Segregation Analysis of Continuous Phenotypes by Using Higher Sample Moments

Hang Lee[1] and Daniel O. Stram[2]

[1]Department of Biostatistics, Harvard University, Cambridge, MA; and [2]Department of Preventive Medicine, University of Southern California, Los Angeles

Summary

The present article discusses the use of computational methods based on generalized estimating equations (GEE), as a potential alternative to full maximum-likelihood methods, for performing segregation analysis of continuous phenotypes by using randomly selected family data. The method that we propose can estimate effect and degree of dominance of a major gene in the presence of additional nongenetic or polygenetic familial associations, by relating sample moments to their expectations calculated under the genetic model. It is known that all parameters in basic major-gene models cannot be identified, for estimation purposes, solely in terms of the first two sample moments of data from randomly selected families. Thus, we propose the use of higher (third order) sample moments to resolve this identifiability problem, in a pseudo-profile likelihood estimation scheme. In principle, our methods may be applied to fitting genetic models by using complex pedigrees and for estimation in the presence of missing phenotype data for family members. In order to assess its statistical efficiency we compare several variants of the method with each other and with maximum-likelihood estimates provided by the SAGE computer package in a simulation study.

Introduction

Much work in population genetics, beginning with Fisher (1918), has described the relationship between parameters in a major-gene model and the resulting phenotype covariances between relatives. Since that time there has been considerable further development of method of moments approaches toward estimating parameters in genetic models, in both segregation analysis (see Whittemore and Gong 1994) and linkage analysis, (see Amos 1994). However, since the introduction of recursive algorithms for evaluating likelihoods for genetics models (Elston and Stewart 1971, 1978), the method of maximum likelihood (ML) has been the most popular approach to parameter estimation.

The generalized estimating equations (GEE) technique (Liang and Zeger 1985; Prentice and Zhao 1991) is a general approach for the estimation of model parameters by using only sample moments of the data, by modeling expectations, rather than maximizing the parameters in the complete distribution of correlated responses as in ML. Recently, Zhao (1994) offered a detailed, but overly optimistic, suggestion for the use of the first and second moments (what is often termed a GEE-2 procedure), for performing segregation analysis of major-gene models for data from randomly selected human pedigrees. As indicated by Zhao (1994), one motivation for a GEE procedure for segregation analysis is that, since GEE-2 requires only the correct specification of the expectations of the first two sample moments, it may be more robust to model misspecification than is the fully parametric ML approach. Unfortunately, however, the three parameters needed to specify even the most basic major-gene model—namely, the allele frequency, dominance mode, and genetic displacement (difference between phenotype means for *AA* vs. *aa* genotypes)—are not mutually distinguishable from only the phenotype means, variances, and covariances used in GEE-2. The implication of the analysis of Lee et al. (1993), for example, in the segregation analysis of continuous phenotypes, is that if the dominance mode is known to be additive (with the phenotype mean of *aA* genotypes midway between *AA* and *aa*), there remains exact aliasing between the allele frequency and the displacement parameters, when using GEE-2. That is, only one of these two parameters can be estimated, even when the dominance mode is known. For other dominance models the aliasing in GEE-2 between the allele frequency and displacement parameters, while not exact, is still close enough so as to make simultaneous estimation of these parameters by using GEE-2 essentially impossible.

Lee et al. (1993) describe a pseudo-profile likelihood method for continuous phenotypes that estimates both

Received November 28, 1994; accepted for publication September 26, 1995.

Address for correspondence and reprints: Dr. Daniel O. Stram, Department of Preventive Medicine, Center for the Health Professions, Room 220I, University of Southern California Medical School, Los Angeles, CA 90033. E-mail: stram@almaak.usc.edu

0002-9297/96/5801-0024$02.00

the allele frequency and genetic displacement in major-gene models still under the assumption that the dominance model is known. This iterative two-stage method uses the first two sample moments, i.e., GEE-2, with one parameter (the allele frequency) fixed, at the first stage of each iteration, and then in the second stage of each iteration estimates the allele frequency by maximizing a very simple pseudo-profile likelihood. In order to allow for the estimation of the dominance parameter and, in addition, to allow for residual, nongenetic correlation between the phenotypes of family members, generally due to unmeasured environmental covariates shared among family members, Stram et al. (1993) suggested that additional higher moments would need to be utilized, beyond those incorporated into GEE-2, in the first stage of the estimation procedure. The present study details the implementation of this suggestion (using first, second, and third sample moments, i.e., GEE-3), and provides a simulation study in which the estimates from the proposed method are compared with the approximate ML estimation (MLE) given by the SAGE package (Elston 1992).

