American Journal of Mathematics and Statistics 2021, 11(1): 18-25
DOI: 10.5923/j.ajms.20211101.03

The Maximum Likelihood Estimates with Wrong Logistic Regression Model and Covariate Assumptions are Violated

Nuri H. Salem Badi*, Mohamed M. Shakandli

Faculty of Science, Statistical Department, University of Benghazi, Benghazi, Libya

Abstract The general method of estimation the logistic regression parameter is maximum likelihood (ML). In a very general sense the ML method yields values for the unknown parameters that maximize the probability of the observed set of data. Much work discusses the behaviour of the distribution of Maximum likelihood estimates (*MLE*) for the logistic regression model under the correct model. However, many issues still need more examination as the relationship between the links of the logistic function and the skew-normal distribution which consider in this work. In this paper, our work considers this behaviour and investigates the *MLE* method under the logistic regression model when the wrong model has been fitted and the assumption on covariates are violated. We will consider this behaivuor and the covariates drawing from a non-normal distribution and evaluate it by simulation.

Keywords Logistic regression model, Maximum likelihood method, Skew-normal distribution, Probit function and expit function

1. Introduction

The subject of the behaviour of maximum likelihood estimation (MLE) method in logistic regression model has attracted the attention of many scientists and researchers. [6] developed the analysis of the binary data and application of the maximum likelihood: see also [7]. [11] introduced the generalized linear model and used special techniques to obtain the maximum likelihood estimates of the parameters, with observations distributed according to some exponential family. [10] discussed the generalized linear model and behaviour of the maximum liklihood (ML) method for binary outcome. [5] discused method used a modified score function to reduce the bias of the maximum likelihood estimates. The ML method under the wrong logistic model has been discussed extensively by [9, p.23]. The idea is to try to find in terms of the true parameters of the model the least false values which are obtained by maximising the incorrect likelihood function. We will use the relationship between $\text{expit}(u) = e^u/(1+e^u)$ function and probit function $\Phi(\cdot)$, and use the properties of the multivariate skew-Normal distribution to compute a good approximation to the least false values under wrong logistic model. The behaviour of MLE for binary outcome has been discussed more extensively by [10]. The logistic model when $Y_i \sim \text{binomial}(m_i, \pi_i)$ with $m_i = 1$ can be fitted using the method of maximum likelihood to estimate the parameters. The first step is to construct the likelihood function which is a function of the unknown parameters and data, then choose those values of the parameters that maximize this function. The log-likelihood function is:

$$l(\pi; Y) = \sum_{i=1}^{n} \left[y_i \log(\pi_i) + (1-y_i)\log(1-\pi_i) \right]$$

Where, in this case we have

$$l(\beta) = \sum_{i=1}^{n} y_i(\alpha + x_i^T\beta) - \sum_{i=1}^{n} \log\left[1 + \exp(\alpha + x_i^T\beta)\right]$$

where β is a p-dimensional vector, x_i are a vector of covariates for i^{th} individual and $i = 1, \cdots, n$. To estimate the parameters α and β_j we differentiate the log-likelihood function with respect to α and β_j. If the fitted model is the true model then, the asymptotic distribution of β_j is $\hat{\beta} \sim N(\beta, I(\beta)^{-1})$ where $I(\beta)$ is the $(p \times p)$ Fisher's information matrix, its $(r,s)^{th}$ element is defined as

$$I_{rs} = \left[-E\left\{ \frac{\partial^2 l}{\partial\beta_r \partial\beta_s} \right\} \right].$$

* Corresponding author:
nuri.badi@uob.edu.ly (Nuri H. Salem Badi)
Received: Dec. 23, 2020; Accepted: Jan. 12, 2021; Published: Jan. 30, 2021
Published online at http://journal.sapub.org/ajms

If is evaluated at MLE $\hat{\beta}$.

2. MLE Under the Wrong Model

[9] discussed how the maximum likelihood method used to estimate the parameters of a given regression model is affected when the assumed model is incorrect. If the data are independent and identically distributed, the log likelihood function in case of the density $f(y_i,\theta)$ for an individual observation, we can write as:

$$\ell_n(\theta)=\sum_{i=1}^{n}\log f(y_i,\theta).$$

The important question here is, if we fit a model for Y as $f(y|\ \theta)$ when true model is $g(y)$, what value do we estimate for θ? We have for each value of θ, by the weak large numbers, in probability, as $n\to\infty$

$$n^{-1}\ell_n(\theta)\to E\left(\log f(Y|\ \theta)\right),$$

The Kullback-Leibler (KL) divergence is

$$KL(g(y),f(y,\theta))=\int g(y)\log\frac{g(y)}{f(y,\theta)}dy. \qquad (1)$$

For model $f(y|\ x,\theta)$, we need to solve likelihood function and find the least false parameter θ^* which minimises $KL(g,f_\theta)$, then

$$E_X\left(E_g\left(\frac{\partial\log f(Y|\ X,\theta)}{\partial\theta}\right)\right)\Big|_{\theta^*}=0. \qquad (2)$$

