Scholars Journal of Physics, Mathematics and Statistics

OPEN ACCESS

Abbreviated Key Title: Sch J Phys Math Stat
ISSN 2393-8056 (Print) | ISSN 2393-8064 (Online)
Journal homepage: https://saspublishers.com

The Appropriate Sample Size to Avoid the Biasness of the Parameters Estimation in Binary Logistic Regression Models

Jalal A. Moaiti[1], Radi A. Othman[1*]

[1]Department of Statistics, Faculty of arts and Science - Al Maraj, University of Benghazi, Benghazi, Libya

DOI: https://doi.org/10.36347/sjpms.2024.v11i12.001 | **Received:** 27.10.2024 | **Accepted:** 03.12.2024 | **Published:** 10.12.2024

***Corresponding author:** Radi A. Othman
Department of Statistics, Faculty of arts and Science - Al Maraj, University of Benghazi, Benghazi, Libya

Abstract

Original Research Article

In general, the maximum likelihood method with Newton-Raphson iteration is used to estimate the parameters of the binary logistic regression models, and also can be estimated using an iterative (re-)weighted least squares (IWLS) whose iterations are equivalent to the Newton-Raphson iterations. However, it is known that these methods produce bias estimates if small sample sizes are used. The main aim of this paper is to determine the appropriate sample size to achieve the unbiasedness of parameters estimates in the binary logistic regression model. To investigate the appropriate sample size three models were suggested and generated with known parameters. The estimates of the suggested models were collected via simulation study, using different sample sizes. The expected values for the collected parameters by the simulation where compared with the actual values of parameters of the suggested models. From the results of the simulation study, it was found that the appropriate sample size to achieve the unbiasness for such models, is 80 or more.

Keywords: Logistic regression, Maximum likelihood estimation, Iterative (re-)weighted least squares, Convergence problems, Bias reduction technique.

1. INTRODUCTION

Great deal of practical data in the medical sciences, social sciences, and other fields need to model binary response variables for which the response outcomes are success or failure. For example, one might be interested in modeling the existence of certain diseases. In this case, the response variable would take the value of 1 if the disease exists and 0 if not. One of the statistical models that can be used to deal with binary response data is the binary logistic regression model. Binary logistic regression models can be used to study relationship between multiple explanatory variables and a single binary response variable or a categorical variable with two categories. In general, to estimate the parameters of the binary logistic regression models, the maximum likelihood parameter estimation method with Newton-Raphson iteration is used [1]. Furthermore, the parameters of binary logistic regression models can be estimated using an iterative (re-) weighted least squares (IWLS) solver. It is straightforward to prove that the Newton-Raphson iterations are equivalent to (IWLS) iterations [2]. These methods are not convergent if the sample size is small and the proportion of success events is small [3]. If the result of parameter estimation through the iteration is not convergent, indicate that the model formed is no suitable for the data being analyzed [4]. Since the logistic model is widely used in the medical sciences, social sciences. The subject of studying the behavior of the maximum likelihood estimates for logistic regression model is very important. Therefore, we need to solve the un-convergence problem. There are many discussions on the un-convergence problem in logistic regression model like [5]. Also, there is a study about the bias reduction of the estimates like [6], and for the effects of the sample size see [7]. One way to resolve this problem is using the score function modification, modification on score function discovered by Firth [8]. Many studies have used this technique like [3, 9]. These studies concentrate to evaluate the behavior and properties of the bias reduction method using the score function modification by simulated data with different sample sizes and parameters.

However, it is well known that the maximum likelihood estimates are asymptotically unbiased, which results in a bias for small samples [8]. Therefore, the convergence or (bias) problem is related to the sample size, so the purpose of this research is to determine the appropriate or optimal sample size to avoid the problem of convergence and obtain the unbiased Parameter estimates of the binary logistic regression model.

