Journal of Mathematics and Statistics 8 (4): 481-488, 2012
ISSN 1549-3644

doi:10.3844/jmssp.2012.481.488 Published Online 8 (4) 2012 (http://www.thescipub.com/jmss.toc)

Parameter Estimation and Determination of Sample Size in Logistic Regression

[1]Adeleke Kazeem Adedayo and [2]Dawud Adebayo Agunbiade

[1]Department of Mathematics, Faculty of Science, Obafemi Awolowo University Ile-Ife, Osun State, Nigeria
[2]Department of Mathematical Sciences, Faculty of Science, Olabisi Onabanjo University, Ago-Iwoye, Nigeria

Received 2012-06-26, Revised 2012-10-02; Accepted 2012-10-02

ABSTRACT

The determination of Sample-Size is often an important step in planning a statistical study and it is usually a difficult task. Among the important hurdles to be surpassed, one must obtain an estimate of one or more error variances and specify an effective sample size of importance. The study was carried out to check for the estimation of parameter and sample sizes in logistic regression, although, there is the temptation to take some shortcuts. We to looked at two methods of obtaining sample sizes having obtain the parameter estimates by varying the response probability. The results from the real life data showed that when the response probabilities are small, an approximation of corrected term Equation 12 performs better than the approximation Equation 8, but it highly over estimates when the response probabilities are large.

Keywords: Logistic Regression, Sample Size, Power, Proportional Odds Model

1. INTRODUCTION

Logistic regression is a type of regression used when the dependent variables are categorical Adeleke and Adepoju (2010). The dependent variable may have two categories (e.g., alive/dead; male/female; Republican/Democrat) or more than two categories. If it has more than two categories they may be ordered or unordered. However, a lot of statistics is concerned with predicting the value of a continuous variable like Blood pressure, intelligence, oxygen levels, wealth and so on. But this kind of statistics dominates when your response variable is binary. It is highly robust and the independent variables do not have to be normally distributed, or have equal variance in each group. Logistic regression is useful in some situations when assumptions of linear regression fail. It requires a different type of data and its coefficient have different interpretations. Like linear regression, logistic regression allows results to be graphed with regression lines and prediction to be made given a set of conditions. In this study, our interest focuses on the determination and parameter estimation of sample size using logistic regression analysis. Literature reviews have shown many studies aimed at determining whether a particular variable has an effect on a binary response. Agresti (2007) argued that the study design should determine the sample size needed to provide a good chance of detecting an effect of a given size. He used simple logistic regression as a case study. His study did not provide much result for the multiple logistic regressions. This study therefore considers a thorough analysis on the multiple cases that enhances better approach to sample size determination. We begin by given background information on the related terms like power analysis.

Power analysis can optimize the resource usage and design of a study, improving chances of conclusive results with maximum efficiency. Power analysis is the most effective when performed at the study planning stage and as such it encourages early collaboration between researcher and statistician. Muller and Benignus (1992); O'Brien and Muller (1993) and Russell (2001), provide cogent discussions of these and related concepts.

Power analysis is often problematic in practice, being performed infrequently or improperly. There are several reasons for this: it is technically complicated, usually under-represented in statistical curricula and often not perform early enough to be effective. Good software tools for power analysis can alleviate these difficulties and help you to benefit from these techniques.

Corresponding Author: Adeleke Kazeem Adedayo, Department of Mathematics, Faculty of Science, Obafemi Awolowo University Ile-Ife, Osun State, Nigeria

We propose to develop sample size calculation methods within the proportional odds model structure. Such a sample size is needed to construct a test of hypothesis in Ordinal Logistic Regression (OLR) having desired power. The use of logistic regression has widely been accepted in scientific fields (biostatistics, epidemiology, engineering). This is because it is a simple and effective method to describe the effect of some explanatory variables on a categorical response variable.