Application to Logistic Regression

Now, we will apply the MLE method under wrong model on logistic regression model. The idea is to use this method to obtain equations which determine the least false value θ^* for a logistic regression. To explain the behaviour of the MLE in this case we will partition of the vector covariates X, as previous (X_f,X_a). The fitted model is

$$\pi=\text{expit}(\alpha+\beta_f X_f)$$

However, this model is wrong because the true model is.

$$\pi=\text{expit}(\alpha+\beta_f X_f+\beta_a X_a)$$

So, expectation of the ML equations are zero when $\theta=\theta^*=(\alpha^*,\beta^*)$. From the above equations where Y is binary, the expectation in this case becomes

$$E_X(E_{Y|\ X}(Y))=E_X(\text{expit}(\alpha^*+\beta_f^* X_f)),$$

and

$$E_X(X_f E(Y|\ X))=E_X(X_f\,\text{expit}(\alpha^*+\beta_f^* X_f)).$$

The $E(Y|\ X)$ is $\Pr(Y=1|\ X)$ and this is given by the true model

$$\Pr(Y=1|\ X)=\text{expit}(\alpha+\beta_f^T X_f+\beta_a^T X_a).$$

But we fit the model without X_a. The least false values, α^* and β_f^*, can be found in terms of α,β_f and β_a and the parameters of the distribution of the covariates as from

$$E\left[\text{expit}(\alpha^*+\beta_f^{*T}X_f)\right]=E\left[\text{expit}(\alpha+\beta^T X)\right] \qquad (3)$$

$$E\left[X_{fj}\text{expit}(\alpha^*+\beta_f^{*T}X_f)\right]=E\left[X_{fj}\text{expit}(\alpha+\beta^T X)\right]. \qquad (4)$$

where, X_{fj} is the j^{th} element of X_f $(j=1,\ldots,p)$. These equations can be solved approximately if X follows a multivariate normal distribution, by approximating $\text{expit}(\cdot)$ and using the skew-normal distribution. In fact, the previous work considers by [14], discussed the inconsistent treatment estimates from mis-specified logistic regression analyses of randomized trials. In this paper, we are interested in more investigate the behaviour of MLE under the wrong mode when covariate assumptions are violatedl and assessment by simoulation.

Skew-Normal Distribution

The skew-Normal distribution has been discussed by [3] and [4]. More discussion and numerical evidence of the presence of skewness in real data by [17] and [2]. Other discussion for quadratic forms and flexible class of skew-symmetric distribution discussed by [12] and [8] also, by [16]. A random variable U is called skew normal with parameter λ, so $U\sim SN(\lambda)$, if its density function is :

$$f(u;\lambda)=2\phi(u)\Phi(\lambda u) \qquad (5)$$

where $u\in R$, $\phi(\cdot)$ and $\Phi(\cdot)$ are the density and distribution function of standard normal distribution respectively, that defined by [3]. In general case, [1], discussed extends the skew normal distribution and properties of this family. We can defined the extend multivariate skew-normal distribution as; a p-dimensional random variable U has extended skew-normal distribution, $ESN(\vartheta,\Omega,\lambda,\nu)$, if it has density:

$$\frac{\phi_p(u;\vartheta,\Omega)\Phi(\lambda^T(u-\vartheta)+\nu)}{\Phi(\nu/\sqrt{1+\lambda^T\Omega\lambda})}$$

where ν is a scalar, Ω is dispersion matrix has $p\times p$ dimensional and parameters ϑ and λ are p-dimensional. The $\phi_p(.;\vartheta,\Omega)$ is the density of a p-dimensional normal variable with mean θ and dispersion matrix Ω where $\Phi(.)$ is the cumulative distribution function of a univariate standarad normal variable.

Probit Function and expit Function

We are consider the approximation of $\text{expit}(\cdot)$ by $\Phi(\cdot)$, the distribution function of a standard normal variable. [15] reported that, the logistic distribution closely resembles the normal distribution which discussed the shape of distribution both are symmetrical and noted some properties. [13, p.5], point out the comparison of logistic and normal cumulative distribution function. The approximation form defined as: $\text{expit}(u) \approx \Phi(ku)$, where $k = (16\sqrt{3})/(15\pi)$.