2. BINARY LOGISTIC REGRESSION MODEL

Binary logistic regression is an existing causes and effects analysis for such binary response variables for which the response outcomes are success or failure. The binary random response can be defined as,

$$y = \begin{Bmatrix} 1 \; if \; the \; outcome \; is \; success \\ 0 \; if \; the \; outcome \; is \; failure \end{Bmatrix}.$$

The above binary random response can be considered as a bernoulli random variable with probability of success p and probability of failure $(1 - p)$. Then, the sum of the responses over a sample n will have a binomial distribution. The general form for the multiple logistic regression model can be written as [10],

$$\boldsymbol{logit\ (p) = \log\left(\frac{p}{1-p}\right) = X\beta}$$

The right-hand side of the above equation is called the systematic component, where X is a (n x k) matrix and β is a (k x 1) vector of parameters.

The term $logit\ (p) = \log\left(\frac{p}{1-p}\right)$ is a logit transformation from probabilities to a continuous random response, and it is called the link function. From the above equation, we can write the probability of success vector as,

$$\boldsymbol{p = \frac{e^{X\beta}}{1 + e^{X\beta}}}$$

3. THE MAXIMUM LIKELIHOOD METHOD TO ESTIMATE THE PARAMETERS OF THE MODEL

Let $y_1, \dots, y_n$ be independent random variables such that y_i is the number of successes in the group or class i n_i is the number of trials in the i^{th} class, and p_i is the probability of success in the i^{th} class. In this case y_i will have a binomial distribution with parameters (n_i, p_i). The likelihood function for the i^{th} observation can be written [10],

$$\boldsymbol{l(p_i; y_i) = \binom{n_i}{y_i} p_i^{y_i}(1 - p_i)^{n_i - y_i}}$$

For the independent observations, the likelihood function will be,

$$\boldsymbol{L(P; Y) = \prod_{i=1}^{m} \binom{n_i}{y_i} p_i^{y_i}(1 - p_i)^{n_i - y_i}}$$

Therefore, the log likelihood function is,

$$\boldsymbol{L(P; Y) = \sum_{i=1}^{m} \left[log \binom{n_i}{y_i} + y_i \log\left(\frac{p_i}{1 - p_i}\right) + n_i \log(1 - p_i) \right]}$$

We can write the log likelihood function in terms of the $X_{i's}$ and $\beta_{i's}$ as follows,

$$\boldsymbol{L(\beta; Y) = \sum_{i=1}^{m} \left[log \binom{n_i}{y_i} + y_i \sum_{j=1}^{k} X_{ij}\beta_j - n_i log \left(1 + \exp \sum_{j=1}^{k} X_{ij}\beta_j \right) \right]}$$

To find the estimators for the coefficients $\underset{\sim}{\boldsymbol{\beta}}$, we derive the log likelihood function with respect to $\underset{\sim}{\boldsymbol{\beta}}$ and maximize this function. First the derivative of the log likelihood function with respect to p_i is,

$$\boldsymbol{\frac{\partial L}{\partial p_i} = \sum_{i=1}^{m} \frac{y_i - n_i p_i}{p_i(1 - p_i)}}$$

Using the relation $p_i = \frac{e^{\beta_0 + \beta_1 X_{i1} + \dots + \beta_k X_{ik}}}{1 + e^{\beta_0 + \beta_1 X_{i1} + \dots + \beta_k X_{ik}}}$, we can find $\frac{\partial p_i}{\partial \beta_j}$

By applying the chain rule, we can find the derivative of the log likelihood function with respect to β_j as follows,

$$\boldsymbol{\frac{\partial L}{\partial \beta_j} = \frac{\partial L}{\partial p_i}\frac{\partial p_i}{\partial \beta_j}}$$

Where

$$\boldsymbol{\frac{\partial L}{\partial p_j} = \sum_{i=1}^{m} \frac{y_i - n_i p_i}{p_i(1 - p_i)}}$$

And

$$\frac{\partial p_i}{\partial \beta_j} = \frac{e^{\beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik}}}{[1 + e^{\beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik}}]^2} X_{ij}$$

Therefore, the maximum likelihood estimator for the j^{th} coefficient can be obtained by solving the following score equations,

$$\frac{\partial L}{\partial \beta_j} = \sum_{i=1}^{m} \frac{y_i - n_i p_i}{p_i(1-p_i)} \frac{\partial p_i}{\partial \beta_j} = 0$$

We are not going to be able to set this to zero and solve exactly. (That's a transcendental equation, and there is no closed-form solution). We can however approximately solve it numerically.

