely $M > 0$, α is some vector, then $M - \alpha\alpha' \geq 0$ if and only if $\alpha' M^{-1}\alpha \leq 1$.*

Lemma 3.2 (Baksalary and Trenkler, [5])**.** *Let $C_{n\times p}$ be the set of complex matrices and $H_{n\times n}$ be the Hermitian matrices. Further, given $L \in C_{n\times p}$, L^*, $R(L)$ and $\kappa(L)$ denote the conjugate transpose, the range and the set of all generalized inverses, respectively of L. Let $A \in H_{n\times n}$, $a_1 \in C_{n\times 1}$ and $a_2 \in C_{n\times 1}$ be linearly independent, $f_{ij} = a_i'A^- a_j$, $i,j = 1,2$ and $A \in \kappa(L)$, $a_1 \notin R(A)$. Let*

$$s = \Big[a_1'\big(I - AA^-\big)'\big(I - AA^-\big)\,a_2\Big]\Big/\Big[a_1'\big(I - AA^-\big)'\big(I - AA^-\big)\,a_1\Big]\,.$$

Then $A + a_1a_1' - a_2a_2' \geq 0$ if and only if one of the following sets of conditions holds:

(**a**) $A \geq 0,\ a_i \in R(A),\ i = 1,2,\ (f_{11}+1)(f_{22}-1) \leq |f_{12}|^2,$

(**b**) $A \geq 0,\ a_1 \notin R(A),\ a_2 \in R(A\!:\!a_1),\ (a_2 - sa_1)'A^-(a_2 - sa_1) \leq 1 - |s|^2,$

(**c**) $A = U\Delta U' - \lambda vv',\ a_i \in R(A),\ i = 1,2,\ v'a_1 \neq 0,\ f_{11}+1 \leq 0,\ f_{22}-1 \leq 0,$ $(f_{11}+1)(f_{22}-1) \geq |f_{12}|^2,$

where $(U\!:\!v)$ shows a sub-unitary matrix, λ is a positive scalar and Δ is a positive definite diagonal matrix. Further, the conditions (a), (b) and (c) denote all independent of the choice of A^-, A^- stands for the generalized inverse of A.

To compare the estimators, we use the mean squared error matrix (MSEM) criterion which is defined for an estimator $\check{\beta}$ as follows:

$$\mathrm{MSEM}\big(\check{\beta}\big) \;=\; \mathrm{Cov}\big(\check{\beta}\big) + \mathrm{Bias}\big(\check{\beta}\big)\;\mathrm{Bias}\big(\check{\beta}\big)' ,$$

where $\mathrm{Cov}\big(\check{\beta}\big)$ is the covariance matrix of $\check{\beta}$, and $\mathrm{Bias}\big(\check{\beta}\big)$ is the bias vector of $\check{\beta}$. Moreover, scalar mean squared error (SMSE) of an estimator $\check{\beta}$ is also given as

$$\mathrm{SMSE}\big(\check{\beta}\big) \;=\; \mathrm{tr}\big\{\mathrm{MSEM}\big(\check{\beta}\big)\big\} \,.$$

3.1. Comparison of the new estimator (PCLTL) to the ML estimator

From (2.4), we can compute the asymptotic variance of the new estimator as follows:

$$\mathrm{Cov}\big(\hat{\beta}_r(k,d)\big) \;=\; T_r\, S_r(k)^{-1}\, \Lambda_r^{-1} S_r(d)\; \Lambda_r\, S_r(d)\; \Lambda_r^{-1} S_r(k)^{-1}\, T_r' \,, \tag{3.1}$$

where $S_r(k) = \Lambda_r + k\, I_r,\;\; S_r(d) = \Lambda_r - d\, I_r$.