3. Least False Values Under Wrong Logistic Model

The main point is, suppose that we model a binary outcome, Y, using a logistic regression, i.e.

$$\Pr(Y = 1 | X_f) = \text{expit}(\alpha + \beta_f^T X_f),$$

but that the true model includes more covariates, i.e.

$$\Pr(Y = 1 | X) = \text{expit}(\alpha + \beta_f^T X_f + \beta_a^T X_a).$$

Now, to find the least false values in terms of parameters of the true logistic model, use the approximation form $\text{expit}(\cdot) \approx \Phi(\cdot)$, properties of the skew-normal distribution and we use the two equations which determine the *MLEs*, as we have discussed in section 2 about MLE under the wrong model to find the least false values. Let us assume that X has $(p+q)$-dimensional multivariate Normal distribution, where p and q denote the dimensions of X_f and X_a respectively. The presence of an intercept in the above models means that we may assume, wlog, that $E(X) = 0$. If $\text{var}(X) = \Omega$, then also suppose that the partition of this matrix corresponding to X_f and X_a is:

$$\Omega = \begin{pmatrix} \Omega_{ff} & \Omega_{fa} \\ \Omega_{af} & \Omega_{aa} \end{pmatrix},$$

then we can apply the approximation to (3) and (4) using $\text{expit}(u) \approx \Phi(ku)$, which this leads to

$$E_X\left(\Phi\left(k[\alpha^* + \beta_f^{*T} X_f]\right)\right) = E_X\left(\Phi\left(k[\alpha + \beta^T X]\right)\right) \quad (6)$$

Now we use the properties of skew-normal distribution, in this case the density function of skew-normal distribution where $E(X) = 0$ is

$$f(X, \alpha, \beta) = \frac{\Phi(k(\alpha + \beta^T X))\phi(X)}{\Phi\left(\dfrac{k\alpha}{\sqrt{1 + k^2 \beta^T \Omega \beta}}\right)}.$$

Then we can write the (6) as

$$\Phi\left(\frac{k\alpha^*}{\sqrt{1 + k^2 \beta_f^{*T} \Omega_{ff} \beta_f^*}}\right) = \Phi\left(\frac{k\alpha}{\sqrt{1 + k^2 \beta^T \Omega \beta}}\right),$$

which is

$$\frac{\alpha^*}{\sqrt{1 + k^2 \beta_f^{*T} \Omega_{ff} \beta_f^*}} = \frac{\alpha}{\sqrt{1 + k^2 \beta^T \Omega \beta}}. \quad (7)$$

Turning our attention to (4) and using the results for the expectation of a SN distribution, we obtain

$$\begin{aligned} &\frac{\Omega_{ff} \beta_f^*}{\sqrt{1 + k^2 \beta_f^{*T} \Omega_{ff} \beta_f^*}} \phi\left(\frac{\alpha^*}{\sqrt{1 + k^2 \beta_f^{*T} \Omega_{ff} \beta_f^*}}\right) \\ &= \frac{(\Omega\beta)_1}{\sqrt{1 + k^2 \beta^T \Omega \beta}} \phi\left(\frac{\alpha}{\sqrt{1 + k^2 \beta^T \Omega \beta}}\right), \end{aligned} \quad (8)$$

where $\phi(\cdot)$ is the standard Normal density, and $(\Omega\beta)_1$ denotes the first p elements of $\Omega\beta$, which is $\Omega_{ff}\beta_f + \Omega_{fa}\beta_a$. Using the result in (7), we can simplify (8) and finally, we can follows that

$$\beta_f^* = \frac{1}{\sqrt{1 + k^2 \beta_a^T \Omega \beta_a}} (\beta_f + \Omega_{ff}^{-1} \Omega_{fa} \beta_a) \quad (9)$$

and

$$\alpha_f^* = \frac{\alpha}{\sqrt{1 + k^2 \beta_a^T \Omega \beta_a}}. \quad (10)$$

where $\Omega = \Omega_{aa} - \Omega_{af} \Omega_{ff}^{-1} \Omega_{fa}$. From this we get Note that (9) includes a denominator, such that $\beta_f^* \neq \beta_f$ even when $\Omega_{fa} = 0$, although, of course, $\beta_f^* = \beta_f$ if $\beta_a = 0$.

4. Simulation Study when the Covariates Follow Normal Distribution

The goal of this simulation, is to assess the approximation computed for the least false values for logistic regression model. We are interested to application on case of the covariates generated by multivariate Normal distribution. Applied on different cases with different variance and different correlation to check on the behaviour of the formulae of the least false values under wrong model. We looking in this simulation for check the approximation of the last false values for a true logistic regression model has five covariates $p = 5$ is

$$\pi_i = \text{expit}(\alpha + \beta^T X)$$

where, $\beta^T = (\beta_1, \beta_2, \ldots, \beta_5)$, $X = (x_{i1}, \ldots, x_{i5})$ and in the fitted model there are two covariates. We designed the simulation as follows:

- We choose X as a draw from the multivariate normal distribution $X \sim N_5(0, \Omega)$.
- We consider the 5×5 covariance matrix Ω is

$$\Omega = \sigma^2 \begin{bmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{bmatrix},$$

where,

$$\Omega_{11} = \begin{bmatrix} 1 & \rho_{12} \\ \rho_{21} & 1 \end{bmatrix}, \Omega_{21} = \begin{bmatrix} \rho_{31} & \rho_{32} \\ \rho_{41} & \rho_{42} \\ \rho_{51} & \rho_{52} \end{bmatrix},$$

$$\Omega_{22} = \begin{bmatrix} 1 & \rho_{34} & \rho_{35} \\ \rho_{43} & 1 & \rho_{45} \\ \rho_{53} & \rho_{54} & 1 \end{bmatrix}, \Omega_{21}^T = \Omega_{12}.$$

- Use two different variance $\sigma^2 = 0.1,1.5$.
- We consider 3 different cases of correlation which is each case of Ω_{ij} has same ρ_{ij} designed as: (0.2,0.2,0.4), (0.8,0.7,0.9), (0.2,-0.2,-0.2). Values are chosen to assume Ω is positive definite.
- We choose the parameters $\beta_1,\ldots,\beta_5$ and α to give us two cases $\Pr(Y=1) \simeq 10\%$ and $\Pr(Y=1) \simeq 60\%$. As calculate the unconditional $\Pr(Y=1)$ by properties skew-normal distribution we get,

$$\Pr(Y=1) \approx \Phi\left(\frac{k\alpha}{\sqrt{1+k^2\beta^T\Omega\beta}}\right).$$

Choose $\beta_1 = 0.25, \beta_2 = 0.35, \beta_3 = 0.40, \beta_4 = 0.3, \beta_5 = 0.2$ and adjust α, so that over the covariates $\Pr(Y=1) \simeq 10\%$ $(\alpha = -2.2)$ and $\Pr(Y=1) \simeq 60\%$ $(\alpha = 0.4)$.

- The sample size has been used $n = 500$, $n = 10000$ and $N = 1000$ number of simulation.

4.1. Results and Discussion

We report the accuracy of the estimation parameters of the wrong logistic regression model has two covariates when the true model has five covariates. Tables shows comparison between the least false values which is computed by approximation of $\text{expit}(u) \approx \Phi(ku)$ and skew-Normal distribution properties and values of estimated parameters by fitted logistic regression model. R_1, R_2, R_3 denote the ratios of the mean of the simulated fits to the comuted last false value.

Table 1 and Table 2, shows the results of simulation of data generated by multivariate Normal distribution in cases of $\Pr(Y=1) \simeq 60\%$ and $\Pr(Y=1) \simeq 10\%$ respectively with sample size $n = 500$. Table 3 and Table 4, shows the results of simulation with sample size $n = 10000$. We can see clearly the results show ratios close to one. The same behaviour results found in both cases of $\Pr(Y = 60\%)$ and $\Pr(Y = 10\%)$, where is the ratio found close to one. That is meaning the approximation form of the least false values works well, although the probability of outcome Y is very low about 10%, but a good results and reasonable behaviour have been found. Some issues of low ratio a raised in case of sample size $n = 500$, that there are some estimated values were very small close to zero which affect on ratio. Moreover, the parameter selection and correlation selection may be having slightly affected in a few cases.

Table 1. Simulation results of last false values by multivariate Normal distribution in case $\Pr(Y=1) \simeq 60\%$, $n = 500$ and R_i denote to the Ratio

$\sigma^2 = 0.1$			Parameters estimated, Least false values and Ratio								
Ω_{11}	Ω_{12}	Ω_{22}	α	α^*	R_1	β_1	β_1^*	R_2	β_2	β_2^*	R_3
0.2	0.2	0.4	0.4048	0.3969	1.01	0.4212	0.3969	1.06	0.4999	0.4962	1.01
0.8	0.7	0.9	0.4025	0.3978	1.01	0.6908	0.5967	1.15	0.6381	0.6961	0.92
0.2	-0.2	-0.2	0.4068	0.3955	1.01	0.0756	0.0998	0.75	0.2481	0.1995	1.24
$\sigma^2 = 1.5$											
0.2	0.2	0.4	0.3993	0.3606	1.10	0.3693	0.3606	1.02	0.4298	0.4507	0.95
0.8	0.7	0.9	0.3689	0.3706	0.99	0.5579	0.5560	1.00	0.6239	0.6486	0.96
0.2	-0.2	-0.2	0.3495	0.3869	0.90	0.0996	0.0967	1.03	0.1995	0.1934	1.03

Table 2. Simulation results of last false values by multivariate Normal distribution in case $\Pr(Y=1) \simeq 10\%$, $n = 500$ and R_i denote to the Ratio

