Logistic Regression of Family Data From Retrospective Study Designs

Alice S. Whittemore and Jerry Halpern

Stanford University School of Medicine, Stanford, California

We wish to study the effects of genetic and environmental factors on disease risk, using data from families ascertained because they contain multiple cases of the disease. To do so, we must account for the way participants were ascertained, and for within-family correlations in both disease occurrences and covariates. We model the joint probability distribution of the covariates of ascertained family members, given family disease occurrence and pedigree structure. We describe two such covariate models: the random effects model and the marginal model. Both models assume a logistic form for the distribution of one person's covariates that involves a vector β of regression parameters. The components of β in the two models have different interpretations, and they differ in magnitude when the covariates are correlated within families. We describe ascertainment assumptions needed to estimate consistently the parameters β_{RE} in the random effects model and the parameters β_M in the marginal model. Under the ascertainment assumptions for the random effects model, we show that conditional logistic regression (CLR) of matched family data gives a consistent estimate $\hat{\beta}_{RE}$ for β_{RE} and a consistent estimate for the covariance matrix of $\hat{\beta}_{RE}$. Under the ascertainment assumptions for the marginal model, we show that unconditional logistic regression (ULR) gives a consistent estimate for β_M , and we give a consistent estimator for its covariance matrix. The random effects/CLR approach is simple to use and to interpret, but it can use data only from families containing both affected and unaffected members. The marginal/ULR approach uses data from all individuals, but its variance estimates require special computations. A C program to compute these variance estimates is available at <http://www.stanford.edu/dept/HRP/epidemiology>. We illustrate these pros and cons by application to data on the effects of parity on ovarian cancer risk in mother/daughter pairs, and use simulations to study the performance of the estimates. *Genet Epidemiol* 25:177–189, 2003. © 2003 Wiley-Liss, Inc.

Key words: ascertainment; estimating equation; family-based association; odds-ratio; retrospective data

Grant sponsor: NIH; Grant numbers: CA 94069, CA 69417.

*Correspondence to: Alice S. Whittemore, Stanford University School of Medicine, Department of Health Research and Policy, Redwood Building, Room T204, Stanford, CA 94305-5405. E-mail: alice@stanford.edu

Received 28 March 2003; Accepted 29 May 2003

Published online in Wiley InterScience (www.interscience.wiley.com)

DOI: 10.1002/gepi.10267

INTRODUCTION

We wish to evaluate associations between disease risk and one or more genetic or environmental covariates, using data from families ascertained because of the disease statuses of some of their members. We seek consistent estimates of an r -component column vector β of regression parameters that reflect the strength of association between disease status y and an r -component row vector z of covariates. Here y is a disease status indicator for an individual, with $y = 1$ denoting disease presence and $y = 0$ denoting its absence. We assume that the population consists of nonoverlapping families each containing one or more individuals. Then a family of n members is characterized by a vector $\mathbf{y} = (y_1, \dots, y_n)$ of length n , a vector $\mathbf{z} = (z_1, \dots, z_n)$ of length nr , and a

pedigree structure S . Both disease occurrences and covariates may be correlated within families.

To analyze such data, we shall model the family covariate distribution $pr(\mathbf{z}|\mathbf{y}, S)$ rather than the family disease distribution $pr(\mathbf{y}|\mathbf{z}, S)$. The reason for doing so is that each family is selected because some members had the disease, so that $\mathbf{y}$ is determined by the study design. The families were not selected because they had given covariates $\mathbf{z}$. Of course the two distributions are related. In particular, when $pr(\mathbf{y}|\mathbf{z}, S)$ involves a vector β of regression parameters, then $pr(\mathbf{z}|\mathbf{y}, S)$ also involves β . Consequently β can be estimated by utilizing $pr(\mathbf{z}|\mathbf{y}, S)$.

We shall present two different models for $pr(\mathbf{z}|\mathbf{y}, S)$, which are analogues of two well-known models for $pr(\mathbf{y}|\mathbf{z}, S)$. One is called the random effects model and the other the marginal model

[see Diggle et al., 1994 for a review of these models for $pr(y|z, S)$]. We shall describe the family ascertainment assumptions needed for consistent estimates of the regression parameters in these models. We will show that conditional logistic regression (CLR) [Breslow and Day, 1980] can be used to estimate consistently the regression parameter β_{RE} and its asymptotic covariance matrix in the random effects model. We also shall present estimators for the corresponding quantities in the marginal model, and we shall prove that these estimators are consistent under the ascertainment assumptions. The estimate for β_M can be obtained using unconditional logistic regression (ULR). We shall apply the marginal covariate model to data on ovarian cancer in mothers and daughters ascertained because of the daughters' disease status, and we shall present simulation results evaluating the performance of the estimates.

COVARIATE MODELS

In formulating covariate models, we must accommodate possible dependence of family covariates z on family pedigree structure S . For example, in a study of the effects of parity on ovarian cancer risk in mother-daughter pairs, the distribution of parity in daughters differs from that of their mothers, since by definition all mothers are parous. We also must accommodate possible dependence of an individual's covariates on the disease status of his relatives. For example, in a study of lung cancer and smoking in families containing one of more cases of lung cancer, the smoking histories of cases with a strong family history may differ from those of cases with no family history. Accordingly, we shall classify each family member into one of H strata, labelled $h = 1, \dots, H$, according to his position in the pedigree and the disease statuses of his relatives. Individuals in stratum h are assigned the $(H - 1)$ -dimensional row vector $x = e_h$, where e_h contains 1 at position h and zero's elsewhere, $h = 1, \dots, H - 1$. Individuals in stratum H are assigned the $(H - 1)$ -dimensional zero vector $x = e_H = 0$. For an n -member family, we let $\mathbf{x} = (x_1, \dots, x_n)$ denote the $n(H - 1)$ -dimensional row vector of stratum membership indicators. We assume that the family covariate distribution depends on the pedigree structure only through $\mathbf{x}$:

$$pr(\mathbf{z}|\mathbf{y}, \mathbf{x}, S) = pr(\mathbf{z}|\mathbf{y}, \mathbf{x}).$$

RANDOM EFFECTS COVARIATE MODEL

In this model, the covariates are distributed independently among family members, conditionally on $\mathbf{y}$, on $\mathbf{x}$, and on the value of a family-specific random effect v . There is a distribution $f(v)$ of random effects across families. The explicit form of $pr(\mathbf{z}|\mathbf{y}, \mathbf{x})$ is

$$pr(\mathbf{z}|\mathbf{y}, \mathbf{x}) = \int f(v) \left[\prod_{\ell=1}^n pr(z_{\ell}|y_{\ell}, x_{\ell}; v) \right] dv. \quad (1)$$

Equation (1) states that family members' covariates are conditionally independent, given the family-specific random effect v and their disease statuses y and their stratum membership indicators x . This assumption would be violated if members' covariates were their genotypes for a polymorphism of interest, and if pairs of members with given y and h had different kinship coefficients. The violation would occur because the genotypes of closely related members would be more correlated than those of more distantly related members.

The individual covariate distributions $pr(z|y, x; v)$ are assumed to depend on y and x through a logistic form. Specifically, we assume that

$$\begin{aligned} \varphi(z, e_h, v) &\equiv \frac{pr(z|y = 1, x = e_h; v)}{pr(z|y = 0, x = e_h; v)} \\ &= e^{\zeta(v) + \gamma_h + z\beta_{RE}}, \quad \gamma_H = 0. \end{aligned} \quad (2)$$

Assumption (2) is essential for odds-ratio estimation using retrospective data from unrelated individuals. It is implicit in any retrospective logistic regression [Anderson, 1972; Prentice and Pyke, 1979].

