MODEL-BASED ESTIMATES OF POPULATION ATTRIBUTABLE RISKS FOR ORDINAL DATA

Basu, Srabashi
Theoretical Statistics and Mathematics Unit, Indian statistical Institute

Landis, J. Richard
Center for Clinical Epidemiology and Biostatistics, University of Pennsylvania, School of Medicine

https://doi.org/10.5109/12580

出版情報：Bulletin of informatics and cybernetics. 36, pp.73-90, 2004-12. Research Association of Statistical Sciences
バージョン：
権利関係：

MODEL-BASED ESTIMATES OF POPULATION ATTRIBUTABLE RISKS FOR ORDINAL DATA

by

Srabashi Basu and **J. Richard Landis**

Reprinted from the Bulletin of Informatics and Cybernetics
Research Association of Statistical Sciences, Vol.36

FUKUOKA, JAPAN
2004

Bulletin of Informatics and Cybernetics, Vol. 36, 2004

MODEL-BASED ESTIMATES OF POPULATION ATTRIBUTABLE RISKS FOR ORDINAL DATA

By

Srabashi Basu* and **J. Richard Landis†**

Abstract

Threshold-specific population attributable risk measures are developed to account for ordinal disease classifications. A cumulative logit model is utilized to formulate the threshold-specific risks as functions of the underlying model parameters. Covariate-adjusted and overall attributable risk measures are proposed to quantify the impact of exposure to an ordinal risk factor on an ordinal disease classification, in the presence of a confounding variable. These methods are developed under prospective and cross-sectional sampling designs. The asymptotic dispersion matrices of the risk estimates are obtained using multivariate Taylor series expansions which incorporate the sampling variation of the estimated model parameters and the appropriate estimates of risk factor prevalences. These methods are illustrated within the context of a health examination data, investigating the potential influence of body mass index, adjusted for race, on the prevalence distribution of diastolic blood pressure among adult women in the U.S.

Key Words and Phrases: cumulative logit modeling, implicit function theorem, ordinal data, population attributable risk.

1. Introduction

Measures of relative risk (RR) and population attributable risk (PAR) are two major etiologic concepts used to quantify the association between a putative risk factor and a selected disease in a target population. Quite frequently, both the risk factor and the disease response are reported on an ordinal measurement scale. Even when the underlying response variable is continuous, (*e.g.* an individual's blood pressure measurements indicating his/her hypertension status) it is a common practice to categorize the response variable for clinicians' benefits (Archives of Internal Medicine 1993). Consider situations where the disease classification includes $J > 2$ ordered response categories, (*e.g.*, none, mild, moderate and severe stages of disease) and the risk factor has $I \geq 2$ ordinal levels, (*e.g.*, none, low, medium and high exposure). The important concern here is to quantify the extent of disease reduction in the target population, relative to each increasing level of the ordinal classification, which (theoretically) could be realized if the risk factor were eliminated.

To date, the methods in the research literature for quantifying PAR are limited to binary disease classifications (Levin 1953; Walter 1975, 1976, 1978, 1980; Walker

* Theoretical Statistics and Mathematics Unit, Indian Statistical Institute, 203 B. T. Road, Kolkata 700 108, INDIA srabashi@isical.ac.in

† Center for Clinical Epidemiology and Biostatistics, University of Pennsylvania, School of Medicine, Philadelphia, PA 17104–6021, U.S.A. jrlandis@cceb.upenn.edu

1981; Whittemore 1982, 1983; Greenland 1984, 1987; Kuritz and Landis 1987, 1988a, 1988b). Several authors have proposed logistic model-based approaches to estimate PAR, but keeping it limited to binary disease levels (Deubner et al 1980; Bruzzi et al 1985; Benichou and Gail 1989, 1990; Drescher and Schill 1991; Greenland and Drescher 1993; Basu and Landis 1995). In contrast to such methods, measures of relative risk and population attributable risk for ordinal disease outcomes require special consideration to accomodate multiple levels of disease. A naive approach to this problem is to collapse the $I \times J$ contingency table at each of the $J-1$ thresholds and then estimate the risk parameters assuming a binary disease classification at that threshold. In this article we propose a more efficient method based on a cumulative logit modeling to formulate estimates of RR and PAR directly for the ordinal disease classifications. Modeling ordinal response data, in contrast to analyzing $J-1$ separate binary partitions, provides considerable gains in efficiency when an appropriate model is specified, as well as a simplified parametrization and ease of interpretation.

The new estimates of PAR are compared against classical estimates obtained by collapsing the disease levels at each ordinal threshold. It is empirically observed that the former have smaller asymptotic variances as well as smaller MSEs in finite samples.

In the next section, the threshold-specific relative risks and attributable risks are defined and formulated in terms of the parameters of a cumulative logit model. In Section 3 maximum likelihood estimation procedure is described and the asymptotic dispersion matrices of the threshold-specific risk estimators are obtained. In Section 4 a series of simulation studies are described to demonstrate the efficacy of model-based estimation over the estimation procedure based on collapsing. In Section 5 a second set of simulation studies is considered to show validity of the asymptotic procedures. In the final section, the model-based procedure is applied to a health survey data to study the effect of body mass index on hypertension in the presence of a covariate.

2. Measures of threshold-specific risk

Consider the disease to be present at J levels in the target population, $0, \ldots, J-1$. At each increasing threshold indexed by j, let $D^{(j)}$ be the observed number of individuals classified as diseased. For example, suppose $J = 4$, corresponding to disease levels labeled as none, mild, moderate and severe. At threshold level 1, all ordinal levels of mild, moderate and severe are classified as diseased; whereas, at threshold level 2, only moderate and severe are classified as diseased. Threshold-specific measures of both relative risk and attributable risk are defined in such a way that, for each threshold, these parameters for the $2 \times J$ table are identical to the ordinary risk measures for the 2×2 table, obtained by collapsing the original $2 \times J$ table at that threshold.

Let $\boldsymbol{x_l}$, $l = 1, \ldots, v$ be a row vector having the primary risk factor for the disease, other covariates associated with the risk factor and their product terms as its constituent elements. Further let $\boldsymbol{z_l}$, $l = 1, \ldots, v$ be another vector corresponding to $\boldsymbol{x_l}$, such that $\boldsymbol{z_l}$ is the value of the predicting variables a subject with actual predicting value $\boldsymbol{x_l}$ would have if not exposed to the primary risk factor. Further, let $\boldsymbol{z_0} = \boldsymbol{x_0}$ be the baseline value of the covariate combinations, so that a subject having $\boldsymbol{z_0}$ value of the predicting variables is at the lowest risk of having the disease.