In many cases the use of GEE methods is motivated by a desire to provide consistent estimates of specific model parameters of interest without having to specify the complete distribution of the correlated data. Our focus on the exploration reported here, however, is on fitting exactly the same parametric models that are used in MLE. Thus, in the simulation study we choose the "regressive" model for residual familial correlation (Bonney 1984) that is implemented by the MLE package SAGE. Our aim is to begin to evaluate the statistical efficiency of GEE methods when the segregation model is fully specified. We do not attempt here to compare the robustness of the GEE versus ML when the model is misspecified, but this is clearly a claim that will require further investigation. Besides statistical efficiency, it will be important to compare the computational efficiency of GEE with ML. We begin this evaluation in the simulations. In particular, we address the trade-off in efficiency versus computational gains achieved as the number of higher (third-order) moments used in the "working vector" of the GEE machinery increases.

Models for Familial Association Involving Major Genes and Residual Association

The Major-Gene Model

Consider a continuous phenotype y_{ij} for a member $j(j = 1, 2, \ldots, J_i)$ in a randomly sampled family $i(i = 1, 2, \ldots, I)$. The major-gene model, the parameters of which are to be estimated in the course of the segregation analysis, is specified in terms of a conditional model for the phenotype in light of a major locus genotype g_{ij} and a covariate vector x_{ij} represented by the following linear function:

$$y_{ij} = \mu + x_{ij}^{t}\underline{\alpha} + \beta\{f_{g_{ij}}(d) - m(d, p)\} + e_{ij}\,. \qquad (1)$$

Here, $f_{g_{ij}}(d)$ is the penetrance function of the major locus allele taking values 0, d, and 1 for the major locus genotypes *aa*, *aA*, and *AA*, respectively. d is the degree of dominance of the allele A with respect to the allele a, which takes 1 if allele A is dominant, 0 if recessive, and 0.5 if additive. $m(d,p)$ is the mean of the penetrance function, $\Sigma_{g_{ij}} f_{g_{ij}} P\{g_{ij} \mid p\}$, where $P\{g_{ij} \mid p\}$ is the population genotype probability. The parameter β is the difference in means between genotypes *aa* and *AA* (genetic displacement), and p is the population frequency of the major locus allele A. The e_{ij} is an error term and is assumed to be independent and identically distributed (iid) from a normal distribution with mean 0 and variance σ_e^2.

Mixed Major-Gene Polygenetic Models for Residual Correlation

In a polygenic model for residual correlation between family members a new random variable, γ_{ij}, which we term a polygenotype, is added to the model in equation (1), which denotes either the action of a polygene, that is, the summary result of a large number of genes each of which have a small influence on the phenotype, or an unobserved covariate, which tends to be shared among family members, with the degree of sharing related to the degree of relationship between family members. The joint distribution of the vector of the quantitative polygenotypes for each family, $(\gamma_{i1}, \gamma_{i2}, \ldots, \gamma_{iJ_i})$ is assumed to follow a multivariate normal distribution with mean 0 and covariance matrix

$$\sigma_\gamma^2 \begin{bmatrix} 1 & \rho_{12} & \rho_{13} & \cdot & \cdot & \cdot & \rho_{1J_i} \\ \rho_{21} & 1 & \rho_{23} & \cdot & \cdot & \cdot & \rho_{2J_i} \\ \rho_{31} & \rho_{23} & 1 & \cdot & \cdot & \cdot & \rho_{3J_i} \\ \cdot & & & \cdot & & & \cdot \\ \cdot & & & & \cdot & & \cdot \\ \rho_{J_i1} & \rho_{J_i2} & \rho_{J_i3} & \cdot & \cdot & \cdot & 1 \end{bmatrix} = \sigma_\gamma^2 \Psi_i\,.$$

In a true polygene model, the correlation, ρ_{jk}, between γ_{ij} and γ_{ik}, will be equal to $(1/2)^{R(j,k)}$, where $R(j,k)$ is the degree of relationship between family member j and k. (That is, $(1/2)^{R(j,k)}$ is the expected fraction of shared genes between these two family members.) The model for residual correlation may take other forms, however. For example, it is common to include a nongenetic environmental factor being shared by siblings within common parents or to allow for correlation between spouses (nonrandom mating). Of course, identifiability of multi-

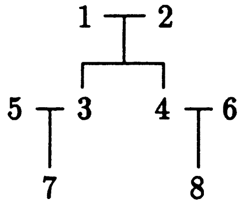

Figure 1 Three-generation family structure pedigree used in simulation of family data for illustration of the method.

levels of residual correlation in a finite data set is problematic for ML methods as well as the methods described here.