The random effects model (1),(2) describes the covariate distributions in diseased and disease-free individuals. But the objective of a study is to determine how disease risk varies with covariates. Fortunately, the model gives the odds-ratios relating disease risk to covariates in terms of the parameter β_{RE} :

$$\begin{aligned} \frac{pr(y = 1|z, x = e_h; v) pr(y = 0|z', x = e_h; v)}{pr(y = 0|z, x = e_h; v) pr(y = 1|z', x = e_h; v)} \\ = e^{(z - z')\beta_{RE}}. \end{aligned} \quad (3)$$

To verify (3), we use the identity

$$pr(y|z, x = e_h; v) = pr(z|y, x = e_h; v) \frac{pr(y|x = e_h; v)}{pr(z|x = e_h; v)} \quad (4)$$

in the left side of (3) to obtain

$$\frac{pr(y = 1|z, x = e_h; v) pr(y = 0|z', x = e_h; v)}{pr(y = 0|z, x = e_h; v) pr(y = 1|z', x = e_h; v)} = \frac{\varphi(z, e_h, v)}{\varphi(z', e_h, v)}, \quad (5)$$

where $\varphi(z, e_h, v)$ is defined by (2). We now substitute the right side of (2) for the φ terms in (5) to obtain (3). Thus $e^{(z-z')\beta_{RE}}$ represents the odds-ratio relating disease risk to covariates z and z' for an individual in stratum h in a given family with random effect v . This odds-ratio, which is independent of h and v , measures association between covariates and disease risk that is conditional on family and stratum membership.

MARGINAL COVARIATE MODEL

This model specifies that the marginal covariate distribution for a subset of $m \leq n$ family members, given the vector $\mathbf{y}$ of family disease status and the family vector $\mathbf{x}$ of stratum memberships, depends on these family attributes only through their own disease status and stratum memberships. In particular, the marginal distribution of z for one individual, given his family attributes $\mathbf{y}, \mathbf{x}$, depends only on his own disease status y and stratum membership x :

$$pr(z|\mathbf{y}, \mathbf{x}) = pr(z|y, x). \quad (6)$$

In addition to assumption (6) for individual family members, the marginal covariate model specifies that the marginal distribution of z_ℓ and $z_{\ell'}$ for two family members ℓ and ℓ' , conditional on $\mathbf{y}, \mathbf{x}$, depends only on y_ℓ, x_ℓ and $y_{\ell'}, x_{\ell'}$:

$$pr(z_\ell, z_{\ell'}|\mathbf{y}, \mathbf{x}) = pr(z_\ell, z_{\ell'}|y_\ell, y_{\ell'}, x_\ell, x_{\ell'}). \quad (7)$$

The marginal distribution (6) of z for an individual is assumed to depend on his disease status and stratum through a logistic form analogous to (2):

$$\begin{aligned} \varphi(z, e_h) &\equiv \frac{pr(z|y = 1, x = e_h)}{pr(z|y = 0, x = e_h)} \\ &= e^{\gamma_H + \gamma_h + z\beta_M}, \quad \gamma_H = 0. \end{aligned} \quad (8)$$

As before, $e^{(z-z')\beta_M}$ represents an odds-ratio for an individual with a given stratum h . Now, however, it represents the odds-ratio relating disease risk to covariates z and z' in a randomly selected individual in stratum h , regardless of the specific family to which he belongs.

PARAMETERS β_{RE} AND β_M IN RANDOM EFFECTS AND MARGINAL MODELS

Now we illustrate the difference in parameter values for the random effects model and the marginal model. Suppose that we have measured continuous covariates (z_1, z_2) for pairs of sibs with disease status (y_1, y_2) . We assume that $pr(z_1, z_2|y_1, y_2)$ is given by the random effects model (1) with $H = 1$, where $pr(z_\ell|y_\ell; v)$ is gaussian with mean $v + y_\ell\Delta$ and variance σ^2 , and $f(v)$ is gaussian with mean 0 and variance τ^2 . This random effects model satisfies (2) with $\beta_{RE} = \Delta/\sigma^2$. These assumptions also imply a bivariate gaussian marginal model (7) for $pr(z_1, z_2|y_1, y_2)$, with mean $(y_1\Delta, y_2\Delta)$ and covariance

$$(\sigma^2 + \tau^2) \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

Here $\rho = \tau^2/(\sigma^2 + \tau^2)$ is the intraclass (i.e., within family) correlation coefficient. This marginal model satisfies (8) with $H = 1$ and with $\beta_M = \Delta/(\sigma^2 + \tau^2)$. Thus $\beta_M = \beta_{RE}(1 - \rho)$. Since $0 \leq \rho \leq 1$, the marginal parameter β_M is less than or equal to the random effects parameter β_{RE} , with equality holding only in two special cases: (1) no covariate effect ($\beta_M = \beta_{RE} = 0$) or (2) covariate independence within families ($\rho = 0$). This difference in parameters underscores the importance of model choice and its implications for the desired inferences.

ASCERTAINMENT ASSUMPTIONS

Suppose we wish to use either the random effects or the marginal covariate model in situations when the covariates $\mathbf{z}$ in families recruited to a study are not a random sample from the distributions $pr(\mathbf{z}|\mathbf{y}, S)$. Using the ascertained data to obtain consistent parameter estimates requires certain assumptions about the types of individuals who were excluded, either by design or because their disease status or covariates were unavailable. We say that an individual's data are *available* if his disease status and covariates can be determined. An individual is *ascertained* if he is sampled and his data are available. Let $a = 1$ if an individual is ascertained and $a = 0$ otherwise, with $\mathbf{a} = (a_1, \dots, a_n)$ denoting the vector of ascertainment indicators for the family. A family is ascertained if at least one of its members is ascertained.

The ascertainment assumptions needed to insure consistent estimates of the regression

parameters in either a random effect or a marginal covariate model are:

- A1. The ascertainment vectors $\mathbf{a}$ for different families are independent;
- A2. The covariate distributions among ascertained members of ascertained families satisfy equation (2) for a random effects covariate model, or equation (8) for a marginal covariate model. Moreover the regression parameters β relating covariates to disease risk in these equations are the same for ascertained and unascertained individuals.

Assumption A2 holds whenever ascertainment is independent of z conditional on y and x , that is $pr(z|x, y, a) = pr(z|x, y)$. However, A2 may hold even when this condition fails, and when the covariates $\mathbf{z}$ in ascertained families are not a random sample of family covariates in the population. For example, the educational levels in families identified by high-risk cancer clinics are likely to differ from those in unascertained families. However, assumption A2 would be violated only if the differences vary with disease status, that is, if ascertainment were an effect-modifier for the association between disease and educational level. As a second example, assumption A2 would be problematic in a case-control study evaluating the effects of tobacco consumption on risk of Parkinson's disease if cigarette-smoking cases were less likely to participate than nonsmoking cases, but control participation were independent of smoking status.

ESTIMATING REGRESSION PARAMETERS IN COVARIATE MODELS

Now we describe methods for estimating β by fitting covariate models for $pr(\mathbf{z}|\mathbf{y}, \mathbf{x})$ to data that satisfy assumptions A1–A2. One option is to completely specify the model and estimate all its parameters by maximizing the full likelihood of the data. However, when several covariates are simultaneously under study, this approach is likely to involve many parameters, to be cumbersome, and to be vulnerable to model misspecification.