We propose a cumulative logit model for the ordinal frequency data under the

assumption of proportional odds by letting

$$\begin{aligned} L_j(\boldsymbol{x}) &= \log \frac{\Pr(D^{(j)}|\boldsymbol{x})}{1-\Pr(D^{(j)}|\boldsymbol{x})} \\ &= \alpha_j + \boldsymbol{x}\mathbf{B}'. \end{aligned} \tag{1}$$

α_j's are the intercept parameters at the j-th threshold of disease. Define $\boldsymbol{\theta^{(j)}}$ to be the vector $(\alpha_j, \mathbf{B})$. Subsequent development of the method is based on this model albeit as an example. Note that the method is completely general and may be used with any sensible model. Rewriting (1), the conditional probability of disease at threshold j can be expressed as

$$\Pr(D^{(j)}|\boldsymbol{x}) = \frac{exp((1,\boldsymbol{x})\boldsymbol{\theta^{(j)}}')}{1+exp((1,\boldsymbol{x})\boldsymbol{\theta^{(j)}}')} = expit((1,\boldsymbol{x})\boldsymbol{\theta^{(j)}}').$$

Let n_l be the number of subjects observed at combination $\boldsymbol{x_l}$ and $\boldsymbol{n}$ be the row vector with l-th element n_l. Moreover define $\boldsymbol{r_z^{(j)}}$ and $\boldsymbol{r_x^{(j)}}$ to be the row vectors with the l-th elements $expit((1,\boldsymbol{z_l})\boldsymbol{\theta^{(j)}}')$ and $expit((1,\boldsymbol{x_l})\boldsymbol{\theta^{(j)}}')$ respectively. Also let $t_z^{(j)} = \boldsymbol{n}\boldsymbol{r_z^{(j)\prime}}$ and $t_x^{(j)} = \boldsymbol{n}\boldsymbol{r_x^{(j)\prime}}, j = 1, \ldots, J-1$ be the expected total number of diseased individuals in the population under the reference and observed covariate levels, as defined at the j-th threshold of disease.

Hence the threshold-specific relative risk at the j-th threshold is defined as

$$R^{(j)}(\boldsymbol{x_l}) = \frac{\Pr(D^{(j)}|\boldsymbol{x_l})}{\Pr(D^{(j)}|\boldsymbol{z_0})},\ j = 1, \ldots, J-1. \tag{2}$$

$R^{(j)}(\boldsymbol{x_l})$ compares the probability that a randomly selected individual exposed to the risk factor combination denoted by $\boldsymbol{x_l}$ has the disease as defined at the j-th threshold, relative to the probability that a randomly selected individual exposed to the baseline combination only has the disease. Further the population attributable risk at the j-th threshold of disease is defined as

$$\lambda^{(j)} = 1 - \frac{t_z^{(j)}}{t_x^{(j)}}. \tag{3}$$

Let us now consider the case of a single ordinal risk factor, a binary covariate and an ordinal disease response in detail. This simple model is chosen on account of its applicability and easy derivation of model-based threshold-specific risk parameters. Extension of the formulation of threshold-specific risk parameters to more complex models is straightforward. Let E denote exposure assumed to be present in I ordinal levels and C denote a binary covariate assumed to be associated with the risk factor. Under this simplistic assumptions the model in (1) reduces to

$$L_j^c(i) = \alpha_j + i\beta + c\gamma\ ,\quad i = 0, \ldots, I-1,\ c = 0, 1, \tag{4}$$

where $L_j^c(i)$ is the j-th cumulative logit at exposure level i and covariate level c, α_j is the intercept of the j-th cumulative logit, β is the trend coefficient of the risk factor and

γ is the coefficient for the covariate. The probability of disease at the j-th threshold is given by

$$\Pr(D^{(j)}|C_cE_i) = \frac{exp(\alpha_j + i\beta + c\gamma)}{1 + exp(\alpha_j + i\beta + c\gamma)}. \tag{5}$$

Hence the relative risk at the j-th threshold of disease at covariate level c and exposure level i is

$$\begin{aligned} R_{ci}^{(j)} &= \frac{\Pr(D^{(j)}|C_cE_i)}{\Pr(D^{(j)}|C_0E_0)} \\ &= exp(i\beta + c\gamma)\left[\frac{1 + exp(\alpha_j)}{1 + exp(\alpha_j + i\beta + c\gamma)}\right]. \end{aligned} \tag{6}$$

Walter (1976) and Whittemore (1982) introduced a covariate-adjusted PAR, λ_A, when one or more confounding factors are present in the target population and the disease is binary. This definition is easily extended to a target population where, at each level of the covariate C, the risk factor is present in I levels and the disease is present in J levels. The attributable risk at the j-th threshold of disease, adjusted for the covariate, is

$$\lambda_A^{(j)} = 1 - \frac{\sum_{c=0}^{1} \Pr(D^{(j)}|C_cE_0)\Pr(C_c)}{\sum_{c=0}^{1}\sum_{i=0}^{I-1} \Pr(D^{(j)}|C_cE_i)\Pr(C_cE_i)}. \tag{7}$$

Walter (1976) also defined an overall PAR, λ_O, when the risk factor has multiple levels, which can be extended readily to the present situation. Therefore, the overall PAR at the j-th threshold of disease may be formulated as

$$\lambda_O^{(j)} = \frac{\sum_{c=0}^{1}\sum_{i=0}^{I-1}(R_{ci}^{(j)} - 1)\Pr(C_cE_i)}{1 + \sum_{c=0}^{1}\sum_{i=0}^{I-1}(R_{ci}^{(j)} - 1)\Pr(C_cE_i)}. \tag{8}$$

Parametric formulations of $\lambda_A^{(j)}$ and $\lambda_O^{(j)}$ are obtained easily utilizing (5) and (6).

