**Int. Journal of Math. Analysis, Vol. 8, 2014, no. 15, 721 - 726**
**HIKARI Ltd, www.m-hikari.com**
**http://dx.doi.org/10.12988/ijma.2014.4394**

# Empirical Distributions of Parameter Estimates in Binary Logistic Regression Using Bootstrap

**Anwar Fitrianto* and Ng Mei Cing**

Department of Mathematics, Faculty of Science
and Institute for Mathematical Research
Universiti Putra Malaysia, Malaysia
*Corresponding author

Department of Statistics, Faculty of Mathematics and Natural Sciences
Bogor Agricultural University, Indonesia

Copyright © 2014 Anwar Fitrianto and Ng Mei Cing. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

**Abstract**

Bootstrapping is a famous statistical tool that involves resampling procedure to select sample from a population. In this study, we applied random-$x$ bootstrap in binary logistic regression for published data set namely Umaru Impact data. We conducted bootstrap for the coefficient by using SAS (Statistical Analysis System). We observe the distribution of the estimated coefficients with different sample sizes. After conducting $B$=10000 bootstrap replications, we found that the distribution of parameters estimates is nearly normal.

**Keywords**: Binary logistic, Bootstrap, Parameter estimates

## 1 Introduction

Bootstrapping is a useful statistical technique used for analyzes and obtain estimate coefficient in regression. We can use parametric bootstrap when the distribution form of the data is known, and we can conduct non parametric bootstrap when the distribution form of the data is unknown. We will conduct non parametric bootstrap for binary logistic regression coefficients on an existing data set where the distributional of the data is unknown. The bootstrap performance is

measured based on the biases and variances of the bootstrap estimates. We also try to investigate effect of sample size on the distribution of the parameter estimates.

Logistic regression is a popular and useful statistical method in modeling categorical dependent variable. Many researchers used it to analyze in all fields. Kleinbaum and Kelvin, [6] proposed that logistic regression is a mathematical modeling approach used to investigate the relationship between the independent variable and dichotomous dependent variable. Suppose we have a set of observation with the outcome is binary, outcome is either 0 or 1. The corresponding model is written as:

$$Y_i = \pi(x_i) + \varepsilon_i \quad , \tag{1}$$

where

$$\pi(x_i) = P(Y = 1 | X) = \frac{\exp(\beta_0 + \beta_1 X_1 + ... + \beta_i X_i)}{1 + \exp(\beta_0 + \beta_1 X_1 + ... + \beta_i X_i)} . \tag{2}$$

The $Y_i$ is the outcome and $\varepsilon_i$ are assumed to be independent identically distributed with zero mean and constant variance, $\sigma^2$.

Because the independent variables $x_i$ is not linear in $\pi(x_i)$, therefore we need to transform the $\pi(x_i)$ using logit function which is written as

$$\text{Logit}\pi(x_i) = \ln \frac{\pi(x_i)}{1 - \pi(x_i)} \tag{3}$$

$$= \ln\left(\frac{\dfrac{\exp(\beta_0 + \beta_1 X_1 + ... + \beta_i X_i)}{1 + \exp(\beta_0 + \beta_1 X_1 + ... + \beta_i X_i)}}{1 - \dfrac{\exp(\beta_0 + \beta_1 X_1 + ... + \beta_i X_i)}{1 + \exp(\beta_0 + \beta_1 X_1 + ... + \beta_i X_i)}}\right)$$

$$= \ln[\exp(\beta_0 + \beta_1 X_1 + ... + \beta_i X_i)] = \beta_0 + \beta_1 X_1 + ... + \beta_i X$$

By using logit function, it provides much easier model to fit because it is in linear form.

In logistic regression, we commonly use maximum likelihood to estimate the parameters which are the best to fit the model. Czepiel, [2] stated that we are not able to use least square estimation to produce the minimum variance unbiased estimators in logistic regression. In binary logistic regression, our outcome is involving value 0 or 1 only.

With regard to bootstrap techniques, Efron, [3] introduced bootstrap approach to estimating unknown parameters and making inference about an unknown population. He also proposed his idea about the bootstrap method can be used to estimate the parameters for multiple regression models. Michael, [7] stated that bootstrap is a method that involves resamples procedure from a data. The basic

idea of bootstrap is when we have a data with unknown population's distribution; we can replace it with the known empirical distribution using bootstrap.

Batmanz et.al [1] defined bootstrap as a computer-intensive statistical method where it treat data as a population and select samples from the population with replacement. In their research, they applied empirical distributions of the parameters to investigate the significant of the parameters in nonparametric regression and Conic Multivariate Adaptive Regression Spines (CMARS). They involve three different bootstrappings in their research which are random-$x$, fixed-$x$ and Wild bootstrap in four data sets with different sample size and scale. The results of their study show that random-$x$ method gives more precise and less complex model with the medium size and medium scale data.