$\sigma^2 = 0.1$			Parameters estimated, Least false values and Ratio								
Ω_{11}	Ω_{12}	Ω_{22}	α	α^*	R_1	β_1	β_1^*	R_2	β_2	β_2^*	R_3
0.2	0.2	0.4	-2.208	-2.183	1.01	0.4305	0.3969	1.08	0.4841	0.4962	0.98
0.8	0.7	0.9	-2.226	-2.188	1.01	0.6183	0.5967	1.03	0.6868	0.6961	0.98
0.2	-0.2	-0.2	-2.193	-2.194	0.99	0.1043	0.0997	1.04	0.1989	0.1995	1.07
$\sigma^2 = 1.5$											
0.2	0.2	0.4	-2.014	-1.983	1.01	0.4165	0.3610	1.15	0.4410	0.4507	0.98
0.8	0.7	0.9	-2.043	-2.039	1.00	0.5435	0.5560	0.98	0.6473	0.6486	0.99
0.2	-0.2	-0.2	-2.123	-2.128	0.99	0.0664	0.0967	0.68	0.1953	0.1934	1.01

Table 3. Simulation results of last false values by multivariate Normal distribution in case $\Pr(Y=1)\simeq 60\%$, $n=10000$ and R_i denote to the Ratio

$\sigma^2=0.1$			Parameters estimated, Least false values and Ratio								
Ω_{11}	Ω_{12}	Ω_{22}	α	α^*	R_1	β_1	β_1^*	R_2	β_2	β_2^*	R_3
0.2	0.2	0.4	0.3966	0.3969	0.99	0.4062	0.3969	1.02	0.4963	0.4962	1.00
0.8	0.7	0.9	0.3985	0.3981	1.00	0.5951	0.5967	0.99	0.6950	0.6961	0.99
0.2	-0.2	-0.2	0.3981	0.3990	0.99	0.1025	0.0997	1.02	0.2014	0.1995	1.01
$\sigma^2=1.5$											
0.2	0.2	0.4	0.3341	0.3606	0.93	0.3517	0.3606	0.98	0.4481	0.4507	0.99
0.8	0.7	0.9	0.3631	0.3706	0.98	0.5592	0.5560	1.01	0.6517	0.6486	1.01
0.2	-0.2	-0.2	0.3669	0.3869	0.95	0.0903	0.0967	0.93	0.1963	0.1934	1.01

Table 4. Simulation results of last false values by multivariate Normal distribution in case $\Pr(Y=1)\simeq 10\%$, $n=10000$ and R_i denote to the Ratio

$\sigma^2=0.1$			Parameters estimated, Least false values and Ratio								
Ω_{11}	Ω_{12}	Ω_{22}	α	α^*	R_1	β_1	β_1^*	R_2	β_2	β_2^*	R_3
0.2	0.2	0.4	-2.183	-2.183	1.00	0.3981	0.3971	1.00	0.5058	0.4962	1.02
0.8	0.7	0.9	-2.191	-2.190	1.00	0.6094	0.5967	1.02	0.6885	0.6961	0.99
0.2	-0.2	-0.2	-2.193	-2.194	0.99	0.1043	0.0997	1.04	0.1989	0.1995	0.99
$\sigma^2=1.5$											
0.2	0.2	0.4	-1.967	-1.983	0.99	0.3610	0.3610	1.00	0.4600	0.4507	1.02
0.8	0.7	0.9	-2.031	-2.038	0.99	0.5518	0.5560	0.99	0.6607	0.6486	1.02
0.2	-0.2	-0.2	-2.129	-2.128	1.00	0.0982	0.0967	1.02	0.2011	0.1934	1.04

5. Covariate Assumptions are Violated

The previous analysis discussed the least false values with multivariate covariates, the simulation providing us reasonable results. That was applied to covariates draw from multivariate normal distribution. In this part we are interested to consider the model with symmetric distribution different from multivariate normal distribution. As we know, the behaviour of the *MLE* maybe affected by the assumption of normality on the covariates. So we will consider two of symmetric multivariate distribution, say, t-distribution and multivariate uniform distribution.

5.1. Simulation of Multivariate t and Multivariate Uniform Distribution

The goal of this simulation is to use the same computed formulae of the last false value which used in the previous analysis, to assess the approximation computed for the least false values for logistic regression model and with multivariate t and uniform distribution. We use the same assumption which used in previous simulation, let consider we have a true logistic regression model which has five covariates $p=5$ is

$$\pi_i = \text{expit}(\alpha+\beta^T X)$$

- We choose X as a draw from one of two multivariate distribution; either
  - Multivariate Uniform distribution, or Multivariate t-distribution.
- We are generating multivariate Uniform covariates by related with standard Normal distribution as:
  - $Z \sim MVN(0,R)$ where R is the correlation matrix.
  - $U=\Phi(Z)\rightarrow[0,1]$, (element wise) and

$$X_U \sim 5\sigma(U-\frac{1}{2})\rightarrow[-2\frac{1}{2}\sigma, 2\frac{1}{2}\sigma].$$

- We consider the 5×5 covariance matrix Ω is

$$\Omega=\begin{bmatrix}\Omega_{11} & \Omega_{12}\\ \Omega_{21} & \Omega_{22}\end{bmatrix}.$$