Using (2.4), we get:

$$E\big(\hat{\beta}_r(k,d)\big) \;=\; T_r\, S_r(k)^{-1}\, \Lambda_r^{-1} S_r(d)\; \Lambda_r T_r' \beta \,. \tag{3.2}$$

By

$$T_r\, S_r(k)^{-1}\, \Lambda_r T_r' - I_p \;=\; -\Big(T_{p-r}\, T_{p-r}' + k\, T_r\, S_r(k)^{-1}\, T_r'\Big) \,, \tag{3.3}$$

then we get the asymptotic bias of the new estimator as follows:

$$\mathrm{Bias}\big(\hat{\beta}_r(k,d)\big) \;=\; \Big(-T_{p-r}\, T_{p-r}' - (d+k)\, T_r\, S_r(k)^{-1}\, T_r'\Big)\, \beta \,.$$

Now, we can get the asymptotic mean squared error matrix of the new estimator as follows:

$$\begin{aligned} \mathrm{MSEM}\big(\hat{\beta}_r(k,d)\big) \;=\;& T_r\, S_r(k)^{-1}\, \Lambda_r^{-1} S_r(d)\; \Lambda_r\, S_r(d)\; \Lambda_r^{-1} S_r(k)^{-1}\, T_r' \\ &+ \Big(-T_{p-r}\, T_{p-r}' - (d+k)\, T_r\, S_r(k)^{-1}\, T_r'\Big)\, \beta \\ &\times\, \beta'\Big(-T_{p-r}\, T_{p-r}' - (d+k)\, T_r\, S_r(k)^{-1}\, T_r'\Big) \,. \end{aligned} \tag{3.4}$$

Theorem 3.1. *Assume that $d<k$ and $d+k>0$ then the new estimator is superior to the ML estimator under the asymptotic mean squared error matrix criterion if and only if*

$$\beta' T_r (k+d)^2 \Big[2(k+d)\, I_r + (k^2-d^2)\, \Lambda_r^{-1}\Big]^{-1} T_r' \beta \;+\; \beta' T_{p-r}\, \Lambda_{p-r}\, T_{p-r}' \beta \;\le\; 1 \,.$$

Proof: The asymptotic mean squared error matrix of MLE is given by

$$\mathrm{MSEM}\big(\hat{\beta}\big) = \big(X' V X\big)^{-1}. \tag{3.5}$$

By $\Lambda = \begin{pmatrix} \Lambda_r & O \\ O & \Lambda_{p-r} \end{pmatrix}$ and $T = (T_r, T_{p-r})$, we may obtain

$$(X'VX)^{-1} = T\Lambda^{-1}T' = T_r\Lambda_r^{-1}T_r' + T_{p-r}\Lambda_{p-r}^{-1}T_{p-r}' .$$

Let us consider the difference $\Delta_1 = \text{MSEM}(\hat{\beta}) - \text{MSEM}(\hat{\beta}_r(k,d))$ such that

$$\begin{aligned}\Delta_1 &= T_r S_r(k)^{-1}\Big[2(k+d)I_r + (k^2-d^2)\Lambda_r^{-1}\Big]S_r(k)^{-1}T_r' \\ &\quad + T_{p-r}\Big[\Lambda_{p-r} - T_{p-r}'\beta\beta' T_{p-r}\Big]T_{p-r}' - (k+d)^2 T_r S_r(k)^{-1} \\ &\quad \times T_r'\beta\beta' S_r(k)^{-1}T_r' + (k+d)T_r S_r(k)^{-1}T_r'\beta\beta' T_{p-r}T_{p-r}' \\ &\quad + (k+d)T_{p-r}T_{p-r}'\beta\beta' T_r S_r(k)^{-1}T_r' .\end{aligned} \tag{3.6}$$

Let

$$S^* = \begin{pmatrix} \frac{S_r(k)}{k+d} & 0 \\ 0 & \Lambda_{p-r} \end{pmatrix}$$

and

$$(\Lambda^*)^{-1} = \begin{pmatrix} \frac{2(k+d)I_r + (k^2-d^2)\Lambda_r^{-1}}{(k+d)^2} & 0 \\ 0 & \Lambda_{p-r} \end{pmatrix} . \tag{3.7}$$

Now we can write (3.6) as

$$\Delta_1 = T(S^*)^{-1}\Big[(\Lambda^*)^{-1} - T'\beta\beta' T\Big](S^*)^{-1}T' . \tag{3.8}$$

Thus Δ_1 is a nonnegative definite matrix if and only if $(\Lambda^*)^{-1} - T'\beta\beta' T$ is a nonnegative definite matrix. Using Lemma 3.1, $(\Lambda^*)^{-1} - T'\beta\beta' T$ is a nonnegative definite matrix if and only if $\beta' T\Lambda^* T'\beta \leq 1$. Invoking the notation of Λ^* in (3.7), we can prove Theorem 3.1. □