The above equation can be reduced in vectors notation, over all parameters as,

$$\frac{\partial L}{\partial \beta} = X^T(Y - \mu)\,, \qquad \mu = np$$

The fisher information for β is,

$$-E(\frac{\partial^2 L}{\partial \beta_r\, \partial \beta_s}) = \sum_i \frac{n_i}{p_i(1-p_i)} \frac{\partial p_i}{\partial \beta_r} \frac{\partial p_i}{\partial \beta_s}$$

$$= \frac{n_i}{p_i(1-p_i)} (\frac{\partial p_i}{\partial \eta_i})^2 x_{ir} x_{is}\,, \qquad \eta_i = \sum_{j=1}^{k} x_{ij}\beta_j = \{X^T W X\}_{rs}$$

Where W is a diagonal matrix of weights giving by,

$$W = diag\left\{n_i\, (\frac{\partial p_i}{\partial \eta_i})^2 / \, p_i(1-p_i)\right\}$$

And it can be reduced to,

$$W = diag[n_i p_i(1-p_i)]$$

The response variable,

$$\eta_i = log\left(\frac{p_i}{1-p_i}\right) = \sum_{j=1}^{k} x_{ij}\beta_j$$

Can be approximated, using first order Taylor series approximation as,

$$z_i = \hat{\eta}_i + \frac{y_i - n_i \hat{p}_i}{n_i} \frac{\partial \eta_i}{\partial p_i}$$

In matrix notation, the above quantities being computed at the initial estimate $\hat{\beta}_0$ can maximize the likelihood function if,

$$X^T W X \hat{\beta} = X^T W Z$$

Which can be solved iteratively using iterative (re-)weighted least squares (IWLS) method,

$$\hat{\beta} = (X^T W X)^{-1} X^T W Z$$

However, due to the approximation on the response vector Z,

$$E(Z) \neq X\beta \;\rightarrow\; E(\hat{\beta}) \neq \beta$$

In other words, the iteration process results in biased estimates, especially if the sample size is small.

4. SIMULATION STUDY

Simulation studies were conducted to investigate the appropriate sample size to avoid the problem of convergence and obtain the unbiased parameter estimates of the binary logistic regression model. The simulation plan used are based on three binary logistic models were generated according to following model equations.

- First model: $\boldsymbol{logit = 2 + 0.58\, x}$ $(i.e\ \beta_0 = 2\ and\ \beta_1 = 0.58)$
- Second model: $\boldsymbol{logit = 0.5 + x}$ $(i.e\ \beta_0 = 0.5\ and\ \beta_1 = 1)$

- Third model: $\boldsymbol{logit = 1 - 1.2\,x}$ $(i.e\ \beta_0 = 1\ and\ \beta_1 = -1.2)$

The response variable, which takes the value 0 or 1, was generated using a random bernoulli variable, simulation data for the three models were generated using different sample sizes, the smallest sample $n = 10$ and the largest sample $n = 10{,}000$. For each sample size we perform 10,000 simulations. We will examine the precision of the estimation for these models with different sample sizes by compare expected values of parameters $E(\beta)$ obtained by simulation for (IWLS) method with known values of parameters β to each the true model.

5. RESULTS AND DISCUSSION

A simulation study conducted to observe the behavior of the bias of (IWLS) parameter estimates in binary logistic regression for different sample sizes. The results of 10,000 simulations for the three true models were as follows.

Table 1 shows results for true first model, $\boldsymbol{logit = 2 + 0.58\,x}$ $(i.e\ \beta_0 = 2\ and\ \beta_1 = 0.58)$.

Table 1: Results the simulated first model

Sample size	$E(\beta_0)$	$E(\beta_1)$
10	10.74	3.112
20	3.150	0.901
40	2.284	0.671
80	2.120	0.615
160	2.047	0.589
320	2.025	0.588
640	2.011	0.583
1280	2.009	0.582
2560	2.006	0.581
5120	2.001	0.580
10000	2.000	0.580

Table 2 shows results for true second model, $\boldsymbol{logit = 0.5 + x}$ $(i.e\ \beta_0 = 0.5\ and\ \beta_1 = 1)$.