As we know, the mean of the uniform distribution is $U=1/2$ and the variance is $\text{var}(U)=1/12$. So, in this case we have $\text{var}(X_U)=25/12$ and $\text{cov}(X_{Ui},X_{Uj})=25\,\text{cov}(U_i,U_j)$, where the covariance is

$$\text{cov}(U_i,U_j)=\frac{\arcsin(\frac{\rho_{ij}}{2})}{2\pi}.$$

Then, the components of covariance matrix Ω are

$$\Omega_{11}=25\begin{bmatrix}\frac{1}{12} & \text{cov}(U_1,U_2)\\ \text{cov}(U_2,U_1) & \frac{1}{12}\end{bmatrix},$$

$$\Omega_{22}=25\begin{bmatrix}\frac{1}{12} & \text{cov}(U_3,U_4) & \text{cov}(U_3,U_5)\\ \text{cov}(U_4,U_3) & \frac{1}{12} & \text{cov}(U_5,U_4)\\ \text{cov}(U_5,U_3) & \text{cov}(U_5,U_4) & \frac{1}{12}\end{bmatrix},$$

$$\Omega_{21} = 25\begin{bmatrix} \text{cov}(U_3,U_1) & \text{cov}(U_3,U_2) \\ \text{cov}(U_4,U_1) & \text{cov}(U_4,U_2) \\ \text{cov}(U_5,U_1) & \text{cov}(U_5,U_2) \end{bmatrix}, \Omega_{21}^T = \Omega_{12}.$$

- We generating multivariate t-distribution with various value of degrees of freedom df, which changes the shape of the distribution, we choose two cases $df = (5, 200)$ and use two different variance $\sigma^2 = 0.1, 1.5$ in each case.
- Use the same assumption on which used in the priviouse simulation: correlation and variance, also use $\Pr(Y = 1) \simeq 10\%$ and $\Pr(Y = 1) \simeq 60\%$ with sample size $n = 500$, $n = 10000$ and $N = 1000$ number of simulation.

5.2. Results and Discussion

The results concerning two simulation data generated by multivariate Uniform distribution and multivariate t-distribution. The results of this simulation with Uniform distribution, showed in Table 5 and Table 6, in cases of $\Pr(Y = 1) \simeq 60\%$ and $\Pr(Y = 1) \simeq 10\%$ respectively with two sample size $n = 500$, $n = 10000$. The same results appeared, the ratio found nearly close to one in almost cases. A few cases appeared low ratio in case of sample size $n = 500$, which there are some estimated value were very small (i.e, when $\Omega_{11} = 0.2, \Omega_{12} = -0.2, \Omega_{22} = -0.2$ the parameter estimated was $\beta_1 = 0.0756, \beta_1^* = 0.0998$ and the ratio was $R_2 = 0.75$). In general we found the least false values in this case have the same behaviour of the multivariate normal covariates. The results of the second part of this simulation, concerning for results of data generated by multivariate t-distribution which showed in Table 7 and Table 8 in cases of $\Pr(Y = 1) \simeq 60\%$ and $\Pr(Y = 1) \simeq 10\%$ respectively with sample size $n = 500$. Table 9 and Table 10 shows the results in case of sample size $n = 10000$. The results of four cases with different degree of freedom $df = 200, 5$ and one case of variance has been used $\sigma^2 = 0.5$. Comparing these results with case of Normal distribution, more clearly when the degree of freedom larger enough we can reported that the results have the same behaviour. Moreover, we can say that the ratio appeared nearly close to one in all cases of correlation and degree of freedom, some slightly differences with low ratio appeared in few cases when degree of freedom is $df = 5$ and $n = 500$, which have the same behaviour found in case of the normal multivariate covariates when the estimated value was very small.

Overall, if we assume normality on covariates, but the covariates are drawn from a multivariate t-distribution with variety of degree of freedom and multivariate Uniform distribution, which use large sample size $n = 10000$. We found that, for different combination of correlations and variances, are appeared the results from (9) and (10) still appear to hold.

Table 5. Simulation results of last false values using different values of ρ_{ij} by generated variables from multivariate Uniform distribution in case $\Pr(Y = 1) \simeq 60\%$, $n = 500$, $n = 10000$ and R_i denote to the Ratio

$n = 500$			Parameters estimated, Least false values and Ratio								
Ω_{11}	Ω_{12}	Ω_{22}	α	α^*	R_1	β_1	β_1^*	R_2	β_2	β_2^*	R_3
0.2	0.2	0.4	0.3275	0.3485	0.94	0.3433	0.3437	1.00	0.4454	0.4308	1.03
0.8	0.7	0.9	0.3294	0.3594	0.92	0.5420	0.5339	1.01	0.5640	0.6238	0.90
0.2	-0.2	-0.2	0.3971	0.3811	0.93	0.1142	0.1005	1.13	0.1825	0.3811	0.93
$n = 10000$											
0.2	0.2	0.4	0.3309	0.3485	0.95	0.3299	0.3437	0.96	0.4129	0.4308	0.96
0.8	0.7	0.9	0.3350	0.3594	0.93	0.5262	0.5339	0.99	0.6194	0.6238	0.99
0.2	-0.2	-0.2	0.3830	0.3811	1.00	0.0971	0.1005	0.97	0.1935	0.1957	0.99