These difficulties are not specific to family data, but apply also to data on unrelated individuals. In early applications of logistic regression to data

from individuals with and without heart disease, for example, the covariates were assumed to have multivariate gaussian distributions in each disease category [Truett et al., 1967]. While this approach yields consistent estimates of regression parameters when the assumptions of multivariate normality hold, severe bias can result when they do not [Efron, 1975; Press and Wilson, 1978]. In most practical applications, the covariates are discrete and/or distinctly nonnormal. Thus, as noted by Breslow and Day [1980] for data from unrelated individuals, it is prudent to make as few assumptions about the covariate distributions as possible. For unrelated individuals, this can be accomplished either by (1) eliminating the covariate distributions from the likelihood by using an appropriate conditional likelihood, or (2) by using the full likelihood but allowing the covariate distributions either to be unspecified or to depend on many parameters that are functionally independent of β . If option (2) is chosen, Anderson [1972] and Prentice and Pyke [1979] noted a remarkable simplification: consistent estimates for β and its covariance matrix can be obtained by ordinary unconditional logistic regression of disease status against covariates. Here we show that these two options extend to the retrospective analysis of family data under either the random effects or marginal covariate models.

RANDOM EFFECTS COVARIATE MODEL

The full likelihood option would require specifying a model for the distribution $f(\mathbf{v})$ of the random effects and a model for the individual covariate distributions $pr(\mathbf{z}_\ell|\mathbf{y}_\ell, \mathbf{x}_\ell; \mathbf{v})$. Appropriate conditioning allows us to avoid this extensive modelling and its vulnerability to misspecification. By conditioning on family membership, we avoid having to model $f(\mathbf{v})$. By further conditioning on a family's covariates, strata, and total number of disease occurrences, we avoid having to model the covariate distributions $pr(\mathbf{z}_\ell|\mathbf{y}_\ell, \mathbf{x}_\ell; \mathbf{v})$. A family's contribution to this conditional likelihood is the conditional probability that covariates and strata are linked to disease status as observed, given the set of members' covariates and strata, their ascertainment indicators, and the family disease count d . Letting D denote an arbitrary set of d ascertained family members, and letting D^* denote the set of affected family

members, this probability can be written

$$L_v = \frac{\prod_{\ell \in D^*} pr(z_\ell | y_\ell = 1, x_\ell, a = 1; v) \prod_{m \notin D^*} pr(z_m | y_m = 0, x_m, a = 1; v)}{\sum_D \left\{ \prod_{\ell' \in D} pr(z_{\ell'} | y_{\ell'} = 1, x_{\ell'}, a = 1; v) \prod_{m' \notin D} pr(z_{m'} | y_{m'} = 0, x_{m'}, a = 1; v) \right\}} \quad (9)$$

[Breslow and Day, 1980]. Here $\sum_D$ denotes summation over all sets D containing d family members. We now divide the numerator and denominator of (9) by $\prod_{\ell=1}^n pr(z_\ell | y_\ell = 0, x_\ell; v)$ to obtain

$$L_v = \frac{\prod_{\ell \in D^*} \tilde{\varphi}(z_\ell, x_\ell, v)}{\sum_D \left\{ \prod_{\ell' \in D} \tilde{\varphi}(z_{\ell'}, x_{\ell'}, v) \right\}}, \quad (10)$$

where

$$\tilde{\varphi}(z, x, v) = \frac{pr(z | y = 1, x, a = 1; v)}{pr(z | y = 0, x, a = 1; v)}.$$

By assumption A2 and (2) we may write

$$\tilde{\varphi}(z, x, v) = e^{\tilde{\zeta} + x\tilde{\gamma} + z\beta_M}, \quad (11)$$

where $\tilde{\zeta}$ and $\tilde{\gamma} = (\tilde{\gamma}_1, \dots, \tilde{\gamma}_{H-1})^T$ are parameters relevant to the ascertained population. Next we use (11) in the numerator and denominator of (10) to obtain the family's contribution to the conditional likelihood as

$$L_v = \frac{\exp(x_{D^*}\tilde{\gamma} + z_{D^*}\beta_{RE})}{\sum_D \exp(x_D\tilde{\gamma} + z_D\beta_{RE})}. \quad (12)$$

Here $z_D = \sum_{\ell \in D} z_\ell$ and $x_D = \sum_{\ell \in D} x_\ell$ denote sums of covariates and stratum indicators for the members in D . The conditional likelihood, which is the product of terms (12) for all the families, is independent of the intercept ζ and the family-specific random effect v . The estimate $\hat{\beta}_{RE}$ is obtained by maximizing this likelihood. When ascertainment assumptions A1 and A2 are met, standard likelihood theory [Cox and Hinkley, 1974] can be used to show the consistency and asymptotic normality of $\hat{\beta}_{RE}$, as the number K of families becomes infinite. Estimates for the asymptotic covariance matrix of $\hat{\beta}_{RE}$ can be obtained using standard software.

The CLR likelihood just described is attractive for estimating β_{RE} because it depends only on β_{RE} and γ , and its estimates are consistent under both random effects covariate models and random effects disease models [Kraft and Thomas, 2000], which facilitates parameter interpretation. However a family's contribution (12) degenerates to unity in the absence of at least one pair of family members with discordant disease status, which in some applications can result in considerable information loss.

MARGINAL COVARIATE MODEL

We now consider the problem of estimating the regression parameters β_M in the marginal model (6),(8). For this model, we cannot sidestep the full likelihood pitfalls by conditioning on family membership as we did for the random effects model. This is because the regression parameters β_M are log odds-ratios relating the distributions of disease status and covariates for an individual in the general population, and not these distributions conditional on the family to which he or she belongs. Instead we shall avoid the pitfalls by estimating β_M as the solution to an estimating equation.

To motivate this approach, we review the results of Prentice and Pyke [1979]. We can relate our family-based design to the one considered by Prentice and Pyke by assuming that each family contains a single individual in a single stratum $H = 1$. Then, omitting the stratum indicator, the probability of the observed covariate data for the n_1 diseased cases and n_0 disease-free controls is

$$L = \prod_{i=0}^1 \prod_{j \in S_i} pr(z_j | y_j = i), \quad (13)$$

where S_i is the set of n_i individuals with disease status i , $i = 0, 1$. Our objective is to find distributions $pr(z | y = 1)$ and $pr(z | y = 0)$ to maximize L , subject to the constraint (8), which involves β_M . However we want to avoid specifying the two distributions $pr(z | y = 1)$ and $pr(z | y = 0)$ parametrically, for the reasons discussed at the start of this section. Instead, we let these two distributions be unrestricted, subject only to (8). For this situation, Prentice and Pyke showed that maximizing L is equivalent to maximizing the function

$$F(\theta) = \prod_{i=0}^1 \prod_{j \in S_i} [p(z_j, \theta)]^i [q(z_j, \theta)]^{1-i}, \quad (14)$$

where $\theta = (\alpha, \beta_M^T)^T$ and

$$p(z, \theta) = 1 - q(z, \theta) = \frac{e^{\alpha + z\beta_M}}{1 + e^{\alpha + z\beta_M}}.$$

Prentice and Pyke also derived the asymptotic sampling distribution of the maximum likelihood estimator for β_M .

We shall extend these results to obtain consistent estimates for β_M from our ascertained families. To do so, we introduce the set S_{kih} , which consists of the members of family k having disease status i and stratum membership h . Also we let

$$n_{kih} = |S_{kih}| \quad (15)$$

denote the number of members of family k with disease status i and stratum h , $k = 1, \dots, K$, $i = 0, 1$, $h = 1, \dots, H$. Then we proceed as follows.