Sometimes, random-$x$ bootstrap is also called as pairs bootstrap or observation bootstrap which is firstly proposed by [4]. Let say we have a set of observation $Z_i^{'} = [Y_i, x_{ik}, ..., x_{nk}]$ where $i$=1,2,…,$n$ and $k$ is the number of variables. Then, we do the following steps:

Step 1 : Select $n$ observation $Z_1^{'}, Z_2^{'}, ..., Z_n^{'}$ as sample,

Step 2 : Fit the selected sample into the model and calculate the estimates of parameters of the logistic regression,

Step 3 : Repeated the above procedure $B$ times to obtain the bootstrap estimates of parameters.

## 2 Data and Method

### Data

In this article, we are using the Umaru Impact data. This data is taken from [5]. The data consist of seven independent variables which are age at enrollment, beck depression score at admission, IV drug use history at admission, number of prior drug treatment, subject's race, treatment randomization assignment, and treatment site. The dependent variable is remained drug free for 12 months, which is binary with 1 means the participant remained drugging free for 12 months and 0 for otherwise.

### Method

We use PROC LOGISTIC in SAS to analyze this data in logistic model. to obtain coefficients of the logistic regression, $\beta_i$ where $i = 1,2,\ldots,7$ for the original dataset. In bootstrapping stage, random-$x$ bootstrap will be employed. First, we bootstrap a sample size $n$ with replacement by using PROC SURVEYSELECT in SAS. Then we fit it into logistic regression model and obtain the each estimate of coefficient, $\hat{\beta}_i$. We replicate the bootstrap $B$ times.

Then we compute the mean of each parameter estimates which is denoted as

$\overline{\hat{\beta}}_i$. Then, we observe the distribution of the estimated coefficient. We plot histograms of the estimated coefficients for each sample size with same scale, so that we can compare the location of the histogram.

In this paper, bootstrap replications, *B*, is 10000. The simulation will be conducted at four different sample sizes which are 25, 50, 250 and 575 . Sample sizes 25 is referred as small sample, sample size 50 is referred as medium sample, sample sizes 250 is referred as large sample and sample size 575 is exactly the number of the observation in the original data. Therefore, we can justify the performance of bootstrap using different sample size.

## 3 Results and Discussion

Table 1: Maximum Likelihood Estimates for Original Umaru Impact Data

| Parameter | df | Estimate | Standard Error | Wald Chi-Square | *p* value | Odd Ratio Estimate | 95% Wald Confidence Limits | |
|---|---|---|---|---|---|---|---|---|
| Intercept | 1 | 1.9992 | 0.5879 | 11.5622 | 0.0007 | | | |
| $x_1$ | 1 | -0.0482 | 0.0172 | 7.8498 | 0.0051 | 0.953 | 0.921 | 0.986 |
| $x_2$ | 1 | -0.00042 | 0.0108 | 0.0015 | 0.9691 | 1.000 | 0.979 | 1.021 |
| $x_3$ | 1 | 0.3700 | 0.1283 | 8.3100 | 0.0039 | 1.448 | 1.126 | 1.862 |
| $x_4$ | 1 | 0.0618 | 0.0257 | 5.7685 | 0.0163 | 1.064 | 1.011 | 1.119 |
| $x_5$ | 1 | -0.2296 | 0.2230 | 1.0606 | 0.3031 | 0.795 | 0.513 | 1.230 |
| $x_6$ | 1 | -0.4293 | 0.1985 | 4.6749 | 0.0306 | 0.651 | 0.441 | 0.961 |
| $x_7$ | 1 | -0.1213 | 0.2152 | 0.3177 | 0.5730 | 0.886 | 0.581 | 1.351 |

Table 1 is the results that we have obtained for the logistic regression analysis of the data. From the result, we obtained all the true value of parameters from our original data regardless significance of the coefficients. By applying all coefficients to the logistic regression model, we obtain the following predicted full model:

$$\hat{\pi}(x_i) = \frac{-0.4293x_6 - 0.1213x_7)}{1+\exp(1.9992 - 0.0482x_1 - 0.00042x_2 + 0.3700x_3 + 0.0618x_4 - 0.2296x_5 - 0.4293x_6 - 0.1213x_7)}.$$

And the corresponding logit function can be written as follows:

$$\text{Logit } \pi(x_i) = 1.9992 - 0.0482x_1 - 0.00042x_2 + 0.3700x_3 + 0.0618x_4 - 0.2296x_5 - 0.4293x_6 - 0.1213x_7.$$

Variable $x_3$ and $x_4$ have positive coefficients which imply that $P(Y=1)$

increases as $x_3$ and $x_4$ increase. Meanwhile, variables $x_1$ , $x_2$ , $x_5$ , $x_6$ and $x_7$ are having negative coefficients to indicate that $P(Y=1)$ will increase if those variables at lower values and vice versa. After we obtained the logistic regression coefficient, $\beta_i$ where *i*=0, 1, 2,…,7. We start our bootstrap and obtain the estimated coefficient $\hat{\beta}_i$.

