Biostatistics (2001), **2**, 4, *pp.* 473–483
Printed in Great Britain

The quadratic cumulative odds regression model for scored ordinal outcomes: application to alcohol dependence

DANIEL O. SCHARFSTEIN*, KUNG-YEE LIANG

Department of Biostatistics, Johns Hopkins Bloomberg School of Hygiene and Public Health, 615 N. Wolfe Street, Baltimore, MD 21205, USA
dscharf@jhsph.edu

WILLIAM EATON

Department of Mental Hygiene, Johns Hopkins Bloomberg School of Hygiene and Public Health, 615 N. Wolfe Street, Baltimore, MD 21205, USA

LI-SHIUN CHEN

Department of Psychiatry, Washington University School of Medicine, 660 S. Erclid Avenue, St. Louis, MO 63110, USA

SUMMARY

In this paper, we develop new regression models for the analysis of scored ordinal data (i.e. ordinal outcomes where the categories are assigned numeric values). The novel feature of these models is that they enable one to capture and identify nonlinear aspects of the relationship between an ordinal clinical measurement (used for disease diagnosis) and risk factors. These nonlinearities may be useful in generating hypotheses about the risk factor's role in the etiologic process as well as suggesting how to design future studies of the risk factor. We apply our model to study the effects of race, gender, and family history on alcohol dependence among a cohort of lifetime drinkers from the 1992 National Longitudinal Alcohol Epidemiologic Survey.

Keywords: Constrained maximum likelihood; Diagnostic threshold; Proportional odds model.

1. INTRODUCTION

Diagnosis of alcohol dependence is made by identifying constellations of dependence symptoms set forth in the American Psychiatric Association's Diagnostic and Statistical Manual of Mental Disorders—DSM-IV (American Psychiatric Association, 1994). Specifically, an individual is characterized as alcohol dependent by identifying the presence or absence of at least three of the following seven exchangeable dependence criteria: tolerance; withdrawal or drinking for relief or avoidance of withdrawal; frequent drinking in larger amounts or more often than intended; persistent desire or unsuccessful attempts to cut down on drinking or stop; spending much time getting alcohol, drinking, or getting over the effects of alcohol; giving up or reducing occupational, social, or recreational activities in favor of drinking; and

*To whom correspondence should be addressed

Table 1. *Race, gender, and family history composition within levels of number of dependence symptoms*

	Number of dependence symptoms							
	0	1	2	3	4	5	6	7
#	10013	4939	3620	1681	1464	779	547	561
% White	83.7	85.0	85.4	89.2	86.8	85.0	88.8	86.3
% Male	42.3	47.7	51.4	57.8	61.8	64.8	64.0	68.4
% Family history	25.5	34.2	39.2	44.0	49.4	58.3	70.6	77.2

continuing to drink despite physical or psychological problems caused or exacerbated by the alcohol/drug use. This *diagnostic threshold* has been somewhat arbitrarily defined by clinicians as a way of identifying subgroups of the population who (1) are at high risk for future morbidity or mortality and (2) can benefit from treatment.

The 1992 National Longitudinal Alcohol Epidemiologic Survey (Grant *et al.*, 1994) conducted in-person interviews of 42 862 subjects, who were 18 years or older and resided in the contiguous United States. To evaluate the above alcohol dependence for the prior 12 months and for the past, the Alcohol Disorder and Associated Interview Schedule (AUDADIS; Grant *et al.*, 1995) was administered to each subject. Hasin and Paykin (1999) analysed these data, with a view towards using risk factors to validate the three-symptom diagnostic threshold. Toward this end, they used established risk factors to assess differences among subjects with a DSM-IV diagnosis of alcohol dependence: subjects who manifest one or two dependence symptoms ('diagnostic orphans'), and subjects who exhibit no symptoms. They conducted a subset analysis of 23 604 subjects who had 12 or more drinks during any year in the past and who did not have a current or past diagnosis of alcohol abuse. In their analysis, they used race, gender, and family history as risk factors. Race was defined as white or nonwhite and family history was considered positive if the subject had a biological mother, father, sister, or brother who had a history of alcohol-related problems. These risk factors were considered because past studies have shown that white males with a family history are at highest risk of alcohol dependence. The authors found differences in risk factor profile between subjects with an alcohol diagnosis and diagnostic orphans and differences between subjects with an alcohol diagnosis and subjects with no symptoms. They used these results as support for the three-symptom threshold.

In Table 1 we present the race, gender and family history composition within levels of number of dependence symptoms. Note that these ordered categorical levels are naturally scored by the number of symptoms. In studying the association between alcohol dependence and risk factors, the standard epidemiologic approach is to use the diagnostic threshold to define a subject as alcohol dependent and run logistic regressions to estimate the crude (i.e. ignoring other risk factors) and adjusted (i.e. holding other risk factors constant) odds ratios. The odds ratios serve as measures of association between risks factors and alcohol dependence. In Table 2 we present these odds ratios, along with their 95% confidence intervals. As with past studies, we see that being male and having a family history are strong risk factors of alcohol dependence diagnosis, while being white is a milder risk factor.