We consider the estimate $\hat{\beta}$ for β_M which maximizes the function

$$L = \prod_{h=1}^H \prod_{i=0}^1 \prod_{k=1}^K \prod_{j \in S_{kih}} pr(z_j | y_j = i, e_h, a = 1).$$

L is the product of each individual's covariate probability, conditional on his or her disease status, stratum membership, and ascertainment indicator $a = 1$. This function, which is the multi-strata analogue of (13), would be the actual likelihood function for the data, if family members' covariates were statistically independent. We shall show that it can be used to estimate β_M consistently, even when family members' covariates are correlated.

We now use the result of Prentice and Pyke to maximize instead the product over the H strata of stratum-specific terms (14). This product is

$$F(\theta) = \prod_{h=1}^H \prod_{i=0}^1 \prod_{k=1}^K \prod_{j \in S_{kih}} [p(e_h, z_j, \theta)]^i [q(e_h, z_j, \theta)]^{1-i}, \quad (16)$$

where

$$p(e_h, z, \theta) = 1 - q(e_h, z, \theta) = \frac{e^{\alpha + \delta_h + z\beta}}{1 + e^{\alpha + \delta_h + z\beta}}, \quad (17)$$

and $\theta = (\alpha, \delta^T, \beta_M^T)^T$, with $\delta^T = (\delta_1, \dots, \delta_{H-1})$. The corresponding estimating equation for θ is $U(\theta) = 0$, where

$$\begin{aligned} U(\theta) &= \frac{d}{d\theta} \log F(\theta) \\ &= \sum_{h=1}^H \sum_{i=0}^1 \sum_{k=1}^K \sum_{j \in S_{kih}} [s_i(e_h, z_j, \theta)]^T. \end{aligned} \quad (18)$$

Here $s_i(e_h, z, \theta) = \frac{\partial}{\partial \theta} \ln \{ [p(e_h, z, \theta)]^i [q(e_h, z, \theta)]^{1-i} \}$ is the score for an individual in stratum h with disease status i . The scores s_1 and s_0 are the $(H+r)$ -dimensional row vectors $(1, e_h, z)$ multiplied by the scalars q and $-p$ respectively:

$$\begin{aligned} s_1(e_h, z, \theta) &= q(e_h, z, \theta)(1, e_h, z) \\ s_0(e_h, z, \theta) &= -p(e_h, z, \theta)(1, e_h, z). \end{aligned} \quad (19)$$

The estimating equation $U(\theta) = 0$, where $U(\theta)$ is given by (18), is the likelihood equation we would have obtained had we sampled our study subjects as unrelated individuals from a prospective cohort study, and modelled their disease risks as logistic functions of their covariates. However our estimating equation is not the likelihood equation for the data at hand. Nevertheless, we shall show in the next section that its solution $\hat{\theta} = (\hat{\alpha}, \hat{\delta}^T, \hat{\beta}^T)^T$ provides a consistent estimate $\hat{\beta}$ for the parameter β_M of (8). We also provide a consistent estimate for the asymptotic variance matrix of $\hat{\beta}$. These results extend to family data the results of Anderson [1972] and Prentice and Pyke [1979] for retrospective data from unrelated individuals. The consistency of $\hat{\beta}$ in the presence of correlated covariate data, and the estimate for the covariance matrix of $\hat{\beta}$, are analogous to the results of Liang and Zeger (1986).

ASYMPTOTIC DISTRIBUTION OF $\hat{\beta}$ FOR MARGINAL COVARIATE MODELS

Let $\hat{\theta} = (\hat{\alpha}, \hat{\delta}^T, \hat{\beta}^T)^T$ satisfy $U(\theta) = 0$, where $U(\theta)$ is given by (18), and let $\theta_0 = (\alpha_0, \delta_0^T, \beta_0^T)^T$, where β_0 denotes the true value of β_M , and α_0 and δ_0 denote the values of α, δ induced by the family ascertainment. From (18) we can write the expected value of $U(\theta_0)$ with respect to the covariates, conditional on family members' disease statuses, stratum memberships, and ascertainment indicators, as

$$\begin{aligned} E[U(\theta_0)] &= \sum_{k=1}^K \sum_{i=0}^1 \sum_{h=1}^H n_{kih} \mu_{ih}(\theta) \\ &= \sum_{i=0}^1 \sum_{h=1}^H n_{.ih} \mu_{ih}(\theta). \end{aligned} \quad (20)$$

Here $n_{.ih} = \sum_k n_{kih}$ where n_{kih} is given by (15), and

$$\mu_{ih}(\theta) = \int s_i(e_h, z, \theta) pr(z | y = i, x = e_h, a = 1) dz \quad (21)$$

is the expected score for an ascertained individual in stratum h with disease status i . Now we subtract (20) from (18), and write the result as

$$U(\theta_0) - E[U(\theta_0)] = \sum_{k=1}^K U_k(\theta_0).$$

Here $U_k(\theta_0)$ is defined by

$$U_k(\theta) = \sum_{i=0}^1 \left[\left\{ \sum_{j \in S_{ki}} s_i(x_j, z_j, \theta) \right\} - \sum_{h=1}^H n_{kih} \mu_{ih}(\theta) \right], \quad (22)$$

where $S_{ki} = \cup_{h=1}^H S_{kih}$ is the set of members of family k with disease status i .

The following theorem gives the limiting distribution of $\sqrt{K}(\hat{\beta} - \beta_0)$ as the number K of families becomes infinite.

THEOREM

Consider a sample of K families whose covariate distributions satisfy the marginal covariate model (6)–(8) and the ascertainment assumptions A1–A2. Let $\hat{\theta} = (\hat{\alpha}, \hat{\delta}^T, \hat{\beta}^T)^T$ be the solution of the estimating equation $U(\theta) = 0$, where $U(\theta)$ is given by (18). Assume that $\lim_{K \rightarrow \infty} K^{-1} \sum_k n_{kih} = \zeta_{hi} < \infty$ and that $\lim_{K \rightarrow \infty} K^{-1} \sum_k n_{kih} n_{ki'h} = v_{ih i' h'}$, with $0 < v_{ih i' h'} < \infty$, $h, h' = 1, \dots, H$, $i, i' = 0, 1$. Then, under the mild regularity conditions described by Prentice and Pyke [1979], $\hat{\beta}$ is consistent for β_0 , and $\sqrt{K}(\hat{\beta} - \beta_0)$ is asymptotically multivariate Gaussian as $K \rightarrow \infty$ with zero mean and variance matrix

$$\hat{V} = [G^{-1} \Sigma G^{-1}]_{22}. \quad (23)$$

Here

$$G(\theta) = \lim_{K \rightarrow \infty} K^{-1} E \left[-\frac{\partial}{\partial \theta} U(\theta) \right], \quad (24)$$

$$\Sigma(\theta) = \lim_{K \rightarrow \infty} K^{-1} \sum_k E[U_k^T(\theta) U_k(\theta)],$$

and the subscript 22 on a matrix denotes the submatrix obtained by deleting its first H rows and columns. $\hat{V}$ can be estimated consistently by

$$\hat{V} = (I^{-1} \hat{\Sigma} I^{-1})_{22} \quad (25)$$

evaluated at $\theta = \hat{\theta}$, where

$$I(\theta) = -K^{-1} \frac{\partial}{\partial \theta} U(\theta), \text{ and}$$

$$\hat{\Sigma}(\theta) = K^{-1} \sum_{k=1}^K \hat{U}_k^T(\theta) \hat{U}_k(\theta). \quad (26)$$

Here $\hat{U}_k(\theta)$ is given by (22) with $\mu_{ih}(\theta)$ replaced by $\hat{\mu}_{ih}(\theta)$, which is the overall empirical mean score for individuals in stratum h with disease status i :

$$\hat{\mu}_{ih}(\theta) = n_{ih}^{-1} \sum_{j \in S_{ih}} s_i(x_j, z_j, \theta).$$

The Theorem is proved in the Appendix. When covariates of family members are conditionally independent given their disease statuses and

stratum memberships, the results of Prentice and Pyke show that $U(\theta) = 0$ is the likelihood equation for the data. In this case, standard asymptotic likelihood theory shows that $\Sigma(\theta) = G(\theta)$ in (24). Thus, the asymptotic covariance matrix V of (23) reduces to the usual covariance matrix $V = [G^{-1}]_{22}$ for ULR of case-control data, as it should.