This section onwards present results of bootstrapping the logistic regression coefficients. We arbitrarily choose one of 8 parameters out of eight parameter estimates, namely $\hat{\beta}_3$ since the other parameters will have approximately the same behavior.

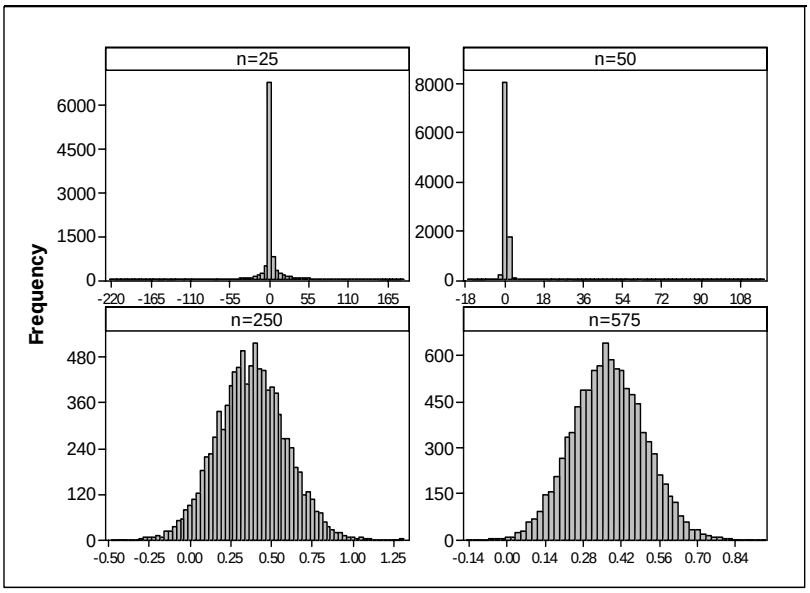


Figure 1: Histogram of $\hat{\beta}_3$ when *B*=10000 with Different Sample Size

We look for patterns of the distribution of the estimated coefficient when we increase the sample size. We also concern about the pattern of the distribution when *B* is getting larger. Firstly, we look at the distribution of $\hat{\beta}_3$. The distribution of $\hat{\beta}_3$ on with different sample sizes is shown in the Figure 1.

We noticed that when the sample size is small and medium, *n*=25 and *n*=50, there is no clear pattern for the distribution of the $\hat{\beta}_3$. But when the sample size 250 and 575, obviously we can observe that the distribution of $\hat{\beta}_3$ is a bell shape and symmetric. It is approximate to the normal distribution. For $\hat{\beta}_3$ coefficient, we may conclude that as the sample increases, the distribution of $\hat{\beta}_3$ is approximate normal.

## 4 Conclusion

In this study, we used random-$x$ bootstrap to estimate the coefficient using different sample sizes. After estimated the logistic regression coefficients, we investigated the distribution of the estimated coefficient. We found that when the sample size becomes larger, the distribution of the estimated coefficient is approximate to normal distribution.

## References

[1] I. Batmanz, Yazıcı, C., Yerlikaya-Özkurt, F. *Bootstrapping Conic Multivariate Adaptive Regression Splines (Bcmars).* Middle East Technical University: Turkey, 2012.

[2] S. A. Czepiel, *Maximum Likelihood Estimation of Logistic Regression Models: Theory and Implementation*, 2002, http://czep.net/stat/mielr.pdf (accessed 16 June 2013).

[3] B. Efron, Bootstrap Methods: Another Look at the Jacknife. *The Annals of Statistics*. **7** (1979), 1-26.

[4] D. A. Freedman, Bootstrapping regression models. *Annals of Statistics*. **9** (1981), 1218–1228.

[5] D. W. Hosmer and Lemeshow, S., *Applied Logistic Regression*. Wiley, NewYork, 2000.

[6] D. G. Kleinbaum, and Klein, M. *Logistic Regression:* Statistics for Biology and Health: Springer Science Business Media, 2010.

[7] R. C. Michael, *Bootstrap Method, A Guide for Practitioners and Researchers*: John Wiley & Sons, Canada, 2008.

**Received: March 1, 2014**