Biostatistics and Biometrics
Open Access Journal
ISSN: 2573-2633

Juniper Publishers
key to the Researchers

Review Article
Volume 4 Issue 5 - February 2018
DOI: 10.19080/BBOAJ.2018.04.555649

Biostat Biometrics Open Acc J

Be Wary of Using Poisson Regression to Estimate Risk and Relative Risk

Zhu C[1], Blizzard L[1], Stankovich J[2], Wills K[1] and Hosmer DW[3]

[1]*Menzies Institute for Medical Research, University of Tasmania, Australia*

[2]*School of Medicine of Tasmania, Australia*

[3]*Department of Public Health, University of Massachusetts, USA*

Submission: November 17, 2017; **Published:** February 09, 2018

***Corresponding author:** Leigh Blizzard, Professor, Senior Member and Principal Researcher Menzies Institute for Medical Research, University of Tasmania, Private Bag 23, Hobart TAS 7001, Tel: 61 3 6226-7719, Fax: 61-3-6226-7704, Email: Leigh.Blizzard@utas.edu.au

Abstract

Fitting a log binomial model to binary outcome data makes it possible to estimate risk and relative risk for follow-up data, and prevalence and prevalence ratios for cross-sectional data. However, the fitting algorithm may fail to converge when the maximum likelihood solution is on the boundary of the allowable parameter space. Some authorities recommend switching to Poisson regression with robust standard errors to approximate the coefficients of the log binomial model in those circumstances. This solves the problem of non-convergence, but results in errors in the coefficient estimates that may be substantial particularly when the maximum fitted value is large. The paradox is that the circumstances in which the modified Poisson approach is needed to overcome estimation problems are the same circumstances when the error in using it is greatest. We recommend that practitioners should be wary of using modified Poisson regression to approximate risk and relative risk.

Keywords: Relative risk; Log binomial model; Poisson regression; Boundary point

Abbreviations: GLM: Generalized Linear Model; LBM: Log Binomial Model; ML: Maximum Likelihood

Introduction

Direct estimation of risk and relative risk for prospective studies requires the fitting of a generalized linear model (GLM) with a binomial error distribution and logarithmic link function. This is the log binomial model (LBM). It provides estimates of probabilities and conditional probabilities that are directly interpretable and are preferred as measures of occurrence and association [1]. An added benefit is that the model provides interpretable estimates of prevalence and prevalence ratios for cross-sectional studies.

The drawback of the LBM is that the logarithmic link function maps the probability of the event onto the negative real line. This imposes bounds on the allowable parameter space for the model coefficients. Estimation subject to boundedness is problematic, but standard methods for fitting GLMs may fail to converge to the maximum likelihood (ML) estimates for a LBM if the fitted probabilities are allowed to equal or exceed unity. Even if the iterations converge and the approximate solution is reasonably accurate, there will be difficulties interpreting and applying the fitted values if one or more of them exceeds unity.

There are several work-around methods to approximate the solution of a LBM and circumvent the problems inherent in its estimation. Other than substituting estimates from a logistic regression model, the modified Poisson regression method has gained the most traction [2]. This method involves fitting a GLM with a Poisson error distribution and logarithmic link [3], and using the sandwich estimator to obtain variance estimates that are robust to the error misspecification [4]. Carter et al. [5] showed that the coefficient estimates from a Poisson regression model consistently estimate the coefficients from the LBM, and that the information sandwich estimator of the covariance matrix of the Poisson regression fit is a consistent estimator of the covariance matrix of estimated coefficients from a log binomial fit.

This approach requires no data modification and can be easily performed using widely available software. It seemingly resolves the convergence issues because Poisson regression maps the logarithm of the count of events to the entire real line. Thus, estimation can proceed even if the linear predictor is non-negative. This means that the resulting coefficient estimates may yield fitted values for the LBM that are inadmissible as probabilities because they exceed unity. Some authors have suggested that these can safely be ignored [6]. However, the approximate solution may be subject to considerable error. Our eyes to this were opened by example data in a recent paper by Williamson et al. [2] exploring sources of failed convergence of

the LBM (Table 1). The authors attempt to fit the single covariate LBM

$$\Pr(Y=1|X)=\beta_0+\beta_1 X$$

Table 1: Example data from Williamson et al. [5].

Exposure	Event	No event	Total
x = –1	10	8	18
x = 0	18	9	27
x = 1	5	0	5

Where, Y is a binary (0/1) outcome indicator with $Y=1$ denoting an event, and X is a covariate taking values $x=\{-1,0,1\}$. The maximum likelihood (ML) solution for a LBM model of these data is $\hat{\beta}_0=-0.344616$ and $\hat{\beta}_1=0.344616.$ This solution is on a boundary $\beta_0+\beta_1=1$ of the allowable parameter space. The authors estimate the LBM with SAS (version 9.2), R (version 2.12.1), Stata (version 11.1) and SPSS (version 19) and report that only SAS is successful in finding the ML solution, though a warning is given in the SAS output that the convergence is questionable because the solution appears to be on the boundary.