While the above analyses are informative, they may be obscuring a more important relationship between the risk factors and the number of dependence symptoms. It is this relationship which may be useful in generating comprehension of the etiology of alcohol dependence. If a risk factor is involved in the pathologic process, then the crude (adjusted) conditional distribution of the number of symptom groups given the risk factor(s) should differ over levels of the risk factor. The comparison of these conditional distributions may yield some clues or generate hypotheses about the risk factor's etiologic contribution. It also may be helpful in identifying subgroups of the population who should be sampled for a future study of the risk factor.

Table 2. *Crude and adjusted odds ratios (with 95% confidence intervals) using a three-symptom threshold for diagnosis of alcohol dependence. Risk factors include race, gender, and family history*

	Crude		Adjusted	
Risk factor	Odds ratio	Confidence interval	Odds ratio	Confidence interval
Race (white)	1.30	[1.18, 1.42]	1.29	[1.17, 1.41]
Gender (male)	1.95	[1.83, 2.07]	2.18	[2.04, 2.33]
Family history	2.72	[2.55, 2.90]	2.95	[2.76, 3.15]

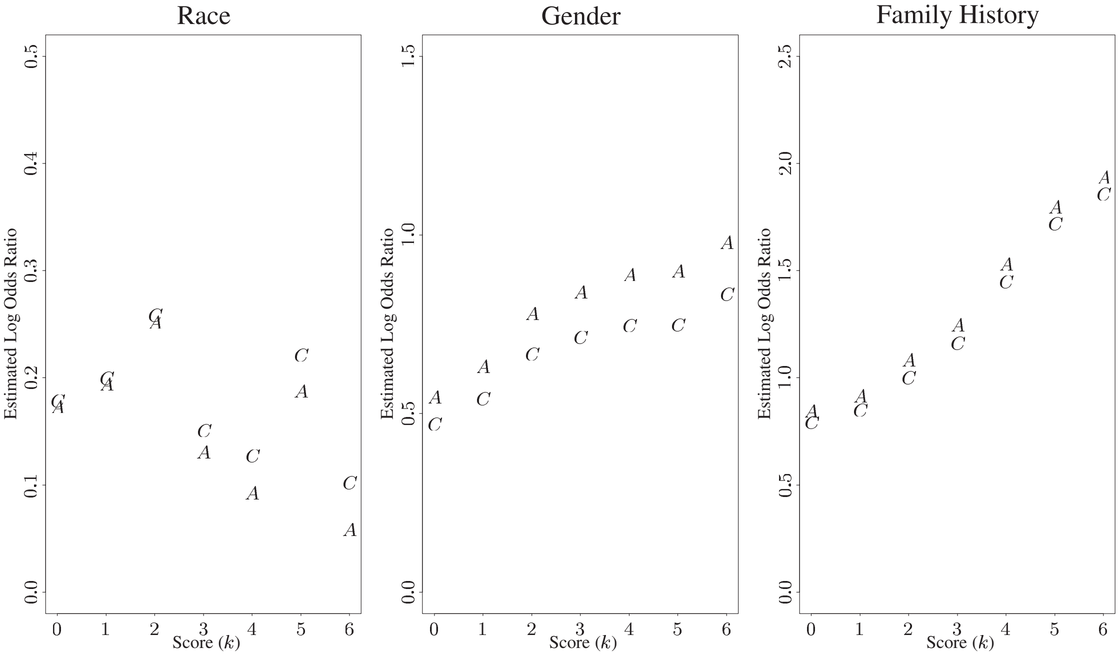

Fig. 1. Crude (C) and Adjusted (A) log odds ratios for varying thresholds of alcohol dependence diagnosis (lifetime drinker cohort). Risk factors include race, gender, and family history.

One way of comparing the crude (adjusted) conditional distributions is by computing the crude (adjusted) odds ratios for all possible thresholds for alcohol dependence diagnosis. For each threshold k, a subject is considered as alcohol dependent if he/she reports more than k of the dependence criteria ($k = 2$ corresponds to the current diagnostic standard). Figure 1 presents the log of these odds ratios as a function of k, the score assigned to level k. The C and A symbols in these figures denote the crude and adjusted estimates, respectively. Note that the vertical scales of these figures differ by risk factor. The proportional odds model for ordinal data (McCullagh, 1980) smooths out these log odds ratios by assuming that they are a constant function of k. By inspection of Figure 1, we see that this model fails, in general, to capture the nonlinear nature of the relationship between the outcome and the risk factors. A more natural, yet parsimonious, model would smooth out these log odds ratios by assuming that they are quadratic in k. The estimated shape of this quadratic function will provide more information about the risk factor's potential role in the etiologic process than the analyses based on the diagnostic threshold. This idea is the motivation for the models discussed in this paper.

The paper is organized as follows. We discuss the single risk factor quadratic cumulative odds regression model in Section 2 and its multiple risk factor extension in Section 4. Section 3 is devoted to issues of estimation in the single risk factor model, which directly carries over to the multiple risk factor model. Model fits to the alcohol dependence dataset and their interpretation are presented in Section 5. The final section is a summary and discussion.

2. SINGLE RISK FACTOR MODELS

Let Y denote a $(K + 1)$-level, scored ordinal response. Without loss of generality, we assume that Y takes on values in the range $0, \dots, K$ and that disease severity increases with Y. Let s_k be the score associated with level k ($s_0 <, \dots, < s_K$). The scores serve as a distance metric between categories. In the analysis of the alcohol dependence data, we let $s_k = k$, denoting equal distances between adjacent levels of the outcome. Let X be a risk factor.

