JACKKNIFE ALGORITHM FOR THE ESTIMATION OF LOGISTIC REGRESSION PARAMETERS

H.O.Obiora-Ilouno[1*] and J.I.Mbegbu[2]

1. Department of Statistics, Nnamdi Azikiwe University, Awka
2. Department of Mathematics, University of Benin, Benin City, Edo State, Nigeria.

[*]E-mail of the corresponding author: obiorailounoho@yahoo.com

Abstract

This paper proposes an algorithm for the estimation of the parameters of logistic regression analysis using Jackknife. Jackknife delete-one and delete-d algorithm was used to provide estimates of logistic regression coefficient. The Jackknife standard deviation provides an estimate of variability of the standard deviation of sample and it is a good measure of precision. The method was illustrated with real life data; and the results obtained from the Jackknife samples was compared with the result from ordinary logistic regression using the maximum likelihood method and results obtained reveals that the values from the jackknife algorithm for the parameter estimation, standard deviation and confidence interval were so close to the result from ordinary logistic regression analysis, this provides a good approximation to the result which shows that there is no bias in the jackknife coefficients.

Keywords: Jackknife algorithm, Logistic regression, dichotomous variable, maximum likelihood

INTRODUCTION

Sometimes one may be interested in situations where one is trying to predict whether something happens or not. For example a patient survives a treatment or not, a person contracts a disease or not, and a student passes a course or not. These are binary measures. Logistic regression regresses a dichotomous dependent variable on a set of independent variables, especially where the data set is very large, and the predictor variables do not behave in orderly ways, or obey the assumptions required of ordinary linear regression or discriminant analysis (Michael, 2008; Russell and Chritine, 2009; Ryan, 1997). Logistic regression applies maximum likelihood estimation after transforming the dependent variable into a logit variable which can be used to determine the effect of the independent variables on the dependent variable using the maximum likelihood estimation method (Russell and Chritine,2009).This is accomplished using iterative estimation algorithm. Jackknifing is used in statistical inference to estimate the bias and standard error when a random sample of observation is used. The basic idea behind the jackknife estimators lies in systematically recomputing the statistic estimate leaving out one or more observations at a time from the sample set. From this new set of replicates of the statistic, an estimate for the bias and an estimate for the variance of the statistic can be calculated. (Efron,1982; Efron and Tibshirami,1993)

MATERIALS AND METHOD

Material: The aim of this paper is to illustrate the Jackknife logistic regression parameter estimation. The data used is a secondary data collected from the delivery ward of general hospital Onitsha, Anambra state, Nigeria. Here, Maternal age, Parity, and babies Sex were considered as independent variables in order to determine their effect on the gestation period of n = 256 mothers. R programming language was used for the statistical analysis of these data.

1 JACKKNIFE DELETE-ONE ALGORITHM

The jackknife delete-one procedure is as follow:

Step1: Given a randomly drawn n sized sample from population consisting of a dichotomous dependent variable and label the element of the vector $Z_i = (Y_i, X_{ji})'$

where $Y_i = (y_1, y_2, y_3, \ldots, y_n)'$ and the matrix $X_{ji} = (x_{j1}, x_{j2}, x_{j3}, \ldots, x_{jn})'$; $j = 1, 2, 3, \ldots, k$, and $i = 1,2,3,\ldots, n.$

Omit first row of the vector $Z_i = (Y_i, X_{ji})'$ and label the remaining $n-1$ sample sized observation sets $Y_i^{(J)} = \left(y_2^{(J)}, y_3^{(J)}, \ldots, y_n^{(J)}\right)'$ and $X_{ji}^{(J)} = \left(x_{j2}^{(J)}, x_{j3}^{(J)}, \ldots, x_{jn}^{(J)}\right)'$ as the first delete-one jackknife sample $Z_1^{(J)}$, and estimate $\hat{\beta}^{(J_1)}$ coefficient from $Z_1^{(J)}$ using the maximum likelihood estimate of the logistic regression model in the Jackknife sample (Efron,1982; Sahinler and Topuz, 2007). The maximum likelihood estimate of β_i in the logistic regression model are those values of β_i that maximize the log-likelihood function

$$E(Y_i) = \frac{\exp(\beta' X_i)}{1 + \exp(\beta' X_i)} \qquad (1)$$

Where $\beta_{xi}' = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_{p-1} X_{i(p-1)}$