Studies on parameter estimation in logistic regression revealed that the power and sample size estimation of different statistical approach within logistic regression model. Whittemore (1989) considered a test for a single parameter with other parameters treated as nuisance parameters. Much literature exists on approximations to the power and sample size of different statistical tests within logistic regression model (Mehta and Tsiatis, 1984; Hilton and Mehta 1993; Lui, 1993). Whittemore (1989) considered sample size approximations in the case of standard logistic regression with small response probability. At present, sample size issues in ordinal logistic regression setting do not appear to have been studied in depth in the literature. Sample size determination in multilevel designs requires attention to the fact that statistical power depends on the total sample sizes for each level. It is usually desirable to have as many units as possible at the top level of the multilevel hierarchy (Snijders, 2005). Russell (2001) offers some suggestions for successful and meaningful sample-size determination and also discussed is the possibility that sample size may not be the main issue; that the real goal is to design a high-quality study. Lin *et al*. (2010) discussed some crucial issues in the problem formulation, parameter specifications and approaches that are commonly proposed for sample size estimation in microarray experiments. Roy *et al*. (2007) consider the problem of sample size determination for three-level mixed-effects linear regression models for the analysis of clustered longitudinal data. Three-level designs are used in many areas, but in particular, multicenter randomized longitudinal clinical trials in medical or health-related research.

Power analysis most effective when performed at the study planning stage and as such it encourages early collaboration between researcher and statistician. It also focuses attention on effect sizes and variability in the underlying scientific process, concept that both researcher and statistician should consider carefully at this stage. Muller and Benignus (1992) and O'Brien and Muller (1993) provide cogent discussions of these and related concepts. These references also provide a good general introduction to power analysis. Our focuses in this study is therefore to fit a suitable model and check the reliability of the model using logistic regression and to suggest sample size and power calculation methods for ordinal logistic regression to test statistical hypothesis.

2. MATERIALS AND METHODS

We deal with studies in which a random samples is drawn from the joint distribution of (Y, X) where Y is an ordinal response and X= (x_1, x_2, x_3,…, x_p) is a vector of covariates Equation 1:

$$\text{Let} \prod\left[\prod_1(x'),\prod_2(x'),\prod_3(x'),...,\prod_k(x')\right] \quad \text{to the predictor } X' \tag{1}$$

Since our response categories have a natural ordering, we use the proportional odds model that is Equation 2:

$$\text{logit}[P_r(Y \le j/X)] = a_j + \gamma' X' \quad j = 1,2,...k-1 \tag{2}$$

where, a is a vector of the intercept parameters and γ' = (γ_1, γ_2,…, γ_p) is the slope parameter vector without intercept term. If $a_j < a_{j+1}$ holds this model fits a common slope cumulative model based on cumulative probabilities of the response categories Equation 3 and 4:

$$\text{Let } \varphi_j(X') = \prod_1(X') + \prod_2(x') + ... + \prod_k(x') \tag{3}$$

$$\left.\begin{aligned} \varphi_1(X') &= \prod_1(X') \\ \varphi_2(X') &= \prod_1(X') + \prod_2(X') \\ &\vdots \\ &\vdots \end{aligned}\right\} \tag{4}$$

$$\varphi_j(X') = \prod_1(X') + \prod_2(X') + \prod_3(X') + ... + \prod_k(x') = 1$$

The OLR follows that Equation 5 and 6:

$$\left.\begin{aligned} &\text{Logit}(\varphi_i) = \log\left(\frac{\varphi_i}{1-\varphi_i}\right) \\ &= a_j + \gamma_1 X_1 + \gamma_2 X_2 + ... + \gamma_p X_p \\ &\text{Logit}(\varphi_1) = \log\left(\frac{\varphi_1}{1-\varphi_1}\right) \\ &= a_1 + \gamma_1 X_1 + \gamma_2 X_2 + ... + \gamma_p X_p \\ &\text{Logit}(\varphi_2) = \log\left(\frac{\varphi_2}{1-\varphi_2}\right) \\ &= a_2 + \gamma_1 X_1 + \gamma_2 X_2 + ... + \gamma_p X_p \\ &\quad : \; : \; : \\ &\quad . \quad . \quad . \\ &\text{Logit}(\varphi_{k-1}) = \log\left(\frac{\varphi_{k-1}}{1-\varphi_{k-1}}\right) \\ &= a_{k-1} + \gamma_1 X_1 + \gamma_2 X_2 + ... + \gamma_p X_p \end{aligned}\right\} \tag{5}$$

Where:

$$\varphi_j(X') = \prod_1(X') + \prod_2(X') + \prod_3(X') + \ldots + \prod_k(X') = \frac{e^{a_j+\gamma}'X'}{1+e^{a_j+\gamma'}X'} \quad (6)$$

This model is known as Proportional odds Model because, the odds ratio of the event (Y≤ j) is independent of category indication.