To motivate our models for the conditional law of Y given X, note that the random variable Y is identically equivalent to the K-dimensional random vector $(Y^{(0)}, Y^{(1)}, \dots, Y^{(K-1)})$, where $Y^{(k)} = I(Y > k)$. This is because $Y = k$ if and only if $Y^{(0)} = \cdots = Y^{(k-1)} = 1$ and $Y^{(k)} = \cdots = Y^{(K-1)} = 0$. This implies that a model for the law of Y given X can be fully specified by simply modeling the marginal laws of $Y^{(k)}$ given X. One natural modeling strategy is to assume that each of the K marginal laws follow a logistic model of the form

$$\text{logit } P[Y^{(k)} = 1|X] = \alpha_k + \beta_k X. \tag{1}$$

Here, $\exp(\alpha_k)$ is the odds (baseline) that Y is greater than k for subjects with $X = 0$ and $\exp(\beta_k)$ represents the multiplicative factor by which the odds that Y is greater than k changes as X is increased by one unit. Since we allow these parameters to vary with k, we refer to this model as the single risk factor saturated cumulative odds model. The model is saturated with respect to the coefficients. It is also saturated with respect to the model when X is a binary indicator. This model has $2K$ parameters; however, the parameter space is constrained. To preserve the ordering of the conditional probabilities ($P[Y^{(0)} = 1|X] \geqslant P[Y^{(1)} = 1|X] \geqslant \cdots \geqslant P[Y^{(K-1)} = 1|X]$), it is required that

$$\alpha_k - \alpha_{k+1} + \min_{x \in \text{support}(X)} \{(\beta_k - \beta_{k+1})x\} \geqslant 0 \tag{2}$$

for $k = 0, \dots, K - 2$.

Estimates of β_k are the points labeled C in Figures 1 and 2. As we saw, these points tend to form a nonlinear function of s_k. We argued that a parsimonious, yet flexible, way of capturing and smoothing these patterns is to further parametrize β_k via a quadratic model, i.e. $\beta_k = \gamma_0 + \gamma_1 s_k + \gamma_2 s_k^2$. We refer to this as the single risk factor quadratic cumulative odds model. If $\gamma_2 < 0$ or $\gamma_2 > 0$, then the association is concave or convex, respectively. In this model, restriction (2) reduces to

$$\alpha_k - \alpha_{k+1} + \min_{x \in \text{support}(X)} \{[\gamma_1(s_k - s_{k+1}) + \gamma_2(s_k^2 - s_{k+1}^2)]x\} \geqslant 0 \tag{3}$$

for $k = 0, \dots, K - 2$. Note that our quadratic model has $K + 3$ parameters.

When $\gamma_1 = \gamma_2 = 0$, the quadratic cumulative odds model is equivalent to the well known proportional odds model. In contrast to the proportional odds model, our model is not closed under aggregation of adjacent response levels. That is, the interpretation and estimates of the parameters, $\gamma_0, \gamma_1, \gamma_2$, changes if adjacent response levels are collapsed and assigned a new intermediate numeric scale. This is because our model utilizes the ordinal score, whereas the proportional odds model does not.

We assume that we observe n independent and identically distributed copies $\{(Y_i, X_i) : i = 1, \dots, n\}$ of (Y, X). Formally, our model assumes that for $k = 0, \dots, K - 1$,

$$\text{logit}\, P[Y > k|X] = \alpha_k + \gamma_0 X + \gamma_1 s_k X + \gamma_2 s_k^2 X. \tag{4}$$

Let $\boldsymbol{\beta} = (\boldsymbol{\alpha}', \boldsymbol{\gamma}')'$ where $\boldsymbol{\alpha} = (\alpha_0, \dots, \alpha_{K-1})'$ and $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \gamma_2)'$. Let $\boldsymbol{\Omega}$ denote the set of $\boldsymbol{\beta}$ satisfying restriction (3). Let $\boldsymbol{\beta}_0 = (\boldsymbol{\alpha}_0', \boldsymbol{\gamma}_0')'$, where $\boldsymbol{\alpha}_0 = (\alpha_{0,0}, \dots, \alpha_{K-1,0})$ and $\boldsymbol{\gamma}_0 = (\gamma_{0,0}, \gamma_{1,0}, \gamma_{2,0})$ denote the true values of $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$, respectively. We assume that $\boldsymbol{\beta}_0$ lies in the interior of a compact subset of $\boldsymbol{\Omega}$. Let $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\alpha}}', \widehat{\boldsymbol{\gamma}}')'$ denote an estimator of $\boldsymbol{\beta}_0$. Finally, let

$$Q(k; \boldsymbol{\gamma}) = \gamma_0 + \gamma_1 s_k + \gamma_2 s_k^2$$

be a characterization of the relationship between Y and X. We refer to $\hat{Q}(\cdot) = Q(\cdot; \hat{\boldsymbol{\gamma}})$ and $Q_0(\cdot) = Q(\cdot; \boldsymbol{\gamma}_0)$ as the estimated and true crude association function for X, respectively.

