Computational Statistics and Data Analysis 56 (2012) 4180–4189

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Model-based estimation of the attributable risk: A loglinear approach

Christopher Cox *, Xiuhong Li

Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, United States

ARTICLE INFO

Article history:
Received 5 November 2010
Received in revised form 24 April 2012
Accepted 24 April 2012
Available online 7 May 2012

Keywords:
Adjusted attributable risk
Case-control study
Cohort study
Poisson regression
Delta method
Model-based estimate
Bootstrap methods

ABSTRACT

This paper considers model-based methods for estimation of the adjusted attributable risk (*AR*) in both case-control and cohort studies. An earlier review discussed approaches for both types of studies, using the standard logistic regression model for case-control studies, and for cohort studies proposing the equivalent Poisson model in order to account for the additional variability in estimating the distribution of exposures and covariates from the data. In this paper, we revisit case-control studies, arguing for the equivalent Poisson model in this case as well. Using the delta method with the Poisson model, we provide general expressions for the asymptotic variance of the *AR* for both types of studies. This includes the generalized *AR*, which extends the original idea of attributable risk to the case where the exposure is not completely eliminated. These variance expressions can be easily programmed in any statistical package that includes Poisson regression and has capabilities for simple matrix algebra. In addition, we discuss computation of standard errors and confidence limits using bootstrap resampling. For cohort studies, use of the bootstrap allows binary regression models with link functions other than the logit.

1. Introduction

The attributable risk (*AR*) represents the relative amount by which the prevalence of a disease D would be reduced if an exposure E was eliminated, taking account of both the relative risk and the prevalence of the exposure. As noted by a reviewer, the data on which an estimate is based must be representative of the population to which the results will be applied. A refinement of the basic definition is to adjust for the effects of covariates. Adjustment is based on the prevalence of disease in a population as a function of a risk factor E with I, usually ordered, levels (with reference level $E = 1$ representing no exposure) and a set of categorical covariates x_j $(1 \le j \le J)$, which typically represent a compound index generated by the combined levels of two or more factors (Benichou, 2001). The adjusted *AR* is defined as follows (Basu and Landis, 1995; Eide and Gefeller, 1995; Lehnert-Batar et al., 2006).

$$\begin{aligned} AR &= 1 - \Pr(D|E=1)/\Pr(D) = 1 - \sum_j^J \Pr(x_j)\Pr(D|E=1,x_j)/\Pr(D) \\ &= \left\{ \sum_j \sum_i \Pr(E=i,x_j,D) - \sum_j \Pr(x_j)\Pr(D|E=1,x_j) \right\} \Big/ \Pr(D) \\ &= \sum_j \sum_{i>1} \Pr(E=i,x_j) \left\{ \Pr(D|E=i,x_j) - \Pr(D|E=1,x_j) \right\} \Big/ \Pr(D). \end{aligned} \tag{1}$$

* Correspondence to: 615 N. Wolfe Street, E7642, Baltimore, MD 21205, United States. Tel.: +1 410 955 4320; fax: +1 410 955 7587.
E-mail address: ccox@jhsph.edu (C. Cox).

0167-9473/$ – see front matter
doi:10.1016/j.csda.2012.04.017

The final expression in (1) is used to define the *AR* for the *j*th level of the covariates or, reversing the order of summation, the *i*th level of exposure ($i > 1$). Eide and Gefeller (1995) and Eide and Heuch (2006) refer to the latter as components of the *AR* due to a particular level of exposure and note that these components sum to the total *AR*.

We consider model-based approaches for estimation of the adjusted attributable risk. Model-based methods use a regression model to estimate the probabilities in Eq. (1). For both cohort and case-control studies the standard approach is based on a logistic regression model for disease status as a function of exposure and covariates. For cohort studies Cox (2006) previously proposed the loglinear model equivalent to the standard logistic regression model for the case of cross-sectional sampling, in order to account for the additional variability in estimating the joint distribution of exposure and covariates (Basu and Landis, 1995). The Poisson regression model can also be used for stratified cohort studies when additional data are available to estimate the exposure distribution.