Regressive Models

Regressive models for residual dependence replace the two terms, e_{ij} and γ_{ij}, with a single term, ω_{ij}, where ω_{ij} is assumed to follow an autoregressive type process, and the rest of the assumptions are the same as those in equation (1). If random assortment is assumed, ω_{ij} can be expressed as

$$\omega_{ij} = \rho\omega_{ik} + \varepsilon_{ij} ,$$

where the member k is a first-degree relative to the member j, ρ is a correlation between any first-degree relative pair, and ε_{ij} is an iid normal random variable with mean 0 and variance σ_ε^2. Therefore, the joint distribution of $(\omega_{i1},\omega_{i2}, \ldots, \omega_{ij_i})$ is multivariate normal with mean 0. The covariance matrix of this distribution, in the case of an eight-member sample pedigree structure (see fig. 1), will be

$$\frac{\sigma_\varepsilon^2}{1-\rho^2}\begin{bmatrix} 1 & 0 & \rho & \rho & 0 & 0 & \rho^2 & \rho^2 \\ 0 & 1 & \rho & \rho & 0 & 0 & \rho^2 & \rho^2 \\ \rho & \rho & 1 & \rho & 0 & 0 & \rho & \rho^2 \\ \rho & \rho & \rho & 1 & 0 & 0 & \rho^2 & \rho \\ 0 & 0 & 0 & 0 & 1 & 0 & \rho & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & \rho \\ \rho^2 & \rho^2 & \rho & \rho^2 & \rho & 0 & 1 & \rho^3 \\ \rho^2 & \rho^2 & \rho^2 & \rho & 0 & \rho & \rho^3 & 1 \end{bmatrix} = \sigma^2\Omega_i , \tag{2}$$

where each position of row or column of this matrix represents the corresponding member index in the pedigree. When ρ is zero, the residual covariance structure $\sigma_\varepsilon^2 I_i$ describes the pure random error structure of the simple major-gene mechanism of equation (1). Like the mixed model, the regressive model may be modified to allow for sibling correlations different than correlations between parents and offspring, for example, to give what Bonney (1986) terms a class D model in which additional sibling effects, given common parentage, may be present equally among all siblings. While both the mixed model and the regressive model can be thought of as providing alternatives to the major-gene model, which allows for additional familial correlation beyond what is attributable to the action of a major gene alone, these models are distinct in that neither can be regarded as a special case of the other. In general, ML computation under the regressive model is considerably simpler than under the mixed major-gene polygene model (Elston 1992), and the regressive model forms the basis for the SAGE program.

Estimating Equations Approach to Fitting a Simple Major-Gene Model

In general, as described in detail in Lee et al. (1993), it is not possible to estimate all parameters in the model of equation (1) by using only either the sample means in a GEE-1 procedure or the sample means and covariances (GEE-2). In particular, Lee et al. (1993) finds that, of the three genetic parameters, β, d, and p, in the model of equation (1), only one of these at a time is readily identifiable from the first two sample moments; no two can be simultaneously estimated from their effects upon the expectations of only these moments. For example, in order to estimate β, using GEE-2, both the degree of dominance, d, and the population allele frequency, p, must be assumed known. In order to estimate both β and p, with d still remaining fixed, Lee et al. (1993) proposed a pseudo-profile likelihood approach to estimating p. For a discussion of pseudo-profile procedures, see, for example, Gong and Samaniego (1981).

In order to justify the application of two-stage GEE-based methods toward fitting the parameters in equation (1), consider MLE of p by maximizing the profile likelihood, $\ell_\omega(\hat{\Theta}_p,p)$ for the entire sample (founder and nonfounders). Here, $\hat{\Theta}_p$ denotes the MLE vector, with p fixed, for all the other parameters in the simple major-gene model. If $\ell_f(\Theta,p)$ denotes the likelihood for the founders only, then the following score equation holds at the MLE of p:

$$\begin{aligned} 0 &= \frac{d}{dp}\,\ell_\omega(\hat{\Theta}_p, p) \\ &= \frac{\partial}{\partial\hat{\Theta}_p}\,\ell_\omega(\hat{\Theta}_p, p) \times \frac{\partial}{\partial p}\,\hat{\Theta}_p + \frac{\partial}{\partial p}\,\ell_\omega(\hat{\Theta}_p, p) \\ &= \frac{\partial}{\partial p}\,\ell_\omega(\hat{\Theta}_p, p) \,. \end{aligned}$$

