Published in final edited form as:

Comput Stat Data Anal. 2012 December 1; 56(12): 4180–4189. doi:10.1016/j.csda.2012.04.017.

Model-Based Estimation of the Attributable Risk: A Loglinear Approach

Christopher Cox, Ph.D. and Xiuhong Li, M.A.S.

Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health

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.

Keywords

adjusted attributable risk; case-control study; cohort study; Poisson regression; delta method; model-based estimate; bootstrap methods

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 were 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 $x_j (1 \leq j \leq J)$, 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, Pfahlberg and Gefeller, 2006).

© 2012 Elsevier B.V. All rights reserved.

Christopher Cox, 615 N. Wolfe Street, E7642, Baltimore, MD 21205, Telephone: 410-955-4320, FAX: 410-955-7587, ccox@jhsph.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

$$\begin{aligned}
 AR &= 1 - P(D|E=1)/P(D) = 1 - \sum_j P(x_j)P(D|E=1, x_j)/P(D) \\
 &= \left\{ \sum_j \sum_i P(E=i, x_j, D) - \sum_j P(x_j)P(D|E=1, x_j) \right\} / P(D) \quad (1) \\
 &= \sum_{j>1} \sum_i P(E=i, x_j) \{P(D|E=i, x_j) - P(D|E=1, x_j)\} / P(D)
 \end{aligned}$$

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 equation (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.

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

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 $g(i|k)$ ($1 \leq i \leq I$), defined for each level of exposure, k ($1 \leq k \leq I$).

$$\Pr^*(E=i, x_j) = \sum_k g(i|k) \Pr(E=k, x_j) \quad (3)$$

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 ($0 < q_1 + q_2 < 1$) 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 .

2. Cohort studies

Consider a logistic regression model with binary outcome $D = d(0/1)$, denoted by a disease indicator, z_d , and model (design) matrix $Z_{m \times p}$, which includes an intercept. The model matrix consists of one or more exposure variables (indicators for categories of exposure) and categorical covariates, specified by $(E = i, x_j | 1 \leq i \leq I, 1 \leq j \leq J)$. The usual case for exposure is an ordinal, categorical exposure E having I levels, with a saturated $(I-1)$ parameters) model. We assume that each row z_k' of Z , together with the corresponding element of z_d , specifies a unique combination of the disease outcome, exposure and covariates, so that $m = 2IJ$. Note that in this formulation we require separate rows for subjects with and without disease, each with the same pattern of exposure and covariates. Let $n_k = 0$ (some of which may be zero) denote the total number of observations with a given pattern specified by z_k' and the corresponding element of z_d and define the vector $n = (n_k)$. For a given pattern of exposure and covariates, the sums of the separate totals for

diseased and non-diseased subjects are fixed in the likelihood for the logistic regression model, although the individual totals themselves are not (they could be used as frequency weights in fitting the logistic model). Denote the maximum likelihood estimates of the unknown parameters by $\hat{\beta}_z$, so that the vector of predicted probabilities of disease based on the model is $(P(D|E=i, x_j)) = p(\hat{\beta}_z) = \text{expit}(Z\hat{\beta}_z) = (\text{expit}(z_k'\hat{\beta}_z))$, where $\text{expit}(z) = 1/[1 + \exp(-z)]$. (Note that each predicted probability is repeated twice, once for diseased and once for non-diseased subjects having the same pattern of exposure and covariates.) To compute the adjusted AR we define the matrix Z_1 as the matrix Z with all exposure columns set to zero (so that in every case the exposure is the reference value, $E = 1$), in which case $p_1(\hat{\beta}_z) = \text{expit}(Z_1\hat{\beta}_z) = \hat{P}(D|E = 1, x_j)$. The adjusted AR (1) is then (Greenland and Drescher, 1993)

$$AR = 1 - \frac{n' \hat{p}_1}{n' \hat{p}} = \frac{n' (\hat{p} - \hat{p}_1)}{n' \hat{p}} \quad (4)$$

Because the predicted probabilities are repeated for diseased and non-diseased subjects, the AR depends only on the sums of the corresponding totals, not the totals themselves. Since these combined totals are fixed in the logistic regression, a variance computed from (4) using the delta method will not take account of variation in these totals, which determine the joint distribution of exposure and covariates. Similarly, $p_1(\hat{\beta}_z)$ is independent of the particular level of exposure in a given row of Z , so that the values of n_k corresponding to different levels of exposure for a given pattern of covariates are effectively summed over all levels of exposure, as well as over disease status. For this reason,

$$n' \hat{p}_1 = n_+ \sum_j P(x_j) \hat{P}(D|E=1, x_j), \quad \text{where the plus notation denotes summation.}$$

Although the standard model involves logistic regression, the formulation (4) can clearly be used with other link functions. For an arbitrary link function $g(z)$ we can compute $p(\hat{\beta}_z) = g^{-1}(Z\hat{\beta}_z)$ and $p_1(\hat{\beta}_z) = g^{-1}(Z_1\hat{\beta}_z)$, and then use (4).