In these circumstances, the analyst might follow well-intentioned advice to fit a modified Poisson model [5,7]. If so, the coefficient estimates $\hat{\beta}_0^{poi}=-0.3596015$ and $\hat{\beta}_1^{poi}=0.2713417$ would be obtained. The percent error in the slope estimate $\hat{\beta}_1^{poi}$ is a staggering 21.3%. In fitting a modified Poisson model with four categorical covariates, Marschner & Gillett [8] also found errors greater than 20% for some categories of the risk factors. To investigate, we simulated data from a LBM with a single continuous covariate and for a range of values of the parameters β_0 and β_1. Values of the continuous covariate were drawn at random from a distribution uniform on the range

$$\left[\frac{\ln(0.01)-\beta_0}{\beta_1},\frac{\ln(1)-\beta_0}{\beta_1}\right].$$

A value at the top end of that range would lie on the boundary of the parameter space. For each observation $i=1,2,...n,$ values of $Y_i=1$ were assigned to the outcome indicator if a random drawing from a distribution uniform on was less than the design value $\pi(x_i)=\exp(\beta_0+\beta_1 x_i)$ of the LBM probability. Datasets of size $n=500$ were chosen as representative of many encountered in practice, and 10,000 replications were drawn for each setting. Table 2 shows percentiles of absolute percent error in the Poisson estimate $\hat{\beta}_1^{poi}$ of the LBM slope parameter β_1 relative to the ML estimate $\hat{\beta}_1$. The results were similar irrespective of the design values β_0 and β_1. For brevity, they are given for the setting $\beta_0=\ln(0.3)$ and $\beta_1=\ln(1.5)$ only. There was a moderate correlation $(r=0.48)$ between absolute percent error and the maximum fitted value from the modified Poisson model. This deterioration in the performance of the modified Poisson model with the size of the fitted value has been identified previously [8]. Relative error was at least 10% on 7.6 percent of simulations overall, on 11.3 percent of simulations when the ML solution was on the boundary, on 14.3 percent of simulations when a Poisson fitted value exceeded unity, and on 15.6 percent of simulations when the ML solution was on the boundary and additionally the Poisson fitted value exceeded unity.

Conclusion

We recommend that practitioners be wary of using the modified Poisson approach to estimate a LBM. Whilst errors greater than 20% may be a rarity, the estimates are subject to substantial bias. In the context of confounding, one authority has nominated 10% as the threshold for bias than cannot be ignored [9]. Based on that standard, the modified Poisson method failed on one-in-nine occasions when the ML solution was on the boundary, and on almost one-in-six occasions when additionally the Poisson fitted value exceeded unity. The relevance of a boundary solution is that it brings about the failure of standard fitting algorithms. The paradox is that this is the circumstance that prompts practitioners to switch to the modified Poisson approach. There are substantial error rates even when the solution is not on the boundary, but the modified Poisson approach is not required in those circumstances because standard software for fitting the LBM should be successful in iterating to the ML solution.

Table 2: Percentiles of absolute percent error in $\hat{\beta}_1^{poi}$ as an estimate of ML $\hat{\beta}_1$.

Maximum Poisson fitted value*	Number of boundary points	Number of replicates	Percentile			
			50th	90th	95th	100th
<1	0,1	5145	2.5	6.4	7.6	15
≥1	0	1052	5.1	9.3	10.7	19.3
≥1	1	3803	4.7	10.9	12.8	23.9

How to cite this article: Zhu C, Blizzard L, Stankovich J, Wills K, Hosmer DW. Be Wary of Using Poisson Regression to Estimate Risk and Relative Risk. Biostat Biometrics Open Acc J. 2018; 4(5): 555649. DOI: 10.19080/BBOAJ.2018.04.555649

References

1. Greenland S (1987) Interpretation and choice of effect measures in epidemiologic analyses. Am J Epidemiol 125(5): 761-768.
2. Williamson T, Eliasziw M, Fick GH (2013) Log-binomial models: exploring failed convergence. Emerg Themes Epidemiol 10(1): 14.
3. McNutt LA, Wu C, Xue X, Hafner JP (2003) Estimating the Relative Risk in Cohort Studies and Clinical Trials of Common Outcomes. Am J Epidemiol 157(10): 940-943.
4. Zou G (2004) A modified poisson regression approach to prospective studies with binary data. Am J Epidemiol 159(7): 702-706.
5. Carter RE, Lipsitz SR, Tilley BC (2005) Quasi-likelihood estimation for relative risk regression models. Biostatistics 6(1): 39-44.
6. Lumley T, Kronmal R, Ma S (2006) Relative risk regression in medical research: models, contrasts, estimators, and algorithms. UW Biostatistics Working Paper Series, Working Paper 293.
7. Spiegelman D, Hertzmark E (2005) Easy SAS calculations for risk or prevalence ratios and differences. Am J Epidemiol 162(3): 199-200.
8. Marschner IC, Gillett AC (2012) Relative risk regression: reliable and flexible methods for log-binomial models. Biostatistics 13(1): 179-192.
9. Greenland S (1989) Modeling and variable selection in epidemiologic analysis. Am J Public Health 79(3): 340-349.

DOI: 10.19080/BBOAJ.2018.04.555649

How to cite this article: Zhu C, Blizzard L, Stankovich J, Wills K, Hosmer DW. Be Wary of Using Poisson Regression to Estimate Risk and Relative Risk. Biostat Biometrics Open Acc J. 2018; 4(5): 555649. DOI: 10.19080/BBOAJ.2018.04.555649