3.2. Comparison of the new estimator (PCLTL) to the PCLR estimator

Theorem 3.2. *Suppose that $d < k$ and $d + k > 0$ then the new estimator is better than the PCLR estimator under the asymptotic mean squared error matrix criterion if and only if $T_r'\beta = 0$.*

Proof: Suppose that $k = d$ in Equation (3.4), then we get

$$\text{MSEM}(\hat{\beta}_r) = T_r\Lambda_r^{-1}T_r' + (T_rT_r' - I_p)\beta\beta'(T_rT_r' - I_p) . \tag{3.9}$$

Now let us consider the difference $\Delta_2 = \text{MSEM}(\hat{\beta}_r) - \text{MSEM}(\hat{\beta}_r(k,d))$ such that

$$\begin{aligned}\Delta_2 &= T_r S_r(k)^{-1}\Big[2(k+d)I_r + (k^2-d^2)\Lambda_r^{-1}\Big]S_r(k)^{-1}T_r' \\ &\quad + (T_rT_r' - I_p)\beta\beta'(T_rT_r' - I_p) \\ &\quad + \Big(-T_{p-r}T_{p-r}' - (d+k)T_rS_r(k)^{-1}T_r'\Big)\beta \\ &\quad \times \beta'\Big(-T_{p-r}T_{p-r}' - (d+k)T_rS_r(k)^{-1}T_r'\Big) .\end{aligned} \tag{3.10}$$

To apply Lemma 3.2, let $A = T_r B T_r'$, where

$$B \,=\, S_r(k)^{-1}\Big[2(k+d)\,I_r + (k^2-d^2)\,\Lambda_r^{-1}\Big]\,S_r(k)^{-1}$$

and $a_1 = \big(T_r T_r' - I_p\big)\,\beta$, $a_2 = \big(-T_{p-r}T_{p-r}' - (d+k)\,T_r\,S_r(k)^{-1}\,T_r'\big)\,\beta$.

When $d < k$ and $d+k > 0$, B is a positive definite matrix. Then we get the Moore–Penrose inverses of A which is $A^+ = T_r B^{-1} T_r'$, and $AA^+ = T_r T_r'$. Thus $a_1 \in R(A)$ if and only if $a_1 = 0$. Since $a_1 \neq 0$, we cannot use part (a) and (c) of Lemma 3.2, we can only apply part (b) of Lemma 3.2. Using the definition of s, we may obtain that $s = 1$. On the other hand, $a_2 - a_1 = A\eta$, where

$$\eta \,=\, (d+k)\,T_r\,S_r(k)\Big[2(k+d)\,I_r + (k^2-d^2)\,\Lambda_r^{-1}\Big]^{-1} T_r'\beta\,.$$

Thus, we can easily obtain $a_2 \in R(A\!:\!a_1)$. Then Using Lemma 3.2, we can get that the new estimator is superior to the PCLR estimator under the asymptotic mean squared error matrix criterion if and only if $(a_2-a_1)\,A^-(a_2-a_1) \leq 0$ or $\eta' A\eta \leq 0$. In fact, $(a_2-a_1)\,A^-(a_2-a_1) \geq 0$, so the new estimator is better than the PCLR estimator under the asymptotic mean squared error matrix criterion if and only if $\eta' A\eta = 0$, that is

$$\beta'\,T_r\Big[2(k+d)\,I_r + (k^2-d^2)\,\Lambda_r^{-1}\Big]^{-1} T_r'\beta \,=\, 0$$

and $\beta'\,T_r\big[2(k+d)\,I_r + (k^2-d^2)\,\Lambda_r^{-1}\big]^{-1} T_r'\beta = 0$ if and only if $T_r'\beta = 0$. Thus, the proof is finished. □

3.3. Comparison of the new estimator (PCLTL) to the Liu-type logistic estimator

Theorem 3.3. *The new estimator is superior to the Liu-type logistic estimator under the asymptotic mean squared error matrix criterion if and only if* $T_{p-r}'\beta = 0$.