Generalizing the approach of Walter (1976) for the two-by-two table, Cox (2006) proposed a Poisson regression model for the case of cross-sectional sampling, in order to account for estimation of the joint distribution of exposure and covariates. As described below the loglinear model is equivalent to the standard logistic regression model in the sense that the corresponding regression coefficients and variance matrix for exposure and covariates are identical to those from the logistic model. The Poisson regression model uses the counts n_k

0, as the dependent variable and has augmented model matrix $W = [Z_d, U]$, where $Z_d = z_d * Z$ (z_d is the indicator for disease status), and U is a model matrix saturated in the covariates and exposure (corresponding to a categorical variable denoting all possible combinations of levels in the data), including the overall intercept. That is, Z_d represents the interaction between the outcome of disease status and the Z terms in the original logistic model, and U contains Z and all distinct pairwise products of columns of Z (with the exception of the intercept).

For example, consider a logistic model with four categories of exposure and one binary covariate. One possible way to set up the required matrices is the following. Columns 2–4 of Z denote the level of exposure and column 5 is the covariate.

$$\begin{aligned}
z_d = & \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \quad Z = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 \end{bmatrix} \quad Z_d = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 \end{bmatrix} \quad U = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 \end{bmatrix}
\end{aligned}$$

The first 5 columns of U replicate Z , and the last three are interactions between the covariate and each of the three indicator variables for exposure. To obtain Z_1 we set columns 2–4 of Z to zero. The matrix $W = [Z_d, U]$ has 16 rows and 13 columns.

It is well known that for this loglinear model the estimated marginal totals corresponding to the independent variables are equal to the observed totals (McCullagh and Nelder, 1989, Section 6.4.2). The parameter vector may also be partitioned, $\beta = (\beta_z, \beta_u)$, so that the vector β_z corresponds to the parameters of the logistic model, and the m.l.e., $\hat{\beta}_z$, and corresponding sub-matrix of the estimated variance matrix, $\Sigma(\hat{\beta})$ are identical to the estimates from the logistic regression. The remaining parameters account for variation in the marginal totals for exposure and covariates. These parameters are used to estimate the joint distribution of exposure and covariates, and sufficient data must be available to estimate this distribution if (1) is to be applied. Denote the vector of predicted means from the Poisson regression by $\hat{\mu} + \hat{n} = \exp(W\hat{\beta}) = (\exp(w'_j \hat{\beta}))$, where w'_j denotes the j^{th} row of W . Because of the form of the likelihood equations for the Poisson model (Agresti, 2002, Section 4.4.7), the number of diseased subjects can be estimated directly by μz_d , which actually equals the observed number.

To compute the adjusted AR we define the matrix $V_1 = [Z_1, 0]$, having the same dimensions as W but replacing Z by Z_1 and U by 0, so that $p_1(\hat{\beta}) = \text{expit}(V_1 \hat{\beta})$. Then because the estimated marginal totals for the exposure and covariates equal the observed, it follows from (4) that (Cox, 2006),

$$AR = 1 - \frac{\hat{\mu}' \hat{p}_1}{\hat{\mu}' z_d} = \frac{\hat{\mu}' (z_d - \hat{p}_1)}{\hat{\mu}' z_d} \quad (5)$$

To apply the delta method we have $\hat{\mu}' \hat{\beta} = \hat{\mu}'^* W$, so that

$$\frac{\partial}{\partial \beta} \hat{\mu}' z_d = z_d' (\hat{\mu}^* W)$$

and similarly

$$\frac{\partial}{\partial \beta} \widehat{\mu} \widehat{p}_1 = \widehat{p}_1' (\widehat{\mu} * W) + \widehat{\mu}' \{ \widehat{p}_1 * (1_m - \widehat{p}_1) * V_1 \}$$

The standard approach is now to use the delta method to compute the asymptotic variances and covariance of the numerator and denominator of the AR (or $1 - AR$), and then a second application yields the approximate variance for a ratio (Greenland and Drescher, 1993). To avoid possible numerical difficulties, however, we take a direct approach based on the quotient rule.

$$\frac{\partial}{\partial \beta} AR = \left\{ \frac{\partial}{\partial \beta} (\widehat{\mu}' \widehat{p}_1) - (1 - AR) \frac{\partial}{\partial \beta} (\widehat{\mu}' z_d) \right\} / \widehat{\mu}' z_d \quad (6)$$

The delta method then gives the estimated variance.

$$\widehat{\text{var}}(AR) = \frac{\partial AR'}{\partial \beta} \sum (\widehat{\beta}) \frac{\partial AR}{\partial \beta}$$

We can also apply the delta method to obtain an approximation for $\text{var}[\log(1 - AR)]$, a transformation that has been recommended to produce a more normally distributed estimate. Alternative transformations have also been proposed, including the logistic (Lehnert-Batar, Pfahlberg and Gefeller, 2006).