3. Estimation

To estimate $\boldsymbol{\beta}_0$ in model (4), we maximize the log-likelihood over $\boldsymbol{\Omega}$. The log-likelihood has the following multinomial form:

$$\sum_{i=1}^{n} \sum_{k=0}^{K} I(Y_i = k) \log p_k(X_i; \boldsymbol{\beta})$$

where

$$p_k(X; \boldsymbol{\beta}) = P[Y = k|X; \boldsymbol{\beta}] = \begin{cases} 1 - \pi_0(X; \boldsymbol{\beta}) & k = 0 \\ \pi_{k-1}(X; \boldsymbol{\beta}) - \pi_k(X; \boldsymbol{\beta}) & k = 1, \dots, K - 1 \\ \pi_{K-1}(X; \boldsymbol{\beta}) & k = K \end{cases}$$

and

$$\pi_k(X; \boldsymbol{\beta}) = \frac{\exp(\alpha_k + \gamma_0 X + \gamma_1 s_k X + \gamma_2 s_k^2 X)}{1 + \exp(\alpha_k + \gamma_0 X + \gamma_1 s_k X + \gamma_2 s_k^2 X)}.$$

To carry out the maximization subject to constraint (3), we follow a two-step approach. First, we perform unconstrained maximization via a Newton–Raphson (NR) algorithm and check to see if the solution can be found and if so whether it is feasible. If the solution is infeasible or unattainable, we then move onto the second stage where we perform constrained maximization via a sequential quadratic programming (SPQ) algorithm (Fletcher, 1980; Gill *et al.*, 1981). We adopt this two-step strategy because the NR algorithm is much faster than the SPQ algorithm. In the analyses performed in Section 5, we utilized MATLAB's NR function *fsolve* and the SPQ function *fmincon*.

The NR algorithm requires the first derivative of the log-likelihood with respect to $\boldsymbol{\beta}$. The first derivative can be expressed as

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} B(X_i; \boldsymbol{\beta}) A(Y_i, X_i; \boldsymbol{\beta})$$

where

$$B(X;\beta) = \left[\frac{\partial p_0(X;\beta)}{\partial\beta}, \frac{\partial p_1(X;\beta)}{\partial\beta}, \ldots, \frac{\partial p_K(X;\beta)}{\partial\beta}\right]$$

$$\frac{\partial p_0(X;\beta)}{\partial\beta} = \begin{bmatrix} -\pi_0(X;\beta)(1-\pi_0(X;\beta)) \\ z_{K-1} \\ -\pi_0(X;\beta)(1-\pi_0(X;\beta))X \\ -\pi_0(X;\beta)(1-\pi_0(X;\beta))s_0X \\ -\pi_0(X;\beta)(1-\pi_0(X;\beta))s_0^2X \end{bmatrix}$$

$$\frac{\partial p_k(X;\beta)}{\partial\beta} = \begin{bmatrix} z_{k-1} \\ \pi_{k-1}(X;\beta)(1-\pi_{k-1}(X;\beta)) \\ -\pi_k(X;\beta)(1-\pi_k(X;\beta)) \\ z_{K-k-1} \\ \pi_{k-1}(X;\beta)(1-\pi_{k-1}(X;\beta))X - \pi_k(X;\beta)(1-\pi_k(X;\beta))X \\ \pi_{k-1}(X;\beta)(1-\pi_{k-1}(X;\beta))s_{k-1}X - \pi_k(X;\beta)(1-\pi_k(X;\beta))s_kX \\ \pi_{k-1}(X;\beta)(1-\pi_{k-1}(X;\beta))s_{k-1}^2X - \pi_k(X;\beta)(1-\pi_k(X;\beta))s_k^2X \end{bmatrix}$$

for $k = 1, \ldots, K-1$, and

$$\frac{\partial p_K(X;\beta)}{\partial\beta} = \begin{bmatrix} z_{K-1} \\ \pi_{K-1}(X;\beta)(1-\pi_{K-1}(X;\beta)) \\ \pi_{K-1}(X;\beta)(1-\pi_{K-1}(X;\beta))X \\ \pi_{K-1}(X;\beta)(1-\pi_{K-1}(X;\beta))s_{K-1}X \\ \pi_{K-1}(X;\beta)(1-\pi_{K-1}(X;\beta))s_{K-1}^2X \end{bmatrix}.$$

$A(Y, X;\beta)$ is the $(K+1)$-dimensional vector whose kth element is equal $I(Y = k-1)/p_{k-1}(X;\beta)$, and z_k is a $k \times 1$ vector of zeros. MATLAB's NR algorithm does not require analytic second derivatives. Instead, it uses finite difference numerical derivatives.

Under mild regularity conditions and the assumption that β_0 lies in the interior of a compact subset of $\mathbf{\Omega}$, we know that asymptotically the maximum likelihood estimator, $\hat{\beta}$, will be in the interior of $\mathbf{\Omega}$. Thus, we can apply standard asymptotic theory and prove that $\hat{\beta}$ is consistent and asymptotically normal (CAN). The asymptotic variance of $\hat{\beta}$ is equal to the inverse of the Fisher information matrix, which can be estimated by inverting the observed information matrix. By an application of the multivariate delta method, we can show that $(\sqrt{n}(\hat{Q}(0) - Q_0(0)), \ldots, \sqrt{n}(\hat{Q}(K-1) - Q_0(K-1)))$ converges in distribution to $\mathbf{Z} = (Z_0, \ldots, Z_{K-1})$, where $\mathbf{Z}$ is a K-dimensional multivariate normal random vector with mean zero. Then the continuous mapping theorem tells us that $\max_{k=0,\ldots,K-1} |\sqrt{n}(\hat{Q}(k) - Q_0(k))|$ converges in distribution to $\max_{k=0,\ldots,K-1} |Z_k|$. These latter facts are useful for constructing a confidence band for $Q_0(\cdot)$.