2.1. Maximum Likelihood Function

When more observation on Y occurs at a fixed X^t value, it is sufficient to record the number of observations n_j^t and the number of j outcome, for j = 1,…, k.

Thus we let Y^t, t=1,…, n, be an independent multinomial random (response) variable, then Y^t is ~ multinomial $n_1^t,\ldots,n_k^t$ with $E(Y^t) = n_j^t\varphi_j(X^t)$ Where Equation 7:

$$n_1^t + \ldots + n_k^t = 1$$

We define:

$$\left.\begin{array}{c} S_1^t = n_1^t \\ S_2^t = n_1^t + n_2^t \\ \vdots \quad \vdots \quad \vdots \\ S_k^t = n_1^t + n_2^t + \ldots + n_k^t = 1 \end{array}\right\} \quad (7)$$

Since we are dealing with cumulative probabilities, in term of the parameters of the cumulative transformations, the likelihood can be written as the product of k-1 quantities. The joint probability mass function of $(Y_1,\ldots Y_n)$ is proportional to the product of multinomial functions.

For a given sample size n, the likelihood of the observations y^t, x^t, t = 1,2,…, n is:

$$L(\underline{a}',\underline{\gamma}') = \prod_{t=1}^{n} f(x^t) f(y^t / x^t)$$

$$\prod_{t=1}^{n}\left\{\left(\frac{\varphi_1^t}{\varphi_2^t}\right)^{s_1^t}\left(\frac{\varphi_2^t-\varphi_1^t}{\varphi_2^t}\right)^{s_2^t-s_1^t}\right\}*$$

$$\left\{\left(\frac{\varphi_2^t}{\varphi_3^t}\right)^{s_1^t}\left(\frac{\varphi_3^t-\varphi_2^t}{\varphi_3^t}\right)^{s_3^t-s_2^t}\right\}*\ldots*$$

$$\left\{\left(\frac{\varphi_{k-1}^t}{\varphi_k^t}\right)^{s_1^t}\left(\frac{\varphi_k^t-\varphi_{k-1}^t}{\varphi_k^t}\right)^{s_k^t-s_{k-1}^t}\right\}$$

where, F(x)is a joint p.d. of x. it is assumed that f (x) does note depend on unknown parameters $(\underline{a}',\underline{\gamma}')$.

The validity of this model shows that MLE $(\underline{\hat{a}}',\underline{\hat{\gamma}}')$ satisfy approximately $(\underline{\tilde{a}}',\underline{\hat{\gamma}}) \sim N[(\underline{\hat{a}}',\underline{\tilde{\gamma}}'),I^{-1}(\underline{a}',\underline{\gamma}')]$.

2.2. Sample Size Estimation

One of the main objectives of this write up is estimation of sample size and this is achieved by obtain a sample size that is just sufficiently large enough to be confidence of being able to achieve an inference with required precision. It is directly related to the cost and time involved in a survey or data collection.

Let us test the null hypothesis:

$$H_0 : \gamma = 0 \text{ Vs } H_a : \gamma = \tilde{\gamma}$$

At a level with power ≥1-β when the distribution of $\hat{\gamma}$ is treated as normal with mean γ and variance σ^2, the critical region is:

$$\begin{cases} \hat{\gamma} > Z_a\left(\frac{\sigma_0}{\sqrt{n}}\right), & \text{if } \tilde{\gamma} > 0 \\ \hat{\gamma} < -Z_a\left(\frac{\sigma_0}{\sqrt{n}}\right), & \text{if } \tilde{\gamma} < 0 \end{cases}$$

where, Z_a is 100(1-a)% of the standard normal distribution. The sample size n will be found so that the test has a specified power (1-β) at the alternative $H_a : \gamma = \tilde{\gamma}$, the sample size n is thus chosen so that:

$$\Pr\left(\hat{\gamma} > \frac{z_a\left(\frac{\sigma_0}{\sqrt{n}}\right)}{\gamma} = \hat{\gamma} > 0, \sigma = \sigma_0\right) = (1-\beta)$$

$$\text{or } \Pr\left(\hat{\gamma} < -\frac{z_a\left(\frac{\sigma_0}{\sqrt{n}}\right)}{\gamma} = \hat{\gamma} < 0, \sigma = \sigma_0\right) = (1-\beta)$$

