AFRICAN JOURNAL OF MEDICINE and medical sciences

VOLUME 43 NUMBER 3 SEPTEMBER 2014

ISSN 1116–4077

Afr. J Med med. Sci. (2014) 43, 195-204 *Review Article*

An appraisal of convergence failures in the application of logistic regression model in published manuscripts

OB Yusuf, EA Bamgboye, RF Afolabi and MA Shodimu
Department of Epidemiology and Medical Statistics, College of Medicine, University of Ibadan, Ibadan, Nigeria

Abstract

Background: Logistic regression model is widely used in health research for description and predictive purposes. Unfortunately, most researchers are sometimes not aware that the underlying principles of the techniques have failed when the algorithm for maximum likelihood does not converge. Young researchers particularly postgraduate students may not know why separation problem whether quasi or complete occurs, how to identify it and how to fix it.
Objective: This study was designed to critically evaluate convergence issues in articles that employed logistic regression analysis published in an African Journal of Medicine and medical sciences between 2004 and 2013.
Methods: Problems of quasi or complete separation were described and were illustrated with the National Demographic and Health Survey dataset. A critical evaluation of articles that employed logistic regression was conducted.
Results: A total of 581 articles was reviewed, of which 40(6.9%) used binary logistic regression. Twenty-four (60.0%) stated the use of logistic regression model in the methodology while none of the articles assessed model fit. Only 3 (12.5%) properly described the procedures. Of the 40 that used the logistic regression model, the problem of convergence occurred in 6 (15.0%) of the articles.
Conclusion: Logistic regression tends to be poorly reported in studies published between 2004 and 2013. Our findings showed that the procedure may not be well understood by researchers since very few described the process in their reports and may be totally unaware of the problem of convergence or how to deal with it.

Keywords: *Logistic regression, convergence, quasi, complete separation, maximum likelihood estimates*

Résumé

Introduction: Le model de régression logistique est largement utilisé en recherche de santé pour description et buts prédictif. Malheureusement, plusieurs chercheurs ne sont pas souvent au courant que les principes supposés des techniques ont échoué quand l'algorithme pour une probabilité maximum ne converge pas. Les jeunes chercheurs particulièrement les étudiants au cycle supérieur peuvent ne pas savoir pourquoi le problème de séparation soit quasi ou complet survient, comment l'identifier et comment le fixer.
Objective: Cette étude était désignée pour délicatement évaluer les problèmes de convergence dans les articles qui employaient l'analyse de régression logistique publiés dans un Journal Médical Africain entre 2004 et 2013.
Méthode : Les problèmes de séparation quasi ou complet étaient décrites et étaient illustrer avec les données de l'étude nationale démographique et de santé. Une délicate évaluation des articles qui employaient la régression logistique était conduite.
Résultats : Un total de 581 articles étaient revus, desquels 40 (6,9%) utilisaient la régression logistique binaire. Vingt-quatre (60,0%) énonçaient l'usage du model de régression logistique dans la méthodologie tandis qu'aucun des articles n'imposait la convenance du model. Seulement 3 (12,5%) décrivaient proprement les procédures. Des 40 articles qui employaient le model de régression logistique, le problème de convergence apparu dans 6 (15,0%) de ces articles.
Conclusion : La régression logistique tend à être pauvrement reportée dans les études publiées entre 2004 et 2013. Nos résultats montraient que la procédure peut ne pas être bien comprise par les chercheurs puisque très peu décrivait le procès dans leurs exposés et peuvent être totalement sans connaissance du problème de convergence ou comment s'en occuper de ceci.

Mots clé: *Régression logistique, convergence, quasi, séparation complète, estimations de probabilité maximum*

Introduction

Use of logistic regression modelling in epidemiological research is very common because most outcome variables are categorical, utilising the disease present or absent dichotomy or event: yes or no category. The wide use of this model is also facilitated by its facility to explain a specific outcome using observed variables in the presence of confounding variables which could be categorical or continuous. However, the majority of users do not consider the assumptions underlying the use of this technique to examine if assumptions are satisfied

Correspondence: Dr. O.B. Yusuf, Department of Epidemiology and Medical Statistics, College of Medicine, University of Ibadan, Ibadan, Nigeria. E-mail: boyusuf@comui.edu.ng; bidemiyusuf1@gmail.com

by their data [1]. Such assumptions include multicollinearity among the independent variables and the non-convergence of the algorithm for maximum likelihood. In this work, we present a review of logistic regression analysis, with emphasis on the problem of convergence using real life data.