Table 2: Results the simulated second model

Sample size	$E(\beta_0)$	$E(\beta_1)$
10	3.135	5.287
20	0.813	1.617
40	0.571	1.142
80	0.510	1.049
160	0.520	1.029
320	0.505	1.015
640	0.505	1.007
1280	0.502	1.006
2560	0.499	0.999
5120	0.500	1.000
10000	0.500	1.000

Table 3 shows results for true third model, $\boldsymbol{logit = 1 - 1.2\,x}$ $(i.e\ \beta_0 = 1\ and\ \beta_1 = -1.2)$.

Table 3: Results the simulated third model

Sample size	$E(\beta_0)$	$E(\beta_1)$
10	5.300	-6.210
20	1.751	-2.062
40	1.147	-1.380
80	1.055	-1.273
160	1.010	-1.227
320	1.001	-1.213
640	1.010	-1.215
1280	1.007	-1.208
2560	1.004	-1.202
5120	1.000	-1.201
10000	0.999	-1.200

Results presented above give biased estimates (convergence problems) for known values of parameters when the sample sizes ($n = 10$,$n = 20$, and $n = 40$) under the three true models. Therefore,

$$E(\boldsymbol{\beta_i}) \neq \boldsymbol{\beta_i}\ , i = 0,1$$

But when sample sizes ($n \geq 80$), the expected values of parameters $E(\beta)$ are unbiased estimates for known values of parameters β. Therefore,

$$E(\boldsymbol{\beta_i}) \cong \boldsymbol{\beta_i}\ , i = 0,1$$

6. CONCLUSION

The simulation study shows that the empirical results of computations to estimate the parameters of the logistic regression model, Clearly the bias of the parameter estimates obtained from the (IRLS) method depends on the sample size, if the sample size less than 80 we found convergence problem, unless the sample sizes are equal to 80 or more. However, from this study we can conclude that,

- If sample size (n ≥ 80), the estimate values obtained from the (IWLS) method in the binary logistic regression model are approximate the values of actual parameters (unbiased estimates).
- If sample size (n < 80), the estimate values obtained from the (IWLS) method in the binary logistic regression model are not approximate the values of actual parameters (biased estimates).

REFERENCES

1. Larsen, P. V. (2003). In All Likelihood: Statistical Modelling and Inference Using Likelihood. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *52*(3), 416-417. doi: 10.1111/1467-9884.00369_20.
2. Hastie, T., Tibshirani, R., & Friedman, J. (2009). Springer Series in Statistics The Elements of Statistical Learning Data Mining, Inference, and Prediction, 2nd edition., Springer, New York, NY, ch. 4.1.
3. Badi, N. H. S. (2017). Properties of the Maximum Likelihood Estimates and Bias Reduction for Logistic Regression Model. *OAlib*, *4*(5), 1-12. doi: 10.4236/oalib.1103625.
4. Czepiel, S. A. Maximum Likelihood Estimation of Logistic Regression Models: Theory and Implementation. [Online]. Available: http://czep.net/contact.html
5. Cox, D. R., & Hinkley, D. V. (1974). *Theoretical Statistics*. Chapman and Hall, London. doi: 10.1007/978-1-4899-2887-0.
6. Anderson, J. A., & Richardson, S. C. Logistic Discrimination and Bias Correction in Maximum Likelihood Estimation. *Technometrics*, *21*(1), 71-78, 1979, doi: 10.2307/1268582.
7. McCullagh, P. (1986). The conditional distribution of goodness-of-fit statistics for discrete data. *Journal of the American Statistical Association*, *81*(393), 104-107. doi: 10.1080/01621459.1986.10478244.
8. Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, *80*(1), 27-38. Available: http://www.jstor.orgURL:http://www.jstor.org/stable/2336755Accessed:22/05/200814:38
9. Febrianti, R., Widyaningsih, Y., & Soemartojo, S. (2021). The parameter estimation of logistic regression with maximum likelihood method and score function modification. In *Journal of physics: Conference series* (Vol. 1725, No. 1, p. 012014). IOP Publishing. doi: 10.1088/1742-6596/1725/1/012014.
10. McCullagh, P. N. (1989). *Generalized linear models*, 2nd Edition. Chapman and Hall, London, 1989.