$$\beta_{p\times1} = \begin{bmatrix} \beta_0 \\ \beta_0 \\ \vdots \\ \beta_{p-1} \end{bmatrix}; X_{ip\times1} = \begin{bmatrix} 1 \\ x_{i1} \\ x_{i2} \\ \vdots \\ x_{i(p-1)} \end{bmatrix}$$

$$E(Y) = [1 + \exp(-\beta' x_i)]^{-1} \qquad (2)$$

Where Y$_i$ are ordinary Bernoulli random variables with expected values $E(Y_i) = \pi_i$,
Where

$$E(Y) = \pi_i = \frac{\exp(\beta' X_i)}{1 + \exp(\beta' X_i)}$$

Since each Y$_i$ observation in an ordinary Bernoulli random variable, where
P (Y$_i$ = 1) = π_i
P (Y$_i$ = 0) = 1 - π_i
then, its probability distribution can be represented as

$$f_i(Y_i) = \pi_i^{Y_i}(1 - \pi_i)^{1-Y_i}, \quad \text{Y}_i = 0,1; i = 1 \cdots n$$

Since Y$_i$ observations are independent, their joint probability function is

$$g(Y_i,\ldots,Y_n)=\prod_{i=1}^{n} f_i(Y_i)=\prod_{i=1}^{n}\pi_i^{Y_i}(1-\pi_i)^{1-Y_i} \qquad (3)$$

To find the maximum likelihood estimates we take the log of both sides

$$\log\left[g(Y_i,\ldots,Y_n)\right]=\log\left[\prod_{i=1}^{n}\pi_i^{Y_i}(1-\pi_i)^{1-Y_i}\right]$$

$$=\sum_{i=1}^{n}Y_i\log(\pi_i)+\sum_{i=1}^{n}(1-Y_i)\log(1-\pi_i)$$

$$=\sum Y_i\log(\frac{\pi_i}{1-\pi_i})+\sum_{i=1}^{n}\log(1-\pi_i) \qquad (4)$$

Therefore, the log likelihood function for multiple logistic regression is

$$\log L(\beta)=\sum_{i=1}^{n}Y_i(\beta'X_i)-\sum_{i=1}^{n}\log(1+\exp(\beta'X_i)) \qquad (5)$$

Step2:

Computer intensive numerical search procedures are employed to find the values of $\beta_0,\beta_1,\ldots,\beta_{p-1}$ that maximize log (β) using the Gauss Newton method. These maximum likelihood estimates will be denoted by $b_0,b_1,b_2,\ldots,b_{p-1}$,.

$$\beta_{p\times1}=\begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_{p-1} \end{bmatrix}$$

The fitted logistic response function and fitted value can be expressed as

$$E(Y_i)=\hat{\pi}_i=\frac{\exp(b'x_i)}{1+\exp(b'x_i)}=\left[1+\exp(-b'x_i)\right]^{-1} \qquad (6)$$

Here standard statistical R codes and minitab package were used for logistic regression to conduct the numerical search procedure by iteratively re-weighted least squares for the maximum likelihood estimates.

step3:

Iterative Procedures:

a. Obtain starting basic values for the regression parameters denoted by b(0). This can be obtained by ordinary least squares regression of Y on the predictor variables using a first-order linear model.

b. Using these starting values, obtain

$$\hat{\pi}_i'(0)=\left[b(0)\right]'x_i \qquad (7)$$

$$\hat{\pi}_i'(0) = \frac{\exp[\hat{\pi}_i'(0)]}{1+\exp[\hat{\pi}_i'(0)]} \qquad (8)$$

C. Calculate the new response variable

$$Y_{(0)}' = \hat{\pi}_i(0) + \frac{Y_i - \hat{\pi}_i(0)}{\hat{\pi}_i(0)[1-\pi_i(0)]} \qquad (9)$$

And the weights $W_i(0) = \hat{\pi}_i(0)\left[1-\hat{\pi}_i(0)\right]$

d. Regress $Y_{(0)}'$ in (equ. 10) on the predictor variables $X_i,\dots,X_{p-1}$ using re-weighted least squares for the maximum likelihood estimates to obtain b (1).Repeat step (a) through (d) using the latest revised estimated regression coefficient until there is little if any change in the estimated coefficients which leads to convergence (Neter et al,1996; Hamadu, 2010; Ryan,1997)