As defined in (7), $\lambda_A^{(j)}$ is the measure of the theoretical reduction in the disease prevalence (or incidence, if the design is prospective) at the j-th threshold, adjusted for the covariate, presuming that the risk factor is completely eliminated from the population. This will be achieved only if subjects exposed to any of the multiple levels of the risk factor revert to the baseline risk corresponding to no exposure. On the other hand, $\lambda_O^{(j)}$ measures the overall proportionate reduction in the disease rate at the j-th threshold, theoretically possible only if the risk factor and all the covariates are eliminated from the target population. The underlying assumption here is that the higher category of the risk factor, as well as the presence of the covariates, increase the probability of disease at each threshold.

3. Estimation procedure

In this section the maximum likelihood estimates for the threshold-specific covariate-adjusted PAR and the overall PAR, together with their asymptotic variances, under the model given in (4), are derived in detail. These results are directly extendable to the general model given in (1). A cross-sectional sampling design is assumed for variance

formulations. Results under the prospective sampling design follow very easily from those under the cross-sectional survey and they have been indicated wherever appropriate. Under the cross-sectinal sampling scheme, n individuals are cross-classified into a $2\times I\times J$ table according to their covariate status, risk factor exposure and ordinal disease levels. Our goal is to derive maximum likelihood estimates of $\mathbf{\Lambda_A} = \left(\lambda_A^{(1)},\ldots,\lambda_A^{(J-1)}\right)$ and $\mathbf{\Lambda_O} = \left(\lambda_O^{(1)},\ldots,\lambda_O^{(J-1)}\right)$ and their asymptotic variance–covariance matrices.

Maximum likelihood estimates of the logit model parameters can be obtained readily from any standard software package (*e.g.* PROC LOGIST in SAS) that fits cumulative logit models. Let the resulting estimated parameter vector be denoted by $\mathbf{b} = (\widehat{\alpha}_1,\ldots,\widehat{\alpha}_{J-1},\widehat{\beta},\widehat{\gamma})$ and its estimated asymptotic variance–covariance matrix by $\widehat{V}_b$, which is an estimate of the inverse information matrix. The covariate–exposure prevalences are estimated by the observed cell proportions, which are the maximum likelihood estimates of the theoretical probabilities. Note that the exposure rates in different categories of the risk factor do not change with the thresholds of disease. Replacing the parameters by their maximum likelihood estimates in (7) and (8), the MLEs of $\mathbf{\Lambda_A}$ and $\mathbf{\Lambda_O}$ are obtained. Let the estimates be denoted by $\widehat{\mathbf{\Lambda}}_\mathbf{A}$ and $\widehat{\mathbf{\Lambda}}_\mathbf{O}$ respectively.

The multivariate delta method is used to derive the asymptotic dispersion matrices of $\widehat{\mathbf{\Lambda}}_\mathbf{A}$ and $\widehat{\mathbf{\Lambda}}_\mathbf{O}$, denoted by $\widehat{\Delta}_r(\widehat{\mathbf{\Lambda}}_\mathbf{A})$ and $\widehat{\Delta}_r(\widehat{\mathbf{\Lambda}}_\mathbf{O})$, the subscript r denoting cross-sectional sampling. Since it is more convenient to work with the complements of $\widehat{\lambda}_A^{(j)}$ on the natural logarithm transform, let

$$\bar{\mathbf{\Lambda}}_\mathbf{A} = \left(1-\log_e(\lambda_A^{(1)}),\ldots,1-\log_e(\lambda_A^{(J-1)})\right)$$

and, define

$$\bar{\mathbf{\Lambda}}_\mathbf{O} = \left(\log_e(\lambda_O^{(1)}),\ldots,\log_e(\lambda_O^{(J-1)})\right).$$

Also note that, under this sampling scheme, the vector of estimated covariate-adjusted and overall threshold-specific risk parameters are functions of two dependent sets of random variables, $\mathbf{b}$, the estimated logit model parameters and $\mathbf{p}$, the sample proportions. Define J_A and J_O to be the Jacobian matrices of $\widehat{\bar{\mathbf{\Lambda}}}_\mathbf{A}$ and $\widehat{\bar{\mathbf{\Lambda}}}_\mathbf{O}$ with respect to the elements of $\mathbf{b}$. Similarly, define B_A and B_O to be the Jacobian matrices of $\widehat{\bar{\mathbf{\Lambda}}}_\mathbf{A}$ and $\widehat{\bar{\mathbf{\Lambda}}}_\mathbf{O}$ with respect to the elements of $\mathbf{p} = (p_{1,I-1,J-1},\ldots,p_{1,I-1,0},p_{1,I-2,J-1},\ldots,p_{000})$, where p_{cij} denotes the observed proportion in the c-th level of the covariate, i-th level of the risk factor and j-th level of the ordinal disease classification.

Following Benichou and Gail (1989, 1990), the covariance matrix between $\mathbf{p}$ and $\mathbf{b}$ is obtained as

$$C = \widehat{V}_b H \widehat{\Sigma}\ , \tag{9}$$

where $\widehat{\Sigma}$ is the estimated dispersion matrix of $\mathbf{p}$ and H is the Hessian matrix under the cross-sectional sampling scheme. $\widehat{\Sigma}$ is the estimated variance-covariance matrix of a single multinomial random vector $\mathbf{p}$. The (k,l)–th element of H is given by $\frac{\partial^2(l)}{\partial(b_k)\partial(p_l)}$, where l is the log likelihood under the cross-sectional sampling design, b_k is the k–th element of $\mathbf{b}$ and p_l is the l–th element of $\mathbf{p}$. The whole thing is evaluated at the MLEs of the parameters. Define $\bar{A}_A$ and A_O to be diagonal matrices with j–th diagonals given by $\frac{\partial(1-\lambda_A^{(j)})}{\partial\log(1-\lambda_A^{(j)})}$ and $\frac{\partial(\lambda_O^{(j)})}{\partial\log(\lambda_O^{(j)})}$, respectively. Thus, applying the delta method twice in

succession

$$\widehat{\Delta}_r(\widehat{\mathbf{\Lambda}}_{\mathbf{A}}) = \bar{A}_A \left[J_A \widehat{V}_b J_A' + B_A \widehat{\Sigma} B_A' + 2 B_A C' J_A' \right] \bar{A}_A' \tag{10}$$

and

$$\widehat{\Delta}_r(\widehat{\Lambda}_O) = A_O \left[J_O \widehat{V}_b J_O' + B_O \widehat{\Sigma} B_O' + 2 B_O C' J_O' \right] A_O' \ . \tag{11}$$