A C program to compute the asymptotic covariance matrix $\hat{V}$ is available at <http://www.stanford.edu/dept/HRP/epidemiology>.

MODELING THE BIVARIATE MARGINALS IN MARGINAL COVARIATE MODELS

The variance matrix of the ULR estimate $\hat{\beta}$ for the parameter β_M in a marginal covariate model depends on $\Sigma(\theta)$, which is a sum of $3H(H+1)/2$ covariance matrices

$$E \left\{ [s_i(e_h, z) - \mu_{ih}]^T [s_{i'}(e_{h'}, z) - \mu_{i'h'}] \right\} \quad (27)$$

for the scores of an arbitrary pair of individuals with disease status i and i' in strata h and h' , $i, i' = 0, 1$, $h, h' = 1, \dots, H$. The entries of the summands (27) are based on the bivariate marginals $pr(z, z' | y, x, y', x')$ for a pair of individuals, given their disease status and stratum memberships. The precision of estimates for the variance of $\hat{\beta}$ may be impaired by small numbers of pairs of individuals contributing to the estimate for a given summand, and some modeling of these bivariate marginals may be needed. For example, it may be highly likely that certain covariates are uncorrelated within families, regardless of disease status or stratum membership. Then the corresponding entries of the matrices (27) could be set to zero. If none of the covariates are correlated within families, then $\Sigma = I$ and the asymptotic variance matrix of $\hat{\beta}$ is $K^{-1}(I^{-1})_{22}$. In this case, the variance estimates obtained from standard logistic regression packages can be used.

An interesting example of bivariate marginal modeling in genetic epidemiology is provided by Slager and Schaid [2001], who propose variance estimates for odds-ratios relating disease risk to genotypes of a single diallelic polymorphism, using data from related cases and controls. The authors apply their method to genotype data from prostate cancer cases ascertained in a linkage study of multiple-case families, their unaffected relatives, and a set of unrelated cancer-free controls. In this application, covariates are genotypes. They

assume a marginal covariate model of the form (6),(8), and implicitly make ascertainment assumptions A1–A2 for their data. In particular, they assume that odds-ratios pertaining to the families ascertained in their linkage study equal those for all such multiple-case families (assumption A2). In their marginal covariate model, the univariate marginal distribution of genotype for a single family member, given his disease status and family characteristics (size, structure, and disease count), is independent of the family characteristics. Moreover, the bivariate marginal genotype distribution for two men depends only on their disease status and genealogical relationship, independent of their other family characteristics. The stratum labels h index the various pedigree positions that determine the covariances of genotypes in pairs of family members. However, because the authors assume that the univariate marginal genotype distributions depend only on disease status and not on pedigree position, the scores $s_i(z)$ depend only on case-control status i . Here z denotes the number of variant alleles of the diallelic polymorphism in an individual's genotype. The probability $pr(z_\ell, z_{\ell'} | y_\ell = y_{\ell'} = 1, \mathcal{R})$ that two cases with relationship $\mathcal{R}$ have genotypes z_ℓ and $z_{\ell'}$ is modeled in terms of the probabilities that they share $t = 0, 1$ or 2 alleles identical-by-descent (IBD):

$$pr(z_\ell, z_{\ell'} | y_\ell = y_{\ell'} = 1, \mathcal{R}) = pr(z_\ell | y_\ell = 1) \\ pr(z_{\ell'} | z_\ell, y_\ell = y_{\ell'} = 1, \mathcal{R}),$$

where

$$pr(z_{\ell'} | z_\ell, y_\ell = y_{\ell'} = 1, \mathcal{R}) = \sum_{t=0}^2 pr(IBC = t | \mathcal{R}) \\ \times pr(z_{\ell'} | z_\ell, y_\ell = y_{\ell'} = 1, IBC = t).$$

The univariate marginal $pr(z_\ell | y_\ell = 1)$ and the conditional probabilities $pr(z_{\ell'} | z_\ell, y_\ell, y_{\ell'}, IBC = t)$ are estimated from the case genotypes in all the families, as described by Slager and Schaid [2001]. Analogous modeling is used to estimate the bivariate marginals for pairs of controls and for case-control pairs, which are then used to estimate the score covariances in Σ needed for the variance of $\hat{\beta}$. Work is needed to evaluate the extent to which such modelling improves the precision of variance estimates.

APPLICATION TO OVARIAN CANCER DATA

Table I gives data from a case-control study of ovarian cancer described by Whittemore [1995]

and Zhao et al. [1998]. The cases and controls (hereafter called probands) are women with and without a diagnosis of ovarian cancer. Each proband reported her mother's ovarian cancer status, and her own and her mother's reproductive histories. The covariate of interest is parity, defined as a woman's total number of childbirths. Parity is classified in three categories: zero births, one or two births, and three or more births. The objective is to determine the relation between cancer risk and parity, allowing for mother-daughter correlation in disease risk and covariates. These family data are a random sample from the distribution $pr(y_2, z_1, z_2 | y_1, \mathbf{x})$, where $\mathbf{x} = (x_1, x_2)$ is a vector of indicators for mother-daughter status, and the subscripts 1 and 2 denote a proband and her mother, respectively. Thus, we may estimate association parameters relating parity to disease risk by modelling either the conditional distribution $pr(z_1, z_2 | y_1, y_2, \mathbf{x})$ given the mother's disease status, or the distribution $pr(y_2, z_1 | y_1, z_2, \mathbf{x})$ given the mother's parity. For either option, fitting a random effects model and estimating its parameters by family-matched CLR is problematic, because only those 381 families in Table I discordant for ovarian cancer can contribute to such an analysis. Instead, we consider a marginal model. Specifically, we factor the distribution generating the data as

$$pr(y_2, \mathbf{z} | y_1, \mathbf{x}) = pr(y_2 | y_1, \mathbf{x}) pr(\mathbf{z} | \mathbf{y}, \mathbf{x}).$$

We assume $pr(y_2 | y_1, \mathbf{x})$ is uninformative for β_M and fit a marginal model to $pr(\mathbf{z} | \mathbf{y}, \mathbf{x})$.

In contrast, Whittemore [1995] and Zhao et al. [1998] assumed a marginal model for $pr(\mathbf{y} | \mathbf{z}, \mathbf{x})$. In particular, they assumed

$$pr(y_j | \mathbf{z}, \mathbf{x}) = pr(y_j | z_j, \mathbf{x}), \quad j = 1, 2.$$

This assumption implies that y_1 and z_2 are conditionally independent given $(z_1, \mathbf{x})$. Then the distribution generating the data may be factored as

$$pr(y_2, \mathbf{z} | y_1, \mathbf{x}) = pr(z_1 | y_1, \mathbf{x}) pr(z_2 | z_1, \mathbf{x}) pr(y_2 | y_1, \mathbf{z}, \mathbf{x}).$$

The term $pr(z_2 | z_1, \mathbf{x})$ is assumed independent of β_M , and estimates are obtained from the product $pr(z_1 | y_1, \mathbf{x}) pr(y_2 | y_1, \mathbf{z}, \mathbf{x})$.

