Send Orders for Reprints to reprints@benthamscience.net

18 *The Open Epidemiology Journal,* **2013,** *6,* **18-20**

Open Access

Binary Regression Models with Log-Link in the Cohort Studies

Katri Jalava[*,1], Sirpa Räsänen[2], Kaija Ala-Kojola[2], Saara Nironen[3], Jyrki Möttönen[4] and Jukka Ollgren[1]

[1]*Department of Infectious Disease Surveillance and Control, National Institute for Health and Welfare, Helsinki, Finland*

[2]*Health Services, City of Tampere, Tampere, Finland*

[3]*Health Services, Rauma Town, Rauma, Finland*

[4]*Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland*

Abstract: Regression models have been used to control confounding in food borne cohort studies, logistic regression has been commonly used due to easy converge. However, logistic regression provide estimates for OR only when RR estimate is lower than 10%, an unlikely situation in food borne outbreaks. Recent developments have resolved the binary model convergence problems applying log link. Food items significant in the univariable analysis were included for the multivariable analysis of two recent Finnish norovirus outbreaks. We used both log and logistic regression models in R and Bayesian model in Winbugs by SPSS and R. The log-link model could be used to identify the vehicle in the two norovirus outbreak datasets. Convergence problems were solved using Bayesian modelling. Binary model applying log link provided accurate and useful estimates of RR estimating the true risk, a suitable method of choice for multivariable analysis of outbreak cohort studies.

Keywords: Cohort studies, linear models, regression analysis, risk, outbreak, foodborne illnesses, norovirus.

BACKGROUND

Regression models have been used in analytical outbreak studies when several variables are significant in the univariable analysis. More specifically, they have been used to control confounding in food borne analytical studies [1]. Logistic regression has been commonly used in cohort settings due to its ability to converge in most situations [2-4]. However, logistic regression provides estimates for odds ratio (OR) which only can be used as risk ratio (RR) estimates when the incidence is lower than 10%, an unlikely situation in many food borne outbreaks [5]. From theoretical point of view, log regression models should be used in cohort settings, but often the estimates are on the boundary of the parameter space producing convergence difficulties [6]. To overcome this problem, several different methods have been recently published to enable the convergence of binary models with log link, like maximum likelihood estimation [6] or modified Poisson regression [7]. Attempts have also been made to use mathematical equations to convert OR values to RR values [8]. However, the validity of these equations has been questioned [9] and recent developments in Bayesian modelling have resolved convergence problems of binary log link models [10, 11]. We used a simple binary model with log link Bayesian framework applying an algorithm ensuring confinement of the estimates within the parameter space using Winbugs with applicable results.

The aim of this study was to show the applicability of binary model with log link in outbreak situations. Furthermore, we showed empirically that the estimates from binary models with log link are appropriate estimates for RR unlike those obtained from logistic regression by applying the method to two recent Finnish norovirus outbreaks where multivariable analysis was needed in the study. We further confirmed theoretically these results by showing that the commonly used transformation equation for converting OR's to RR's was invalid.

MATERIALS AND METHODS

We used the Bayesian log regression method for two recent real outbreak datasets with high attack rates requiring log regression due to several significant variables identified in the univariable analysis. Briefly, the first outbreak (outbreak 1) was caused by a norovirus in a working place canteen with attack rate of 53%, total cohort size was 175. In the univariable analysis, pesto chicken with potatoes and salmon in indie sauce were significant (Table **1**). Second outbreak (outbreak 2) was also a norovirus outbreak in a working place canteen with an attack rate of 57%, total cohort size was 74. In the univariable analysis, cold fish items and pasta salad were significant (Table **1**). We used both log and logistic regression models in R (packages glm and glm2) and created a binary Bayesian model with log-link in Winbugs by SPSS and R. R is a free statistical and mathematical software with a number of application

*Address correspondence to this author at the Department of Infectious Disease Surveillance and Control, National Institute for Health and Welfare, Helsinki, Finland; Tel: 00-358-29-524 8914; Fax: 00-358-29-524 8468;
E-mail: katri.jalava@thl.fi

Table 1. Univariable Analysis of the Food Exposures for the "Outbreak 1" and "Outbreak 2" Data