When the sampling design is prospective, the covariance–exposure prevalences are not estimable from the data. To get an estimate of the covariate-adjusted and overall threshold-specific PAR parameters, the prevalences must be presumed to be known from some other source. The ML estimates of the logit model parameters are the same under both prospective and cross-sectional designs. However, since in this case the risk estimates are functions of **b** only (and not of **p** any more), the asymptotic variance–covariance matrices have simplified forms. These are obtained by equating B_A and B_O to null matrices in (10) and (11). Specifically

$$\widehat{\Delta}_s(\widehat{\mathbf{\Lambda}}_{\mathbf{A}}) = \bar{A}_A \left[J_A \widehat{V}_b J_A' \right] \bar{A}_A'$$

and

$$\widehat{\Delta}_s(\widehat{\mathbf{\Lambda}}_{\mathbf{O}}) = A_O \left[J_O \widehat{V}_b J_O' \right] A_O' \ ,$$

where the subscript s denotes the prospective sampling design.

4. Advantages of modeling

Note that the j-th threshold-specific attributable risk, as defined in Section 2, coincides with that of PAR in the binary disease situation if the $I \times J$ contingency table is collapsed at the j-th threshold, $j = 1, \ldots, J-1$. In this section, we attempt to put forward the advantages of the model-based procedures through a set of Monte Carlo simulations designed to show improvement due to modeling ordinal response data with small to moderate cell sizes.

It is well known that model-based estimation leads to increased asymptotic precision when the assumed model adequately describes the data (Bishop, Fienberg and Holland 1975; Altham 1984). The simpler the model, the better the performance of the model-based estimator. Note, however, that in order to have improved precision, the assumed model has to fit the data. If the model gives a poor fit to the data, the model-based estimates still have lower asymptotic variance but the bias does not go to 0 with increasing sample size. In this article, a p-value for the lack-of-fit statistic that is greater than 0.05 is used as evidence that the model is adequate for the data. However, in the model building process a higher p-value, *e.g.* 0.20 or 0.25 may be preferred.

Comparable improvements are also possible in small samples when the stipulated model holds. A set of simulation studies is described next to illustrate that the model-based estimates of threshold-specific PAR can outperform the estimates of PAR from collapsed tables in terms of a smaller mean square error. We consider three different scenarios – (i) a 2×4 contingency table with the disease at 4 ordinal levels and a single binary risk factor; (ii) a 4×4 contingency table with both the risk factor and the disease at 4 ordinal levels; and (iii) a $2 \times 4 \times 4$ contingency table where in addition a binary covariate is present.

Consider a multinomial frequency distribution of dimension $C \times I \times J$ which can be modeled by (4). When the values of the model parameters, and consequently the

underlying multinomial probabilities, are known, the true values of the threshold-specific covariate-adjusted and overall PAR parameters may be directly determined.

Random samples of size n were generated from the multinomial distribution with $C \times I \times J$ known cell probabilities using the IMSL random number generator RNMTN in FORTRAN. A cumulative logit model was fitted to each of the observed contingency tables. MLEs of the logit model parameters were estimated in PROC LOGIST in SAS. These MLEs in turn were used to obtain MLEs of the threshold-specific covariate-adjusted and overall PARs.

Alternatively, the observed contingency table was collapsed at each increasing threshold of the disease and covariate-adjusted and overall PAR were estimated corresponding to each binary disease response, as defined at that threshold. Let these estimates be called classical estimates. The complete procedure was repeated 5000 times and the MSE of the classical and the model-based estimates were calculated as the average of the squared deviation of the estimates from the true value of the PAR parameters. These results are presented in Tables 1–3. The last columns of these tables are Ratio $= \frac{\text{Model-based MSE}}{\text{Classical MSE}}$.

Case I

Consider a sample distributed in a 2×4 contingency table, where the risk factor is present at 2 levels, 0 (not exposed) and 1 (exposed), and the disease is present at 4 ordinal levels. Disease risk at the j-th threshold is given by

$$\Pr(D^{(j)}|i) = [\exp(\alpha_j + i\beta)]\,[1 + \exp(\alpha_j + i\beta)]^{-1}, \quad i = 0, 1 \text{ and } j = 1, 2, 3.$$

$\alpha_1 = -1.2$, $\alpha_2 = -2.8$ and $\alpha_3 = -3.5$ are effects for increasing threshold levels and $\beta = 1.25$ is the exposure effect. The risk factor prevalence is assumed to be 25% in the population. In the absence of any covariate and multiple levels of risk, only one set of threshold-specific relative risks and corresponding PARs are defined. The true values of the threshold-specific PAR are

$$\lambda^{(1)} = 0.233,\ \lambda^{(2)} = 0.339,\ \lambda^{(3)} = 0.360\ .$$

Table 1 presents a series of cross-sectional study simulations from this population.

Case II

Now consider a sample distributed in a 4×4 contingency table with the risk factor at 4 ordinal levels, 0, 1, 2 and 3 and the disease at 4 ordinal levels also. Disease risk at the j-th threshold is given by

$$\Pr(D^{(j)}|i) = [\exp(\alpha_j + i\beta)]\,[1 + \exp(\alpha_j + i\beta)]^{-1}, \quad i = 0, 1, 2, 3 \text{ and } j = 1, 2, 3.$$

$\alpha_1 = -1.4$, $\alpha_2 = -2.9$ and $\alpha_3 = -4.3$ are effects for increasing threshold levels and $\beta = 1.5$ is a trend effect for exposure. The risk factor prevalence is assumed to be 35%, 30%, 20% and 15% at levels 0, 1, 2 and 3 respectively. The threshold-specific relative risk at each of the $i = 1, 2, 3$ levels of the risk factor is defined compared to the baseline exposure. In absence of any covariate, only threshold-specific overall PARs are defined and they are

$$\lambda^{(1)} = 0.631,\ \lambda^{(2)} = 0.830,\ \lambda^{(3)} = 0.909\ .$$

Table 2 presents a series of cross-sectional study simulations from this population.

