PARAMETER ESTIMATION AND DETERMINATION OF SAMPLE SIZE IN LOGISTIC REGRESSION

Adeleke Kazeem Adedayo and Dawud Adebayo Agunbiade

1Department of Mathematics, Faculty of Science, Obafemi Awolowo University Ile-Ife, Osun State, Nigeria
2Department 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 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 focus 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 \(X=(x_1, x_2, x_3, \ldots, x_p)\) is a vector of covariates Equation 1:

\[
\begin{align*}
\text{Logit}(r_j) &= \log\left(\frac{\phi_j}{1-\phi_j}\right) \\
&= a_j + \beta_j x_1 + \beta_j x_2 + \ldots + \beta_j x_p
\end{align*}
\]

(3)

where, \(a_j < a_{j+1}\) holds this model fits a common slope cumulative model based on cumulative probabilities of the response categories Equation 3 and 4:

\[
\begin{align*}
\phi_j &= \prod_{i=1}^{j} (X') \\
\phi_j &= \prod_{i=1}^{j} (X') + \prod_{i=1}^{j} (X') + \cdots + \prod_{i=1}^{j} (X') = 1
\end{align*}
\]

(4)

Let \(\phi_j(X') = \prod_{i=1}^{j} (X') \) to the predictor \(X'\).

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

\[
\text{Logit}(r_j) = \log\left(\frac{\phi_j}{1-\phi_j}\right)
\]

(2)

(5)

The OLR follows that Equation 5 and 6:

\[
\begin{align*}
\text{Logit}(r_j) &= \log\left(\frac{\phi_j}{1-\phi_j}\right) \\
&= a_j + \beta_j x_1 + \beta_j x_2 + \cdots + \beta_j x_p
\end{align*}
\]

(5)

482
Where:
\[ \varphi_j(X) = \prod_i(X_i^j) + \prod_i(X_i^\prime) + \prod_i(X_i^\prime) + \ldots + \prod_i(X_i^\prime) = \frac{e^{\gamma_i^j} X_i}{1 + e^{\gamma_i^j} X_i} \]  

(6)

This model is known as Proportional odds Model because, the odds ratio of the event \((Y \leq j)\) 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 \(j = 1, \ldots, k\). Thus we let \(Y^t, t = 1, \ldots, n\) be an independent multinomial random (response) variable, then \(Y^t\) is \(\sim\) multinomial \(n_1^t, \ldots, n_k^t\) with \(E(Y^t) = n_\varphi(X^t)\) Where Equation 7:

\[ n_1^t + \ldots + n_k^t = 1 \]

We define:

\[ S_1 = n_1^t \]
\[ S_2 = n_1^t + n_2^t \]
\[ \vdots \]
\[ S_k = n_1^t + n_2^t + \ldots + n_k^t = 1 \]

(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 \(y_{1t}, x^t, t = 1, 2, \ldots, n\) is:

\[ L(a^t, \gamma) = \prod_{t=1}^{n} f(x) f(y^t / x^t) \]

\[ \prod_{i=1}^{k} \left( \varphi_i^t \varphi_i^\prime \right)^{y_i^t} \left( \varphi_i^t - \varphi_i^\prime \right)^{\gamma_i^t - y_i^t} \cdot \ldots \cdot \left( \varphi_i^t \varphi_i^\prime \right)^{y_i^t} \left( \varphi_i^t - \varphi_i^\prime \right)^{\gamma_i^t - y_i^t} \]

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

The validity of this model shows that MLE \((\hat{a}, \hat{\gamma})\) satisfy approximately \((\hat{a}, \hat{\gamma}) \sim N((a, \gamma), \Gamma^{-1}(a, \gamma))\).

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:

\[ H_0 : \gamma = 0 \quad Vs \quad H_1 : \gamma = \ddot{\gamma} \]

At a level with power \(\geq 1-\beta\) when the distribution of \(\hat{\gamma}\) is treated as normal with mean \(\gamma\) and variance \(\sigma^2\), the critical region is:

\[ \hat{\gamma} > Z_{1-\alpha} \frac{\sigma_\gamma}{\sqrt{n}} \text{ if } \gamma > 0 \]
\[ \hat{\gamma} < -Z_{\alpha} \frac{\sigma_\gamma}{\sqrt{n}} \text{ if } \gamma < 0 \]

where, \(Z_{\alpha}\) is 100(1-\(\alpha\))% of the standard normal distribution. The sample size \(n\) will be found so that the test has a specified power \((1-\beta)\) at the alternative \(H_1 : \gamma = \ddot{\gamma}\), the sample size \(n\) is thus chosen so that:

\[ \Pr \left( \hat{\gamma} > Z_{1-\alpha} \frac{\sigma_\gamma}{\sqrt{n}} \right) = (1-\beta) \]