Proof: Putting $r = p$ into (3.4), we get

$$\begin{aligned} \text{MSEM}\big(\hat\beta(k,d)\big) \,&=\, TS(k)^{-1}S(d)\,\Lambda^{-1}S(d)\,S(k)^{-1}\,T' \\ &\quad + \,(k+d)^2\,TS(k)^{-1}\,T'\beta\beta'\,TS(k)^{-1}\,T'\,, \end{aligned} \tag{3.11}$$

where $S(k) = \Lambda + k\,I_p$ and $S(d) = \Lambda - d\,I_p$. Now we study the following difference $\Delta_3 = \text{MSEM}\big(\hat\beta(k,d)\big) - \text{MSEM}\big(\hat\beta_r(k,d)\big)$ where

$$\begin{aligned} \Delta_3 \,=\, & TS(k)^{-1}S(d)\,\Lambda^{-1}S(d)\,S(k)^{-1}\,T' \\ & -\,T_r\,S_r(k)^{-1}\Lambda_r^{-1}S_r(d)\,\Lambda_r\,S_r(d)\,\Lambda_r^{-1}S_r(k)^{-1}\,T_r' \\ & +\,(k+d)^2\,TS(k)^{-1}\,T'\beta\beta'\,TS(k) \\ & -\Big(-T_{p-r}T_{p-r}' - (d+k)\,T_r\,S_r(k)^{-1}\,T_r'\Big)\,\beta \\ & \times\,\beta'\Big(-T_{p-r}T_{p-r}' - (d+k)\,T_r\,S_r(k)^{-1}\,T_r'\Big)\,. \end{aligned}$$

Suppose that $C = T_{p-r} D T'_{p-r}$, where

$$D = S_{p-r}(k)^{-1} S_{p-r}(d)\, \Lambda_{p-r}^{-1} S_{p-r}(d)\, S_{p-r}(k)^{-1}$$

and $a_3 = (d+k)\, T S(k)^{-1} T' \beta$, $a_2 = \left(-T_{p-r} T'_{p-r} - (d+k)\, T_r S_r(k)^{-1} T'_r\right) \beta$. We can apply part (b) of Lemma 3.2. The Moore–Penrose inverse of C is $C^+ = T_{p-r} D^{-1} T'_{p-r}$, and $CC^+ = T_{p-r} T'_{p-r}$. So $a_3 \notin R(C)$, $a_2 \in R(C : a_3)$, $s = 1$ and $a_2 - a_3 = C\eta_1$, where

$$\eta_1 = -T_{p-r} S_{p-r}(k)^{-1} S_{p-r}(d)\, \Lambda_{p-r}^{-1} T'_{p-r} \beta\,.$$

Then by Lemma 3.2, we obtain that the new estimator is superior to the Liu-type logistic estimator under the asymptotic mean squared error matrix criterion if and only if $(a_2 - a_3)\, C^-(a_2 - a_3) \leq 0$ or $\eta'_1 C \eta_1 \leq 0$. In fact, $(a_2 - a_3)\, C^-(a_2 - a_3) \geq 0$, so the new estimator is better than the Liu-type logistic estimator under the asymptotic mean squared error matrix criterion if and only if $\eta'_1 C \eta_1 = 0$, that is $\beta' T_{p-r} \Lambda_{p-r} T'_{p-r} \beta = 0$. □

4. A MONTE CARLO SIMULATION STUDY

In this simulation study, we study the logistic regression model. In this section, we present the details and the results of the Monte Carlo simulation which is conducted to evaluate the performances of the MLE, PCLR, LTL and PCLTL estimators. There are several papers studying the performance of different estimators in the binary logistic regression. Therefore, we follow the idea of Lee and Silvapulle ([11]), Månsson *et al.* ([15]), Asar ([3]) and Asar and Genç ([4]) generating explanatory variables as follows:

$$x_{ij} = \left(1 - \rho^2\right)^{1/2} z_{ij} + \rho\, z_{iq}\,, \tag{4.1}$$

where $i = 1, 2, ..., n$, $j = 1, 2, ..., q$ and z_{ij}'s are random numbers generated from standard normal distribution. Effective factors in designing the experiment are the number of explanatory variables q, the degree of the correlation among the independent variables ρ^2 and the sample size n.