Then, omit second row of the vector $Z_i = (Y_i, X_{ji})'$ and label the remaining $n-1$ sample size observation sets as $Z_2^{(J)}$, estimate $\hat{\beta}^{(J_2)}$ coefficient from $Z_2^{(J)}$ using the maximum likelihood estimate of the logistic regression also. Alternatively, omit each row of the observation set and estimate the $\hat{\beta}^{(J_i)}$ coefficient using maximum likelihood estimate of the logistic regression. Where $Y_i = (y_1, y_2, y_3, \dots, y_n)$ and the matrix of the independent variable $X_{ji} = (x_{j1}, x_{j2}, x_{j3}, \dots, x_{jn})$ and $\hat{\beta}^{(J_i)}$ is jackknife logistic regression coefficient vector estimated after deleting the ith observation sets from Z_i.

Step4:
Obtain the probability distribution F ($\hat{\beta}^{(J_i)}$) of jackknife estimates $\hat{\beta}^{(J_1)}$, $\hat{\beta}^{(J_2)}$, …, $\hat{\beta}^{(J_n)}$

Step5:
Calculate the jackknife regression coefficient estimate which is the mean of the F ($\hat{\beta}^{(J_i)}$) distribution as;

$$\hat{\beta}^{(J)} = \frac{\sum_{i=1}^{n} \hat{\beta}^{(J_i)}}{n} = \overline{\beta}^{(J_i)} \qquad (10)$$

Step 6

The delete-one jackknife logistic regression equation is thus, $\hat{Y} = \left[1+\exp(-b^{(J)})X\right]^{-1}$ where $b^{(J)}$ is the unbiased estimator of β.

2 JACKKNIFE DELETE-D ALGORITHM

Steps to the jackknife delete-d are as follows:
Step1:
Given a randomly drawn n sized sample $(Z_1, Z_2, \cdots, Z_n)$ from a population,
divide the sample into "S" independent group of size d.

Step2:
Omit first d observation set from full sample at a time and estimate the logistic regression coefficient using

the maximum likelihood estimate $\hat{\beta}^{(J_2)}$ from $(n-d)$ sized remaining observation set.

Step3:

Omit second d observation set from full sample at a time and estimate the logistic regression coefficient $\hat{\beta}^{(J_2)}$ from ($n-d$) sized remaining observation set.

Step4:

Alternately omit each d of the n observation sets and estimate the coefficients as $\hat{\beta}^{(J_k)}$, where $\hat{\beta}^{(J_k)}$ is the jackknife regression coefficient vector estimated after deleting of kth d observation set from full sample.
Thus,

$$S = \binom{n}{d}$$

delete-d jackknife sample are obtained , K=1,2,...,S.

Step 5:

Obtain the probability distribution F($\hat{\beta}^{(J)}$) of delete-d of jackknife estimates $\hat{\beta}^{(J_1)}, \hat{\beta}^{(J_2)}, \hat{\beta}^{(J_3)}, ..., \hat{\beta}^{(J_s)}$.

Step 6:

Calculate the jackknife regression coefficient estimate which is the mean of the F ($\hat{\beta}^{(J)}$) distribution as;

$$\hat{\beta}^{(J)} = \frac{\sum_{k=1}^{n} \hat{\beta}^{(J_k)}}{s} = \overline{\beta}^{(J_k)} \tag{11}$$

The jackknife confidence interval

$$\overline{X} \pm Z_{\alpha} \frac{\delta}{\sqrt{n}} \tag{12}$$

The jackknife delete-one estimate of the standard error is defined by

$$\hat{s}e_{jeck} = \left[\frac{n-1}{n} \sum \left(\theta\text{垐}_{(i)} - \theta_{(.)} \right)^2 \right]^{\frac{1}{2}} \tag{13}$$

where $\hat{\theta}_{(.)} = \frac{\sum \hat{\theta}_i}{n}$

The jackknife delete-d estimate of the standard error is

$$\left\{ \frac{r}{\binom{n}{d}} \sum \left(\theta\text{垐}_{(s)} - \theta_{(.)} \right)^2 \right\}^{\frac{1}{2}} \tag{14}$$

(Efron, 1982; Sahinler and Topuz, 2007)

Illustrative example

The logistic regression model was fitted in the data in the table 1 regressing gestation period on mother age, parity and babies' sex.