Case III

Lastly, consider a sample distributed in a $2 \times 4 \times 4$ contingency table with the risk factor present at 4 levels, 0, 1, 2 and 3, a covariate present at 2 levels, 0 and 1, and the disease present at 4 ordinal levels. Disease risk at the j-th threshold is given by

$$\Pr(D^{(j)}|i) = [\exp(\alpha_j + i\beta + c\gamma)]\,[1 + \exp(\alpha_j + i\beta + c\gamma)]^{-1},$$

$i = 0, 1, 2, 3$, $c = 0, 1$, and $j = 1, 2, 3$. Let $\alpha_1 = -1.4$, $\alpha_2 = -2.9$ and $\alpha_3 = -4.3$ be effects for increasing threshold levels, $\beta = 1.5$ be a trend coefficient for exposure and $\gamma = 0.75$ be the coefficient for the covariate effect. The risk factor prevalence is assumed to be 20%, 16%, 9% and 6% at covariate level 0 and 18%, 14%, 10% and 7% at covariate level 1 for levels 0, 1, 2 and 3 respectively. Both the threshold-specific covariate adjusted PAR and the overall PAR are defined for this case, and their true values are

$$\lambda_A^{(1)} = 0.533, \quad \lambda_A^{(2)} = 0.771, \quad \lambda_A^{(3)} = 0.881$$

and

$$\lambda_O^{(1)} = 0.656, \quad \lambda_O^{(2)} = 0.847, \quad \lambda_O^{(3)} = 0.922.$$

Table 3 presents a series of cross-sectional study simulations for this population.

Several features of these three sets of empirical studies are noteworthy. The most important feature is that the Ratio is always less than 1, implying that even in small samples the model-based method performs better than the classical method. When the model is appropriate for data, the point estimates obtained under both methods are very close, as is evident from the columns called 'Classical Mean' and 'Model-based Mean'. However, the classical MSE is always greater than the Model-based MSE, implying that the true variance of model-based estimates are always lower than that of classical estimates.

Secondly, the amount of savings, defined as 1−Ratio, is increasing at each increasing threshold. This could be easily explained by the structure of the $C \times I \times J$ contingency tables. Typically as disease threshold increases, the disease prevalence (or incidence) rate as defined at that threshold decreases. This in turn makes the classical estimates of the thresold-specific population attributable risks less stable at the higher thresholds. Application of model-based inference procedure guards against that disadvantage of the classical estimation methodology by borrowing strength from the overall distribution of disease prevalence, not merely from the disease prevalence at that threshold.

Moreover, note that the amount of savings at each threshold is increasing with the number of cells in the contingency table. For a 2×4 table, the maximum amount of savings is approximately 76%, obtained at threshold level 3. For a 4×4 contingency table, the maximum savings is approximately 86% and for a $2 \times 4 \times 4$ table it is over 90%. This apparently overwhelming performance of the model-based methodology is due to the sharply increasing ratio of number of parameters in the saturated model for the contingency table over that in the specific model in use. In the 2×4 table number of parameters in the saturated model is 8 and the number of parameters in the cumulative logit model adopted is 4. In Case II, the ratio is 16:4, and in Case III the ratio is 32:5. The improvement due to modeling is a function of the number of parameters. When the

data are generated from a stipulated model the relevent information is contained only in the parameters of the model, rendering the extra parameters in the saturated model totally superfluous.

However, in case of real data the savings may not be as high. It is highly unrealistic to presume that all information regarding an observed contingency table is contained solely in the model parameters, even when the model does fit the data.

5. Empirical validity of asymptotic procedures

In this section a set of simulations are presented paralleling those given by Benichou and Gail (1990) and Greenland and Drescher (1993) to demonstrate the validity of asymptotic model-based procedures. The sampled distribution is obtained based on a study of hypertension, described in detail in the next section. To keep things simple, a single risk factor at four levels and an ordinal disease outcome at four levels are assumed. The scenario considered here is identical to that of Case II described in Section 4.

The results of the Monte Carlo study are presented in Table 4, each column of which summarizes 5000 simulation trials on a series of cross-sectional studies. Two maximum likelihood confidence intervals are considered: the untransformed interval is given by $\widehat{\lambda}^{(j)} \pm 1.96\widehat{se}(\widehat{\lambda}^{(j)})$ and the log transformed interval is given by $exp\big(\log\widehat{\lambda}^{(j)} \pm 1.96\widehat{se}(\log\widehat{\lambda}^{(j)})\big)$.

Contrary to the previous observations (Benichou and Gail 1990, Greenland and Drescher 1993) the threshold-specific population attributable risk estimates are not downwardly biased even at a moderate sample size. The average estimate of $\widehat{se}(\widehat{\lambda}^{(j)})$ is within 10% of the sample standard deviation of $\widehat{\lambda}^{(j)}$ in each case. Both the untransformed and the log transform confidence intervals report slight undercoverage for considerably large sample sizes. Moreover, both intervals are equivalent in terms of the mean length.

6. Examples and results

National health examination survey data are used to illustrate the parallel methods developed for prospective and cross-sectional study designs fitting a cumulative logit model. The data in Table 5, obtained from the second National Health and Nutrition Examination Survey (NHANES II), conducted from 1976–1980 (DHHS 1976-1980), were selected for illustration due to keen public health interest in the potentially confounding roles of race and body mass index, defined as $\frac{(weight\ in\ kg.)}{(height\ in\ meters)^2}$, in affecting the distribution of diastolic blood pressure. These unweighted frequency data for women, ages 18–24 years at the time of examination, summarize the distribution of diastolic blood pressure (DBP) classified into 4 ordinal levels determined by the stages of hypertension published as a clinical guideline by American Medical Association. For convenience, we have defined the disease levels as DBP ≤ 89 (normal and high normal), $90 \leq$ DBP ≤ 99 (mild hypertension), $100 \leq$ DBP ≤ 109 (moderate hypertension) and $110 \leq$ DBP (severe and very severe hypertension), all measurements in mm/Hg. Collapsing of the categories normal and high normal, and severe and very severe hypertension is done from a clinician's viewpoint as well as to avoid a large number of empty cells.

The primary risk factor for elevated DBP is body mass index (BMI), classified at four ordinal levels, with race, classified as black or white, considered as a potentially

confounding factor. A key question in cardiovascular epidemiology is the extent to which observed racial differences in blood pressure distributions are attributable to risk factor differences such as BMI. For women in the U.S. population, a BMI of 23 is approximately the median, so for these purposes, BMI < 23 is considered as exposure level 0 (E_0). Using the notation developed previously in this article, race white is labeled C_0 and DBP less than 90 mm/Hg as level 0 of disease, and the higher DBP measurements labeled 1, 2 and 3, respectively, even though these ordinal thresholds of DBP are not intended to imply 'disease' at any level.