Table 6. Simulation results of last false values using different values of ρ_{ij} by generated variables from multivariate Uniform distribution in case $\Pr(Y = 1) \simeq 10\%$, $n = 500$, $n = 10000$ and R_i denote to the Ratio

$n = 500$			Parameters estimated, Least false values and Ratio								
Ω_{11}	Ω_{12}	Ω_{22}	α	α^*	R_1	β_1	β_1^*	R_2	β_2	β_2^*	R_3
0.2	0.2	0.4	-1.901	-1.916	0.99	0.6911	0.3437	1.04	0.7496	0.4308	1.03
0.8	0.7	0.9	-2.009	-1.977	1.01	0.5728	0.5339	1.07	0.5876	0.6238	0.94
0.2	-0.2	-0.2	-1.927	-1.919	1.00	0.0892	0.0809	1.10	0.1871	0.1682	1.11
$n = 10000$											
0.2	0.2	0.4	-1.900	-1.916	0.99	0.3551	0.3437	1.03	0.4483	0.4308	1.04
0.8	0.7	0.9	-1.963	-1.977	0.99	0.5217	0.5339	0.98	0.6557	0.6238	1.05
0.2	-0.2	-0.2	-2.089	-2.096	0.99	0.1022	0.1005	1.01	0.1938	0.1957	0.99

Table 7. Simulation results of last false values by multivariate t-distribution in case $\Pr(Y=1)\simeq 60\%$, $n=500$ and R_i denote to the Ratio

$df=200$			Parameters estimated, Least false values and Ratio								
Ω_{11}	Ω_{12}	Ω_{22}	α	α^*	R_1	β_1	β_1^*	R_2	β_2	β_2^*	R_3
0.2	0.2	0.4	0.389	0.385	1.01	0.409	0.385	1.06	0.422	0.481	0.88
0.8	0.7	0.9	0.393	0.389	1.01	0.549	0.584	0.94	0.682	0.681	1.00
0.2	-0.2	-0.2	0.395	0.395	1.00	0.098	0.099	0.99	0.174	0.197	0.88
	$df=5$										
0.2	0.2	0.4	0.402	0.385	1.04	0.332	0.385	0.86	0.466	0.481	0.97
0.8	0.7	0.9	0.375	0.389	0.96	0.564	0.584	0.97	0.722	0.681	1.06
0.2	-0.2	-0.2	0.379	0.395	0.96	0.086	0.098	0.88	0.186	0.198	0.94

Table 8. Simulation results of last false values by multivariate t-distribution in case $\Pr(Y=1)\simeq 10\%$, $n=500$ and R_i denote to the Ratio

$df=200$			Parameters estimated, Least false values and Ratio								
Ω_{11}	Ω_{12}	Ω_{22}	α	α^*	R_1	β_1	β_1^*	R_2	β_2	β_2^*	R_3
0.2	0.2	0.4	-2.100	-2.120	0.99	0.369	0.385	0.96	0.529	0.482	1.09
0.8	0.7	0.9	-2.165	-2.142	1.01	0.575	0.584	0.99	0.738	0.681	1.08
0.2	-0.2	-0.2	-2.212	-2.175	1.01	0.110	0.098	1.11	0.200	0.197	1.01
	$df=5$										
0.2	0.2	0.4	-2.118	-2.120	0.99	0.430	0.385	1.11	0.498	0.482	1.03
0.8	0.7	0.9	-2.114	-2.142	0.99	0.709	0.584	1.21	0.550	0.682	0.81
0.2	-0.2	-0.2	-2.180	-2.175	1.00	0.096	0.098	0.97	0.234	0.197	1.18

Table 9. Simulation results of last false values by multivariate t-distribution in case $\Pr(Y=1)\simeq 60\%$, $n=10000$ and R_i denote to the Ratio

$df=200$			Parameters estimated, Least false values and Ratio								
Ω_{11}	Ω_{12}	Ω_{22}	α	α^*	R_1	β_1	β_1^*	R_2	β_2	β_2^*	R_3
0.2	0.2	0.4	0.381	0.385	0.99	0.394	0.385	1.02	0.478	0.481	0.99
0.8	0.7	0.9	0.390	0.389	1.00	0.570	0.584	0.98	0.688	0.681	1.01
0.2	-0.2	-0.2	0.394	0.395	0.99	0.103	0.098	1.05	0.187	0.197	0.95
	$df=5$										
0.2	0.2	0.4	0.380	0.385	0.99	0.357	0.385	0.93	0.458	0.481	0.95
0.8	0.7	0.9	0.380	0.389	0.98	0.557	0.584	0.96	0.670	0.681	0.98
0.2	-0.2	-0.2	0.391	0.395	0.99	0.093	0.098	0.94	0.185	0.198	0.94