This can be written as

\[ 1 - \Phi\left( Z_{1-\alpha} \frac{\sigma_\gamma}{\sqrt{n}} / \gamma \right) = (1-\beta) \text{ if } \gamma > 0 \]
\[ \Phi\left( -Z_{\alpha} \frac{\sigma_\gamma}{\sqrt{n}} / \gamma \right) = (1-\beta) \text{ if } \gamma < 0 \]

\[ \Phi\left( -Z_{1-\alpha} \frac{\sigma_\gamma}{\sqrt{n}} / \gamma \right) = (1-\beta) \text{ if } \gamma > 0 \]

\[ \Phi\left( Z_{\alpha} \frac{\sigma_\gamma}{\sqrt{n}} / \gamma \right) = (1-\beta) \text{ if } \gamma < 0 \]
If $\gamma > 0$ and an appropriate $n$, a solution of the above formula satisfies:

$$Z_{\sigma_0} - \frac{\gamma}{\sigma_0} = Z_{\sigma_0}$$

Otherwise if $\gamma < 0$:

$$-Z_{\sigma_0} - \frac{\gamma}{\sigma_0} = Z_{\sigma_0}$$

Hence:

$$n = \frac{Z_{\sigma_0} + Z_0}{\sigma_0^2}$$ (8)

For both cases $\gamma > 0$ and $\gamma < 0$

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

$$\logit (\pi) = \alpha + \gamma X$$

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 $\pi$ at the mean of $x$. the size of this effect is the odds ratio $\theta$ comparing $\pi$ to the probability of success one standard deviation above the mean of $x$. Let $k = \log(\theta)$ An approximate sample size is Equation 9:

$$n = \left[ \frac{Z_{\gamma} + Z_{\gamma} e^{\left(\frac{k}{2}\right)}}{(1 + 2\pi\delta) / (4\pi k^2)} \right]^2$$ (9)

Where Equation 10:

$$\delta = 1 + (1 + k^2) e^{\left(\frac{k}{2}\right)} /[1 + e^{-x^2/2}]$$ (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:

$$f(z) = \frac{Z}{1 + Z} = \sum_{i=1}^{m} (-1)^{z_i} z_i, \text{if } 0 < z < 1$$

$$\sum_{i=1}^{m} (-1)^{z_i} z_i, \text{if } 1 < z < \infty$$

And is simply approximated in Equation 6 by:

$$f(z) = \varphi(z) = \frac{e^{x_i X_i}}{1 + e^{x_i X_i}}$$

$$\begin{cases} e^{x_i X_i} + O(e^{z_i}), & \text{if } a_i + \gamma X_i \leq 0 \\ e^{-x_i X_i} + O(e^{z_i}), & \text{if } a_i + \gamma X_i \geq 0 \end{cases}$$

where, $e^{z_i}$ is small when $a_i + \gamma X_i \leq 0$ (or $e^{-z_i}$ is small when $a_i + \gamma X_i \geq 0$).

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

$$\frac{e^{x_i X}}{1 + e^{x_i X}} < 0.5 \text{ and } \frac{e^{x_i X}}{1 + e^{x_i X}} > 0.5$$

Then:

$$e^{x_i X} = e^{x_i X} + O(e^{2z_i})$$

And:

$$\frac{e^{x_i X}}{1 + e^{x_i X}} = e^{x_i X} + O(e^{2z_i})$$

Where Equation 10:

$$\begin{cases} e^{x_i X} & e^{-x_i X} \\ e^{-x_i X} & e^{x_i X} \end{cases}$$

$$\begin{cases} e^{x_i X} & e^{-x_i X} \\ e^{-x_i X} & e^{x_i X} \end{cases}$$

$$\approx \frac{e^{x_i X}}{1 + e^{x_i X}}$$

$$\approx = \gamma$$

$$\approx = \gamma$$

$$\approx = \gamma$$

Therefore variance is Equation 11:

$$\text{Var}(\gamma) = \frac{m(-\gamma)}{nc^{-\gamma}} [m(-\gamma) - m_{1}(-\gamma)]$$ (11)