If both of these marginal models are valid, their estimates $\hat{\beta}_M$ should be similar, since they are consistent for the same parameter β_M . However, the two marginal models require different estimates for the covariance matrix of $\hat{\beta}$, since $pr(\mathbf{z} | \mathbf{y}, \mathbf{x})$ concerns mother-daughter correlation in parity, while $pr(\mathbf{y} | \mathbf{z}, \mathbf{x})$ concerns mother-daughter correlation in ovarian cancer risk. Table II

shows estimates $\hat{\beta}$ and their standard errors (SE's) based on each of these models. Parity is coded with an indicator for 1–2 births and an indicator for nulliparity, so that the two logistic regression parameters represent the logit of risk relative to that of women with 3+ births. The model for $pr(y|z, x)$ includes an indicator x with value one for probands and zero for mothers [Whittemore, 1995; Zhao et al., 1998]. The model for $pr(z|y, x)$ includes a vector $x = (x_1, x_2, x_3)$ of indicators for pedigree position and family history (FH), with $x_1 = 1$ (FH+ proband), $x_2 = 1$ (FH- proband), and $x_3 = 1$ (FH+ mother). A woman was designated family-history-positive (FH+) if her paired relative was diagnosed with ovarian cancer.

It is evident from Table II that the two methods give similar estimates for the regression parameters and for the SE's of the estimates. Moreover, comparison with the naive SE's shows that ignoring mother–daughter correlation has little impact on the SE's, despite the significantly positive mother–daughter correlation coefficient in disease risk obtained for the marginal disease model by both Whittemore [1995] and Zhao et al. [1998]. A limitation of this example is the lack of information on β_M contributed by the mothers, among whom only eight ovarian cancers occurred.

In this example, the families were ascertained because of the affected or unaffected disease statuses of the probands, with cases ascertained from cancer registries, and controls ascertained by randomly sampling the populations from which the cases arose. The ascertainment assumption A2 would be violated (and the estimated regression coefficients for parity would be biased) if, say, case probands with many siblings were more likely to respond than case probands without siblings, while the response of control probands were independent of family size. Such differential ascertainment would be particularly likely to occur if the case probands had been ascertained because they had a sibling with ovarian cancer.

SIMULATIONS

We used simulations to compare the performance of CLR and ULR estimates when applied to data ascertained according to ascertainment assumptions A1–A2. We generated a single continuous covariate z_ℓ for each member of 500 “families” of size two (sib pairs) or one (singletons) with various disease phenotypes. To facilitate comparison of the CLR and ULR estimates, we generated covariates according to a distribution that satisfies the assumptions of both the random effects and marginal covariate models. However, the regression parameters differ: the CLR estimates are consistent for β_{RE} , while the ULR estimates are consistent for $\beta_M = (1 - \rho)\beta_{RE}$, where ρ is the correlation coefficient for sibs' covariates. For sib pairs, we assumed $H = 2$ pedigree strata, with stratum $h = 1$ containing individuals whose siblings were diseased, and stratum $h = 2$ containing those whose siblings were disease-free. The stratum membership vector x for a sib is the scalar e_h , with $e_1 = 1$ and $e_2 = 0$. Singletons were assigned the value $x = 0$.

In the random effects model, the family-specific covariate distribution for a sib with disease status y and stratum h is gaussian with mean $v + \gamma x + \beta y$ and variance $\sigma^2 = 1$. We generated data for 500 discordant sib pairs, and call this Design A. Then, we generated data for 300 discordant sib pairs, 100 pairs concordant for disease, and 100 singleton controls, and call this Design B.

Table III shows results based on 1,000 trials for each combination of study design and true parameter value $(\beta_{RE}, \gamma, \rho)$. There are $2 \times 8 = 16$ combinations, corresponding to two designs and eight parameter values. Table III gives for each combination the mean of the 1,000 estimates $\hat{\beta}$, the mean of the 1,000 model-based standard deviation (SD) estimates for $\hat{\beta}$, and the empirical SD of $\hat{\beta}$, based on the 1,000 repetitions. For UCL, Table III gives both the robust SD estimate given by

TABLE I. Distribution of 769 mother–daughter pairs according to ovarian cancer status and parity

		Number of childbirths							
		Proband 0		Proband 1–2		Proband 3+			
Ovarian cancer		Mother		Mother		Mother			
	Proband	Mother	1–2	3+	1–2	3+	1–2	3+	Total
a	Yes	Yes	1	3	1	1	0	1	7
b	Yes	No	29	85	36	105	41	84	380
c	No	Yes	0	1	0	0	0	0	1
d	No	No	22	50	33	103	38	135	381

TABLE II. Estimated regression coefficients and standard errors (SE's) for data in Table I

Number of births	Marginal disease model ^a			Marginal covariate model		
	Coeff	SE	Naive SE ^b	Coeff	SE	Naive SE ^b
1-2	0.347	0.165	0.163	0.342	0.163	0.164
0	0.761	0.188	0.189	0.760	0.189	0.189
ρ for mother-daughter disease status	0.017	0.008	—	—	—	—

^aFrom Whittemore [1995] and Zhao et al. [1998].

^bIgnoring mother-daughter correlation.

(25),(26), and the usual logistic SD estimate (called the “naive” estimate) that is consistent when the covariates of family members are uncorrelated [Prentice and Pyke, 1979].

The first half of Table III presents results for Design A. Rows 1 and 2 give results when sibs’ covariates are uncorrelated ($\rho = 0$). They show that estimates including family history strata are unbiased. In contrast, as seen in row 2, the unstratified estimates are biased when the covariate distribution depends on the family history indicators x ($\gamma > 0$). The bias is the same for both ULR and CLR estimates. As expected, stratified estimates are more variable than unstratified ones.

As shown in rows 3 and 4 of Table III, the same pattern of bias and variability is evident in the presence of covariate correlation ($\rho = 0.67$). The CLR unstratified estimates appear to be more biased than the ULR ones, but when the CLR bias is scaled by the factor $1 - \rho$, the magnitudes of the two biases are similar. The two methods perform similarly under the null hypothesis $\beta_{RE} = \beta_M = 0$, with similar (scaled) variances and, when biased, similar (scaled) bias. Thus, in this simple example the CLR and ULR estimates have similar coefficients of variation (defined as the mean divided by the SD). This similarity suggests that the random effects/CLR and marginal/ULR approaches have similar power to test the null hypothesis, provided all families contribute to the CLR analysis.

The second half of Table III gives results when the data were generated according to Design B, which includes disease-concordant sib pairs and singleton controls that do not contribute to the CLR analysis. The patterns are similar to those seen for Design A, with some exceptions. As expected, the CLR estimates, which use only the discordant sib pairs, are more variable than the ULR ones. Moreover, the unstratified CLR estimates show greater (scaled) bias than do the unstratified ULR estimates.

Comparison of the naive and robust ULR SD’s with the correct empirical SD shows that, for both

designs, the naive estimate is biased upward in the presence of covariate correlation within families. The direction of this bias is surprising and counterintuitive, since the naive estimate, which does not account for covariate correlation, assumes an artificially large “effective size.” The upward bias indicates that confidence intervals obtained using the naive variance estimates will cover the true β more often than its nominal coverage probability. However, the bias typically is small, at least for the level of covariate correlation examined in these limited simulations.