Table 10. Simulation results of last false values by multivariate t-distribution in case $\Pr(Y=1)\simeq 10\%$, $n=10000$ and R_i denote to the Ratio

$df=200$			Parameters estimated, Least false values and Ratio								
Ω_{11}	Ω_{12}	Ω_{22}	α	α^*	R_1	β_1	β_1^*	R_2	β_2	β_2^*	R_3
0.2	0.2	0.4	-2.119	-2.121	0.99	0.395	0.385	1.03	0.504	0.482	1.05
0.8	0.7	0.9	-2.141	-2.142	0.99	0.589	0.584	1.01	0.676	0.681	0.99
0.2	-0.2	-0.2	-2.181	-2.180	1.00	0.099	0.0988	1.01	0.205	0.197	1.03
	$df=5$										
0.2	0.2	0.4	-2.079	-2.120	0.98	0.379	0.385	0.98	0.466	0.482	0.97
0.8	0.7	0.9	-2.119	-2.142	0.99	0.552	0.584	0.95	0.688	0.682	1.01
0.2	-0.2	-0.2	-2.150	-2.175	0.99	0.109	0.098	1.10	0.186	0.198	0.94

6. Conclusions

The goal of this paper considers to investigate the behaviour of the MLE and find a formula to compute the least false values when the wrong logistic model has been fitted. Moreover, we examined the behaviour of the model when the assumption on covariates are violated. Corresponding to the simulation analysis, we found a good results in all cases when the covariates are drawn from the multivariate normal. The results appeared the *MLE* has reasonable behaviour with the least false values in case of wrong model, which computed in terms of the true parameters. On the other hands, we have applied the results defined in (9) and (10), which assumed covariates were multivariate normal, when the covariates do not follow this distribution. Again the results derived in (9) and (10) gave accurate answers. We consider five dimensional multivariate uniform and t-variables when only two covariates were fitted.

In fact, we can see clearly that, both the low probability of outcome *Y* and the sample size have affected on the estimated value of parameters. In the case of the large sample size, the standard error will be close to zero and the ratio close to one. In contrast, the standard error will be increases and the ratio will be appear faraway from one in some cases of small sample size. The results showed that for these symmetric non-normal variables, the violation of the assumption of normality made little difference. Some discrepant were noticed when the value of coefficients were very small close to zero.

REFERENCES

[1] Arnold, B.C. and Beaver, R.J. (2000). Hidden Trunction Models. The Indian Journal of Statistics, 62, 23-35.

[2] Azzalini, A. (2005). The Skew-normal Distribution and Related Multivariate Families. Journal of Statistics, 32, 159-188.

[3] Azzalini, A. (1985). A Class of Distribution Which Includes the Normal Ones. Scan- dinavian Journal of Statistics, 12, 171-178.

[4] Azzalini, A. (1986). Further Results on A Class of Distribution Which includes the Normal Ones. Statistica, 46, 199-208.

[5] Badi, N.H.S. (2017). Properties of the Maximum Likelihood Estimates and Bias Reduction for Logistic Regression Model. Open Access Library Journal, 4, e3625.

[6] Cox, D. R. (1970). Analysis of Binary data. Chapman and Hall, London.

[7] Cox, D. R. and Snell, E. J. (1989). Analysis of Binary data. Chapman and Hall, New York.

[8] Chiogna, M. (2004). Ma, Y. and Genton, M. G. Scandinavian Journal of Statistics, 31, 459-468.

[9] Claeskens, G. and Hjort, N. L. (2008). Model Selection and Model Averaging. Cambridge University Press.

[10] McCullagh, P. and Nelder, J. A. (1989). Generalized Linear Models. Chapman and Hall, London, 2nd.

[11] Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized Linear Models. Journal of the Royal Statistical Society, series A, 135, 379-384.

[12] Loperfido, N. (2001). Quadratic Forms of Skew-normal Random Vectors. Statistics and Probability Letters, 54, 381-387.

[13] Johnson, N. L. and Kotz, S. (1970). Continuous Univariate Distributions. Wiley- Interscience Publication, U.S.A.

[14] Matthews, J. N. S. and Badi, N. H. (2015). Inconsistent treatment estimates from mis-specified logistic regression analyses of randomized trials. Statistics in Medicine, 34, 2681-2694.

[15] Gumbel, E. J. (1961). Bivariate logistic distributions. Journal of the American Statistical Association, 56, 335-349.

[16] Hill, Henze, N. (1986). A Probabilistic Representation of the Skew-normal Distribution. Scandinavian Journal of Statistics, 13, 271-275.

[17] Hill, M. A. and Dixon, W. J. (1982). Robustness in Real Life. Biometrics, 38, 377-396.