Overview of the binary logistic regression

Logistic regression is a mathematical method for investigating the association of a quantal dependent variable with one or more independent variables that may be binary, categorical or continuous. It is binary logistic regression when the outcome or dependent variable is binary or dichotomous, a common situation in epidemiological research where one is often interested in the survival or death of a patient, the presence or absence of a disease, the success or failure of a treatment or procedure and so on. In such a situation the data is usually coded as 1 if outcome is: yes, true, success, pregnant, died, smoker, etc., or 0 if outcome is: No, false, failure, non-pregnant, alive, non-smoker, etc. (for easy understanding and interpretation of results).

The goal of the analysis is to find the best fitting model to describe and explain the relationship between the dichotomous outcome variable and a set of independent variables. Logistic regression measures the effects of risk factors on the occurrence of a disease or any binary outcome variable of interest while adjusting for other confounding effects of others covariate or the interrelationships between them. The variables that affect the probability of the outcome are measured as odds ratios which are called adjusted odds ratios. Logistic regression generates the coefficients, its standard errors and significant levels in a formula to predict a logit transformation of the probability of occurrence of the dependent variable. This is the well-known assumption that the risk of developing a disease or occurrence of any outcome variable of interest is linearly and additively related to the risk factors (X_i) on the logit scale:

$$\text{logit}(p) = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + \ldots + b_k X_k$$

where p is the probability of occurrence of the characteristic of interest or outcome. The logit transformation is written as the log odds:

$$\text{logit}(p) = \ln\left[\frac{p}{1-p}\right]$$

$$\text{odds} = \frac{p}{1-p} = \frac{\textit{probability of presence of characteristic}}{\textit{probability of absence of characteristic}}$$

Assumptions of the logistic regression analysis

The following assumptions underline the use of the logistic regression:

1. The sample from which data was collected is representative of the population to which generalisation will be made
2. The data was collected at the time when the relationship between the dependent variable and the explanatory variables remain constant.
3. All necessary explanatory variables have been considered and included.
4. The sample size is appropriately determined to support the model.
5. The degree of collinearity of the explanatory variables with one another is not high (measured by the variance inflation factor)
6. Other alternate outcome variables are not included as explanatory variables.
7. Assumptions of the chi square test must also be met.

The logistic regression model

Recall the simple linear regression model expressed by the equation:

$$Y=\beta_0+\beta_1 X_1+\varepsilon \qquad (1)$$

In which Y is an arbitrary observed value of the continuous dependent variable, β_0 is the intercept, β_1 is the regression coefficient, and X_1 is the independent variable.

Equation (1) may be written as

$$\mu_{y/x} = \beta_0 + \beta_1 \qquad (2)$$

when the difference between the observed Y and the regression line is zero i.e. $\varepsilon = 0$

the right hand side of equations (1) and (2) may assume any value between minus infinity and plus infinity. This model is not appropriate when Y is a dichotomous variable because the expected value (or mean) of Y is the probability that Y=1 and is therefore limited to the range 0 through 1. Therefore, equations (1) and (2) become inappropriate. However, if p= P(Y=1), then the ratio p/(1-p) can take on values between 0 and plus infinity. Furthermore, the natural logarithm (ln) of p/(1-p) can take on values between minus infinity and plus infinity just as the right hand side of equations (1) and (2). Therefore, a linear additive relation between the occurrence of the event and risk factor x_1 is:

$$\ln(p/(1-p)) = \beta_0 + \beta_1 x_1 \qquad (3)$$

i.e. the logit transformation of p to ln(p/(1-p). Equation (3) is then called the logistic model because the transformation of equation (3) may also be written as:

$$p = \frac{\exp(\beta 0 + \beta 1x)}{1+\exp(\beta 0 + \beta 1x)} \quad (4)$$

This expression is the inverse of the natural logarithm of the odds that some event will occur. In linear regression, methods of least squares are used. This method minimizes the sum of squared deviations of predicted values from observed values; by solving a system of N linear equations each having N unknown variables that can be solved algebraically using the method of least squares. However, the method of least squares cannot be used to solve logistic regression equation and so a maximum likelihood approach is used. This method is capable of producing minimum variance unbiased estimators for the actual parameters of the logit equation. The log likelihood function for the logit model of equation (3) is given by

$$l(\beta) = \sum \beta xiyi - \sum \ln(1 + \exp(\beta xi) \quad (5)$$

The aim of maximum likelihood is to find a set of values for β that maximize this function. The least squares approach is used to differentiate equation (5) with respect to β, set the derivative equal to 0 and then solve the set of equations. These equations look like the normal equations in the least squares linear regression but y here is now a non- linear function of the x_i's rather than a linear function. In some models, these non- linear equations can be solved for the ML estimator for β. However, for some models and data, these equations have no explicit solutions and must be solved by numerical methods one of which is the Newton-Raphson algorithm which has been well described by Alison [2]. A common problem is when the maximum likelihood estimates of some functions do not exist and the iterative process could not find a solution. In these situations, we observe that the model did not reach convergence. This situation arises when the predictors completely predict the outcome variable which is described as complete separation or quasi complete separation. Other data situations that can lead to non-convergence of the model are when the data has a large proportion of empty cells or the numbers of cases are few in relation to the number of variables or there is multicollinearity among the independent variables.

Unfortunately, the new researcher or user of the logistic regression model may not be aware of these problems and may not have a clue about why this has happened. We explain below how to identify this problem of non-convergence and propose solutions using real life data from the Nigerian Demographic and Health Survey; NDHS 2008 survey [3].

Illustration with real life dataset

In a posthoc analysis of the Nigerian Demographic and Health Survey, with the objective of identifying the risk factors for "tobacco use" using logistic regression, the following variables were included: Chewing tobacco (V463), sex of household head (V151), type of place of residence (V102), and highest level of education (V106). The dependent variable is chewing tobacco coded as yes/no, while the following were the independent variables: Sex of household head coded as male and female (1 and 2), type of place of residence: rural versus urban coded as 1 and 2, highest level of education coded as 0 = no education, 1 = primary education, 2 = secondary education.

A logistic regression of the outcome on the predictors was fitted using STATA [4]:

xi: logit v463c i.v151 i.v102 i.v106.

Output is given by:

```
xi: logit v463c  i.v151 i.v106 i.v102, or
i.v151_Iv151_1-2(naturally coded; _Iv151_1 omitted)
i.v106_Iv106_0-3  (naturally coded; _Iv106_0 omitted)
i.v102_Iv102_1-2  (naturally coded; _Iv102_1 omitted)
note: _Iv151_2 != 0 predicts failure perfectly
 _Iv151_2 dropped and 2663 obs not used
note: _Iv106_3 != 0 predicts failure perfectly
_Iv106_3 dropped and 1139 obs not used
Iteration 0:   log likelihood = -204.34159
Iteration 1:   log likelihood = -202.40448
Iteration 2:   log likelihood = -202.25274
Iteration 3:   log likelihood = -202.25196
Logistic regression                Number of obs =     24783
LR chi²(3)     =      4.18 Prob> chi²  =     0.2427
Log likelihood = -202.25196 Pseudo R²  =     0.0102
```

v463c	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_Iv106_1	.7924818	.3836436	-0.48	0.631	.3068451	2.046724
_Iv106_2	.4844566	.2808965	-1.25	0.211	.1554939	1.509373
_Iv102_2	.4434551	.1860791	-1.94	0.053	.1948387	1.009309

In addition, the SPSS [5] output of fitting a logistic regression of the outcome on the predictors is given by:

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	404.504[a]	.000	.028

a. Estimation terminated at iteration number 20 because maximum iterations has been reached. Final solution cannot be found.