DISCUSSION

We have considered the problem of estimating association parameters using data from families ascertained because of their disease phenotypes. We have described random effects and marginal models for the distribution $pr(\mathbf{z}|\mathbf{y}, S)$ of family covariates, given the vector $\mathbf{y}$ of family phenotype indicators and the family pedigree structure S . We have shown that CLR gives consistent estimates for the parameters in a random effects model, while ULR gives consistent estimates for the parameters in a marginal model. However, the odds-ratios in the two models have different interpretations, and they also differ in magnitude when covariates are correlated. Thus, estimates for their odds-ratios are not comparable.

This lack of comparability indicates the need for care in choosing an estimating method appropriate for the data at hand, and care in interpreting the estimates. The random effects disease model for $pr(\mathbf{y}|\mathbf{z}, S)$ has been used for family data when familial correlation in disease risk is thought to reflect the sharing of many genetic variants, each in itself of small effect [Kraft and Thomas, 2000]. When this “polygenic” model is plausible, the best strategy for estimating β_{RE} from retrospective family data is family-matched CLR. This is because the random effects models for $pr(\mathbf{y}|\mathbf{z}, S)$ and for $pr(\mathbf{z}|\mathbf{y}, S)$ involve the same parameter β_{RE} .

TABLE III. Performance of conditional and unconditional logistic regression (CLR and ULR) estimates for β^a

Parameters used to generate data		CLR stratified on family history ^b				ULR stratified on family history			
		Yes		No		Yes		No	
		$\hat{\beta}$	Model ^e	Empirical ^f	$SD\hat{\beta}$	$\hat{\beta}$	Model ^e	Empirical ^f	$SD\hat{\beta}$
γ^c	ρ^d								
A. 500 Disease-discordant sib pairs									
$\beta_{RE}=\beta_M=0$									
0	0	0.001	0.09	0.09	0.06	0.001	0.08	0.08	0.06
0.4054	0	0.000	0.09	0.09	0.06	-0.001	0.08	0.08	0.06
0	0.67	-0.001	0.09	0.09	0.06	-0.001	0.05	0.03	0.02
0.4054	0.67	0.001	0.09	0.09	0.06	0.003	0.05	0.03	0.02
$\beta_{RE}=\beta_M=0.6931$									
0	0	0.698	0.11	0.11	0.08	0.696	0.09	0.09	0.07
0.4054	0	0.704	0.11	0.11	0.07	0.698	0.09	0.09	0.07
$\beta_{RE}=0.6931, \beta_M=0.2310$									
0	0.67	0.702	0.11	0.11	0.08	0.232	0.05	0.04	0.02
0.4054	0.67	0.696	0.11	0.11	0.07	0.232	0.05	0.04	0.02
B. 300 Disease-discordant sib pairs, 100 disease-concordant sib pairs, 100 unrelated controls									
$\beta_{RE}=\beta_M=0$									
0	0	0.004	0.12	0.12	0.08	-0.001	0.07	0.07	0.06
0.4054	0	-0.000	0.12	0.12	0.08	-0.001	0.07	0.07	0.06
0	0.67	-0.003	0.12	0.12	0.08	0.000	0.04	0.03	0.03
0.4054	0.67	0.000	0.12	0.12	0.08	-0.001	0.04	0.03	0.03
$\beta_{RE}=\beta_M=0.6931$									
0	0	0.703	0.14	0.14	0.10	0.697	0.08	0.08	0.07
0.4054	0	0.709	0.14	0.15	0.09	0.695	0.08	0.07	0.07
$\beta_{RE}=0.6931, \beta_M=0.2310$									
0	0.67	0.709	0.14	0.14	0.10	0.232	0.04	0.03	0.03
0.4054	0.67	0.700	0.14	0.15	0.09	0.231	0.04	0.04	0.03

^aBased on data simulated from a random-effects gaussian covariate model, as described in text.^bMatched on sibship.^cRegression coefficient for family history.^dCorrelation coefficient for sibs' covariates.^eBased on information matrix.^fEmpirical standard deviation of estimates $\hat{\beta}$ obtained from 1,000 iterations.^gDefined in text.

Family-matched CLR is problematic, however, if a substantial proportion of the ascertained families drop out of the analysis because they lack a pair of members discordant for the disease. In this case, one might consider a marginal covariate model and use ULR to estimate its parameter β_M . The resulting odds-ratio estimates also can be interpreted as measures of covariate effects on disease risk, because the marginal models for $pr(\mathbf{y}|\mathbf{z}, S)$ and for $pr(\mathbf{z}|\mathbf{y}, S)$ involve the same parameter β_M . However, the estimate for β_M must be interpreted differently from that for β_{RE} , and when the covariates are correlated, the ULR logistic regression parameters differ in value from those obtained by CLR.

The random effects and marginal covariate models considered here ignore the complexities that arise in analyzing time-to-failure data when families are ascertained solely because one or more members have developed the disease in a given calendar period. Langholz et al. [1999] provide a thoughtful discussion of these issues in relation to a study of genetic factors in juvenile diabetes. The issues are particularly relevant to the analysis of time-dependent covariates. The hypothetical situations considered by Langholz et al. provide instructive examples of how A2 might be violated. Similar potential violations of A2 arise in studies of covariates with strong regional variations, when cases are ascertained in one geographic area and unaffected relative controls are ascertained elsewhere [Siegmund and Langholz, 2002].

In summary, if both random effects and marginal models are plausible for a given application, the choice of method might be guided by the proportion of sampled families containing pairs of members discordant for both disease and covariates. If this proportion is high, the random effects/CLR approach will provide efficient estimates that are simple to obtain and interpret. If this proportion is not high, the marginal/ULR approach will provide more precise parameter estimates and consistent variance estimates. This approach to modeling $pr(\mathbf{z}|\mathbf{y}, S)$ parallels the strategy proposed by Liang and Zeger [1986] for applying marginal disease models to $pr(\mathbf{y}|\mathbf{z}, S)$.

ACKNOWLEDGMENTS

The authors thank Joseph B. Keller, Duncan Thomas, Susan Slager, Peter Kraft, Bryan Langholz, and John Hopper for helpful discussion.

REFERENCES

- Anderson JA. 1972. Separate sample logistic discrimination. *Biometrika* 59:19–35.
- Breslow NE, Day NE. 1980. Statistical methods in cancer research: Volume I. The analysis of case-control studies. IARC Sci Publ (32).
- Cox DR, Hinkley DV. 1974. Theoretical Statistics. London: Chapman and Hall.
- Diggle PJ, Liang KY, Zeger SL. 1994. Analysis of longitudinal data. Oxford: Clarendon Press.
- Efron B. 1975. The efficiency of logistic regression compared to normal discriminant analysis. *JASA* 70:892–898.
- Kraft P, Thomas DC. 2000. Bias and efficiency in family-based gene-characterization studies: conditional, prospective, retrospective, and joint likelihoods. *Am J Hum Genet* 66: 1119–1131.
- Langholz B, Ziogas A, Thomas DC, Faucett C, Huberman M, Goldstein L. 1999. Ascertainment bias in rate ratio estimation from case-sibling control studies of variable age at onset diseases. *Biometrics* 55:1129–1136.
- Liang KY, Zeger SL. 1986. Longitudinal data analysis using generalized linear models. *Biometrika* 73:13–22.
- Prentice RL, Pyke R. 1979. Logistic disease incidence models and case-control studies. *Biometrika* 66:403–411.
- Press SJ, Wilson J. 1978. Choosing between logistic regression and discriminant analysis. *JASA* 70:699–705.
- Siegmund KD, Langholz B. 2002. Ascertainment bias in family-based case-control studies. *Am J Epidemiol* 155: 875–880.
- Slager SL, Schaid DJ. 2001. Evaluation of candidate genes in case-control studies: a statistical method to account for related subjects. *Am J Hum Genet* 68:1457–1462.
- Truett J, Cornfield J, Kannel W. 1967. A multivariate analysis of the risk of coronary heart disease in Framingham. *J Chron Dis* 20:511–524.
- Whittemore AS. 1995. Logistic regression of family data from case-control studies. *Biometrika* 82:57–67; Addendum: Logistic regression of family data from case-control studies. 1997. *Biometrika* 84:989–990.
- Zhao LP, Hsu L, Holte S, Chen Y, Quiaoit F, Prentice RL. 1998. Combined association and aggregation analysis of data from case-control family studies. *Biometrika* 85: 299–315.