Four different values of the correlation ρ corresponding to 0.8, 0.9, 0.99 and 0.999 are considered. Moreover, four different values of the number of explanatory variables consisting of $q = 6$, 8 and 12 are considered in the design of the experiment. The sample size varies as 50, 100, 200, 500 and 1000. Moreover, we choose the number of principal components using the method of percentage of the total variability which is defined as

$$\text{PTV} = \frac{\sum_{j=1}^{r} \lambda_j}{\sum_{j=1}^{p} \lambda_j} \times 100\,.$$

In the simulation, PTV is chosen as 0.75 for $q = 8$ and 12 and 0.83 for $q = 6$ (see Aguilera *et al.* ([1])).

The coefficient vector is chosen due to Newhouse and Oman ([16]) such that $\beta'\beta = 1$ which is a commonly used restriction, for example see Kibria ([10]). We generate the n observations of the dependent variable using the Bernoulli distribution $\text{Be}(\pi_i)$ where $\pi_i = \frac{e^{x_i\beta}}{1+e^{x_i\beta}}$ such that x_i is the i-th row of the data matrix X.

The simulation is repeated for 10 000 times. To compute the simulated MSEs of the estimators, the following equation is used respectively:

$$\text{MSE}(\tilde{\beta}) = \frac{\sum_{c=1}^{10000} (\tilde{\beta}_c - \beta)' (\tilde{\beta}_c - \beta)}{10000}, \tag{4.2}$$

where $\widetilde{\beta}_c$ is MLE, PCLR, LTL, and PCLTL in the c-th replication. The convergence tolerance is taken to be 10^{-6} in the IRLS algorithm.

We choose the biasing parameter as follows:

1. LTL: We refer to Asar ([3]) and choose $d_{\text{LTL}} = \frac{1}{2} \min\left\{\frac{\lambda_j}{\lambda_j+1}\right\}_{j=1}^{p}$ where min is the minimum function and $k_{\text{AM}} = \frac{1}{p} \sum_{j=1}^{p} \frac{\lambda_j - d(1+\lambda_j \hat{\alpha}_j^2)}{\lambda_j \hat{\alpha}_j^2}$.

2. PCLTL: We propose to use the modifications of the methods given above as follows:

$$d_{\text{PCLTL}} = \frac{1}{2} \min\left\{\frac{\lambda_j}{\lambda_j+1}\right\}_{j=1}^{r}$$

and

$$k_{\text{PCLTL}} = \frac{1}{r} \sum_{j=1}^{r} \frac{\lambda_j - d_{\text{PCLTL}}(1+\lambda_j \hat{\alpha}_j^2)}{\lambda_j \hat{\alpha}_j^2}.$$

Table 1: Simulated MSE values of the estimators when $q = 6$.

n	Estimator	ρ			
		0.8	0.9	0.99	0.999
50	MLE	0.9942	0.8060	4.8571	41.7598
	LTL	0.8645	0.7561	2.2694	14.5135
	PCLR	0.9441	0.7780	1.9253	10.1568
	PCLTL	**0.8619**	**0.7480**	**0.9481**	**2.2783**
100	MLE	0.7050	0.7478	4.1961	38.1615
	LTL	0.7328	0.7613	2.6520	15.3411
	PCLR	**0.6913**	**0.7342**	1.6385	14.3406
	PCLTL	0.7169	0.7460	**1.0849**	**2.5414**
200	MLE	0.7286	0.8308	1.2978	5.8428
	LTL	0.7784	**0.7862**	**0.8506**	**1.6886**
	PCLR	**0.7221**	0.8223	1.1428	4.2635
	PCLTL	0.7668	0.7879	0.8571	1.7816
500	MLE	0.7428	0.7620	1.3640	4.0893
	LTL	0.8043	0.7665	1.0366	1.7055
	PCLR	**0.7417**	**0.7551**	0.9309	2.3118
	PCLTL	0.7895	0.7595	**0.8193**	**1.2168**
1000	MLE	0.7325	0.7512	0.9295	1.3950
	LTL	0.7550	0.7930	0.8030	0.8265
	PCLR	**0.7317**	**0.7463**	0.8421	1.1389
	PCLTL	0.7449	0.7878	**0.7766**	**0.8130**