Note that conditional on parents' genotypes, the offsprings' genotype distributions depend only on Mendelian transmission probabilities and do not depend on p. This suggests that it may be possible to restrict our estimation of p, in light of known values of the other parameters, to use only the data concerning the phenotypes of the founders and ignore the phenotypes for the nonfounders. That is, we approximate the partial derivative

$$\frac{\partial}{\partial p}\,\ell_{\omega}(\hat{\Theta}_p, p) \tag{3}$$

as

$$\frac{\partial}{\partial p}\,\ell_f(\hat{\Theta}_p, p)\ . \tag{4}$$

and estimate p and Θ by finding values that satisfy the condition

$$\frac{\partial}{\partial p}\,\ell_f(\hat{\Theta}_p, p) = 0\ . \tag{5}$$

This procedure first picks an initial value of p, then maximizes $\ell_f(\Theta_p, p)$ to estimate the remaining parameters as $\hat{\Theta}_p$, and finally updates the initial value of p by solving the equation (5). The further suggestion of Lee et al. (1993) is, with the degree of dominance parameter d still regarded as fixed, to replace $\hat{\Theta}_p$ in the profile likelihood with GEE-2–based estimates, $\tilde{\Theta}_p$, of Θ_p (stage 1 of the procedure) and to perform a second maximization in stage 2 to solve

$$\frac{\partial}{\partial p}\,\ell_f(\tilde{\Theta}_p, p) = 0\ . \tag{6}$$

Since the calculation of partial derivatives for $\ell_f(\Theta,p)$ is less difficult than evaluating $\ell_{\omega}(\Theta,p)$, it is suggested that the combined procedure may be computationally more efficient than maximizing the full likelihood, while, it is hoped, maintaining a reasonable level of statistical efficiency. The statistical efficiency of the proposed procedure will depend on the validity of the two approximations used; the first is the approximation of $\hat{\Theta}_p$ with $\tilde{\Theta}_p$, and the second is the approximation of the partial derivative in equation (3) with the partial derivative in equation (4).

If only the first two sample moments are used in estimating $\tilde{\Theta}_p$, Lee et al. (1993) indicate that d remains unidentifiable in the first stage. Lee et al. (1993) explored the possibility of estimating d as well as p in the second stage. We suggest here, however, that a more direct approach toward estimating the degree of dominance is to include the estimation of d in stage 1 by incorporating model-based expectations for the third sample moments, in a GEE-3 procedure. We detail this suggestion below.

Estimating Equations: Stage 1

First, consider the simple major-gene model. The general approach of the estimating equations technique is to estimate parameters in the model without resort to maximizing the full likelihood for y but rather by dealing with these parameters only as they affect the marginal means, variance, and covariances of y_{ij} or (as here) higher moments of y_{ij}. For the simple major-gene model the phenotype mean vector of $\mathbf{y}_i$ for the ith pedigree may readily be seen to be equal to

$$\mathbf{E}_i = \mu\mathbf{1} + \mathbf{X}_i\underline{\alpha}\ .$$

The variance covariance matrix of the phenotype vector $\mathbf{y}_i$ for the ith pedigree equals

$$\Sigma_i = \beta^2\mathbf{C}_i(d, p) + \sigma_e^2\mathbf{I}_i\ ,$$

where the first part is the major genetic component and the second part is the nongenetic component. In the major-gene model the nonzero phenotype covariances, $\mathbf{C}_i(d,p)$ are functions of the allele frequency p and the degree of dominance d in the penetrance function. The covariance of founders (those family members whose parents are not in the pedigree) with each other, and with all other members except their descendants, are assumed zero (no assortative mating). Under the random-mating assumption, these major genetic covariances are defined by

$$\mathbf{C}_i(d, p)_{(jk)} = \sum_{g_j, g_k} [f_{g_j} - m(d, p)] \times [f_{g_k} - m(d, p)]P\{g_j, g_k \mid p\}\ .$$

By subtracting off $m(d,p)$ in the conditional model equation (1), the phenotype means contain neither the allele frequency p nor the genetic displacement β, so that it is only the variances and covariances, and higher moments, that contain information about either p or β.