Variables in the equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	V151(1)	14.080	761.726	.000	1	.985	1302383.456
	V106			1.588	3	.662	
	V106(1)	14.689	1067.670	.000	1	.989	2394747.962
	V106(2)	14.456	1067.670	.000	1	.989	1897794.157
	V106(3)	13.964	1067.670	.000	1	.990	1160151.549
	V102(1)	.813	.420	3.755	1	.053	2.255
	Constant	-35.694	1311.544	.001	1	.978	.000

a. Variable(s) entered on step 1: V102.

Explanations of computer outputs

Consider the exact output produced from STATA. Note that STATA gives clear warning messages such as: "lv151_2 != 0 predicts failure perfectly" and also note: "lv106_3 != 0 predicts failure perfectly" i.e. sex of household = "female" and highest level of education= "higher" predicts the outcome "chewing tobacco" perfectly; It therefore dropped all cases where sex = female and highest level of education = higher and so were omitted in the output.

Examination of the output from SPSS, revealed that the problem was not mentioned precisely. It tried to iterate and stopped the process when it couldn't reach a solution after a number of iterations. So the researcher needs to find out why the computation didn't converge. This information can be seen in the extremely large parameter estimates and their standard errors produced. Let's examine the cross tabulation of these variables with the outcome: "chewing tobacco".

*Sex of household head * Chewing tobacco Cross tabulation*

Count

		Chewing tobacco No	Chewing tobacco Yes	Total
Sex of household head	Male	25896	26	25922
	Female	2663	0	2663
Total		28559	26	28585

*Highest educational level * Chewing tobacco Cross tabulation*

Count

		Chewing tobacco No	Chewing tobacco Yes	Total
Highest educational level	No Education	14387	16	14387
	Primary	6532	6	6538
	Secondary	6322	4	6326
	Higher	1334	0	1334
Total		28559	26	28585

Note that this problem of separation does not occur with place of residence (rural /urban) as there is no

Type of place of residence * Chewing tobacco Cross tabulation

Count

		Chewing tobacco No	Chewing tobacco Yes	Total
Type of place of residence	Urban	7583	10	7593
	Rural	20976	16	20992
Total		28559	26	28585

zero in the contingency table with outcome has shown below.

This problem can be fixed by doing any of the proposed options below.

First, variables causing this separation problem may be deleted. Second, collapse categories. This is not possible for the variable sex but it is possible for the variable "highest level of education". Third option is to do nothing and report the likelihood ratio chi squares since the maximum likelihood for other predictor variables are still valid. For the STATA output, the LR was 4.18, p= 0.2427 and the Log likelihood = -202.25196. The chi-square test is used to indicate how well the logistic regression model fits the data. For the SPSS, the LR which was reported as -2log likelihood was 404.504. The fourth option, exact inference is not possible because it only applies to small sample sizes [6, 7]. The last option, using the Bayesian techniques by using the software, BUGGS was also not considered as the software may not be readily available for use by most researchers.

In summary, we propose that the variable "level of education" should be collapsed and the variable "sex of household head" should be removed. When this was done the output below is the results of the new analysis.

For STATA, the new output is

```
xi: logit v463c i.v106new i.v102, or
i.v106new _Iv106new_0-2 (naturally coded; _Iv106new_0 omitted)
i.v102 _Iv102_1-2 (naturally coded; _Iv102_1 omitted)
```

For SPSS, output is given as below:

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	v106new			3.277	2	.194	
	v106new(1)	1.050	.582	3.251	1	.071	2.857
	v106new(2)	.747	.653	1.311	1	.252	2.111
	V102(1)	.811	.421	3.712	1	.054	2.250
	Constant	-8.025	.575	194.947	1	.000	.000

a. Variable(s) entered on step 1: V102.

```
Iteration 0:  log likelihood = -208.05423
Iteration 1:  log likelihood = -205.48383
Iteration 2:  log likelihood = -205.28378
Iteration 3:  log likelihood = -205.28232
Iteration 4:  log likelihood = -205.28232

Logistic regression                 Number of obs =     28585
                                    LR chi2(3)    =      5.54
Prob> chi2     =     0.1360
Log likelihood = -205.28232         Pseudo R2     =    0.0133

------------------------------------------------------------------------------
v463c | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
------+-----------------------------------------------------------------------
_Iv106new_1 |   .739   .3573    -0.63   0.532     .2864     1.9067
_Iv106new_2 |   .3500  .2037    -1.80   0.071     .1118     1.0955
_Iv102_2 |    .4444   .1870    -1.93   0.054     .1947     1.014
------------------------------------------------------------------------------
```

Explanation of separation

Complete separation

A complete separation occurs when the two categories of the outcome variable can be separated perfectly by values of one of the independent variables. Allison described this with hypothetical data [8]. Complete separation or perfect prediction can occur for several reasons. For example, if one is studying an age-related disease (presence/absence) and age is one of the predictors, there may be subgroups (e.g., women over 65yrs all of whom have the disease. Complete separation may also occur if there is a coding error or one mistakenly includes another version of the outcome as a predictor. For example, if a continuous variable X is dichotomized into a binary variable Y. If interest is in the relationship between Y and some predictor variables and if X is included as a predictor variable, the problem of perfect prediction will occur, since by definition, Y separates X completely.

Quasi separation

Quasi-complete separation in logistic regression happens when the outcome variable separates a predictor variable or a combination of predictor variables to a certain degree. Allison has also described this problem with hypothetical data. [8].

Diagnosis of the problem

Of the 2 methods, quasi complete separation is the most common. The problem occurs whenever there is a zero in any cell of a 2x2 table then, the maximum likelihood estimate of the logistic slope coefficient does not exist. So for any dichotomous independent variable in a logistic regression, if there is a zero in the 2x2 table formed by that variable and the dependent variable, the ML estimate for the regression coefficient will not exist. Software such as SPSS and STATA do give clear warning messages when these problems occur so it is very easy to detect. However, even if clear messages are not given, the problem can be identified by examining the coefficients and their standard errors. Variables with non-existence coefficient will definitely have large parameter estimates, usually greater than 5 and very large standard errors, producing Wald statistics that are close to zero. If any of these variables is an indicator variable, then produce a 2x2 table for each variable and the dependent variable. A frequency of zero in any single cell of the table means quasi complete separation. If there are two zeros in the table, then it is complete separation. So the variables that are causing the separation problems would have been identified.

Solutions to the problem of Quasi separation

The following solutions have been proposed by Altman [2].

1. Variables causing the problems may be deleted from the model i.e. those variables whose coefficients do not converge. However this may not

Table 1: Assessment of logistic regression analysis in published articles.

Year of Publication	Number of Journal editions	Number of Articles	Logistic Regression used	Logistic Regression Stated	Logistic Regression described	Convergence
2004	2	36(6.2)	2(5.0)	1	0	1
2005	4	70(12.0)	6(15.0)	2	0	1
2006	4	63(10.8)	1(2.5)	0	0	0
2007	2	31(5.3)	0(0)	0	0	0
2008	4	56(9.6)	5(12.5)	4	1	0
2009	5	65(11.2)	6(15.0)	2	0	1
2010	5	72(12.4)	2(5.0)	2	0	1
2011	4	53 (9.1)	7(17.5)	5	1	1
2012	5	87(15.0)	6(15.0)	4	1	1
2013	4	48(8.3)	5(12.5)	4	0	0
Total	39	581(100.0)	40(100.0)	24	3	6

Table 2: Multivariate logistic regression of ever smoked tobacco and determinant of tobacco smoking in HIV patients.