This can be written as $1-\varphi\left(Z_a\frac{\sigma_0}{\sigma_a} - \tilde{\gamma}/\frac{\sigma_a}{\sqrt{n}}\right) = (1-\beta)$ if $\tilde{\gamma} < 0$:

$$\varphi\left(-Z_a\frac{\sigma_0}{\sigma_a} - \tilde{\gamma}/\frac{\sigma_a}{\sqrt{n}} = \right) = (1-\beta) \text{ if } \tilde{\gamma} > 0$$

If $\tilde{\gamma} > 0$ and an appropriate n, a solution of the above formula satisfies:

$$Z_a \frac{\sigma_o}{\sigma_a} - \frac{\tilde{\gamma}}{\frac{\sigma_a}{\sqrt{n}}} = Z_\beta$$

Otherwise if $\tilde{\gamma} < 0$:

$$-Z_a \frac{\sigma_0}{\sigma_a} - \frac{\tilde{\gamma}}{\frac{\sigma_a}{\sqrt{n}}} = Z_\beta$$

Hence:

$$n = \frac{z_a\sigma_o + z_\beta\sigma_a}{\tilde{\gamma}^2} \quad (8)$$

For both cases $\tilde{\gamma} > 0$ and $\tilde{\gamma} < 0$

For model of the form in Equation (2), (5) and (6) with one predictor i.e.:

$$\log it(\pi) = a + \gamma_1 X$$

Hsieh (1989) uses an approximate sample size formula to obtain the sample size needed for testing H_0: $\gamma = 0$. Here we need to guess the probability of success $\bar{\pi}$ at the mean of x. the size of this effect is the odds ratio θ comparing $\bar{\pi}$ to the probability of success one standard deviation above the mean of x. Let k = log(θ) An approximate sample size is Equation 9:

$$n \approx \left[Z_{\frac{a}{2}} + Z_{\frac{\beta}{2}} e^{\left(\frac{k^2}{4}\right)} \right]^2 (1 + 2\bar{\pi}\delta) / (\bar{\pi}k^2) \quad (9)$$

Where Equation 10:

$$\delta = \left[1 + (1 + k^2)\, e^{\left(\frac{sk^2}{4}\right)} \right] / [1 + e^{(-k^2/4)}] \quad (10)$$

In the case of Proportional Odds Model (POM), estimation of sample size with general response probabilities where we have more than two categories which can either small or large, then:

$$f(z) = \frac{z}{1+z} = \begin{cases} \sum_{i=1}^{\infty} (-1)^{i+1} z^i, \text{ if } 0 < z < 1 \\ \sum_{i=1}^{\infty} (-1)^{i+1} z^{-i}, \text{ if } 1 < z < \infty \end{cases}$$

And is simply approximated in Equation 6 by:

$$f(z) = \varphi_j(X') = \frac{e^{a_j + \gamma}'X'}{1 + e^{a_j + \gamma}'X'} =$$
$$\begin{cases} e^{a_j + \gamma'X'} + O(e^{2a_j}), \text{if } a_j + \gamma'X' \le 0 \\ e^{-(a_j + \gamma'x')} + O(e^{-2a_j}), \text{if } a_j + \gamma'X' \ge 0 \end{cases}$$

where, e^{aj} is small when $a_j + \gamma'X' \le 0$ (or e^{-aj} is small when $a_j + \gamma'X' \ge 0$).