Two sets of threshold-specific PAR risk measures are estimated, one for the impact of BMI, adjusted for race (covariate-adjusted PAR) and the other for the combined impact of BMI and race (overall PAR). The actual design for the NHANES II survey utilized a complex, weighted, multi-stage cluster sampling approach, but for illustration of these PAR estimation methods, these data in Table 5 will be analyzed assuming an unweighted, simple random sample (SRS) design. Both the cross-sectional sample methods (corresponding to the prevalence distribution of DBP) and the prospective sample methods (corresponding to the incidence distribution of DBP, hypothetically assuming a longitudinal design) are implemented to facilitate comparisons between these two sampling schemes. Furthermore, under the prospective sampling design the covariate–exposure prevalences are taken to be equal to their estimates under the cross-sectional sampling design, so that the point estimates of the covariate-adjusted and the overall threshold-specific PARs under both sampling designs are identical.

A cumulative logit model, with equally-spaced scores (0, 1, 2, 3) assigned to the ordinal levels of BMI, and a standard binary effect parameterization for race, was fit to the data in SAS PROC LOGISTIC, resulting in a lack-of-fit statistic of 4.72 with 2 df (p=0.32), suggesting that the assumption of proportional odds is acceptable for these data. The parameter estimates are

$$\widehat{\alpha}_1 = 0.413, \widehat{\alpha}_2 = -1.292, \widehat{\alpha}_3 = -2.515, \widehat{\beta} = 0.478, \widehat{\gamma} = 0.164.$$

Recall that the first three parameters correspond to the three thresholds, β is the trend effect of ordinal levels of BMI and γ is the race effect. Covariate-adjusted (for race) threshold-specific PAR estimates for BMI are summarized in Table 6, together with overall PAR estimates assessing the combined impact of BMI and race on DBP. It is noteworthy, for these data in Table 5, that the race-adjusted estimates of PAR are somewhat smaller than the overall PAR estimates at each threshold of DBP. Furthermore it suggests that nearly 1/3 of the women with DBP exceeding 110 mm/Hg (threshold 3) may be attributable to BMI exceeding 23.

Estimated asymptotic variances also are provided in Table 6 for the threshold-specific PARs, both under the prospective sample and cross-sectional sample assumptions, the primary difference related to the prevalence distribution of levels of BMI assumed known (prospective) or estimated from the data (cross-sectional). At each threshold of the DBP distribution, these estimated variances are slightly larger under the cross-sectional sampling assumptions (due to the additional variability induced by estimating the prevalence of Race $\times$ BMI levels).

It is also to be noted that no non-model based estimate for variance of covariate-adjusted PAR is available for this scenario even when the disease is dichotomized. Hence no comparison between classical and model-based estimates of asymptotic variances of $\widehat{\lambda}_A$ is possible. We do not intend to compute an overall PAR with all the eight categories

of exposure either as that is bound to provide a distorted picture. Instead, just as an illustration to compare the performance of classical and model-based methods in a real data set, let us focus our attention on white women only with two exposure groups, namely BMI ≤ 25 and BMI ≥ 25. The estimated PARs when the table is collapsed at the three threshold levels of the disease are respectively 6%, 17% and 24% while model-based estimates are 5%, 17% and 30%. Under prospective sampling scheme the model-based estimates of the asymptotic s.e.s are $(0.064, 0.695, 1.618) \times 10^{-3}$ whereas non model-based estimates are $(0.112, 0.877, 3.709) \times 10^{-3}$. Similar observations are made in case of cross-sectional sampling scheme also. The former method gives $(0.080, 0.785, 1.760) \times 10^{-3}$ compared to the latter which gives $(0.126, 0.985, 3.936) \times 10^{-3}$ as estimates of asymptotic s.e.s.

In summary, the developments for model-based estimation of threshold-specific PARs, to assess the impact of a risk factor (with adjustments for another covariate) on a disease with ordinal levels, have been illustrated for two important sampling designs. Considerable gains in efficiency for these estimators have been demonstrated after adjusting for the ordinality of disease and the ordinality of the underlying risk factor in cases where non model-based estimates of PAR are also available. Further, in situations where no estimate of asymptotic variances of PARs had previously been proposed, we have formulated variance estimates which are very easily obtained. Adjustment for one or more covariates can now be easily dealt with. With appropriate use of Taylor series expansions, and implicit function methods, the asymptotic variances of these complex ratio estimators have been provided for two alternative sampling situations.

Table 1: Simulation results comparing classical and model-based PAR: Cross-sectional Study in a 2×4 contingency table ($\lambda^{(1)} = 0.233, \quad \lambda^{(2)} = 0.339, \quad \lambda^{(3)} = 0.360$)

Sample Size	Thres-hold	Classical Mean	Classical MSE $\times 10^2$	Model-based Mean	Model-based MSE $\times 10^2$	Ratio
	1	0.234	0.495	0.234	0.470	0.948
200	2	0.340	2.283	0.342	1.010	0.442
	3	0.357	4.616	0.363	1.129	0.245
	1	0.233	0.192	0.233	0.182	0.951
500	2	0.339	0.896	0.340	0.399	0.445
	3	0.362	1.842	0.361	0.447	0.243
	1	0.233	0.132	0.233	0.126	0.954
750	2	0.342	0.628	0.340	0.277	0.441
	3	0.361	1.276	0.361	0.311	0.244
	1	0.233	0.096	0.233	0.092	0.956
1,000	2	0.339	0.486	0.339	0.204	0.420
	3	0.362	0.980	0.360	0.230	0.234
	1	0.232	0.065	0.232	0.062	0.956
1,500	2	0.339	0.306	0.339	0.138	0.449
	3	0.361	0.633	0.359	0.155	0.245
	1	0.233	0.050	0.233	0.048	0.955
2,000	2	0.339	0.239	0.339	0.105	0.439
	3	0.355	0.465	0.360	0.118	0.255

Table 2: Simulation results comparing classical and model-based PAR: Cross-sectional Study in a 4×4 contingency table ($\lambda^{(1)} = 0.631, \quad \lambda^{(2)} = 0.830, \quad \lambda^{(3)} = 0.909$)