The above normality results are asymptotic. In finite samples with a constrained parameter space, the distribution of the standardized parameter estimates may be quite skewed. One way around this difficulty is to use the parametric or nonparametric bootstrap (Efron and Tibshirani, 1993). In the parametric bootstrap, we first construct B datasets, each with n subjects. Then, for each subject, the risk factor is drawn from the empirical distribution of X, and Y is drawn from model (4) with β replaced by $\hat{\beta}$. For each dataset, indexed by $b = 1, \ldots, B$, we estimate β under the quadratic cumulative odds model using constrained maximum likelihood and refer to this estimate as $\hat{\beta}^{(b)}$. From $\hat{\beta}^{(b)}$, we can compute standard errors and confidence intervals for the components of $\hat{\beta}$ and β_0, respectively. The bootstrapped standard error of the jth component of $\hat{\beta}$ is the standard deviation of the corresponding component of the bootstrapped estimates of $\boldsymbol{\beta}$. Similarly, a $(1-\alpha)$ bootstrapped confidence interval for the jth component of $\boldsymbol{\beta}_0$ can be found by selecting the $\alpha/2$ and $(1-\alpha/2)$ percentiles of the histogram of these bootstrapped estimates. In the nonparametric bootstrap, each of the B datasets are formed by drawing a random sample of n subjects with replacement from the observed data.

If we knew the value of c_α such that $P[\max_{k=0,\dots,K-1}|Z_k| \leqslant c_\alpha] = 1-\alpha$, then a $(1-\alpha)$ confidence band for $Q_0(\cdot)$ would take the form $\hat{Q}(\cdot) \pm c_\alpha/\sqrt{n}$. While it is possible to find c_α as the solution to a complicated equation, it is simpler to approximate it using the above bootstrapped samples. For each sample, use $\hat{\boldsymbol{\beta}}^{(b)}$ to estimate the quadratic association function and denote it by $\hat{Q}^{(b)}(\cdot)$. Then, compute $\max_{k=0,\dots,K-1}|\sqrt{n}(\hat{Q}^{(b)}(k) - \hat{Q}(k))|$. Finally, plot a histogram of the resulting maxima and select c_α as the $(1-\alpha)$ percentile of the histogram.

4. Multiple risk factor models

The extension to multiple risk factors is straightforward. Suppose that there are P risk factors, $\mathbf{X} = (X_1,\dots,X_P)$. To model the law of Y given $\mathbf{X}$, it is sufficient to model the marginal laws of $Y^{(k)}$ given $\mathbf{X}$. As before, a natural modeling strategy is to assume that the K marginal laws follow logistic models of the form

$$\operatorname{logit} P[Y^{(k)} = 1|\mathbf{X}] = \alpha_k + \beta_k^{(1)} X_1 + \beta_k^{(2)} X_2 + \cdots + \beta_k^{(P)} X_P \quad k = 0,\dots,K-1.$$

Here, $\exp(\alpha_k)$ is the odds (baseline) that Y is greater than k for subjects with $X_1 = \cdots = X_P = 0$ and $\exp(\beta_k^{(p)})$ represents the multiplicative factor by which the odds that Y is greater than k changes as X_p is increased by one unit and the other risk factors are held fixed. Since we allow these parameters to vary with k, we refer to this model as the multiple risk factor saturated cumulative odds model. As before, this model is saturated with respect to the coefficients. This model has $(P+1)K$ parameters. The restrictions in this model are

$$\alpha_k - \alpha_{k+1} + \sum_{p=1}^{P} \min_{x_p \in \text{support}(X_p)} \{(\beta_k^{(p)} - \beta_{k+1}^{(p)})x_p\} \geqslant 0 \tag{5}$$

for $k = 0,\dots,K-1$.

Estimates of $\beta_k^{(p)}$ are labeled as the points A in Figures 1 and 2. To capture the nonlinearity of these estimates as a function of k, a parsimonious and flexible way of capturing and smoothing these patterns is to assume that $\beta_k^{(p)} = \gamma_0^{(p)} + \gamma_1^{(p)} s_k + \gamma_2^{(p)} s_k^2$ for $p = 1,\dots,P$. This defines the multiple risk factor quadratic cumulative odds model, with $K + 3P$ parameters. In this model, restriction (5) reduces to

$$\alpha_k - \alpha_{k+1} + \sum_{p=1}^{(p)} \min_{x_p \in \text{support}(X_p)} \{[\gamma_1^{(p)}(s_k - s_{k+1}) + \gamma_2^{(p)}(s_k^2 - s_{k+1}^2)]x_p\} \geqslant 0 \tag{6}$$

for $k = 0,\dots,K-1$. When $\gamma_1^p = \gamma_2^p = 0$ for all p, then the model reduces to the proportional odds model.