We now prove for response variable with three categories with ordered probabilities. i.e.:

$$\frac{e^{a_{1+}\gamma'}x'}{1 + e^{a_{1+}\gamma'}X'} < 0.5 \text{ and } \frac{e^{a_{2+}\gamma'}X'}{1 + e^{a_{2+}\gamma'}X'} > 0.5$$

Then:

$$\frac{e^{a_{1+}\gamma'}x'}{(1 + e^{a_{1+}\gamma'}X')} = e^{a_{1+}\gamma'}X' + O(e^{2a_2})$$

And:

$$\frac{e^{a_2 + \gamma}'x'}{(1 + e^{a_2 + \gamma'}x')^2} = e^{-(a_2 + \gamma'X')} + O(e - 2a_2)$$

$$I_{11} \approx n\left\{E(e^{-(a_2 + \gamma'X')} - E(e^{a_1 + \gamma}'X')\right\} \frac{e^{a_1 + a_2}}{(e^{a_2} - e^{a_1})^2}$$

$$= n * \frac{e^{a_1 + a_2}}{(e^{a_2} - e^{a_1})^2} \left\{e^{-a_2} m(-\gamma') - e^{a_1} m(\gamma')\right\}$$

$$I_{12} \approx I_{11}, I_{13} = 0, I_{22} \approx I_{11} + ne^{-a_2} m(-\gamma')$$

$$I_{23} = ne^{-a_2} m_I(-\gamma') \text{ and } I_{33} = ne^{-a_2} m_{II}(-\gamma')$$

Therefore variance is Equation 11:

$$\text{Var}(\gamma') \approx \frac{m(-\gamma)}{ne^{-a_2}\{m(-\gamma)m_{II}(-\gamma') - m_I(-\gamma)\}} \quad (11)$$

If X'~N (0, 1) then $\text{Var}(\gamma') = \frac{e - \gamma_1^2}{ne^{-a_2}}$ and Equation 12:

$$n \ge e^{a_1} \frac{\left\{z_a\sigma_o + z_\beta e - \frac{\gamma_1^2}{4}\right\}^2}{\tilde{\gamma}_1^2} \quad (12)$$

Equation 11 discovered method of obtaining σ_a used in (3.0) which is $\sqrt{\text{Var}}(\gamma')$ obtained in Equation 11 above. Hence, we can generalize it to multiple parameters, where we test the hypothesis of:

$$H_0 : \underline{\gamma}_1 = 0 \text{ Vs } H_1 : \underline{\gamma}_1 = \tilde{\underline{\gamma}}_1$$

And let:

$$\underline{\gamma}' = (\gamma_1, \gamma_2, ..., \gamma_s) = (\underline{\gamma}_1', \underline{\gamma}_2')$$

Where:

$$\underline{\gamma}_1' = (\gamma_1, \gamma_2, ..., \gamma_p)$$

And:

$$\underline{\gamma}_2' = (\gamma_{p+1}, \gamma_{p+2}, ..., \gamma_s)$$

3. RESULTS

We illustrate by using the data on diabetics patients from a University College Hospital Ibadan. The data covers 10678 reported cases of patients with diabetes.

3.1. Table 1a. Logit Diabetes versus Smoking

Logistic regression:

Number of obs = 10678
LR chi2 (1) = 2.52
Prob > chi2 = 0.1126
Log likelihood = 7330.5489
Pseudo R^2 = 0.0002

Estimation of sample size using the method proposed by Hsieh (1989) in Equation 9.

Assume $\bar{\pi} = 0.817013$ if we go by the hypothesis that H_0: $\gamma_1 = 0$ against the alternative H_0: $\gamma_1 \neq 0$ from **Table 1a** then:

$$a = 0.05 \text{ and } \beta = 0.10, Z_{\frac{a}{2}} = 1.96 \text{ and } Z_{\frac{\beta}{2}} = 1.64$$

$$k = \log(\text{odds ratio}) = \log(0.93955205) = -0.02709 \text{ and}$$

$$k^2 = 0.0007$$

$$\delta = 1.0009 \text{ and } n \cong 56,946$$

If we now consider the effects of smoking and drinking of alcohol on induced diabetes patients i.e., 2 predictors, then the above can be seen in output of **Table 1b** above where both the coefficients having negative effect on induced diabetics patients. Although the Log-likelihood ratio for model selection support the full model of two (full model) predictors with 2.54 > 0.2803 value of chi^2 with 1 df.