The adjusted *AR* (1) is based on a comparison of disease risk among exposed individuals to that in the unexposed ($E = 1$) population. A generalization (Drescher and Becher, 1997; Eide and Heuch, 2001, and references therein) is to allow comparison to a population in which the exposure is not entirely absent (present only at the lowest level), but rather has a nondegenerate distribution, which is different from that in the original population. An example is when the exposure is reduced but not eliminated as the result of an intervention or education program. The definition of the generalized attributable risk (generalized impact fraction) given by Drescher and Becher (1997) for such an alternative distribution, $\Pr^*(E = i, x_j)$, can be written as follows.

$$gAR = \left[\sum_j \sum_i \left\{\Pr(E = i, x_j)\Pr(D|E = i, x_j) - \Pr^*(E = i, x_j)\Pr(D|E = i, x_j)\right\}\right] \Big/ \Pr(D). \tag{2}$$

Although defined in more general terms, the alternative distribution would typically involve only the levels of the risk factor. For each value of the covariates it is defined as a re-weighting of the original exposure probabilities by a specified probability density function $g(i|k)$ $(1 \le i \le I)$, defined for each level of exposure, k $(1 \le k \le I)$.

$$\Pr^*(E = i, x_j) = \sum_k g(i|k)\Pr(E = k, x_j). \tag{3}$$

This definition has the intuitively appealing property that $\Pr^*(x_j) = \Pr(x_j)$; the special case of the standard *AR* corresponds to $g(1|k) = 1$. To illustrate this idea we will use an example considered by Drescher and Becher (1997). In this case $I = 4$, and we assume that a proportion q_1 $(0 < q_1 < 1)$ of subjects change to the lowest risk level, while an additional proportion q_2 $(0 < q_1 \le q_1 + q_2 \le 1)$ change from the current level to the next lower level, with subjects already at the lowest level of risk remaining where they are. The two-parameter family of density functions specified in (3) is given in the following table, which we will use to illustrate the generalized *AR*.

$g(i\|k)$ k / i	1	2	3	4
1	1	$q_1 + q_2$	q_1	q_1
2	0	$1-(q_1+q_2)$	q_2	0
3	0	0	$1-(q_1+q_2)$	q_2
4	0	0	0	$1-(q_1+q_2)$

In this paper we consider both cohort and case-control studies. For cohort studies we employ the Poisson regression model. As is standard practice the parameters of the model are estimated by the method of maximum likelihood. For this model we first provide expressions for the large sample variance of the model-based *AR* for cohort studies, using the delta method (Cox, 1998), which is the standard method for finding asymptotic variances for functions of the original parameters of the model estimated by maximum likelihood. These expressions can be easily programmed in statistical packages having matrix capabilities, such as R.

For case-control studies we propose the equivalent loglinear model as well, again in order to account for estimation of the distribution of exposure and covariates. Expressions for the variance of the adjusted *AR* based the loglinear model and the delta method are provided. We include a discussion of the generalized *AR* for both types of studies. We also consider bootstrap methods for computing standard errors and confidence intervals. An advantage of the bootstrap is that for cohort studies, link functions other than the logit can be employed in the binary regression model, and we provide an illustration using a discrete survival model. The approach is not difficult to implement in packages that facilitate resampling methods.

The delta method requires expressions for the partial derivatives of various nonlinear functions of the parameters. There are many ways to write the required vectors of partial derivatives; in addition to ordinary matrix multiplication, we use the Schur product (element-wise multiplication) to simplify the notation. This requires either that both dimensions of the two matrix operands are identical, or that one of the two matrices is a vector whose length equals one of the two dimensions of the other. We will denote this operation with an asterisk; a simple relation that is used repeatedly is $(a * b)'c = b'(c * a) = \sum a_i b_i c_i$ for three m-vectors a, b and c.