We assume that we observe n independent and identically distributed copies $\{(Y_i,\mathbf{X}_i) : i = 1,\dots,n\}$ of $(Y,\mathbf{X})$. Formally, our model assumes that for $k = 0,\dots,K-1$,

$$\operatorname{logit} P[Y > k|X] = \alpha_k + \boldsymbol{\gamma}' \cdot [\mathbf{X}', s_k\mathbf{X}', s_k^2\mathbf{X}']' \tag{7}$$

where $\boldsymbol{\gamma}$ is now redefined as the vector $(\gamma_0^{(1)},\dots,\gamma_0^{(P)},\gamma_1^{(1)},\dots,\gamma_1^{(P)},\gamma_2^{(1)},\dots,\gamma_2^{(P)})'$ and $\boldsymbol{\gamma}_0 = (\gamma_{0,0}^{(1)},\dots,\gamma_{0,0}^{(P)},\gamma_{1,0}^{(1)},\dots,\gamma_{1,0}^{(P)},\gamma_{2,0}^{(1)},\dots,\gamma_{2,0}^{(P)})'$ is the true value of $\boldsymbol{\gamma}$. Now, $\boldsymbol{\Omega}$ is defined as the set of

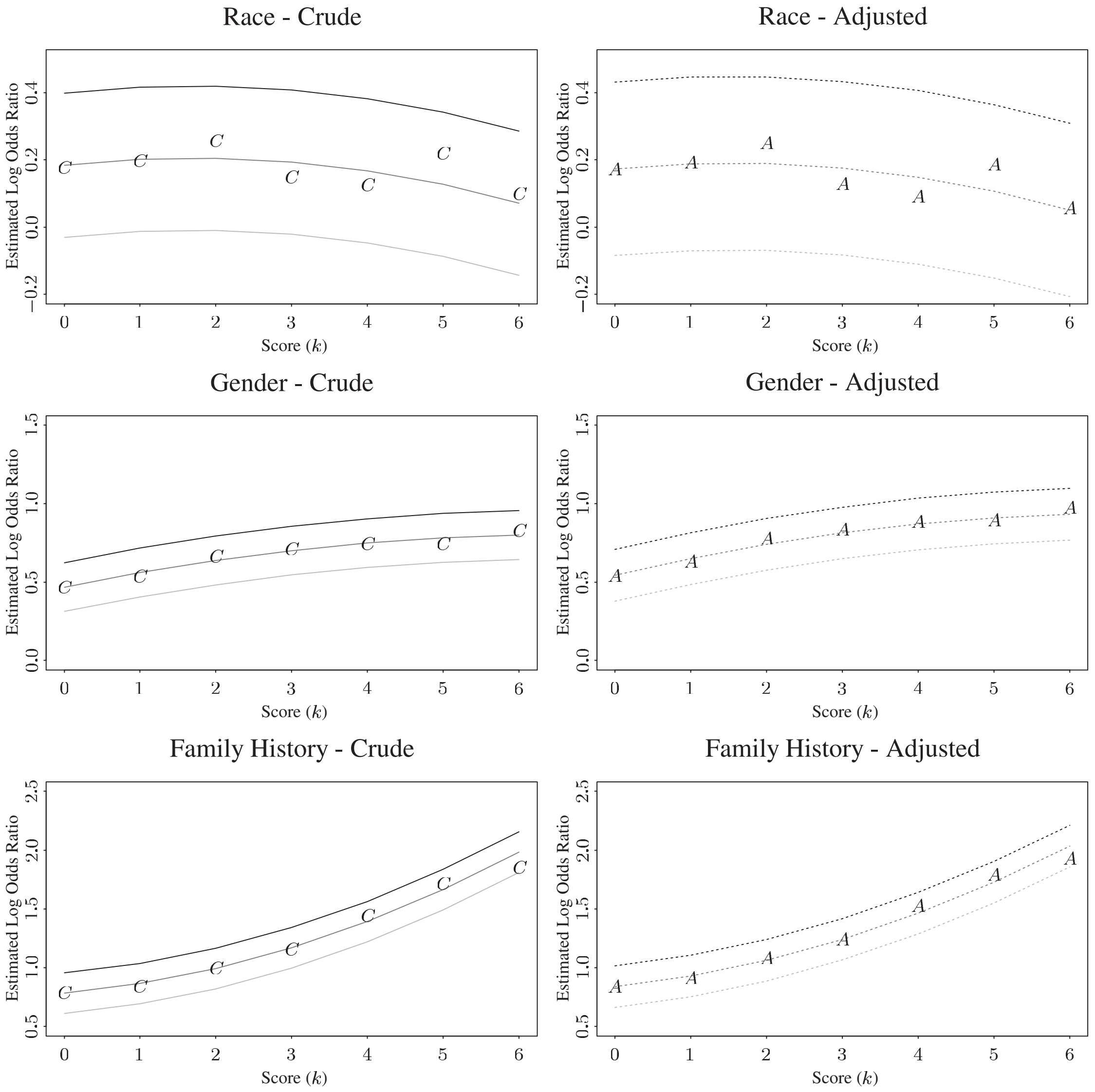

Fig. 2. Crude and adjusted association functions (with confidence bands). Crude (C) and Adjusted (A) log odds ratios for varying thresholds of disease diagnosis.

