NIASRA

NATIONAL INSTITUTE FOR APPLIED STATISTICS RESEARCH AUSTRALIA

National Institute for Applied Statistics Research Australia

The University of Wollongong

Working Paper

02-16

Examining associations in cross-sectional studies

Margo L Barr, Robert Clark and David G Steel

National Institute for Applied Statistics Research Australia, University of Wollongong,
Wollongong NSW 2522. Phone +61 2 4221 5435, Fax +61 2 4221 4845.
Email: anica@uow.edu.au

Examining associations in cross-sectional studies

Margo L Barr, Robert Clark and David G Steel

National Institute for Applied Statistics Research Australia, University of Wollongong, Wollongong, Australia

Email addresses:
MLB: Margo.Barr593@uowmail.edu.au
RC: rclark@uow.edu.au
DGS: david_steel@uow.edu.au

Introduction

There is substantial debate on the most appropriate models and resultant effect measures for cross- sectional studies when using binary outcomes. The most easily interpretable effect measure is the risk ratio or relative risk (RR) reported as so many times more likely and calculated as P_1/P_0 where P is the probability and $_1$ and $_0$ are the exposed and non-exposed groups. However, odds ratios (OR) calculated as $\frac{P_1(1-P_1)}{P_0(1-P_0)}$, are often used because of the ease of using logistic regression models [1]. Because of the difficulty in understanding ORs they are often incorrectly interpreted as a RR. For rare outcomes, RRs and ORs do coincide, but when working with frequent outcomes, which are often collected through cross-sectional surveys, the OR can strongly overestimate the RR [1,2].

Barros and Hirakata 2003 [2] and Lee et al 2009 [3] compared models from which RRs, and the corresponding confidence intervals, are able to be produced including Cox regression with equal times of follow-up assigned to all individuals, log-binomial regression using a generalized linear model with a logarithmic link function and binomial distribution for the residual, and modified Poisson regression models incorporating the robust sandwich variation. They concluded that “the Poisson Regression model incorporating the robust sandwich variance should be used in cross-sectional studies for estimating prevalence ratios (PR)”. Lee et al 2009 also pointed out that “in terms of mathematical properties, the logistic model is undisputedly the best model for binomial Y” [3]. However, neither had examined model fit or non-symmetry in their analysis.

Methods

We compared Poisson regression using GENMOD procedure and logistic regression using SURVEYLOGISTIC procedure in SAS [4]. The model examined the association between any sunburn/no sunburn and season using winter as the reference category adjusting for age group, sex and sun protection index. Given the strong statistical arguments in favour of logistic regression [3] we also sought to compare the goodness of fit to the sunburn data. As a way of doing so, we considered the following blended binomial regression model which smoothly interpolates between a log link and a logistic link:

$$P[Y=1] = \theta \exp(\beta_1 z_1 + ... + \beta_k z_k) + (1-\theta)\ \exp(\beta_1 z_1 + ... + \beta_k z_k) / \left(1 + \exp(\beta_1 z_1 + ... + \beta_k z_k)\right)$$

which is equivalent to log regression when θ=0 and logistic regression when θ=1. We fit this model to the data by maximum likelihood, for values of θ fixed at 0, 0.05, …, 1. This enabled us to find the maximum likelihood estimator of θ, and to test null hypotheses of θ =0 and θ =1 using the asymptotic likelihood ratio test [5] (ignoring for simplicity that 0 and 1 lie on the boundary of the parameter space of θ).

Results and Discussion

As shown in Table 1 the crude RR from Poisson regression model for sunburn was 5.45 and the logistic regression model provides a crude OR of 6.68. In both crude models and adjusted models the differences that were significant were the same. However in the adjusted models the difference between the RR and the OR are even larger (5.60 and 7.42) and if reported incorrectly (ie 7.42 times more likely rather than 7.42 times the odds) as is often done in the literature [6,7] it would imply that the prevalence of sunburn in summer after adjusting for age group, sex and sun protection is 30% higher rather than 22% as measured. When the blended log/logistic binomial model was fitted the maximum likelihood estimator of θ was 1, indicating that the best fitting blended model was in fact the log regression when the outcome was sunburn. The p-value for this value was therefore 1, while the p-value for the null hypothesis of θ =0 was 0.03, indicating the logistic model fits significantly worse at the 5% level.