Study	Variable (Day)	RR Estimate (95% Confidence Intervals)	p-Value
Outbreak 1	Vegetable soup in milk (Mo)	1.1 (0.64-1.84)	1
	Fried baltic herring (Mo)	0.58 (0.36-0.94)	0.04
	Beef in sour cream (Mo)	1.0 (0.63-1.59)	1
	Cheese salad (Mo)	0.61 (0.40-0.95)	1
	Quark with lingonberries (Mo)	1.07 (0.62-1.85)	1
	Green salad (Mo)	1.22 (0.45-3.31)	0.64
	Salad with grated vegetables (Mo)	1.12 (0.61-2.06)	0.73
	Mushed vegetable soup (Tue)	0.81 (0.42-1.55)	0.53
	Pesto chicken with potatoes (Tue)	1.58 (1.0-2.51)	0.059
	Salmon in indie sauce(Tue)	0.62 (0.42-0.90)	0.053
	Fish and shrimp salad (Tue)	0.81 (0.48-1.38)	0.56
	Chocolate foam dessert (Tue)	0.94 (0.57-1.55)	1
	Green salad (Tue)	1.01 (0.49-2.11)	1
	Salad with grated vegetables (Tue)	0.98 (0.60-1.60)	1
	Meat soup (Wed)	0.83 (0.53-1.30)	0.54
	Meat with onion (Wed)	0.98 (0.68-1.41)	1
	Broiler cassarolle (Wed)	0.87 (0.55-1.37)	0.58
	Greek salad (Wed)	0.77 (0.50-1.16)	0.25
	Ice cream portion (Wed)	0.91 (0.63-1.32)	0.76
	Minced meat patties (Wed)	0.47 (0.15-1.48)	0.15
	Broiler sauce (Wed)	0.71 (0.31-1.64)	0.38
	Green salad (Wed)	0.74 (0.50-1.10)	0.41
Outbreak 2	Cold fish	2.28 (1.43-3.64)	0.00014
	Melon salad	1.54 (0.97-2.44)	0.058
	Pasta salad	1.85 (1.22-2.80)	0.0039
	Vegetable salad	2.31 (0.65-8.19)	0.15
	Meat dish with lingonberries	1.17 (0.50-2.73)	1

packages available [12]. The data and code for the respective Bayesian model are presented **in Supplementary Material 1-6**, cases with missing data were excluded. The mathematical proof of the invalidity of the convergence of OR to RR is presented in **Supplementary Material 7**.

RESULTS AND DISCUSSION

Binary model with log link should be used for multivariable analysis of cohort studies in outbreak situations when attack rates are >10% [8]. However, due to convergence problems logistic regression models have been used in practical situations even with higher attack rates [2-4]. Also formulas for converting odds ratios to risk ratios have been suggested [8] but the validity has already previously been questioned [9]. We applied Bayesian binary regression modelling with log-link with good convergence, the data handling was done in SPSS and Winbugs was used through R. Log link model provided accurate and useful estimates of RR estimating the true risk.

In the outbreak 1, the exposure date was determined to be Tuesday based on the epidemic curve and incubation period for norovirus infection (data not shown). Of the food exposures served during that date, pesto chicken with potatoes was identified with higher risk in the univariable analysis, and salmon in indie sauce with a lower risk (Table **1**). The Bayesian log-regression model was used to demonstrate dose response (Table **2**). In the univariable analysis of the outbreak 2, cold fish and pasta salad had high risk ratios (Table **1**). The Bayesian binary model with log link identified cold fish as a vehicle of the outbreak (Table **2**). Both with outbreak 1 and outbreak 2, the logistic OR estimates were higher than the RR estimates from the log-link model.

The most notable difference between the logistic and log regression models was the magnitude of the point estimates. Overall, the logistic regression model gave higher point estimates thus overestimating the risk estimates. However, the point estimates and confidence intervals for the log link models were much lower compared to the logistic regression thus better estimating the true risk. The log regression theoretically estimates risk ratios in the population and the obtained RR's were close to those obtained by the

Table 2. Univariable and Bayesian Log Regression Models for the "Outbreak 1" and "Outbreak 2" Data

Study	Variable	Univariable Analysis, RR Estimate (95% Confidence Intervals)	Bayesian Log Regression, RR Estimate (95% Confidence Intervals)	Logistic Regression, OR Estimate (95% Confidence Intervals)
Outbreak 1	Pesto chicken with potatoes	1.6 (1.0-2.5)	reference (0)	reference (0)
	a small portion	not applicable	0.8 (0.2-1.9)	0.9 (0.1-4.8)
	a large portion	not applicable	1.6 (1.1-2.7)	3.5 (1.1-11.6)
	Salmon in indie sauce	0.62 (0.42-0.90)	not applicable	not applicable
Outbreak 2	Cold fish	2.2 (1.4-3.6)	1.9 (1.1-3.5)	4.9 (1.6-16.5)
	Pasta salad	1.8 (1.2-2.8)	1.3 (0.9-2.2)	2.3 (0.7-7.8)

univariable analysis. Furthermore the point estimates from the logistic regression were unrealistically large in practice. This has a sound theoretical basis [5].