The simplest case is that of a binary exposure and no covariates, considered in detail by Walter (1976). Assuming cross-sectional sampling, let (n_{ij}) denote the cell counts in the cross-classified two-by-two table, where $i = 1, 2$ denotes exposure and $j = 1, 2$ disease, with $i = 2$ indicating exposure and $j = 2$ indicating disease. The logistic model is based on the conditional distributions $n_{i2} \sim b(n_{i+}, p_i)$, $i = 1, 2$, where $p_i = P(D|E = i)$, which are independent, and the maximum likelihood estimates of the probabilities are the sample proportions. The prevalence of exposure is $P(E) = n_{2+}/n_{++}$, which is a fixed quantity in the conditional likelihood, while the probability of disease is $P(D) = (n_{+2})/n_{++}$, which of course is not. The attributable risk is estimated as follows.

$$AR = 1 - \frac{\widehat{P}(D|E=1)}{\widehat{P}(D)} = 1 - \frac{\widehat{p}_1}{(n_{1+}\widehat{p}_1 + n_{2+}\widehat{p}_2)/n_{++}} = 1 - \frac{n_{12}/n_{1+}}{n_{+2}/n_{++}}$$

To compute a large sample variance, we would calculate AR/p_i and apply the delta method using the 2×2 diagonal covariance matrix of $(\widehat{p}_1, \widehat{p}_2)$ with diagonal elements $[p_1(1 - p_1)/n_{1+}, p_2(1 - p_2)/n_{2+}]$, substituting the maximum likelihood estimates for the true parameter values (Lehnert-Batar, Pfahlberg and Gefeller, 2006). For the Poisson model, the 4×4 model matrix W has linearly independent columns and so the model may be reparameterized as four jointly independent Poisson random variables, $n_{ij} \sim \text{Po}(\mu_{ij})$; the maximum likelihood estimates of the parameters of this model are the cell counts. The estimate of the exposure prevalence is the same as for the conditional binomial likelihood, but is now subject to random variation. The estimate of the AR is also identical.

$$AR = 1 - \frac{\widehat{P}(D|E=1)}{\widehat{P}(D)} = 1 - \frac{\widehat{\mu}_{12}/\widehat{\mu}_{1+}}{\widehat{\mu}_{+2}/\widehat{\mu}_{++}} = 1 - \frac{n_{12}/n_{1+}}{n_{+2}/n_{++}}$$

To calculate the asymptotic variance we would compute AR/μ_{ij} and apply the delta method using the 4×4 diagonal covariance matrix $\text{diag}(\mu_{11}, \mu_{12}, \mu_{21}, \mu_{22})$, again with the parameters replaced by their maximum likelihood estimates.

To compute components of the AR for each level of exposure we define 1_i as the indicator variable for $E = i > 1$ (the corresponding column of Z), and from (1) we have

$$AR_i = \frac{(\widehat{\mu} * 1_i)' (z_d - \widehat{p}_1)}{\widehat{\mu}' z_d} \quad (7)$$

Again, because the Poisson model is saturated in the exposure and covariates, we have from

(1) that $\sum_{i>1} (\widehat{\mu} * 1_i)' (z_d - \widehat{p}_1) = \widehat{\mu}' (z_d - \widehat{p}_1)$, so that $AR = \sum_{i>1} AR_i$. The component of the attributable risk for the j^{th} stratum of the covariates is defined similarly. To apply the delta method using (6) we need only the derivative of the numerator of (7),

$$\frac{\partial}{\partial \beta} (\widehat{\mu} * 1_i)' (z_d - \widehat{p}_1) = (z_d - \widehat{p}_1)' (1_i * \widehat{\mu} * W) - (\widehat{\mu} * 1_i)' \{ \widehat{p}_1 * (1_m - \widehat{p}_1) * V_1 \}$$

Finally we consider the generalized AR (2). As discussed above, this is defined by means of an $I \times I$ matrix $G = [g_{ik}] = g(i|k)$ of probabilities ($0 \leq g(i|k) \leq 1$), satisfying $1_I' G = 1_I$, which is used to define the alternative distribution of exposure at each level of the covariates (3). We also need the predicted probabilities for the logistic model, given by $p(\widehat{\beta}) = \text{expit}(V\widehat{\beta})$, using a similar notation for the matrix $V = [Z, 0]$. Depending on how the rows of the data set are arranged, the matrix G can be expanded into an $m \times m$ matrix G^* so that for $\widehat{\mu}^* = G^* \widehat{\mu}$,

$$gAR = 1 - \frac{\sum_j \sum_i \{P^*(E=i, x_j) P(D|E=i, x_j)\}}{P(D)} = 1 - \frac{\widehat{\mu}^*{}' \widehat{p}}{\widehat{\mu}' z_d} \quad (8)$$

If the rows of the Z are sorted by exposure within disease within covariate categories, since we include cases where $n_j = 0$, so that $m = 2JI$, we have $G^* = (I_{2J} \otimes G)$. In our example, $I = 4$ and $J = 2$, so that $G^* = \text{diag}[G, G, G, G]$. To apply the delta method (4) we need only the derivative of the numerator of (8).