Sample Size	Thres-hold	Classical Mean	Classical MSE $\times 10^2$	Model-based Mean	Model-based MSE $\times 10^2$	Ratio
200	1	0.634	0.648	0.633	0.362	0.558
	2	0.831	0.665	0.830	0.210	0.316
	3	0.903	0.741	0.907	0.114	0.153
500	1	0.631	0.259	0.632	0.139	0.538
	2	0.830	0.279	0.830	0.082	0.296
	3	0.908	0.325	0.908	0.043	0.132
750	1	0.631	0.177	0.631	0.094	0.535
	2	0.829	0.187	0.830	0.057	0.303
	3	0.907	0.222	0.908	0.030	0.135
1,000	1	0.631	0.128	0.631	0.069	0.537
	2	0.830	0.136	0.830	0.040	0.295
	3	0.909	0.161	0.908	0.021	0.131
1,500	1	0.631	0.087	0.632	0.047	0.540
	2	0.831	0.088	0.830	0.027	0.305
	3	0.908	0.114	0.909	0.014	0.124
2,000	1	0.631	0.064	0.632	0.036	0.554
	2	0.830	0.072	0.830	0.020	0.285
	3	0.909	0.085	0.909	0.011	0.126

Table 3: Simulation results comparing classical and model-based PAR: Cross-sectional Study in a $2\times4\times4$ contingency table ($\lambda_A^{(1)} = 0.533$, $\lambda_A^{(2)} = 0.771$, $\lambda_A^{(3)} = 0.881$, $\lambda_O^{(1)} = 0.656$, $\lambda_O^{(2)} = 0.847$, $\lambda_O^{(3)} = 0.922$)

		Covariate-adjusted PAR					Overall PAR				
Sample	Thres-	Classical		Model-based			Classical		Model-based		
Size	hold	Mean	MSE $\times10^2$	Mean	MSE $\times10^2$	Ratio	Mean	MSE $\times10^2$	Mean	MSE $\times10^2$	Ratio
	1	0.535	0.589	0.534	0.344	0.584	0.657	1.085	0.657	0.430	0.396
200	2	0.774	0.748	0.771	0.267	0.357	0.847	0.973	0.846	0.199	0.204
	3	0.880	0.889	0.879	0.150	0.169	0.916	1.020	0.921	0.091	0.090
	1	0.533	0.238	0.533	0.141	0.594	0.657	0.422	0.657	0.170	0.401
500	2	0.771	0.297	0.771	0.110	0.371	0.846	0.406	0.847	0.078	0.191
	3	0.881	0.336	0.880	0.060	0.177	0.919	0.397	0.922	0.035	0.088
	1	0.534	0.157	0.533	0.090	0.573	0.656	0.286	0.656	0.112	0.392
750	2	0.772	0.192	0.772	0.070	0.363	0.846	0.275	0.846	0.051	0.186
	3	0.882	0.227	0.881	0.039	0.172	0.921	0.271	0.922	0.023	0.085
	1	0.533	0.121	0.533	0.085	0.577	0.656	0.216	0.657	0.085	0.394
1,000	2	0.772	0.144	0.771	0.039	0.372	0.846	0.205	0.846	0.039	0.189
	3	0.881	0.169	0.881	0.017	0.176	0.922	0.210	0.922	0.017	0.083
	1	0.533	0.073	0.533	0.046	0.623	0.656	0.136	0.656	0.056	0.416
1,500	2	0.772	0.098	0.772	0.036	0.370	0.846	0.136	0.846	0.027	0.199
	3	0.881	0.113	0.881	0.020	0.180	0.921	0.145	0.922	0.012	0.084
	1	0.533	0.061	0.533	0.035	0.571	0.657	0.112	0.657	0.042	0.376
2,000	2	0.772	0.073	0.772	0.027	0.371	0.847	0.100	0.847	0.019	0.193
	3	0.882	0.085	0.881	0.015	0.175	0.923	0.107	0.923	0.009	0.081

Table 4: Simulation results for MLE of threshold-specific PAR: Cross-sectional Study in a 4×4 contingency table, modeled by ordinal risk and ordinal disease ($\lambda^{(1)} = 0.631, \lambda^{(2)} = 0.830, \lambda^{(3)} = 0.909$)

Sample size	Thres-hold	Mean pt est	Sample std dev	Mean SE est	Untransformed coverage	Untransformed mean length	Log Transform coverage	Log Transform mean length
200	1	0.633	0.0601	0.0586	0.932	0.230	0.933	0.231
	2	0.830	0.0458	0.0447	0.931	0.175	0.930	0.176
	3	0.907	0.0337	0.0325	0.920	0.128	0.919	0.128
500	1	0.632	0.0373	0.0372	0.947	0.146	0.949	0.146
	2	0.830	0.0287	0.0284	0.943	0.111	0.943	0.111
	3	0.908	0.0207	0.0206	0.940	0.081	0.940	0.081
750	1	0.631	0.0307	0.0304	0.946	0.119	0.947	0.119
	2	0.830	0.0238	0.0232	0.946	0.091	0.947	0.091
	3	0.908	0.0173	0.0168	0.941	0.066	0.941	0.066
1,000	1	0.631	0.0263	0.0264	0.950	0.103	0.950	0.103
	2	0.830	0.0200	0.0201	0.949	0.079	0.948	0.079
	3	0.908	0.0145	0.0146	0.947	0.057	0.947	0.057
1,500	1	0.632	0.0217	0.0215	0.949	0.084	0.947	0.084
	2	0.830	0.0164	0.0164	0.946	0.064	0.945	0.064
	3	0.909	0.0119	0.0119	0.943	0.046	0.944	0.046
2,000	1	0.632	0.0189	0.0186	0.943	0.073	0.943	0.073
	2	0.830	0.0143	0.0142	0.946	0.056	0.945	0.056
	3	0.909	0.0103	0.0103	0.946	0.040	0.947	0.040

Table 5: Frequency Distribution (Cumulative Row Proportions) of Ordinal Levels of Diastolic Blood Pressure (DBP) by Ordinal Levels of Body Mass Index (BMI) and Race: Women, Ages 18–24: NHANES II, 1976-80