We also attempted to perform the log regression within the frequentistic frame in R using various algorithms designed for multivariable analysis (e.g. fitting with stricter form of step-halving; glm2 or using expectation-maximization algorithm [13]) but these were either difficult to use or did not always converge (data not shown). The drawback of the Bayesian modelling is to construct the model but the one used in the present study is a very simple one. The mathematical proof of the invalidity of the conversion formula as suggested by Zhang *et al.* [8] is presented in **Supplementary Material 7**. Briefly describing, the standard formula for the risk ratio is mathematically formulated to include odds ratio estimates. The obtained end formula is the one presented for converging the OR values to RR [8] However, as the values of the all explanatory variables x_i are used in the formula, it is not generally valid but depends on the data.

CONCLUSIONS

Binary model applying log link provided accurate and useful estimates of RR estimating the true risk, thus proved to be a suitable method for multivariable analysis of cohort studies in outbreak situations. Bayesian modelling was essential to ensure convergence of the model.

ABBREVIATIONS

OR = Odds ratio

RR = Risk ratio

CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

We thank Tiina Laitala for supervising the norovirus outbreak investigation in Rauma which was used in the present study. The study was done and funded in the aforementioned institutions as a part of routine work, no external funding specifically for this project was granted. No ethical approval was needed.

SUPPLEMENTARY MATERIAL

Supplementary material is available on the publisher's web site along with the published article.

REFERENCES

[1] Rothman KJ, Greenland S. Modern epidemiology. 2nd ed. USA: Lippincott Williams & Wilkins 1998.

[2] Vivancos R, Shroufi A, Sillis M, *et al*. Food-related norovirus outbreak among people attending two barbeques: epidemiological, virological, and environmental investigation. Int J Infect Dis 2009; 13(5): 629-35.

[3] Najjar Z, Furlong C, Stephens N, *et al.* An outbreak of Salmonella Infantis gastroenteritis in a residential aged care facility associated with thickened fluids. Epidemiol Infect 2012; 16: 1-9.

[4] Griffiths SL, Salmon RL, Mason BW, Elliott C, Thomas DR, Davies C. Using the internet for rapid investigation of an outbreak of diarrhoeal illness in mountain bikers. Epidemiol Infect 2010; 138(12): 1704-11.

[5] Knol MJ, Le Cessie S, Algra A, Vandenbroucke JP, Groenwold RH. Overestimation of risk ratios by odds ratios in trials and cohort studies: alternatives to logistic regression. CMAJ 2012; 184(8): 895-9.

[6] Petersen MR, Deddens JA. Maximum Likelihood Estimation of the Log-Binomial Model. Commun Stat Theory Methods 2010; 39(5): 874-83.

[7] Zou G. A modified poisson regression approach to prospective studies with binary data. Am J Epidemiol 2004; 159(7): 702-6.

[8] Zhang J, Yu KF. What's the relative risk? A method of correcting the odds ratio in cohort studies of common outcomes. JAMA 1998; 280(19): 1690-1.

[9] McNutt LA, Wu CT, Xue XN, Hafner JP. Estimating the relative risk in cohort studies and clinical trials of common outcomes. Am J Epidemiol 2003; 157(10): 940-3.

[10] Chu H, Cole SR. Estimation of risk ratios in cohort studies with common outcomes: a Bayesian approach. Epidemiology 2010; 21(6): 855-62.

[11] Nurminen M. To use or not to use the odds ratio in epidemiologic analyses? Eur J Epidemiol 1995; 11(4): 365-71.

[12] R_Core_Team. R. A language and environment for statistical computing. R foundation for Statistical Computing, Vienna, Austria. 2013 [cited 29.08.2013]; Available from: http: //www.R-project.org

[13] Marschner IC, Gillett AC. Relative risk regression: reliable and flexible methods for log-binomial models. Biostatistics 2012; 13(1): 179-92.

Received: June 14, 2013 Revised: September 4, 2013 Accepted: September 4, 2013