$$\frac{\partial}{\partial \beta} \widehat{\mu}^*{}' \widehat{p} = \widehat{p}' G^* (\widehat{\mu} * W) + \widehat{\mu}^*{}' \{ \widehat{p} * (1 - \widehat{p}) * V \}$$

3. Case-control studies

For case-control studies we can use Bayes theorem to write the adjusted AR (1) as (Bruzzi et al., 1985, Eide and Heuch, 2001),

$$AR = 1 - \sum_i \sum_j P(E=i, x_j|D) RR_{ij}^{-1} = \sum_{i>1} \sum_j P(E=i, x_j|D) (1 - RR_{ij}^{-1}) \quad (9)$$

where $RR_{ij} = P(D|E=i, x_j)/P(D|E=1, x_j)$ is the relative risk, which will, in general, depend on both the level of exposure and the covariates (Greenland and Drescher, 1993). To compute a model-based estimate, a standard approach is to use a logistic regression model with $m \times p$ model matrix Z and dependent variable denoting case ($d=1$) or control ($d=0$) status, and estimate the relative risks by the corresponding odds ratios. We again assume that each row of Z , together with the corresponding element of the outcome vector, specifies a unique pattern of the independent and dependent variables, and let $\hat{\beta}$ denote the estimated parameters. The m -vector of predicted probabilities, each one repeated twice, is denoted by $\hat{p} = p(\hat{\beta}) = \text{expit}(Z\hat{\beta})$. If $n = (n_k)$ again denotes the vector of counts (frequencies), then $n'\hat{p}$ is a model-based estimate (actually equal to) the total number of cases, whose variability depends only on the estimated parameters (Cox, 2006). (As before, it is true that the numbers of cases and controls are random according to the logistic likelihood, but because the values of $\hat{p}$ depend only on exposure and covariates, they are the same for both cases and controls having the same pattern, so that the estimate $n'\hat{p}$ depends only on the estimated probabilities and the combined (case + control) totals for each distinct combination of exposure and covariates, and these totals are fixed in the logistic model.)

Now let E_x denote an $m \times p$ matrix, each of whose rows is an indicator for the level of exposure in the corresponding row of the matrix Z , with the exception of the reference level, where the entries are all zero, and zero for the covariates as well. In other words E_x corresponds to the matrix Z with all columns except those indicating the various levels of exposure set to zero. In the previous example in Section 2, E_x would be the matrix Z with the first and last columns set to zero. The relative risks are estimated by the odds ratios, $RR = \exp(E_x\hat{\beta})$. Then from (9) we have (Greenland and Drescher, 1993; Cox, 2006)

$$AR = 1 - \frac{(n * \hat{p})' \{\exp(-E_x\hat{\beta})\}}{n'\hat{p}} = \frac{(n * \hat{p})' \{1_m - \exp(-E_x\hat{\beta})\}}{n'\hat{p}} \quad (10)$$

With this formulation the variance of the AR depends only on the variability of the estimated parameters. Thus, although this approach does provide model-based estimates of the probabilities $\Pr(E=i, x_j|D)$, it does not allow variation in the totals, which are fixed in the logistic model, and in this sense does not fully account for the distribution of the exposures and covariates.

In contrast, the expression for the variance of an equivalent expression given by Greenland and Drescher (1993, equation 5) involves the multinomial covariance matrices of the observed counts for cases and controls. Using their notation, we let z'_j ($1 \leq j \leq k$) denote the unique patterns of only the independent variables, and $n_{1j} \geq 0$ denote the number of cases and $n_{0j} \geq 0$ the number of controls with a given pattern, with totals $n_j = n_{1j} + n_{0j}$. Thus multinomial variation is assumed for the counts (n_{0j}) and (n_{1j}). The estimated counts are based on the model rather than the data, and so are model-based. The variance matrix for the totals is then the sum of the two multinomial covariance matrices, reflecting the independent sampling of cases and controls. While this is an interesting approach, it also seems somewhat inconsistent since the total numbers of cases and controls are treated as random by the logistic model but as fixed constants in the multinomial covariance matrices. Overall, however, the approach appears useful, and has been implemented in the post-estimation command `aflogit` in Stata (Brady, 1998). This command calculates the AR and component

AR_j for the individual exposure categories, as well as their standard errors and asymmetric confidence intervals based on the $\log(1 - \cdot)$ transformation using the approach of Greenland and Drescher (1993).