	1	2	3	4	……	254	255	256
Gestation period Y	1	0	0	1	…..	0	1	0
Mother Age	27	30	30	25	…	36	30	31
Parity	5	1	1	2	….	2	5	6
Baby Sex	1	0	1	1	….	1	0	1

Table 1: The data used in calculation of logistic regression and Jackknifes' results with n=256

Let Y_i be the response of the i^{th} randomly selected subject (gestation period) which assumes values of either 1 (positive response) or 0 (negative response) for i= 1, 2, ,n.

$$Y_i = \begin{cases} 1, \mathit{if\ gestation\ period} \geq 39.5\ \mathit{weeks} \\ 0, \mathit{if\ gestation\ period} < 39.5\ \mathit{weeks} \end{cases}$$

Let X_1, X_2 and X_3 be independent variables mothers age, parity and sex of the baby respectively, where 1 represent baby boys and 0 represent baby girls.

Results:

First, logistic regression model was fitted to the data in (Table1) and the results was summarized in the (Table 2) below. The regression model is significant as the P-value is 0.000 that is ($P < 0.05$).

The jackknife samples are generated omitting each d=1 or 2 or 3 sample(s) respectively of the n=256 observation sets and estimated coefficients as $\hat{\beta}^{(J)}$.

Variable	**Ord. J($\hat{\beta}$)**	**S.E($\hat{\beta}$)**	**P-value**	**95% conf. Interval**
Constant ($\hat{\beta}_0$)	0.555092	0.856362	0.517	Lower upper
Mother's age ($\hat{\beta}_1$)	-0.0601875	0.0321786	0.061	0.88 1.00
Parity ($\hat{\beta}_2$)	0.407925	0.0899854	0.00	1.26 1.79
Baby's sex ($\hat{\beta}_3$)	-0.0108895	0.268384	0.968	0.58 1.67

Table 2: The summary statistics of regression coefficients for binary logistics regression

Log-likelihood = -164.996
Test that all slopes are zero: G= 23.335 df= 3 and P-value=0.000

r	Variables	1	2	3	…	256	$\hat{\beta}_0^{\,J}$	$\hat{\beta}_1^{\,J}$	$\hat{\beta}_2^{\,J}$	$\hat{\beta}_3^{\,J}$
1	Gestation mother age Parity Babysex		1 23 3 0	1 38 6 1	… … … …	1 27 5 1	0.4583	-0.0573	0.4139	0.0089
2	Gestation mother age Parity Babysex	1 33 5 1		1 38 6 1	… … … …	1 27 5 1	0.5472	-0.0591	0.4036	-0.0237
4	Gestation mother age Parity Babysex	1 33 5 1	1 23 3 0		… … … …	1 27 5 1	0.5302	-0.0589	0.4017	-0.0018
. . .	. . .	. . .	. . .	. . .	…	. . .	. . .	. . .	. . .	. . .
256	Gestation mother age Parity Babysex	1 33 5 1	1 23 3 0	1 38 6 1	… … … …		0.5334	-0.0589	0.4036	-0.0187
	$\hat{\beta}^{(J_i)}$						0.5552	-0.0602	0.4080	-0.0109

Table 3: The illustration of the Jackknife delete-1 logistic regression procedure from the data in Table 1 for the estimation of the regression parameters

	Constant ($\hat{\beta}_0$)	**Mother's age ($\hat{\beta}_1$)**	**Parity ($\hat{\beta}_2$)**	**Baby's sex ($\hat{\beta}_3$)**
Ordinary logistic coef.	0.55509	-0.0602	0.40793	-0.0109
Delete-1 jackknife coef.	0.55519	-0.0602	0.40797	-0.0109
Delete-2 jackknife coef.	0.5553	-0.0607	0.40809	-0.0109
Delete-3 jackknife coef.	0.5554	-0.0602	0.40805	-0.0109

Table 4: The summaries of ordinary logistic regression and the jackknife results of delete-1, delete-2 and delete-3 values of logistic regression coefficients

var.	ord.S.E	del-1 S.E	del-2 S.E	del-3 S.E	Ord.P-value	del-1	del-2	del-3
const.	0.8564	0.8581	0.8599	0.8619	0.517	0.548	0.563	0.5387
M.age	0.03218	0.03224	0.03226	0.03236	0.061	0.067	0.0734	0.067
parity	0.08999	0.09030	0.0901	0.09078	0.00	0.000	0.000	0.000
B.sex	0.2684	0.2689	0.02692	0.26997	0.968	0.946	0.9508	0.929

Table 5: The summary statistics of the logistic regression standard errors and their P-values and that jackknife delete-1, delete-2 and delete-3 standard errors and their P-values results.