When we inverted the outcome, ie using no-sunburn as the outcome, then the OR for summer compared to winter was 0.15 and the RR was 0.82 as shown in Table 1. So if the OR was interpreted as an RR then it would imply that 14% of the population has no-sunburn in summer rather than 78% as measured. When the blended log/logistic binomial model was also fitted for the outcome of no-sunburn the maximum likelihood estimator of θ was 0, indicating that the best fitting blended model was the logistic model because of probabilities greater than 1 in the Poisson models. This highlights the need for a model that allows for the calculation of PR which does not have issues with non-symmetry and will not produce probabilities greater than 1. Such a measure, the model-adjusted risk ratio has been suggested by Bieler et al (2009) [8], but further exploration in this area is required.

References

1. Viera AJ. Odds ratios and Risk ratios: What's the difference and Why does it matter. *Southern Medical Journal* 2008; 101(7): 730-734.
2. Barros AJD and Hirakata VN. Alternatives for logistic regression in cross-sectional studies: an empirical comparison of models that directly estimate the prevalence ratio. *BMC Med Res Methodol*. 2003; 3: 21
3. Lee J, Tan CS, and Chia KS. A practical guide for multivariate analysis of dichotomous outcomes. *Ann Acad Med Singapore*. 2009; 38:714-9.
4. SAS Institute. The SAS System for Windows version 9.4. Cary, NC: SAS Institute Inc., 2014 www.sas.com
5. Welsh, AH *Aspects of statistical inference*. New York: John Wiley & Sons, 1996. (e.g. pp. 224-226)
6. Young-Ho Khang, Sung-Cheol Yun and John W Lynch. Monitoring trends in socioeconomic health inequalities: it matters how you measure. *BMC Public Health* 2008, 8:66.
7. Tajeu G, Sen B, Allison DB, and Menachemi N. Misuse of Odds Ratios in Obesity Literature: An Empirical Analysis of Published Studies. *Obesity* 2012 August ; 20(8): 1726–1731.
8. Bieler GS, Brown GG, Williams RL and Brogan DJ. Estimating Model-Adjusted Risks, Risk Differences and Risk Ratios from Complex Survey Data. *Am J Epidemiol*. 2010; 171:618-623.

Table 1: Prevalence estimates and crude and adjusted relative risks for the association between sunburn/no sunburn and season. NSW, 2007

Categories	Prevalence	Crude		Adjusted	
		RR (95% CI)	OR (95% CI)	RR (95% CI)	OR (95% CI)
Sunburn					
Spring	7.4 (5.9-8.8)	1.86 (1.29-2.66)	1.92(1.32-2.82)	1.99 (1.38-2.87)	2.13 (1.43-3.15)
Summer	21.6 (18.7-24.6)	5.45 (3.92-7.58)	6.68 (4.67-9.55)	5.60 (4.02-7.80)	7.42 (5.12-10.76)
Autumn	11.7 (10.0-13.3)	2.94 (2.11-4.10)	3.20 (2.25-4.55)	2.91(2.08-4.07)	3.27 (2.27-4.71)
Winter	4.0 (2.8-5.2)	1.0	1.0	1.0	1.0
No sunburn					
Spring	92.6 (94.1-91.2)	0.96 (0.95-0.98)	0.52 (0.76-0.35)	0.96 (0.94-0.98)	0.47 (0.32-0.70)
Summer	78.4 (81.3-75.4)	0.82 (0.78-0.85)	0.15 (0.21-0.10)	0.82 (0.79-0.85)	0.13 (0.09-0.20)
Autumn	88.3 (90.0-86.7)	0.92 (0.90-0.94)	0.31 (0.44-0.22)	0.92 (0.90-0.94)	0.31 (0.21-0.44)
Winter	96.0 (97.2-94.8)	1.0	1.0	1.0	1.0