Rather than pursue this approach, we turn to the following expression for the “case-control likelihood” from Greenland and Drescher (1993 equation 1),

$$\prod_j \frac{\exp(\widehat{z_j' \beta})^{n_{1j}} \pi(z_j)^{n_{1j} + n_{0j}}}{\{1 + \exp(\widehat{z_j' \beta})\}^{n_j} \mu^{n_1} (1 - \mu)^{n_0}}$$

where $\pi(z_j)^{n_j}$ is proportional to the multinomial likelihood for the totals, and $\mu = n_1/n$ is a constant. The parameter estimates, $\widehat{\pi} = (n_1, \dots, n_k)/n$, are based on the saturated model. This is an actual likelihood, which is in fact proportional to the likelihood for the Poisson model, also saturated in the exposures and covariates, corresponding to the logistic model. We propose to use this Poisson model as an alternative to the approach of Greenland and Drescher (1993). The expression for the model-based AR is similar to (10), but the partial derivatives are different since the model now includes $k - 1$ additional parameters for the totals (n_j). The advantage of this approach is that it is based on the (Poisson) likelihood; the disadvantage is that the totals n_1 for cases and n_0 for controls are (consistent with the model) treated as random, which, however, may not be consistent with the sampling design. This might lead one to suspect that the variance calculated from the Poisson model would be larger than the estimate proposed by Greenland and Drescher (1993). Examples show that this is sometimes, but not always, the case.

We now consider the AR based on the loglinear model. To discuss this model we can use the same framework as for cohort studies; that is, the model matrix $Z_{m \times p}$ for the logistic model includes rows for both cases and controls. The Poisson model matrix $W = [Z_d, U]$ is constructed as before using the case indicator z_d . The predicted probabilities are denoted $\widehat{p} = \text{expit}(V\widehat{\beta})$ as before and the predicted counts for cases and controls by $\widehat{\mu} = \exp(W\widehat{\beta})$. For computation of the odds ratios, the matrix E_x must include additional columns of zeroes for the additional parameters, i.e., E_x is based on W rather than Z , with the columns corresponding to U all set to zero. From (10) we then have

$$AR = 1 - \frac{\{\widehat{\mu} * \widehat{p}\}' \exp(-E_x \widehat{\beta})}{\widehat{\mu}' \widehat{p}} = \frac{\{\widehat{\mu} * \widehat{p}\}' \{1_m - \exp(-E_x \widehat{\beta})\}}{\widehat{\mu}' \widehat{p}} \quad (11)$$

For the loglinear model the totals also depend on the parameters, so that the partial derivatives for the numerator and denominator of (11) that are needed for (6) are more complicated than for the logistic model.

$$\frac{\partial}{\partial \beta} \widehat{\mu}' \widehat{p} = \widehat{\mu}' (\widehat{\mu} * W) + \widehat{\mu}' \{\widehat{p} * (1_m - \widehat{p}) * V\}$$

and

$$\frac{\partial}{\partial \beta} (\hat{\mu} * \hat{p})' \{1_m - \exp(-E_x \hat{\beta})\} = [\hat{p} * \{1_m - \exp(-E_x \hat{\beta})\}]' (\hat{\mu} * W) + \hat{\mu}' \left[\{1_m - \exp(-E_x \hat{\beta})\} * \hat{p} * (1_m - \hat{p}) * V - \hat{p} * \exp(-E_x \hat{\beta}) * E_x \right]$$

A simpler expression can be obtained by working with $1 - AR$ (9).

To compute the component of the AR specific to the i^{th} level of exposure, we need exposure matrices E_i which are similar to E_x , but denote only the occurrences of the i^{th} level of

exposure, so that $E_x = \sum_{i>1} E_i$ and $1_m - \exp(-E_x \hat{\beta}) = \sum_{i>1} \{1_m - \exp(-E_i \hat{\beta})\}$. Thus,

$$AR_i = \frac{(\hat{\mu} * \hat{p})' \{1_m - \exp(-E_i \hat{\beta})\}}{\hat{\mu}' \hat{p}} \quad (12)$$

The partial derivatives for (12) are similar to those for the AR (11).

For the generalized AR we again need an expanded matrix G^* , which is based on the transpose of the matrix G (Cox, 2006, equation 12)

$$gAR = 1 - \frac{\{\hat{\mu} * \hat{p} * \exp(-E_x \hat{\beta})\}' G^* \exp(E_x \hat{\beta})}{\hat{\mu}' \hat{p}} \quad (13)$$

If the rows of the data matrix are sorted by exposure within disease within covariate categories, then we have $G^* = (I_{2j} \otimes G^T)$; in our previous example, $G^* = \text{diag}[G^T, G^T, G^T, G^T]$. To apply the delta method using (8) we only need the partial derivatives for the numerator of (13).

$$\begin{aligned} \frac{\partial}{\partial \beta} \{\hat{\mu} * \hat{p} * \exp(-E_x \hat{\beta})\}' G^* \exp(E_x \hat{\beta}) &= \{\hat{p} * \exp(-E_x \hat{\beta}) * G^* \exp(E_x \hat{\beta})\}' (\hat{\mu} * W) + \\ &\{G^* \exp(E_x \hat{\beta})\}' \{\hat{\mu} * \exp(-E_x \hat{\beta}) * \hat{p} * (1_m - \hat{p}) * V - \hat{\mu} * \hat{p} * \exp(-E_x \hat{\beta}) * E_x\} + \\ &\{\hat{\mu} * \hat{p} * \exp(-E_x \hat{\beta})\}' G^* \{\exp(-E_x \hat{\beta}) * E_x\} \end{aligned}$$

4. Illustrations

To compute model-based estimates and their variances, we must be able to fit the Poisson regression model and store the parameter estimates and their covariance matrix for further calculations. Cox (2006) described one approach, involving a general purpose program for maximum likelihood estimation. Here we consider an approach based on statistical software with the ability to perform matrix calculations, so that the computations described above can be programmed directly. Computations for our examples were performed in S-Plus. To compare our results with the approach of Greenland and Drescher (1993) for case-control studies, we include results from the `aflogit`; post-estimation command in Stata. For cohort studies Greenland and Drescher used a logit model that does not account for estimation of the covariate distribution (Basu and Landis, 1995), so that the standard errors will be too small.