If $X \sim N(0, 1)$ then $\text{Var}(\gamma) = \frac{e^{-\gamma} \gamma}{nc^{-\gamma}}$ and Equation 12:

$$n \geq \frac{Z_{\sigma_0} + Z_0 e^{\left(\frac{k}{2}\right)}}{\sigma_0^2}$$ (12)

Equation 11 discovered method of obtaining $\sigma$ 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: \gamma_1 = 0 \text{ Vs } H_1: \gamma_1 = \bar{\gamma}

And let:

\[ Y' = (\gamma_1, \gamma_2, \ldots, \gamma_p) = (\bar{\gamma}_1, \bar{\gamma}_2) \]

Where:

\[ \bar{\gamma}_i = (\gamma_1, \gamma_2, \ldots, \gamma_p) \]

And:

\[ Y'' = (\gamma_{p+1}, \gamma_{p+2}, \ldots, \gamma_n) \]

### 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


Assume \( \pi = 0.817013 \) if we go by the hypothesis that

\( H_0: \gamma_1 = 0 \) against the alternative \( H_1: \gamma_1 \neq 0 \) from Table 1a then:

\[
a = 0.05 \text{ and } \beta = 0.10, Z_a = 1.96 \text{ and } Z_\beta = 1.64
\]

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

\[
k^2 = 0.0007
\]

\[
\delta = 1.0009 \text{ and } n \approx 56,946
\]

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 diabetes 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 addicted 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:

\[
n_2 = 56,946
\]

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 \( e^{10} = 0.05 \) 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

According 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, \( \gamma \) and \( \alpha \), 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 graph of sample size fixed for $k = 2$, $\text{exp(}\alpha_1) = 0.05$

<table>
<thead>
<tr>
<th>Table 1a. Logistic regression: Diabetes vs Smoking</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diabetes</td>
</tr>
<tr>
<td>Smoking</td>
</tr>
<tr>
<td>Cons</td>
</tr>
<tr>
<td>LR</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 1b. Logistic regression: Diabetes versus Smoking and Alcohol</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diabetes</td>
</tr>
<tr>
<td>Smoking</td>
</tr>
<tr>
<td>Alcohol</td>
</tr>
<tr>
<td>Cons</td>
</tr>
<tr>
<td>LR</td>
</tr>
</tbody>
</table>

| Table 2. (Estimates of sample sizes for both equations (3.0) and (3.4)) |
|-----------------|-----------------|-----------------|-----------------|-----------------|
| $\gamma$ | $N_1 (3.4)$ | $N_2 (3.0)$ | $\gamma$ | $N_1 (3.4)$ | $N_2 (3.0)$ | $\gamma$ | $N_1 (3.4)$ | $N_2 (3.0)$ |
| 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 | 921.700 | 415.67 |
| 0.2 | 2176.76 | 2254.37 | 0.2 | 684.48 | 401.00 | 0.2 | 495.190 | 221.56 |
| 0.27 | 1243.02 | 1387.31 | 0.27 | 419.84 | 227.59 | 0.27 | 304.630 | 134.86 |
| 0.35 | 746.88 | 926.55 | 0.35 | 279.21 | 135.44 | 0.35 | 203.340 | 88.78 |
| 0.41 | 452.27 | 652.89 | 0.41 | 195.69 | 80.71 | 0.41 | 143.160 | 61.41 |
| 0.47 | 263.27 | 477.25 | 0.47 | 142.09 | 45.58 | 0.47 | 104.500 | 43.85 |
| 0.53 | 134.93 | 357.88 | 0.53 | 105.66 | 21.70 | 0.53 | 78.210 | 31.91 |
| 0.59 | 120.99 | 273.12 | 0.59 | 79.79 | 4.75 | 0.59 | 59.510 | 23.44 |
| 0.65 | 54.15 | 210.80 | 0.65 | 60.77 | 0.00 | 0.65 | 45.740 | 17.20 |
| 0.71 | 3.70 | 163.68 | 0.71 | 46.38 | 0.00 | 0.71 | 35.310 | 12.49 |
5. CONCLUSION

This study has developed a methodological framework to estimate the parameters of logistic regression and obtain sample sizes at different levels of $\alpha$ and $\beta$. 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