The GEE-2 procedure (Prentice and Zhao 1991) yields estimating equations,

$$\sum_{i=1}^{I} \mathbf{D}_i^t\mathbf{W}_i^{-1}\mathbf{z}_i = 0\ , \tag{7}$$

where $\mathbf{D}_i$ is a matrix of partial derivatives of expectations of means, variances, and covariances with respect to the parameters being estimated; $\mathbf{z}_i$ is a vector of "working variables" (deviations between the observed data and their contributions to the variances and covari-

Table 1

Expectations of the First Three Sample Moments

	Mixed	Regressive
$\mathbf{E}_i$	$\mu 1 + \mathbf{X}_i\alpha$	$\mu 1 + \mathbf{X}_i\alpha$
Σ_i	$\beta^2\mathbf{C}_i(d, p) + \sigma_\gamma^2\Psi_i + \sigma_e^2\mathbf{I}_i$	$\beta^2\mathbf{C}_i(d, p) + \sigma^2\Omega_i$
Γ_i	$\beta^3\mathbf{T}_i(d, p)$	$\beta^3\mathbf{T}_i(d, p)$

ances, and their respective model predictions),

$$\mathbf{z}_i = \begin{pmatrix} \mathbf{y}_i - (\mu 1 + \mathbf{X}_i\underline{\alpha}) \\ \mathrm{Vec}(\mathbf{S}_i - \Sigma_i) \end{pmatrix};$$

and $\mathbf{W}_i^{-1}$ is a weight matrix, which optimally is taken as a generalized inverse of the model-based covariance matrix of the working vector. Denote Vec{**M**}, for matrix **M**, as the vector composed of the columns of **M**.

In order that d as well as β be identifiable in the first stage of the procedure, we consider the addition of sample third moments to the working vector, $\mathbf{z}_i$, and the corresponding enhancements of the other components of the GEE machinery: $\mathbf{D}_i$ and $\mathbf{W}_i^{-1}$.

GEE-3 in Stage 1

For the mixed major-gene polygene model and the regressive model, the expectations of the sample means, variances and covariances, and third-order sample moments ($\mathbf{E}_i$, Σ_i, and Γ_i) of the phenotypes to be used to define the estimating equations are given in table 1.

As we see in table 1, both models have same expectations of the means and third moments. The expected third moments are defined as

$$\Gamma_{i_{(jkl)}} = E\{[y_{ij} - E(y_{ij})]\,[y_{ik} - E(y_{ik})][y_{il} - E(y_{il})]\},$$
$$\text{for all triples } (i,\ k, \text{ and } l)\ ,$$

where the elementary form, for example $[y_{ij} - E(y_{ij})]$, can be written as either $\beta[f_{g_{ij}}(d) - m(d,p)] + \gamma_{ij} + e_{ij}$, for the mixed major-gene polygene model, or $\beta[f_{g_{ij}}(d) - m(d,p)] + \omega_{ij}$, for the regressive model. Since we assumed both γ_{ij} and ω_{ij} have multivariate normal distributions independent with the penetrance function $f_{g_{ij}}$, and the random error term e_{ij} (for the mixed major-gene polygene mixed model), the remaining nonzero terms are

$$\beta^3E\{[f_{g_{ij}}(d) - m(d,\ p)][f_{g_{ik}}(d) - m(d,\ p)]$$
$$\times\ [f_{g_{il}}(d) - m(d,\ p)]\} = \beta^3\mathbf{T}_i(d,\ p)_{(jkl)}\ .$$

These marginal third moments consist of the major genetic three-way covariances multiplied by the cube of the genetic displacement. The major genetic three-way covariances, $\mathbf{T}_i(d,p)$, are defined by

$$\sum_{g_{ij}, g_{ik}, g_{il}} [f_{g_{ij}}(d) - m(d,\ p)][f_{g_{ik}}(d) - m(d,\ p)]$$
$$\times\ [f_{g_{il}}(d) - m(d,\ p)]P\{g_{ij},\ g_{ik},\ g_{ik}\}\ .$$

The GEE-3 estimating equation is now defined as

$$\sum_{i=1}^{I} \dot{\mathbf{D}}_i^t\dot{\mathbf{W}}_i^{-1}\dot{\mathbf{z}}_i = 0\ . \qquad (8)$$