Table 2: Simulated MSE values of the estimators when $q = 8$.

n	Estimator	ρ			
		0.8	0.9	0.99	0.999
50	MLE	0.9148	1.2089	3.8258	54.7686
	LTL	0.8065	0.9414	1.2265	16.7926
	PCLR	0.7669	0.8948	2.3655	12.4373
	PCLTL	**0.7237**	**0.8477**	**1.1400**	**3.2686**
100	MLE	0.7917	0.8182	2.2287	35.6674
	LTL	0.8264	0.7728	1.0472	14.5687
	PCLR	**0.7512**	0.7857	1.5575	17.9893
	PCLTL	0.8204	**0.7573**	**1.0189**	**9.3816**
200	MLE	0.7891	0.8598	1.7860	17.1105
	LTL	0.8293	0.8094	1.0451	6.3030
	PCLR	**0.7710**	0.7962	1.3309	7.4095
	PCLTL	0.8150	**0.7893**	**0.9678**	**3.2204**
500	MLE	0.7359	0.8031	1.2199	3.9098
	LTL	0.7612	0.8043	0.9003	1.4233
	PCLR	**0.7244**	**0.7608**	1.0378	2.4107
	PCLTL	0.7511	0.7959	**0.8712**	**1.2120**
1000	MLE	0.7502	0.7889	0.8576	5.0227
	LTL	0.7873	0.7933	0.7935	2.6289
	PCLR	**0.7462**	0.7516	0.8086	1.8239
	PCLTL	0.7781	**0.7800**	**0.7861**	**1.1132**

Table 3: Simulated MSE values of the estimators when $q = 12$.

n	Estimator	ρ			
		0.8	0.9	0.99	0.999
50	MLE	1.1407	1.2743	11.0452	81.6076
	LTL	0.8948	0.9622	3.4314	15.0524
	PCLR	0.8437	0.9409	3.7290	12.2146
	PCLTL	**0.8157**	**0.8961**	**1.8368**	**2.0556**
100	MLE	0.9247	1.4041	4.8286	22.7269
	LTL	0.8618	1.0687	2.0033	3.4810
	PCLR	**0.7999**	**0.8438**	1.3798	10.3112
	PCLTL	0.8152	0.8585	**0.8643**	**2.7653**
200	MLE	0.8238	1.0956	1.9228	15.8403
	LTL	0.8111	0.9612	0.9866	4.2617
	PCLR	0.7959	**0.8369**	1.4330	5.6484
	PCLTL	**0.7940**	0.8518	**0.9760**	**1.9115**
500	MLE	0.8009	0.8173	3.3133	11.2533
	LTL	0.8387	0.8357	2.3419	4.4671
	PCLR	**0.7809**	**0.8042**	1.0087	6.5955
	PCLTL	0.8331	0.8150	**0.8581**	**3.7834**
1000	MLE	0.7798	0.8081	1.0899	4.4119
	LTL	0.8205	0.8286	**0.8784**	1.7929
	PCLR	**0.7733**	**0.7965**	1.0258	1.7815
	PCLTL	0.8089	0.8212	0.9056	**1.0085**

According to Tables 1–3, the following results are obtained:

1. MSE of the MLE is inflated when the degree of correlation is increased and the sample size is low. On the other hand, the performance of MLE becomes quite well when the sample size is high enough.

2. Similarly, if we consider PCLR and LTL, the MSE values are also inflated for increasing values of the degree of correlation especially when $n = 50$.

3. MLE, PCLR and LTL produce high MSE values when the sample size is low and the degree of correlation is high. However, PCLTL seems to be robust to this situation in most of the cases.

4. Increasing the sample size makes a positive effect on the estimators in most of the situations. However, there is a degeneracy in this property.

5. When the degree of correlation is low, there is no estimator beating all others.

6. Overall, the new estimator PCLTL has the lowest MSE value in most of the situations considered in the simulation.

5. CONCLUSION

In this paper, we develop a new principal component Liu-type logistic estimator as a combination of the principal component logistic regression estimator and Liu-type logistic estimator to overcome the multicollinearity problem. We have proved some theorems showing the superiority of the new estimator over the other estimators by studying their asymptotic mean squared error matrix criterion. Finally, a Monte Carlo simulation study is presented in order to show the performance of the new estimator. According to the results, it seems that PCLTL is a better alternative in multicollinear situations in the binary logistic regression model.