Since the pseudo R^2 is 0.0002 which implies that there is hardly multiple correlation between the predictors and th response variable, the odds ratio in **Table (1a and 1b)** shows that for a smoker, there is approximate value of 6% less times of having diabetis when compared with those who are not smoking, given that all other variable remain constant. The odds of having diabetes for an individual addicted to alcohol is just 0.6% less times those who are not drinking alcohol. Although this results look somehow, but the p-values for smokers (0.113) and individual adicted to alcohol (0.872) are not significant meaning that both factors considered are not really contributing to diabetes problem. These surport the result obtained in R^2.

we compute:

$$n_2 = \frac{n_1}{1 - R^2}$$

where, n_1 is the n obtained when we have one predictor Hence:

$$n_2 \cong 56,946$$

Therefore, we require almost 57000 samples for testing H_0: $\gamma_1 = 0$.

Using the above information we have the following result from our simulation of sample size for both Equation 8 and 12 respectively. Monte Carlo method for selected values of $\alpha = 0.05$, $\beta = 0.1$ and $e^{\alpha 1} = 0.05$ as well as the value of $\gamma > 0$ & $k = 2$ when the explanatory variable has the standard normal distribution. The results in **Table 1** below show us that the approximation (3.0) is suitable when the response probabilities are small but it always under estimates.

4. DISCUSSION

Acording to the results of this study, the estimates of the parameters and sample sizes are obtained from both real life data of diabetes and simulation study, **Table 1 and 2**. Sample size obtained when the predictor is one is approximately the same when the the predictors are two using a real life data. The approximation with corrected term (3.4) performs better than the approximation (3.0) when the response probabilities are small, but it highly over estimates when the response probabilities are large. Also, the graphical representation of the sample sizes for the simulation is given in **Fig. 1-3**. Since the sample sizes depend on the two parameters, γ and α_1, simultaneously, we fixed one parameter to obtain the other. If we change the two parameters simultaneously, the estimated sample sizes fluctuated too much.

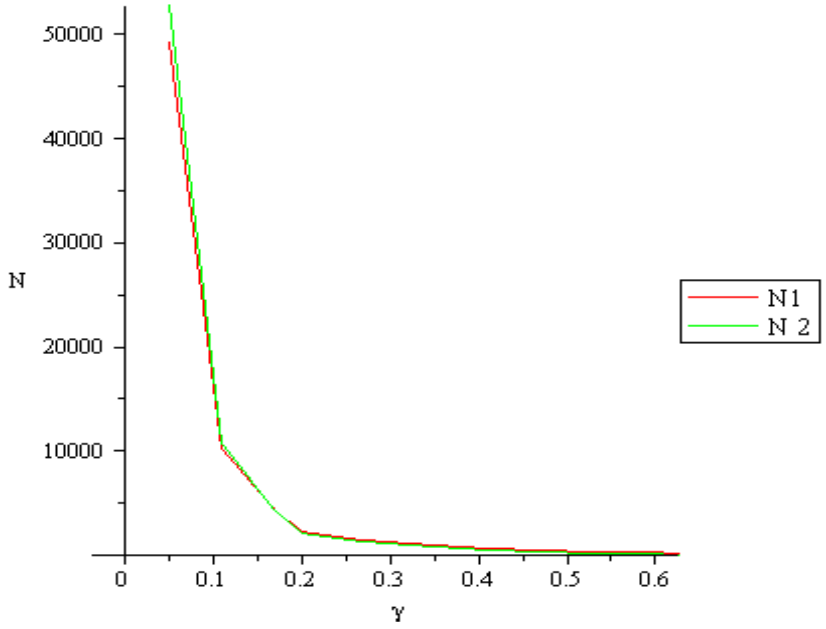

Fig. 1. The gaph of sample size fixed for k = 2, exp (α1) = 0.05

Table 1a. Logistic regression: Diabetes vs Smoking

Diabetes	Coef.	Odds ratio	Std. Err.	z	P>|z|
Smoking	-0.0624	0.9395	0.0393	-1.59	0.113
Cons	-0.2021		0.0258	-7.81	0.000
LR = 2.52, Prob> $(chi)^2$ = 0.2803					