Characteristics	Ever used	Never used	Odd ratio	C.I	P-value
Age range (years)					
15-19	*0	3	1.00		
20-29	18	51	1.01	0.55-1.85	0.979
30-39	21	105	0.42	0.24-0.74	0.002
40-49	33	51	2.43	1.41-4.17	0.001
50-59	9	18	1.48	0.64-3.44	0.361
›60	*0	3	-	-	-
Sex					
Female	24	153	1.00		
Male	57	78	4.66	2.69-8.07	‹0.001
Education					
None formal	6	27	1.00		
Primary	63	165	1.46	0.89-2.48	0.166
Secondary	60	132	0.84	0.50-1.38	0.481
Tertiary	72	177	1.23	0.68-2.22	0.493
Occupational Class					
Group1	27	96	1.00	0.47-1.32	0.361
Group2	48	150	0.79	0.75-2.63	0.285
Group3	18	39	1.41	-	-
Group4	*0	6	-	-	-
Alcohol drinking					
No	36	183	1.00		
Yes	45	48	4.77	2.77-8.19	‹0.001
CD4 Count					
≥500 cell/mm3	3	6	1.00		
200-499	18	114	0.29	0.16-0.53	‹0.001
cell/mm3	60	111	3.09	1.76-5.41	‹0.001
‹200 cell/mm3					
Total = 312	81	231			

Extracted from Desalu *et al* 2009. *complete separation

be a good choice as the variable in question may have a strong relationship or effect with the dependent variable.

2. Categories of the variables may be combined or collapsed if the number of categories is large and sample size is small. However if the variable in question has just 2 categories, this option is not feasible.

Table3: Bivariate analysis showing risk factors for PTB infection by Z-N

Variables	Z-N +ve No(%)	Z-N –ve No(%)	Odds Ratio	95% CI	P-value
Age(yrs)					
‹20(n =17)	01(5.4)	16(94.1)	1.9	(0.23-16.3)	0.37
›20(n= 254)	08(3.1)	246(96.9)			
Sex					
Male	04(3.4)	113(96.6)	1.1	(0.28-4.0)	0.01
Female	05(3.2)	149(96.8)			
Profession					
Skilled(n=117)	**0(0)	73(100.0)	1.05	(1.02-1.08)	***3.43
Unskilled(n= 198)	09(4.5)	189(95.5)			
History of chronic cough					
Yes(n= 17)	**(0.0)	17(100.0)	1.1	(1.04-1.19)	0.62
No(n= 254)	9(3.5)	245(96.5)			
Smoking					
Yes(n= 14)	**(0.0)	14(100.0)	1.1	(1/02-1.17)	0.51
No(n= 257)	09(3.5)	248(96.5)			
Alcohol ingestion					
Yes(n= 59)	06(10.2)	53(89.8)	7.89	(1.91-32-57)	***11.02
No(n= 212)	03(1.4)	209(98.6)			
Contact with patient With chronic cough					
Yes(n= 17)	**0(0)	17(100.0)	1.04	(1.01-1.06)	0.62
No(n= 254)	09(3.5)	245(96.5)			
History of previous Skin test (Mantoux)					
Yes(n= 100)	06(6.0)	94(94.0)	3.57	(0.87-14.62)	***3.54
No(n= 171)	03(1.8)	168(98.2)			
Previous treatment forPTB					
Yes(n= 49)	03(6.1)	46(93.9)	2.35	(0.57-9.73)	***1.46
No(n= 222)	06(2.7)	216(97.3)			
History of BCG vaccination					
Yes(n= 89)	03(3.4)	86(96.6)	1.02	(0.25-4.19)	0.00
No(n= 182)	06(3.3)	176(96.7)			
Period of working inTB unit					
‹2 yrs (n =111)	04(3.6)	107(96.4)	1.16	(0.30-4.42)	0.05
›2 yrs (n= 160)	05(3.1)	155(98.1)			

Extracted from Kehinde *et al* 2010 **Quasi complete separation ***wrong p values

3. Do nothing and report the likelihood ratio chi squares. Other variables will definitely have maximum likelihood estimates which can still be reported. Leave the problem variables in the model but report their coefficients (as -infinity, + infinity). Even though the Standard Error and Wald statistics for these problematic variables are incorrect, the likelihood ratio tests for the null hypotheses that the coefficients are zero are still valid.