Race	Ordinal Levels of BMI	Ordinal Levels of DBP (D=d) 3	2	1	0	Total	Estimated Risk Factor Prevalences
Black	≥ 27	10 (0.3846)	4 (0.5385)	9 (0.8846)	3	26 (1.00)	0.0269
	[25,27)	1 (0.0769)	2 (0.2308)	8 (0.8461)	2	13 (1.00)	0.1035
	[23,25)	3 (0.1429)	6 (0.4286)	7 (0.7619)	5	21 (1.00)	0.0217
	< 23	10 (0.1370)	10 (0.2740)	24 (0.6027)	29	73 (1.00)	0.0756
	Subtotal	24	22	48	39	133	0.1377
White	≥ 27	29 (0.3118)	24 (0.5699)	27 (0.8602)	13	93 (1.00)	0.0963
	[25,27)	7 (0.1094)	16 (0.3594)	25 (0.7500)	16	64 (1.00)	0.0662
	[23,25)	8 (0.0734)	27 (0.3211)	40 (0.6881)	34	109 (1.00)	0.1128
	< 23	40 (0.0705)	79 (0.2099)	229 (0.6138)	219	567 (1.00)	0.5869
	Subtotal	84	146	321	282	833	0.8623
Total		108	168	369	321	966	1.0000

Table 6: Threshold-Specific Race-Adjusted and Overall Population Attributable Risk Ratios for Ordinal Levels of Diastolic Blood Pressure (DBP) by Ordinal Levels of Body Mass Index (BMI) Obtained from a Linear Trend Effects Cumulative Logit Model: Prospective and Cross-sectional Sampling Design

Thres-holds	PAR Estimates $\widehat{\lambda}_A^{(j)}$	$\widehat{\lambda}_O^{(j)}$	Sampling Design: Prospective $\widehat{Var}_p(\widehat{\lambda}_A^{(j)})$ $\times 10^{-3}$	Prospective $\widehat{Var}_p(\widehat{\lambda}_O^{(j)})$ $\times 10^{-3}$	Cross-sectional $\widehat{Var}_c(\widehat{\lambda}_A^{(j)})$ $\times 10^{-3}$	Cross-sectional $\widehat{Var}_c(\widehat{\lambda}_O^{(j)})$ $\times 10^{-3}$
1	0.092	0.100	0.1152	0.1164	0.1343	0.1825
2	0.238	0.251	0.8487	0.9742	0.9525	1.0742
3	0.321	0.336	1.6712	1.7899	1.8382	1.9486

References

Altham, P. M. E. (1984). Improving the precision of estimation by fitting a model *Journal of Royal Statistical Society, Series B*, **46**, 118–119.

Basu, S. and Landis, J. R. (1995). Model-based estimation of population attributable risk under cross-sectional sampling *American Journal of Epidemiology*, **142**, 1338–1343.

Bishop, Y. V. V., Fienberg, S. E. and Holland, P. W. (1975). *Discrete Multivariate Analysis*. MIT Press.

Benichou, J. and Gail, M. H. (1989). A delta method for implicitly defined random variables *The American Statistician*, **43**, 41–44.

Benichou, J. and Gail, M. H. (1990). Variance calculations and confidence intervals for estimates of the attributable risk based on logistic models *Biometrics*, **46**, 991–1003.

Bruzzi, P., Green, S. B., Byar, D. P., Brinton, L. A. and Schairer, C. (1985). Estimating the population attributable risk for multiple risk factors using case-control data *American Journal of Epidemiology*, **122**, 904–914.

Deubner, D. C., Wilkinson, W. E., Helms, M. J., Tyroler, H. A. and Hames C. G. (1980). Logistic model estimation of death attributable to risk factors for cardiovascular disease in Evans county, Georgia *American Journal of Epidemiology*, **112**, 135–143.

Drescher, K. and Schill, W. (1991). Attributable risk estimation from case/control data via logistic regression *Biometrics*, **47**, 1247–1256.

Greenland, S. (1984). Bias in methods for deriving standardized morbidity ratio and attributable fraction estimates *Statistics in Medicine*, **3**, 131–141.

Greenland, S. (1987). Variance estimators for attributable fraction estimates consistent in both large strata and sparse data *Statistics in Medicine*, **6**, 701–708.

Greenland S. and Drescher, K. (1993). Maximum likelihood estimation of the attributable fraction from logistic models *Biometrics*, **49**, 865–872.

Kuritz, S. and Landis, J. R. (1987). Attributable risk ratio estimation from matched-pairs case-control data *American Journal of Epidemiology*, **125**, 324–328.

Kuritz, S. and Landis, J. R. (1988a). Attributable risk estimation from matched case-control data *Biometrics*, **44**, 355–367.

Kuritz, S. and Landis, J. R. (1988b). Summary attributable risk estimation from unmatched case-control data *Statistics in Medicine*, **7**, 507–517.

Levin, M. L. (1953). The occurrence of lung cancer in man *Acta Unio Internationalis Contra Cancrum* **9**, 531–541.

The Fifth Report (1993). The fifth report of the Joint National Committee on detection, evaluation and treatment of high blood pressure (JNC V) January 25, 1993 *Archives of Internal Medicine* **153**, 154–183.

US Dept of Health and Human Services (DHHS) (1976-1980). National Center for Health Statistics, Second national health and nutrition examination survey (NHANES II) Hyattsville, MD, Centers for Disease Control and Prevention.

Walker, A. M. (1981). Proportion of disease attributable to the combined effect of two factors *International Journal of Epidemiology*, **10**, 81–85.

Walter, S. D. (1975). The distribution of Levin's measure of attributable risk *Biometrika*, **62**, 371–375.

Walter, S. D. (1976). The estimation and interpretation of attributable risk in health research *Biometrics*, **32**, 829–849.

Walter, S. D. (1978). alculation of attributable risks from epidemiological data *International Journal of Epidemiology*, **7**, 175–182.

Walter, S. D. (1980). revention for multifactorial diseases *American Journal of Epidemiology*, **112**, 409–416.

Walter, S. D. (1983). Effects of interaction, confounding and observational error on attributable risk estimation *American Journal of Epidemiology*, **117**, 598–604.

Whittemore, A. S. (1982). Statistical methods for estimating attributable risk from retrospective data *Statistics in Medicine*, **1**, 229–243.

Whittemore, A. S. (1983). Estimating attributable risk from case-control studies *American Journal of Epidemiology*, **117**, 76–85.

Received October 15, 2003
Revised June 28, 2004