Table 1b. Logistic regression: Diabetes versus Smoking and Alcohol

Diabetes	Coef	Odds Ratio	Std. Error	Z	P</z/
Smoking	-0.0624	-0.0062	-0.2021	0.9395	0.9937
Alcohol	0.0393	0.0389	0.0258	-1.59	-0.16
Cons	-7.81		0.113	0.872	0.00
LR=2.54, Prob>$(chi)^2$ = 0.2803					

Table 2. (Estimates of sample sizes for both equations (3.0) and (3.4))

(k = 2, $e^{\alpha 1}$ = 0.05)			(k = 2, $e^{\alpha 1}$ = 0.25)			(k = 2, $e^{\alpha 1}$ = 0.5)		
γ	N_1 (3.4)	N_2 (3.0)	γ	N_1 (3.4)	N_2 (3.0)	γ	N_1 (3.4)	N_2 (3.0)
0.05	52911.34	49377.2	0.05	15064.14	9827.57	0.05	10848.02	4933.84
0.11	10663.52	10135.88	0.11	3089.78	1977.30	0.11	2226.890	1009.71
0.17	4267.03	4195.42	0.17	1276.89	789.21	0.17	921.7000	415.67
0.2	2176.76	2254.37	0.20	684.48	401.00	0.20	495.1900	221.56
0.27	1243.02	1387.31	0.27	419.84	227.59	0.27	304.6300	134.86
0.35	746.88	926.55	0.35	279.21	135.44	0.35	203.3400	88.78
0.41	452.27	652.89	0.41	195.69	80.71	0.41	143.1600	61.41
0.47	263.27	477.25	0.47	142.09	45.58	0.47	104.5000	43.85
0.53	134.93	357.88	0.53	105.66	21.70	0.53	78.21000	31.91
0.59	120.99	273.12	0.59	79.79	4.75	0.59	59.51000	23.44
0.65	54.15	210.80	0.65	60.77	0.00	0.65	45.74000	17.20
0.71	3.70	163.68	0.71	46.38	0.00	0.71	35.31000	12.49