4. Make exact inference. Even without separation, Maximum likelihood (ML) estimates don't have good properties with small sample sizes. So one can omit Maximum likelihood and do exact logistic regression. This procedure produces exact p values for the null hypothesis that each predictor variable has a coefficient of 0, conditional on all the other predictors. These p values are not based on large sample chi square approximations but on permutations of the data and are not affected by complete or quasi complete separation. These are computationally feasible only for small sample.

5. Use Bayesian techniques. If all of the above descriptions are not feasible, the Bayesian methods should be done using the BUGGS software.

Table 4: Bivariate analysis showing risk factors for PTB infection by culture

Variables	Culture+ve No(%)	Culture -ve No(%)	Odds Ratio	95% CI	P-value
Age(yrs)					
‹20(n =17)	**0(0)	17(100.0)	1.04	(1.01-1.04)	0.42
›20(n= 254)	06(2.4)	248(97.6)			
Sex					
Male	06(5.1)	111(94.9)	0.95	(0.91-0.99)	***8.08
Female	**0(0)	154(100.0)			
Profession					
Skilled(n=117)	**0(0)	73(100.0)	1.03	(1.00-1.06)	***2.26
Unskilled(n= 198)	06(3.0)	192(97.0)			
History of chronic cough					
Yes(n= 17)	01(5.9)	16(94.1)	3.11	(0.34-28.30)	***1.13
No(n= 254)	05(2.0)	249(98.0)			
Smoking					
Yes(n= 14)	**0(0)	14(100.0)	1.03	(1.01-1.04)	0.33
No(n= 257)	06(0.8)	251(99.2)			
Alcohol ingestion					
Yes(n= 59)	05(8.5)	54(91.5)	19.5	(2.24-170-73)	***13.65
No(n= 212)	01(0.5)	211(99.5)			
Contact with patient With chronic cough					
Yes(n= 17)	01(5.9)	16(94.1)	3.11	(0.34-77.7)	***1.13
No(n= 254)	05(2.0)	249(98.0)			
History of previous Skin test (Mantoux)					
Yes(n= 100)	05(5.0)	95(95.0)	8.94	(1.03-77.7)	***5.68
No(n= 171)	01(0.06)	170(99.4)			
Previous treatment for PTB					
Yes(n= 49)	02(4.1)	47(95.9)	2.32	(041-13.03)	0.96
No(n= 222)	04(1.8)	218(98.2)			
History of BCG vaccination					
Yes(n= 89)	02(2.3)	87(97.7)	1.02	(0.195.70)	0.001
No(n= 182)	04(2.2)	178(97.8)			
Period of working inTB unit					
‹2 yrs (n =111)	03(2.7)	108(97.3)	1.45	(2.29-7.34)	0.21
›2 yrs (n= 160)	03(1.9)	157(98.1)			

Extracted from Kehinde *et al* 2010. **Quasi complete separation, ***wrong p values

Assessment of Journal Articles

We critically evaluated all the articles that were published in the African Journal of Medicine and Medical Sciences between January2004 and December 2013 that employed logistic regression analysis. A total of 581 articles were published, of which 40(6.9%) used binary logistic regression. However, 24 (60.0%) stated the use of logistic regression in the methodology, while only 3 (12.5%) of these properly described the procedures in the methodology. None of the articles assessed model fit while majority presented insufficient details of the procedures. In addition, of the 40 that used the logistic regression, the problem of convergence occurred in 6(15.0%) of the articles. Table 1 shows the distribution of the findings of the assessment of the articles.