The bootstrap provides an alternative to the delta method for the computation of variances. An attractive feature of bootstrap resampling is that the computation of the AR for both

types of studies can be based on the logistic regression model, which simplifies the calculations, particularly since values of $\eta_j = 0$ can be omitted. In addition, for cohort studies other link functions besides the logit can be used, as illustrated in Section 4.1.2.

Computation of the AR for the bootstrap samples can be based on equation (4) for cohort studies, and for case-control studies equation (10) can be used. In our analyses bootstrap computations were performed in SAS 9.2 using the LOGISTIC procedure, which allows a limited number of other link functions, and also has an additional SCORE statement that facilitates computation of the probabilities $p_1(\hat{\beta}_2) = g^{-1}(Z_1\hat{\beta}_2)$ required for (4).

Our first example for cohort studies uses a data set considered by Basu and Landis (1995) who presented data (Table 1) from NHANES II, involving 966 women classified by two racial categories and four categories of exposure based on body mass index, with diastolic blood pressure above the 90th percentile as the outcome. We consider both the original data and a reduced data set with five (rows 1, 5 and 14–16) of the 16 counts set to zero, for a total sample size of 671. A second example used pooled binary regression with the complimentary log-log link ($g(p) = \log[1 - \log(p)]$) to analyze time to seroconversion data from the Multicenter AIDS Cohort Study (MACS), a longitudinal, cohort study of the natural history of HIV infection among MSM (men who have sex with men) in the United States (Ostrow et al., 2009). Briefly, during the follow-up period (1998–2008) there were 57 HIV seroconverters among 1,667 initially HIV-seronegative men. The regression model included all 8 combinations of three sex-drugs (inhaled nitrates or “poppers”, stimulants, and EDDs – erectile dysfunction drugs) used at the current or previous semi-annual visit, adjusting for other risk factors including sexual behavior, alcohol and other drugs used, and depression. For additional details see Ostrow et al. (2009). For case-control studies we consider the Ille-et-Villaine case-control study of esophageal cancer. The data are taken from Appendix I of Breslow and Day (1980), and were discussed extensively by Benichou (1991). There are 200 cases and 775 controls; the exposure variable is alcohol consumption, having four ordered categories with the lowest level of consumption as the reference. Covariates are age and smoking, each with three levels.

4.1 Cohort studies

4.1.1 Example 1—This is the example discussed in Section 2, having four categories of exposure and a single binary covariate. Results using the Poisson model to estimate the AR and the components AR_j for the three categories of exposure above the reference are shown in the first column of Table 1. Bootstrap standard errors based on 2000 samples, in the second column, were consistent with the large sample standard errors from the Poisson model and the delta method. Values from the general purpose MLE program used by Cox (2006, Section 5.2.2) were nearly identical, and are included for reference in the third column. The logistic regression model (using Stata) also gave identical estimates and similar standard errors. As expected the standard errors are smaller than those from the Poisson model, although the differences are slight for this example. We also computed the generalized AR for parameter values of $q_1 = .80$, $q_2 = .20$, which determine the matrix G . The result was $gAR(SE) = .286(.063)$, which is identical to the penultimate row of Table 3 of Cox (2006), based on the MLE program with the Poisson model.

For a further comparison of the Poisson and logistic regression approaches we consider a modification of this example, in which counts for five of the 16 rows in the data were set to zero. Results for this rather extreme example are shown in the second half of Table 1 (Example 1b), and more clearly illustrate the differences between the two approaches, confirming that the Poisson model gives reasonable standard errors compared to the bootstrap, while those based on the logistic model are too small.

4.1.2 Example 2—Results of the pooled binary regression analysis are summarized in Table 2, which includes the total AR as well as the component AR values for all seven active drug combinations. Standard errors were calculated using the bootstrap. In addition we computed a total $AR(SE)$ of .7408 (.1063) when four categories of frequency of unprotected receptive anal sex partners (URASP) were added to the exposure model. In this case the model for the exposure categories was not saturated.

4.2 Case-control studies

4.2.1 Example 1—The first example has four categories of alcohol exposure, and no covariates. The results from the Poisson model in the first column of Table 3 are consistent with the bootstrap in the second column, and with the results from the logistic model using the approach of Greenland and Drescher (1993), but not with the results based solely on the logistic model. The results are also consistent with those obtained using the general purpose MLE program (Cox, 2006, Section 5.1.1). For comparison we also computed the generalized AR for the same parameter values of $q_1 = .80$, $q_2 = .20$, giving the result $AR(SE) = .6809 (.0487)$.