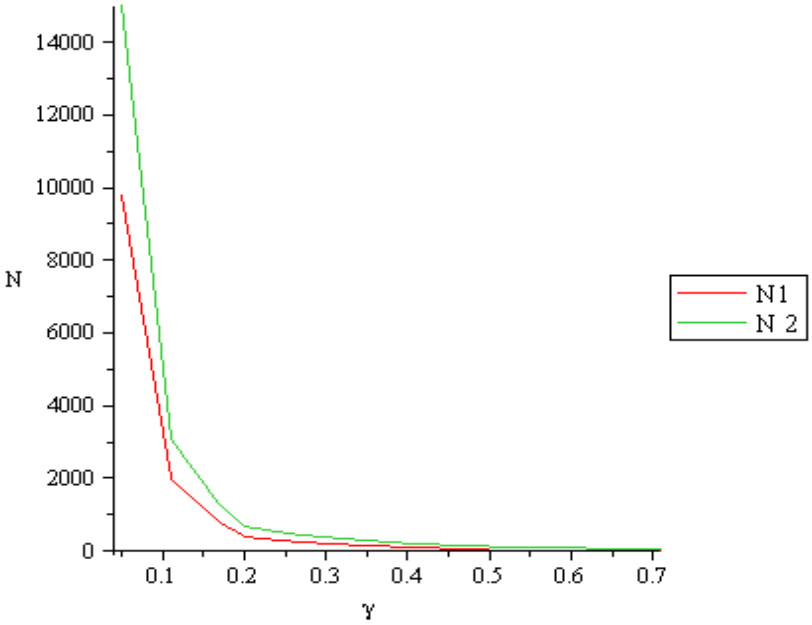

Fig. 2. The gaph of sample size fixed for k = 2, exp (α1) = 0.25

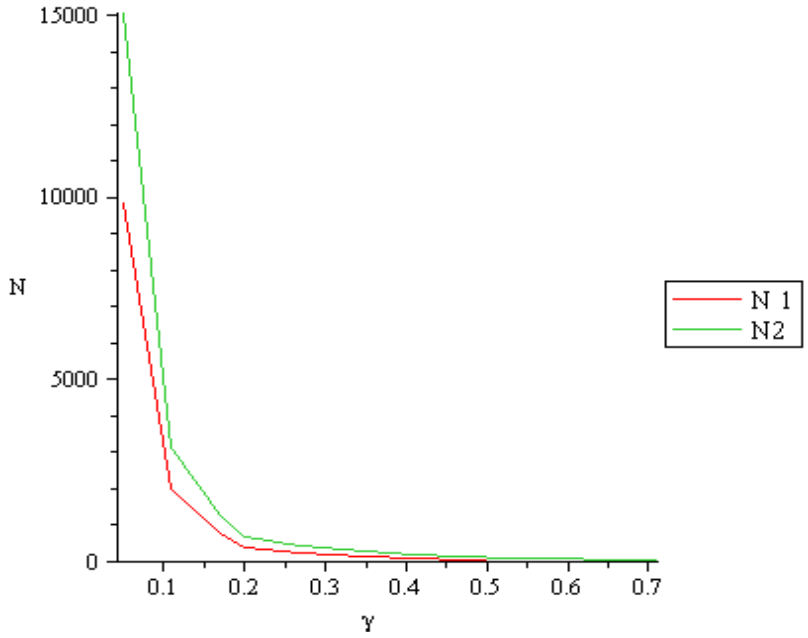

Fig. 3. The gaph of sample size fixed for k = 2, exp (α1) = 0.5

5. CONCLUSION

This study has developed a methodological framework to estimate the parameters of logistic regression and obtain sample sizes at different level of a and β. We have also proposed sample size calculation methods for logistic regression to tests for statistical hypotheses. We have also considered testing the multiple

parameters. We gave a simple closed-form formula for approximated sample sizes when the probabilities of the response categories are small. The results showed that an approximation of corrected term Equation 12 performs better than the approximation Equation 8 when the response probabilities are small, but it highly over estimates when the response probabilities are large.

4. REFERENCES

Adeleke, K.A. and A.A. Adepoju, 2010. Ordinal logistic regression model: An application to pregnancy outcomes. Am. J. Math. Stat., 6: 279-285.

Agresti, 2007. An Introduction to Categorical Data Analysis. 1st Edn., Wiley, New York, ISBN-10: 0471360937, pp: 75-160.

Hsieh, F.Y., 1989. Sample size tables for logistic regression. Stat. Med., 8: 795-802. DOI: 10.1002/sim.4780080704

Hilton, J.F. and C.R. Mehta, 1993. Power and sample size calculations for exact conditional tests with ordered categorical data. Biometrics, 49: 609-616. DOI: 10.2307/2532573

Lin, W.J. H.M. Hsueh and J.J. Chen, 2010. Power and sample size estimation in microarray studies. BMC Bioinform., 11: 48-48. DOI: 10.1186/1471-2105-11-48

Lui, 1993. Sample size determination for cohort studies under an exponential covariate model with grouped data. Int. Biometric Soc., 49: 773-778.

Mehta, P. and Tsiatis, 1984. Exact Significance testing to establish treatment equivalence with ordered categorical data. Biometrics, 40: 819-825. DOI: 10.2307/2530927

Muller, K.E. and V.A. Benignus, 1992. Increasing scientific power with statistical power. Neurotoxicol. Teratol., 14: 211-219. DOI: 10.1016/0892-0362(92)90019-7

O'Brien, R.G. and K.E. Muller, 1993. Unified Power Analysis for t-Tests Through Multivariate Hypotheses. In: Applied Analysis of Variance in Behavioral Science, Edwards, L.K. (Edn.), New York.

Roy, A., D.K. Bhaumik, S. Aryal and R.D. Gibbons, 2007. Sample size determination for hierarchical longitudinal designs with differential attrition rates. J. Int. Biometric Soc., 63: 699-707. DOI: 10.1111/j.1541-0420.2007.00769.x

Russell, V.L., 2001. Some practical guidelines for effective sample size determination. Am. Stat., 55: 187-193. DOI: 10.2307/2685797

Snijders, T.A.B., 2005. Power and sample size in multilevel linear models. Encyclopedia Stat. Behav. Sci., 3: 1570-1573. DOI: 10.1002/0470013192.bsa492

Whittemore, A.S., 1989. Sample size for logistic regression with small response probability. J. Am. Stat. Assoc., 76: 27-32. DOI: 10.1080/01621459.1981.10477597