Each matrix of the estimating equations is defined below. For the regressive model

$$\dot{\mathbf{D}}_i = \begin{bmatrix} \left[\dfrac{\partial}{\partial\Theta_p}\{\mu 1 + \mathbf{X}_i\underline{\alpha}_i\}\right]_{(J_i \times q)} \\ \left[\dfrac{\partial}{\partial\Theta_p}\mathrm{Vec}\{\beta^2\mathbf{C}_i + \sigma^2\Omega_i\}\right]_{(J_i^2 \times q)} \\ \left[\dfrac{\partial}{\partial\Theta_p}\mathrm{Vec}\{\beta^3\mathbf{T}_i\}\right]_{(J_i^3 \times q)} \end{bmatrix}.$$

The vector $\dot{\mathbf{z}}_i$ =

$$\begin{pmatrix} [\{\mathbf{y}_i - \mu 1 - \mathbf{X}_i\underline{\alpha}\}](J_i \times 1) \\ [\mathrm{Vec}\{\{\mathbf{y}_i - \mu 1 - \mathbf{X}_i\underline{\alpha}\}\{\mathbf{y}_i - \mu 1 - \mathbf{X}_i\underline{\alpha}\}^t - \beta^2\mathbf{C}_i - \sigma^2\Omega_i\}]_{(J_i^2 \times 1)} \\ [\mathrm{Vec}\{\{\mathbf{y}_i - \mu 1 - \mathbf{X}_i\underline{\alpha}\} \otimes \{\mathbf{y}_i - \mu 1 - \mathbf{X}_i\underline{\alpha}\} \otimes \{y_i - \mu 1 - X_i\alpha\} - \beta^3\mathbf{T}_i\}]_{(J_i^2 \times 1)} \end{pmatrix}$$

contains the "working variables" (deviations between the contributions to the sample means, variances, covariances, and the third moments, and their respective model predictions). The formulation for the mixed major-gene polygene model is similar to the regressive model except that the $\sigma^2\Omega_i$ is replaced by $\sigma_\gamma^2\Psi_i + \sigma_e^2\mathbf{I}_i$.

The derivatives given in the expression for $\dot{\mathbf{D}}_i$ are calculated by direct summations over the possible values for the genotypes for the family members. Thus, for example, under the regressive model for residual correlation, $\partial/\partial_d\mathbf{C}_i$, the partial derivative of the expected father-offspring major genetic covariance, $\partial/\partial_d\mathbf{C}_i$, is

$$\frac{\partial}{\partial d}\Sigma_{g_F}\Sigma_{g_M}\Sigma_{g_O}\{[f_{g_{iF}}(d) - m(d,\ p)][f_{g_{iO}}(d) - m(d,\ p)]\}$$
$$\times\ P\{g_F\}P\{g_M\}P\{g_O|g_F,\ g_M\}$$
$$= \Sigma_{g_F}\Sigma_{g_M}\Sigma_{g_O}\left\{\frac{\partial}{\partial d}[f_{g_{iF}}(d) - m(d,\ p)]\right.$$
$$\times\ [f_{g_{iO}}(d) - m(d,\ p)] + [f_{g_{iF}}(d) - m(d,\ p)]$$
$$\left.\times\ \frac{\partial}{\partial d}[f_{g_{iO}}(d) - m(d,\ p)]\right\}$$

$$\times P\{g_F\}P\{g_M\}P\{g_O|g_F, g_M\}$$

$$= \Sigma_{g_F}\Sigma_{g_M}\Sigma_{g_O}\ \{[I_{\{g_{iF}=aA\}} - 2p(1-p)][d\cdot I_{\{g_{iO}=aA\}} - p^2 - 2dp(1-p)] + [d\cdot I_{\{g_{iF}=aA\}} - p^2 - 2dp(1-p)][I_{\{g_{iO}=aA\}} - 2p(1-p)]\} \times P\{g_F\}P\{g_M\}P\{g_O|g_F, g_M\}\ .$$

Here, the probabilities, $P\{g_O\,|\,g_F,g_M\}$, reflect the usual Mendelian inheritance rules. The other partial derivatives, in $\dot{\mathbf{D}}_i$, of the expected second and third moments can be calculated similarly.

The weight matrix $\dot{\mathbf{W}}_i^{-1}$ would ideally be chosen as a generalized inverse of the variance-covariance matrix of $\dot{\mathbf{z}}_i$. Since calculating an optimal weight matrix from first principles is quite complex, we consider below several approximations to the optimal weight matrix, for practical computational purposes.