$\boldsymbol{\beta}$ satisfying restriction (6). We again assume that $\boldsymbol{\beta}_0$ lies in the interior of a compact subset of $\boldsymbol{\Omega}$. Let $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\alpha}}', \widehat{\boldsymbol{\gamma}}')'$ denote an estimator of $\boldsymbol{\beta}_0$. Finally, let

$$Q(p)(k; \gamma) = \gamma_0^{(p)} + \gamma_1^{(p)} s_k + \gamma_2^{(p)} s_k^2$$

be a characterization of the adjusted relationship between Y and X_p. We refer to $\hat{Q}^{(p)}(\cdot) = Q^{(p)}(\cdot; \hat{\gamma})$ and $Q_0^{(p)}(\cdot) = Q^{(p)}(\cdot; \gamma_0)$ as the estimated and true adjusted association function for X_p, respectively.

Table 3. *Estimates and confidence for the γ coefficients from the single risk and multiple risk factor quadratic cumulative odds models (intervals)*

		Single risk factor		Multiple risk factor	
Risk factor	Parameter	Estimate	Confidence interval	Estimate	Confidence interval
Race (white)	γ_0	0.1842	[0.1104, 0.2538]	0.1738	[0.1012, 0.2546]
	γ_1	0.0251	[−0.0315, 0.0814]	0.0216	[−0.0333, 0.0798]
	γ_2	−0.0073	[−0.0178, 0.0043]	−0.0070	[−0.0186, 0.0043]
Gender (male)	γ_0	0.4677	[0.4159, 0.5203]	0.5434	[0.4935, 0.5918]
	γ_1	0.0998	[0.0615, 0.1414]	0.1151	[0.0726, 0.1580]
	γ_2	−0.0074	[−0.0162, 0.0006]	−0.0084	[−0.0170, 0.0001]
Family history	γ_0	0.7857	[0.7269, 0.8394]	0.8391	[0.7815, 0.8955]
	γ_1	0.0555	[0.0120, 0.1003]	0.0691	[0.0258, 0.1160]
	γ_2	0.0240	[0.0149, 0.0340]	0.0217	[0.0126, 0.0313]

Parameter estimation in model (7) proceeds in the same fashion as presented in Section 4. The only differences are that $\pi_k(X; \boldsymbol{\beta})$ is replaced by

$$\pi_k(\mathbf{X}; \boldsymbol{\beta}) = \frac{\exp(\alpha_k + \boldsymbol{\gamma}' \cdot [\mathbf{X}', s_k\mathbf{X}', s_k^2\mathbf{X}']')}{1 + \exp(\alpha_k + \boldsymbol{\gamma}' \cdot [\mathbf{X}', s_k\mathbf{X}', s_k^2\mathbf{X}']')},$$

X is replaced by $\mathbf{X}$ and constraint (3) is replaced by constraint (6). Under mild regularity conditions, we can show that $\hat{\boldsymbol{\beta}}$, our estimator for $\boldsymbol{\beta}_0$, is CAN. Finally, we can use the parametric or nonparametric bootstrap to generate confidence bands for $Q_0^{(p)}(\cdot)$, standard errors for the components for $\hat{\boldsymbol{\beta}}$, and confidence intervals for the components of $\boldsymbol{\beta}_0$.

5. Application to alcohol dependence

Table 3 presents the parameter estimates (with 95% confidence intervals) obtained from fitting the single and multiple risk factor quadratic cumulative odds models. The confidence intervals were formed via parametric bootstrap. Figure 2 presents the crude and adjusted association functions (with 95% confidence bands) for each of the risk factors: race, gender and family history.

The single and multiple risk factor analyses tell the same story. We see that the estimated association between race and number of dependence criteria is concave with a maximum odds ratio (1.22, crude; 1.21, adjusted) occurring at $k = 2$. Our model suggests that the role of race as a risk factor decreases as the severity of alcohol dependence increases. Note that due to sampling variability, we cannot rule out a null effect of race. This can be seen by noting, in Figure 2, that a horizontal line at zero fits within the crude and adjusted confidence bands for race. Our estimated association functions indicate that being white may have a mild impact on low levels of alcohol dependence, but severe dependence is due to other factors. This result may be due to racial differences in levels and type of alcohol consumption, drinking norms, or socioeconomic status. However, there may be a true racial component as Jones-Webb *et al.* (1997) found that, after adjusting for potential confounders, 'increases in alcohol consumption were associated with increased drinking consequences for white men, but increased consumption has little effect for black men.' To address this issue, it may be useful for future research on the relationship between race and alcohol to focus on identifying effects that occur at low levels of dependence, after adjusting for the above factors.

For family history, we see that association is convex and increasing. The minimum and maximum crude (adjusted) odds ratios are 2.19 (2.31) and 7.26 (7.65), respectively. Note that the proportional odds and linear models can be rejected at the 0.05 level as no straight line fits within the crude and adjusted confidence bands. Our estimated association functions suggest that family history is influencing all levels of dependence, with the effect increasing with severity. This result is consistent with the literature as (Hill *et al.*, 1994) report, 'family history predisposes individuals to greater alcohol consumption and more severe consequences from use of alcohol and drugs'.