4.2.2 Example 2—In this case we have a binary exposure to any alcohol as well as an interaction between alcohol and age; we consider exposure and the interaction as four categories of exposure. The covariates are age, smoking and their interaction. The results are shown in the right-hand side of Table 3. Of interest is the fact that the bootstrap standard errors are quite consistent with the results from the Poisson model; the results using the approach of Greenland and Drescher (1993) look reasonable as well, although the bootstrap results agree somewhat better with those from the loglinear model.

5. Discussion

The primary goal of this paper is to discuss methods for computing the model-based estimate of the adjusted AR and its standard error using a Poisson regression model and the delta method. The Poisson model is equivalent to the standard logistic model, so that the estimate of the attributable risk is the same, although the standard error is larger. We have argued that the Poisson model is more appropriate since it accounts for estimation of the joint distribution of exposure and covariates using the data. It is hoped that the formulas that were provided will facilitate the use of the loglinear model. For anyone interested in software, we have provided an electronic supplement containing S-Plus programs for the first examples for both cohort and case-control studies; these programs reproduce selected results in Tables 1 and 3, and the text.

In most of the examples in the previous section the model for exposure was saturated, so that the number of exposure parameters was one less than the number of categories. In other words, two different binary exposures would yield four exposure categories. Although this is frequently the case in applications, it is not necessary, as shown by Example 4.1.2. The subsequent calculations are the same, but the results are different; in particular the component AR 's no longer sum to the total.

Graubard and Fears (2005) discussed model-based estimates of the adjusted AR in the sample survey context. They assumed known sampling probabilities and proposed weighted estimates based on a logistic model. Standard errors were calculated using the method of Taylor deviates, which is closely related to the delta method. For case-control studies they showed that their variance estimate is approximately the same as that proposed by Benichou and Gail (1990), which uses the sample proportions to estimate the probabilities $\Pr(E = i, x_j | D)$ and gave similar results to those obtained by Cox (2006) using (logistic) model-based estimates. This is consistent with the fact that the estimator in equation 3 of their paper

involves a weighted average of the binary variable denoting case status. For cohort studies, they also used the logistic model, and the estimator in their equation 4 depends only on the estimated regression coefficients from the model; for this reason one would anticipate results similar to the approach of Greenland and Drescher (1993). Of course their approach could be used with the Poisson model as well, provided that one has a sampling design.

In summary we have proposed a unified approach for both cohort and case-control studies based on the Poisson model and the delta method, which will hopefully encourage the use of model-based estimates. For those so inclined, the bootstrap provides an alternate approach to the computation of the standard error or confidence limits that is in some respects simpler and that appears to work quite well. In fact in our examples, which involved fairly large sample sizes, results using the bootstrap agreed with those from the delta method applied to the Poisson model.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Contract/grant sponsor: National Institute of Allergy and Infectious Diseases; contract/grant numbers: U01-AI-35042, U01-AI-35043, U01-AI-35039, U01-AI-35040, U01-AI-35041, U01-AI-35004, U01-AI-31834, U01-AI-34994, U01-AI-34989, U01-AI-34993, U01-AI-42590.

Contract/grant sponsor: National Institute of Child Health and Human Development: U01-CH-32632.

Comments and additional references from an anonymous reviewer considerably improved the presentation of the proposed method.

References

- Agresti, A. Categorical Data Analysis. 2. New York: John Wiley & Sons; 2002.
- Basu S, Landis JR. Model-based estimation of population attributable risk under cross-sectional sampling. *American Journal of Epidemiology*. 1995; 142:1338–1343. [PubMed: 7503055]
- Benichou J. Methods of adjustment for estimating the attributable risk in case-control studies: a review. *Statistics in Medicine*. 1991; 10:1753–1773. [PubMed: 1845304]
- Benichou J. A review of adjusted estimators of attributable risk. *Statistical Methods in Medical Research*. 2001; 10:195–216. [PubMed: 11446148]
- Benichou J, Gail MH. Variance calculations and confidence intervals for estimates of attributable risk based on logistic models. *Biometrics*. 1990; 46:991–1003. [PubMed: 2085643]
- Brady AR. sbe21: Adjusted population attributable fractions from logistic regression. *Stata Technical Bulletin*. 1998; 42(STB42):8–12.
- Breslow, N.; Day, NE. The Analysis of Case-control Studies. Vol. 1. Lyon: International Agency for Research on Cancer; 1980. Statistical Methods in Cancer Research. Scientific Publications No. 32
- Bruzzi P, Green SB, Byar DP, Brinton LA, Schairer C. Estimating the population attributable risk for multiple risk factors using case-control data. *American Journal of Epidemiology*. 1985; 122:904–914. [PubMed: 4050778]
- Cox, C. Delta Method. In: Armitage, Peter; Colton, Theodore, editors. *Encyclopedia of Biostatistics*. New York: John Wiley & Sons; 1998. p. 1125-1127.
- Cox C. Model-based estimation of the attributable risk in case-control and cohort studies. *Statistical Methods in Medical Research*. 2006; 15:611–625. [PubMed: 17260927]
- Drescher K, Becher H. Estimating the generalized impact fraction from case-control data. *Biometrics*. 1997; 53:1170–1176. [PubMed: 9290235]
- Eide GE, Gefeller O. Sequential and average attributable fractions as aids in the selection of preventive strategies. *Journal of Clinical Epidemiology*. 1995; 48:645–655. [PubMed: 7730921]