Stage 1 of the pseudo-profile likelihood estimation procedure solves the estimating equations for the parameters, $\Theta_p = (\mu_p, \underline{\alpha}_p, \beta_p, d_p, \sigma_p^2\ \rho_p)^t$, by Fisher's scoring with p fixed,

$$\tilde{\Theta}_p^{k+1} = \tilde{\Theta}_p^{k} - \left(\sum_i \dot{\mathbf{D}}_i^t\dot{\mathbf{W}}_i^{-1}\dot{\mathbf{D}}_i\right)^{-1}\left(\sum_i \dot{\mathbf{D}}_i^t\dot{\mathbf{W}}_i^{-1}\dot{\mathbf{z}}_i\right), \quad (9)$$

and stage 2 estimates p by maximizing the founders' profile likelihood explained below.

Estimation of the Allele Frequency: Stage 2

In the stage 2 estimation of p, only founders are used in our approximation. This likelihood function is a mixture of three univariate normal likelihood functions with mixing proportions $(1-p)^2$, $2p(1-p)$, and p^2, which are corresponding probabilities of the three possible genotypes, i.e.,

$$L(p, \tilde{\mu}_p, \underline{\tilde{\alpha}}_p, \tilde{\beta}_p, \tilde{\sigma}_p^2; \mathbf{y}, \mathbf{X})$$

$$= \prod_{i\in founder}\left\{(1-p)^2\cdot\frac{1}{\tilde{\sigma}_p^2} \times \phi\left[\frac{y_i - \tilde{\mu}_p - \mathbf{x}_i^t\underline{\tilde{\alpha}}_p - \tilde{\beta}_p(f_{aa}(d) - m(d, p))}{\tilde{\sigma}_p}\right] + 2p(1-p)\cdot\frac{1}{\tilde{\sigma}_p^2}\cdot\phi\left[\frac{y_i - \tilde{\mu}_p - \mathbf{x}_i^t\underline{\tilde{\alpha}}_p - \tilde{\beta}_p(f_{aA}(d) - m(d, p))}{\tilde{\sigma}_p}\right] + p^2\cdot\frac{1}{\tilde{\sigma}_p^2}\cdot\phi\left[\frac{y_i - \tilde{\mu}_p - \mathbf{x}_i^t\underline{\tilde{\alpha}}_p - \tilde{\beta}_p(f_{AA}(d) - m(d, p))}{\tilde{\sigma}_p}\right]\right\} \quad (10)$$

The entire estimating procedure is an iterative two-stage optimization in which stage 1 of each iteration solves the estimating equations for the parameters in the means, variances, and covariances by Fisher's scoring with p fixed and the stage 2 estimates p, by maximizing expression (10).

Sampling Variances of the Estimators

On the basis of general results (Huber 1981) for M-estimation, the asymptotic variance covariance matrix of the parameter estimates in the two-stage GEE-based pseudo MLE procedure is given as the so-called information sandwich

$$\mathbf{I}_+^{-1}\mathbf{S}_+\mathbf{I}_+^{-1^t}\ .$$

The modified information matrix, $\mathbf{I}_+$, is the nonsymmetric matrix consisting of the partial derivatives of the two equations being solved, equations (8) and (6) with respect to the model parameters

$$\mathbf{I}_+ = \begin{bmatrix} \Sigma_i\dot{\mathbf{D}}_i^t\dot{\mathbf{W}}_i^{-1}\dot{\mathbf{D}}_i & 0 \\ \left(\frac{\partial}{\partial p}\tilde{\Theta}_p\right)^t\Sigma_i\dot{\mathbf{D}}_i^t\dot{\mathbf{W}}_i^{-1}\dot{\mathbf{D}}_i & \frac{\partial^2}{\partial p^2}\ell_f(\tilde{\Theta}_p, p) \end{bmatrix},$$

while $\mathbf{S}_+$ contains the cross-products of the GEE-3 scores, $\dot{\mathbf{D}}_i^t\dot{\mathbf{W}}_i^{-1}\dot{\mathbf{z}}_i$, in equation (8) and the individual score contributions in equation (6).

