An insight on the Misuse of Logistic Regression Model in the face of Non Convergence

Oyindamola B. Yusuf (Ph.D)

Department of Epidemiology and Medical Statistics, College of Medicine, University of Ibadan, Ibadan, Nigeria

Introduction

- Logistic regression is a mathematical method for investigating the association of a binary dependent variable with one or more independent variables that may be binary, categorical or continuous

- Widely used in epidemiological/public health research

- Understanding of the underlying principles is important

- Techniques fail because the algorithm for maximum likelihood does not converge.

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

- Can also occur if there is a coding error or an alternate outcome variable is mistakenly included as a predictor

- Allison (2004) described why numerical algorithm fail to converge

Quasi Complete Separation

- Occurs whenever there is a zero in any cell of a 2x2 table

- 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.

Solutions to the problem

- Delete the variables

- Collapse the number of variable categories

- Do nothing and report likelihood ratio chi-squares

- Make exact inference

- Use Bayesian techniques

Objective

- To evaluate convergence issues in published articles that employed logistic regression analysis.

Methods

- We reviewed published articles between 2004 and 2013 in an African Journal.

Results

- A total of 581 articles were published
- Forty (6.9%) used binary logistic regression.
- Twenty-four (60.0%) stated the use of logistic regression in the methodology,
- Three (12.5%) described the procedures
- None of the articles assessed model fit
- Majority presented insufficient details of the procedures
- Convergence occurred in 6(15.0%) of the articles.

Assessment of Logistic Regression Analysis in Published Articles

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

Multivariate logistic regression of ever smoked tobacco and determinant of tobacco smoking in HIV patients

Variables	Ever used	Never used	Odds 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	-	-	-
Occupational class					
Group 1	27	96	1.00		
Group 2	48	150	0.79	0.47-1.32	0.361
Group 3	18	39	1.41	0.75-2.63	0.285
Group 4	*0	6	-	-	-
Total =312	81	231			

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

Bivariate analysis showing risk factors for PTB infection by Z–N

Variables	Z-N +ve No(%)	Z-N -ve No(%)	Odds ratio	C.I.	P-value
History of chronic cough					
Yes (n=17)	*0(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(0.0)	141(100.0)	1.1	(1.02-1.17)	0.51
No (n-257)	9(3.5)	248(96.5)			
Contact with patient with chronic cough					
Yes (n=17)	*0(0.0)	17(100.0)	1.04	(1.01-1.06)	0.62
No (n=254)	9(3.5)	245(96.5)			

*Extracted from Kehinde et al 2010 *Quasi complete separation*

Bivariate analysis showing risk factors for PTB infection by culture

Variables	Culture +ve No(%)	Culture -ve No (%)	Odds ratio	C.I.	P-value
Age (years)					
<20 (n=17)	*0(0.0)	17(100.0)	1.04	(1.01-1.04)	0.42
>20(n=254)	6(2.4)	248(97.6)			
Smoking					
Yes (n=14)	*0(0.0)	14(100.0)	1.03	(1.01-1.04)	0.33
No (n=257)	6(0.8)	251(99.2)			

*Extracted from Kehinde et al 2010 *Quasi complete separation*

CONCLUSION

- Logistic regression was poorly implemented in studies published between 2004 and 2013

- The procedure may not be well understood by researchers

- Researchers may be 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 the problem of convergence was resolved when it occurred

REFERENCES

- Allison Paul, (2004) Convergence Problems in Logistic Regression In Altman Micah, Gill Jeff and McDonald Michael . Numerical Issues in Statistical Computing for the social scientist . A Wiley-Inter science Publication. John Wiley & Sons, INC. 219- 233

- Desalu O.O, Oluboyo P.O, Olokoba A.B. et al. (2009). Prevalence and determinants of tobacco smoking among HIV patients in North Eastern, Nigeria. Afr. J. Med. Med. Sci. 38:103-108

- Kehinde A.O, Baba A, Bakare R.A. et al. (2010). Risk factors for pulmonary tuberculosis among health-workers in Ibadan, Nigeria. Afr. J. Med. Med. Sci. 39:105-112

Thank you for listening