For gender, we see that the association with number of dependence criteria is concave and monotonically increasing. The minimum and maximum crude (adjusted) odds ratios are 1.60 (1.72) and 2.23 (2.54), respectively. The proportional odds, but not the linear, model can be rejected at the 0.05 level. Like family history, the estimated association function suggests that being male influences all levels of dependence, with the effect increasing with severity, but the effect is less substantial. There are at least two possible explanations for this finding. First, previous research has suggested that, even after accounting for family history, 'women may require a higher susceptibility 'load' before expressing alcohol dependence' (Hill *et al.*, 1994). Second, the gender effect may be due to 'differences in men's and women's exposures to drinking opportunities, in how often, in what quantity and with whom they drink, in their physiological responses to ethanol intake and in the social consequences provoked by their drinking' (Dawson, 1996). More research is required to understand how gender relates to the severity of dependence.

Lastly, we note that, in contrast to the results of Hasin and Paykin (1999), our analysis does not support the three-symptom threshold. However, this does not imply that the diagnostic threshold is not useful, as its purpose is clinical rather than etiologic. From the latter perspective, using the diagnostic threshold involves throwing away information which, as seen above, is potentially useful.

6. DISCUSSION

In this paper, we have developed new regression models for the analysis of scored ordinal data. The novel feature of our models is that they capture and identify nonlinear aspects of the relationship between the scored ordinal outcome and risk factors. These nonlinearities may be useful in generating context-dependent hypotheses about the risk factor's role in the etiologic process. They may also suggest how to design future studies of the risk factor. Fitting these models provides information that is obscured in the standard epidemiologic analysis which uses a diagnostic threshold. In our analysis of cross-sectional data from the 1992 National Longitudinal Alcohol Epidemiologic Survey, we studied the effects of race, gender, and family history on alcohol dependence among lifetime drinkers. We found that race may have an impact on low levels of dependence, while family history and gender have an effect at all levels, with the effect increasing with severity. While the results for family history are well established, the conclusions about race and gender require further study.

In our analysis, we treated all symptom groups as exchangeable. Even more clues could be discovered about the role of a risk factor, by evaluating whether it tends to cluster with a consistent set of dependence criteria. This would provide even more precise hypotheses than the ones generated with our approach. In addition, we have assumed that self-reports of alcohol dependence criteria are a perfect reflection of disease state. However, there is some literature to suggest that there may be reporting differentials between subgroups: for example, gender and family history. That is, two subjects who have the same disease burden but differ with respect to family history (gender) may report a different number of dependence criteria. If this reporting differential is substantial, it could affect the conclusions.

The methodology described here is not limited to a quadratic form for the association function. Depending on the magnitude of K, one can fit a cubic or even a spline model for the sequence of log odds ratios. It is a simple matter to modify the parameter constraints and likelihood score equations. Of course, higher-order models may be more difficult to interpret.

We suggest that the quadratic cumulative odds regression models may be useful in medicine and epidemiology, particularly in situations where the definition of disease or disorder depends on a diagnostic threshold which is arbitrary or controversial. Examples include hypertension, type II diabetes, low birth weight, and depression.

ACKNOWLEDGEMENTS

The authors would like to thank Dr Deborah Hasin and Dr Jim Anthony for feedback on the results of the alcohol dependence analysis, and two anonymous referees for constructive editorial and statistical criticism. This work was supported in part by National Institute of Health grants GM49909, MH47447, and MH56639.

REFERENCES

AMERICAN PSYCHIATRIC ASSOCIATIO (1994). *Diagnostic and Statistical Manual of Mental Disorders*, Vol. 4. American Psychiatric Association.

DAWSON, D. (1996). Gender differences in the risk of alcohol dependence: United States, 1992. *Addiction* **91**, 1831–1842.

EFRON, B. AND TIBSHIRANI, R. J. (1993). *An Introduction to the Bootstrap*. Chapman and Hall.

FLETCHER, R. (1980). *Practical Methods of Optimization*. New York: Wiley.

GRANT, B. F., HARTFORD, T., DAWSON, D. A., CHOU, P. S., DUFOUR, M. AND PICKERING, R. (1996). Prevelence of DSM-IV alcohol abuse and dependence: United States, 1992. *Alcohol Health and Research World* **18**, 243–248.

GRANT, B. F., HARTFORD, T., DAWSON, D. A., CHOU, P. S. AND PICKERING, R. (1994). The Alcohol Disorder and Associated Disabilities Schedule (AUDADIS): reliability of alcohol and drug modules in a general population sample. *Drug and Alcohol Dependence* **39**, 37–44.

GILL, P. E., MURRAY, W. AND WRIGHT, M. H. (1981). *Practical Optimization*. London: Academic.

HASIN, D. AND PAYKIN, A. (1999). Dependence symptoms but no diagnosis: diagnostic ‘orphans’ in a 1992 national sample. *Drug and Alcohol Dependence* **53**, 215–222.

HILL, M. H., BLOW, F. C., YOUNG, J. P. AND SINGER, K. M. (1994). Family history of alcoholism and childhood adversity: joint effects on alcohol consumption and dependence. *Alcoholism: Clinical and Experimental Research* **18**, 1083–1090.

JONES-WEBB, R., HSIAO, C-Y., HANNAN, P. AND CAETANO, R. (1997). Predictors of increases in alcohol-related problems among black and white adults: results from the 1984 and 1992 national alcohol surveys. *American Journal of Drug and Alcohol Abuse* **23**, 281–299.

MCCULLAGH, P. (1980). Regression model for ordinal data. *Journal of the Royal Statistical Society,* Series B **42**, 109–142.

[*Received 18 December, 2000; revised 15 February, 2001; accepted for publication 9 April, 2001*]