Excerpts of tables from articles where convergence occurred

Table 2 illustrates the problem of complete separation, while tables 3-6 shows quasi complete

Table 5: Predictors of PTB infection as diagnosed by Microscopy

Variable	Odd ratio	Confidence Limit	P-value
Age			
‹20yrs	1.00		
›20yrs	2.04	0.20-20.6	0.55
BCG vaccination			
Yes	0.86	0.20-3.6	0.83
No	1.00		
Smoking			
Yes	2.08	0.39-11.1	0.39
No	1.00		
Contact with patient With chronic cough			
Yes	2.58	****0.67-24.9	0.41
No	1.00		
Period of working inTB unit			
‹2 yrs	0.84	0.20-3.5	0.82
›2 yrs	1.00		

*Extracted from Kehinde et al 2010. ****wide CI*

Table 6: Predictors of PTB infection as diagnosed by Culture

Variable	Odd ratio	Confidence Limit	P-value
BCG vaccination			
Yes	0.76	0.13-4.4	0.76
No	1.00		
Smoking			
Yes	1.13	0.12-10.4	0.91
No	1.00		
Contact with patient With chronic cough			
Yes	5.11	****0.48-54.6	0.18
No	1.00		
Period of working inTB unit			
‹2 yrs	1.00	0.22-6.5	0.82
›2 yrs	1.21		

*Extracted from Kehinde et al 2010. ****wide CI*

separation. In table 2, the two categories of the outcome variable (ever smoked tobacco) was separated perfectly by values of one of the independent variables (age and occupational class). Tables 3 and 4 shows the bivariate analyses that illustrates the occurrence of zeros in the 2x2 table formed by the dichotomous independent variables (such as history of chronic cough, smoking, and contact with patient with chronic cough) and the dependent variable (Z-N +ve, Z-N –ve) while tables 5 and 6 show the logistic regression analysis. In tables 5 and 6, we observed that the variables (such as smoking, and contact with patient with chronic cough) that caused the quasi complete separation have unusually wide confidence intervals such as 0.67-24.9, and 0.48-54.6. However, the standard errors were not presented which may have been used to further confirm the occurrence of quasi complete separation. Furthermore, of importance is the p-values reported in tables 3 and 4. We observed that these p values are greater than one and these are errors which should not have been reported.

Conclusion

In this analysis, we have described the problems of convergence in binary logistic regression, how to identify the problem using data from real life studies and also proposed solutions to the problems. We also evaluated published articles reporting the procedures and identified the problem in 6 articles.

It is important to note that these problems can occur in multinomial logistic regression as well. In conclusion, logistic regression tended to be poorly implemented in studies published between 2004 and 2013. Our findings showed that the procedure may not be well understood by researchers since very few described the process in their reports and may be totally unaware of the problem of convergence or how to deal with it. Researchers need to report the type of logistic regression, how variables were entered into the model, how model fit was assessed, and how they dealt with the problem of convergence when it occurred. In addition, we recommend that medical journals should include Biostatisticians in their editorial teams.

References

1. James L. An Insight on the use of Multiple Logistic Regression Analysis to Estimate Association between Risk Factor and Disease Occurrence. International Journal of Epidemiology 1986;15(1): 22-29.
2. Allison P.D. Convergence Failures in Logistic Regression. SAS Global Forum 2008. Statistics and Data Analysis. Paper 360-2008.
3. Nigeria Demographic and Health Survey (NDHS) 2008.
4. StataCorp, Stata Statistical Software: Release 10. College Station, TX: Stata Corp LP, 2007.
5. Statistical Package for the Social Sciences SPSS Inc., Chicago, IL, USA.
6. Hirji KF., Mehta CR. and Patel NR. "Computing Distributions for Exact Logistic Regression," *JASA*, 1987; 82, 1110–1117.

7. Cyrus R M and Nitin R. P. Exact logistic regression: theory and examples. Statistics in Medicine. 2007; 14(19): 2143-2160.
8. Altman M., Jeff G. and McDonald M. Convergence Problems in Logistic Regression In Numerical Issues in Statistical Computing for the social scientist. Allison Paul. A Wiley- Inter-science Publication. 2004 John Wiley and Sons, INC. 219- 233.
9 Desalu O.O, Oluboyo P.O, Olokoba A.B, *et al.* Prevalence and determinants of tobacco smoking among HIV patients in North Eastern, Nigeria.2009; 38:103-108.
10. Kehinde A.O, Baba A, Bakare R.A, *et al.* Risk factors for pulmonary tuberculosis among health-care workers in Ibadan, Nigeria. 2010; 39:105-112.