REFERENCES

[1] Aguilera, A.M.; Escabias, M. and Valderrama, M.J. (2006). Using principal components for estimating logistic regression with high-dimensional multicollinear data, *Computational Statistics and Data Analysis*, **50**(8), 1905–1924.

[2] Alheety, M.I. and Kibria, B.M.G. (2013). Modified Liu-Type Estimator Based on $(r-k)$ Class Estimator, *Communications in Statistics — Theory and Methods*, **2**, 304–319.

[3] Asar, Y. (2017). Some new methods to solve multicollinearity in logistic regression, *Communications in Statistics — Simulation and Computation*, **46**(4), 2576–2586.

[4] Asar, Y. and Genç, A. (2016). New Shrinkage Parameters for the Liu-type Logistic Estimators, *Communications in Statistics — Simulation and Computation*, **45**(3), 1094–1103.

[5] Baksalary, J.K. and Trenkler, G. (1991). Nonnegative and positive definiteness of matrices modified by two matrices of rank one, *Linear Algebra and its Application*, **151**, 169–184.

[6] Farebrother, R.W. (1976). Further Results on the Mean Square Error of Ridge Regression, *Journal of the Royal Statistical Society B*, **38**, 248–250.

[7] Hoerl, A.E. and Kennard, R.W. (1970). Ridge regression: biased estimation for nonorthogonal problems, *Technometrics*, **12**, 55–67.

[8] Huang, W.H.; Qi, J.J. and Huang, N.T. (2009). Liu-type estimator for linear model with linear restrictions, *Journal of System Sciences and Mathematical Sciences*, **29**, 937–946.

[9] Inan, D. and Erdogan, B.E. (2013). Liu-type logistic estimator, *Communications in Statistics — Simulation and Computation*, **42**(7), 1578–1586.

[10] Kibria, B.M.G. (2003). Performance of some new ridge regression estimators, *Communications in Statistics — Simulation and Computation*, **32**(2), 419–435.

[11] Lee, A.H. and Silvapulle, M.J. (1988). Ridge estimation in logistic regression, *Communications in Statistics — Simulation and Computation*, **17**(4), 1231–1257.

[12] Lesaffre, E. and Max, B.D. (1993). Collinearity in Generalized Linear Regression, *Communications in Statistics — Theory and Methods*, **22**(7), 1933–1952.

[13] Li, Y.L. and Yang, H. (2011). Two kinds of restricted modified estimators in linear regression model, *Journal of Applied Statistics*, **38**, 1447–1454.

[14] Liu, K. (1993). A new class of biased estimate in linear regression, *Communications in Statistics — Theory and Methods*, **22**, 393–402.

[15] Månsson, K.; Kibria, B.G. and Shukur, G. (2012). On Liu estimators for the logit regression model, *Journal of Applied Statistics*, **29**(4), 1483–1488.

[16] Newhouse, J.P. and Oman, S.D. (1971). An evaluation of ridge estimators, *Rand Corporation*, **P-716-PR**, 1–16.

[17] Rao, C.R. and Toutenburg, H. (1999). *Linear Models: Least Squares and Alternatives*, Springer-Verlag, New York..

[18] Saleh, A.M.E. and Kibria, B.M.G. (2013). Improved ridge regression estimators for the logistic regression model, *Computational Statistics*, **28**(6), 2519–2558.

[19] Schaefer, R.L.; Roi, L.D. and Wolfe, R.A. (1984). A ridge logistic estimator, *Communications in Statistics — Theory and Methods*, **13**(1), 99–113.

[20] Xu, J.W. and Yang, H. (2011). On the restricted almost unbiased estimators in linear regression, *Communications in Statistics — Theory and Methods*, **38**, 605–617.

[21] Yüzbaşı, B.; Ahmed, S.E. and Güngör, M. (2017). Improved penalty strategies in linear regression models, *REVSTAT — Statistical Journal*, **15**(2), 251–276.