APPENDIX

PROOF OF THEOREM

We first show that $\hat{\theta}$ is consistent for θ_0 by showing that $E[U(\theta_0)] = 0$. From (20)

$$E[U(\theta_0)] = \sum_{h=1}^H [n_{\cdot 0h} \mu_{0h}(\theta_0) + n_{\cdot 1h} \mu_{1h}(\theta_0)]. \quad (28)$$

Here $n_{\cdot ih}$ is the total number of ascertained individuals with disease status i and stratum membership h , and $\mu_{ih}(\theta_0)$, given by (21), is the mean value of the scores $s_i(e_h, \mathbf{z}, \theta_0)$ of (19). From (28) it suffices to show that

$$n_{\cdot 0h} \mu_{0h}(\theta_0) + n_{\cdot 1h} \mu_{1h}(\theta_0) = 0, \quad h = 1, \dots, H. \quad (29)$$

To do so, we substitute (19) into the right side of (21) to obtain

$$\begin{aligned} \mu_{ih}(\theta_0) = & (-1)^{i+1} \int [p(e_h, z, \theta_0)]^{1-i} [q(e_h, z, \theta_0)]^i \\ & \times pr(z|y = i, x = e_h, a = 1)(1, e_h, z) dz. \end{aligned} \quad (30)$$

Here $a = 1$ means that the individual was ascertained, and $p(e_h, z, \theta_0) = 1 - q(e_h, z, \theta_0)$ is given by (17).

Consider the identity

$$\begin{aligned} pr(z|y = i, e_h, a = 1) \\ = \phi_h(z; \eta) \left[\frac{\pi(z, e_h; \eta)}{\eta} \right]^i \left[\frac{1 - \pi(z, e_h; \eta)}{(1 - \eta)} \right]^{1-i}, \end{aligned} \quad (31)$$

where

$$\pi(z, e_h; \eta) = \frac{\eta pr(z|y = 1, e_h, a = 1)}{\phi_h(z; \eta)} \quad (32)$$

and

$$\begin{aligned} \phi_h(z; \eta) = & \eta pr(z|y = 1, e_h, a = 1) \\ & + (1 - \eta) pr(z|y = 0, e_h, a = 1). \end{aligned} \quad (33)$$

This identity holds for any value η in the unit interval. In particular, it holds for

$$\eta = \eta_h^* = \frac{n_{.1h}}{n_{.0h} + n_{.1h}} \equiv \frac{n_{.1h}}{n_{..h}}. \quad (34)$$

Note from (32) and (33) that

$$\frac{\pi(z, e_h; \eta_h^*)}{1 - \pi(z, e_h; \eta_h^*)} = \frac{\eta_h^*}{1 - \eta_h^*} \tilde{\varphi}(z, e_h), \quad (35)$$

where

$$\tilde{\varphi}(z, e_h) = \frac{pr(z|y = 1, e_h, a = 1)}{pr(z|y = 0, e_h, a = 1)}.$$

Using assumptions (8) and A2, we can write

$$\tilde{\varphi}(z, e_h) = e^{\tilde{\zeta} + \tilde{\gamma}_h + z\beta}, \quad (36)$$

where $\tilde{\zeta}$ and $\tilde{\gamma}_h$ are the parameters appropriate for the ascertained population. Substituting (36) into (35) gives

$$\pi(z, e_h; \eta_h^*) = \frac{e^{\alpha_0 + \delta_{0h} + z\beta}}{1 + e^{\alpha_0 + \delta_{0h} + z\beta}}, \quad (37)$$

where

$$\begin{aligned} \alpha_0 = & \tilde{\zeta} + \ln \frac{\eta_H^*}{(1 - \eta_H^*)} \text{ and} \\ \delta_{0h} = & \tilde{\gamma}_h + \ln \frac{\eta_h^*}{(1 - \eta_h^*)} - \ln \frac{\eta_H^*}{(1 - \eta_H^*)}. \end{aligned}$$

Comparison of (37) with (17) shows that

$$\pi(z, e_h, \eta_h^*) = p(e_h, z, \theta_0), \quad (38)$$

where $\theta_0 = (\alpha_0, \delta_0^T, \beta^T)^T$ and $\delta_0^T = (\delta_{01}, \dots, \delta_{0,H-1})$. Substituting (38) into (31) gives

$$\begin{aligned} pr(z|y = i, e_h, a = 1) \\ = \phi_h(z; \eta_h^*) \left[\frac{p(e_h, z, \theta_0)}{\eta_h^*} \right]^i \left[\frac{q(e_h, z, \theta_0)}{1 - \eta_h^*} \right]^{1-i}. \end{aligned} \quad (39)$$

Next we substitute the right side of (39) for $pr(z|y = i, e_h, a = 1)$ in (30) and use (34) to obtain

$$\begin{aligned} \mu_{ih}(\theta_0) = & (-1)^{i+1} \frac{n_{..h}}{n_{.ih}} \int p(e_h, z, \theta_0) q(e_h, z, \theta_0) \\ & \times \phi_h(z; \eta_h^*)(1, e_h, z) dz. \end{aligned} \quad (40)$$

Finally we substitute (40) for the $\mu_{ih}(\theta_0)$ in the left side of (29) to obtain 0, as required.

To show the asymptotic normality of $\sqrt{K}(\hat{\beta} - \beta_0)$, we use a first-order Taylor series expansion of $U(\hat{\theta})$ about $U(\theta_0)$ to obtain

$$0 = U(\hat{\theta}) = U(\theta_0) + \left[\frac{\partial}{\partial \theta} U(\theta_*) \right] (\hat{\theta} - \theta_0), \quad (41)$$

where the components of θ_* lie between those of $\hat{\theta}$ and those of θ_0 . Equation (41) can be written

$$\sqrt{K}(\hat{\theta} - \theta_0) = [I(\theta_*)]^{-1} K^{-\frac{1}{2}} U(\theta_0), \quad (42)$$

where $I(\theta) = -K^{-1} \frac{\partial}{\partial \theta} U(\theta)$. We first verify the asymptotic normality of $K^{-\frac{1}{2}} U(\theta_0)$. To do so, we write

$$K^{-\frac{1}{2}} U(\theta) = K^{-\frac{1}{2}} \sum_{k=1}^K U_k,$$

and apply a multivariate central limit theorem to the mutually independent variables U_k . Also $I(\theta_*)$ is consistent for $G(\theta_0)$ for the reasons outlined by Prentice and Pyke. It follows from (42) that $\sqrt{K}(\hat{\theta} - \theta_0)$ is asymptotically normal with mean zero and variance $G^{-1} \Sigma G^{-1}$. Thus, $\sqrt{K}(\hat{\beta} - \beta_0)$ is asymptotically normal with variance $V = (G^{-1} \Sigma G^{-1})_{22}$. The consistency of $\hat{V}$ as an estimate of V follows from the strong law of large numbers.