3 Discussion and Conclusions:

Jackknife samples were generated by omitting each, one or two or three n observation(s) corresponding to delete-one or delete-two or delete-three jackknife respectively for the estimation of logistic regression coefficients $\hat{\beta}^{(J_i)}$. Ordinary logistic regression on the data in table 1 for $\beta_0, \beta_1, ..., \beta_3$ are

b_0 = 0.55509, b_1= -0.06019, b_2 = 0.40793 and b_3 = -0.01089 respectively, and the estimated precision of these estimates were

$S(b_0)$=0.85636, $S(b_1)$=0.03218, $S(b_2)$=0.08999, $S(b_3)$=0.26838 with 95% confidence intervals as $0.88 \le \beta_1 \ge 1.00$, $1.26 \le \beta_2 \ge 1.79$ and $0.58 \le \beta_3 \ge 1.67$ respectively. The jackknifes results for the estimation of coefficients of logistic regression using the jackknife algorithm , estimation of the precision (standard error) and the confidence interval as shown in table 4 and 5 reveals that the Jackknife delete-one, delete-two, and delete-three estimated coefficients are quite close to analytical result. Also the estimated precisions and the confidence interval of the jackknife delete-one, delete-two and delete-three are also very close to that of analytical result when compared together. This reveals the appropriateness of the algorithm developed to the theoretical method.

Delete d R-Algorithm

```
#This R code defines a function 'jack' for performing delete-d jacknife for logistic regression
#p is the no of cols in the data. p=4 then there is 1 dept. var and 3 indept vars
#d is the no of rows to be deleted
jack=function(data,p,d)
{
n=length(data[,1])  #the sample size
u=combn(n,d) #Assign the matrix of all possible combinations to u
output=matrix(0,ncol=p,nrow=ncol(u))#define the output
y=data[,1] #the response vector
x=data[,2:p] #the matrix of covariates
for (i in 1:(ncol(u)))
{
dd=c(u[,i])
yn=y[-dd] #delete d rows of the independent var
xn=x[-dd,] #delete d rows of the dependent var
logreg=glm(formula=yn~xn[,1]+xn[,2]+factor(xn[,3]),family = binomial(link = "logit"),na.action =
na.pass)#Assuming 3 indpt vars with 1 as a factor
coef=logreg$coef
output[i,]=c(coef) #store the regression coefficients
}
output
}
#This part can be used to obtain a jacknife estimate of the regression
coefficients
u=jack(data,4,2)
beta=c(mean(u[,1]),mean(u[,2]),mean(u[,3]),mean(u[,4]))
```

(Venables and Smith,2007)

References

[1] Efron,B.1982.The Jackknife,the Bootstrap and other Resampling Plans,CBN-NSF Regional COnference Series in Applied Mathematics Philadelphia,Pennsylvania.5-27.

[2] Efron,B. and Tibshirani,R.J.1993.An Introduction to the Boot-strap.Chapman and Hall,New York.

[3] Hamadu,D.2010.A Bootstrap Approach to Bias-Reduction of Non-linear Parameter in Regression Analysis.Journal of science Research Development.Vol.12,110-127.

[4] Michael,P.L.2008.Logistic Regression. Circulation American Heart Association.Doi:10.1161.

[5] Neter,J.,Kutner,M.H.,Nachtsheim,C.J.,andWasserman,W.1996. Regression Analysis.Fourth Edition,Mc-Grow Hill,U.S.A.pp429-536.

[6] Russell,C. and Chritine,C.2009.Resampling Method of Analysis in Simulation Studies.Proceeding of Winter Simulation Conference.45-59.

[7] Ryan,P.T.1997.Modern Regression Method.John Wiley and sons inc,Third aveenue, New York.pp255-308.

[8] Sahinler,S. and Topuz,D.2007.Bootstrap and Jackknife Resampling Agorithm for Estmation of Regression Parameters.Journal of Applied Quantitative Method.vol.2,No.2:188-199.

[9] Venables,W.N.and Smith .2007.An Introduction to R, A programming Environment for Data Analysis and Graphics.Version 2.6.1, 1-100