- Eide GE, Heuch I. Attributable fractions: fundamental concepts and their visualization. *Statistical Methods in Medical Research*. 2001; 10:159–193. [PubMed: 11446147]
- Eide GE, Heuch I. Average attributable fractions: A coherent theory for apportioning excess risk to individual risk factors and subpopulations. *Biometrical Journal*. 2006; 48:820–837. [PubMed: 17094346]
- Graubard BI, Fears TR. Standard errors for attributable risk for simple and complex sample designs. *Biometrics*. 2005; 61:847–855. [PubMed: 16135037]
- Greenland S, Drescher K. Maximum likelihood estimation of the attributable fraction from logistic models. *Biometrics*. 1993; 49:865–872. [PubMed: 8241375]
- Lehnert-Batar A, Pfahlerberg A, Gefeller O. Comparison of confidence intervals for adjusted attributable risk estimates under multinomial sampling. *Biometrical Journal*. 2006; 48:805–819. [PubMed: 17094345]
- McCullagh, P.; Nelder, JA. *Generalized Linear Models*. 2. London: Chapman and Hall; 1989.
- Ostrow DG, Plankey MW, Cox C, Li X, Shoptaw S, Jacobson LP, Stall RC. Specific sex drug combinations contribute to the majority of recent HIV seroconversions among MSM in the MACS. *Journal of Acquired Immune Deficiency Syndromes*. 2009; 51:349–355. [PubMed: 19387357]
- Walter SD. The estimation and interpretation of attributable risk in health research. *Biometrics*. 1976; 32:829–849. [PubMed: 1009228]

Table 1

Values of the total and component attributable risks and their standard errors (second row) from a cohort study with four exposure categories and one binary covariate, using Poisson and logistic regression models. The left hand columns are for the full data set and the right-hand columns are for the same data with five of the 16 counts set to zero. For the third and seventh columns, a general purpose MLE program was used.

Estimate SE	Example 1				Example 1b			
	Poisson	Bootstrap	MLE Poisson	Logistic	Poisson	Bootstrap	MLE Poisson	Logistic
AR	.2920 .0692	.0724	.2919 .0692	.2920 .0673	.5602 .0508	.0518	.5602 .0507	.5602 .0421
AR_2	.005307 .0321	.0322	.005272 .0321	.0053 .0321	.1065 .0304	.0319	.1065 .0305	.1065 .0167
AR_3	.01666 .0260	.0259	.01667 .0260	.0167 .0259	.07659 .0261	.0261	.07659 .0261	.0766 .0109
AR_4	.2700 .0494	.0510	.2700 .0495	.2700 .0464	.3771 .0486	.0496	.3771 .0486	.3771 .0318

Table 2

Component and Total *AR* values for the risk of seroconversion from categories of recreational drug use using a pooled, binary regression model with the complementary log-log link. Standard errors were calculated using the bootstrap.

Drug(s)	<i>AR</i>	<i>SE</i>
Stimulant (S)	.0979	.0498
Popper (P)	.1549	.0670
EDD (E)	.0572	.0397
S and P	.1295	.0558
S and E	.0339	.0265
P and E	.0421	.0341
S and P and E	.1107	.0440
Total	.6262	.1179

Table 3

Values of the total and component attributable risks and their standard errors from two different models for a case-control study, each with four exposure categories, from Poisson and logistic models. For the third and seventh columns, a general purpose MLE program was used. In the third and seventh columns the method of Greenland and Drescher (1993) was used with the logistic model. For the fourth and eighth columns, the standard logistic model was used.

Estimate SE	Example 1 – No Covariates				Example 2 – Covariates			
	Poisson	Bootstrap	Logistic (G-D)	MLE Logistic	Poisson	Bootstrap	Logistic (G-D)	MLE Logistic
AR	.7089 .0511	.0514	.7089 .0508	.7089 .0502	.7234 .0501	.0505	.7234 .0498	.7234 .0491
AR ₂	.2698 .0443	.0440	.2698 .0451	.2698 .0423	.03828 .0163	.0167	.0383 .0298	.03828 .0157
AR ₃	.2223 .0329	.0330	.2223 .0340	.2223 .0287	.2190 .0304	.0300	.2190 .0322	.2190 .0244
AR ₄	.2167 .0299	.0301	.2167 .0287	.2167 .0208	.4662 .0511	.0511	.4662 .0507	.4662 .0482