$$\mathbf{S}_+ = \begin{bmatrix} \Sigma_i\dot{\mathbf{D}}_i^t\dot{\mathbf{W}}_i^{-1}\dot{\mathbf{z}}_i\dot{\mathbf{z}}_i^t\dot{\mathbf{W}}_i^{-1}\dot{\mathbf{D}}_i & (\Sigma_i\dot{\mathbf{D}}_i^t\dot{\mathbf{W}}_i^{-1}\dot{\mathbf{z}}_i)\left(\frac{\partial}{\partial_p}\ell_f(\tilde{\Theta}_p, p)\right) \\ \left(\frac{\partial}{\partial p}\ell_f(\tilde{\Theta}_p, p)\right)^t(\Sigma_i\dot{\mathbf{z}}_i^t\dot{\mathbf{W}}_i^{-1}\dot{\mathbf{D}}_i) & \Sigma_f\left(\frac{\partial}{\partial_p}\ell_f(\tilde{\Theta}_p, p)\right)^2 \end{bmatrix}$$

is the "naive" sampling variances covariances matrix estimated from the data.

Computational Considerations

As given in equation the estimation of parameters, μ, α_1, α_2, β, d, and σ^2, in the first stage of the estimating equation, involve the manipulation of rather large matrices. For example, each working vector, $\dot{\mathbf{z}}_i$, corresponding to the simple three-generation pedigree shown in figure 1 is nominally of length $8 + 8^2 + 8^3 = 584$, with the derivative matrix $\dot{\mathbf{D}}_i$ and the corresponding weight matrix, $\dot{\mathbf{W}}_i^-$ of size 584 × 6 and 584 × 584, respectively. In fact, however, it is not the number of individuals, but rather the number of unique types of relationships between individuals, which determines the number of distinct elements that need to be included in $\dot{\mathbf{z}}_i$ and the other matrices. For example, there are six distinct two-person family relationships among individuals in figure 1, namely, self-self, parent-offspring, sib-sib, grandparent-grandchild, uncle-nephew, and cousin-cousin. Thus the size of the portion of the working vector, $\dot{\mathbf{z}}_i$, which corresponds to the sample covariance matrix, $\mathbf{S}_i$, may be reduced by summing together equivalent contributions to $\mathbf{S}_i$. Similarly, the number of distinct three-person family relationships contained within this pedigree structure is far less than 8^3. In general, we can sum together the jth row of $\dot{\mathbf{z}}_i$, for all j where rows of $\dot{\mathbf{D}}_i$ are identical and where the corresponding rows and columns of $\dot{\mathbf{W}}_i$ are the same, while simultaneously summing together the corresponding rows of $\dot{\mathbf{D}}_i$, and rows and columns of $\dot{\mathbf{W}}_i$. For large extended pedigrees where the number of distinct relationships between family members becomes large, further reductions in computations may be made by simply dropping the contributions corresponding to distant relationships. This computational savings would come at the cost of some reduction in statistical efficiency. The issue of which third-order moments to include in the working vectors for pedigrees given in figure 1, is explored in the simulations below.

The next computational issue we explore in the simulations is a comparison of several alternative forms of the weight matrix, $\dot{\mathbf{W}}_i$. First of all, it appears to be quite important not to include higher-order sample moments in estimates of parameters that are already well estimated on the basis of the first and second moments alone. Thus, in the following, we always break up the updating of the parameters in stage 1 into two parts, so that for all parameters except for d and β the GEE-2 estimates from equation (7) are used rather than the estimates from equation (8). In the simulations described below we compare the use of three different versions of the weight matrix in the updating of the parameter estimates in equation (8).

1. An optimal weight matrix is taken as a generalized inverse of the true covariance matrix of $\dot{\mathbf{z}}_i$. We closely approximate this matrix by simulation. For the simulations described below we generate a large number of $\dot{\mathbf{z}}_i$ from the eight-person pedigree in figure 1 and calculate the observed covariance matrix of these vectors, using the true values of the parameter estimates, when calculating $\dot{\mathbf{z}}_i$.
2. A second weight matrix is based on the (erroneous) assumption that the $\dot{\mathbf{z}}_i$ are distributed as the first, second, and third sample moments of a multivariate normal random vector having mean $\mathbf{E}_i = \mu\mathbf{1} + \mathbf{X}_i\underline{\alpha}$ and variance covariance matrix $\Sigma_i = \beta^2\mathbf{C}_i(d,p) + \sigma^2\Omega_i$. That is, we use a generalized inverse of a matrix made up of the second to sixth moments of the multivariate normal distribution, in forming $\dot{\mathbf{W}}_i^{-1}$.
3. Last, we use the empirical covariance matrix of the $\dot{\mathbf{z}}_i$ calculated for each simulated data set.

Because the weight matrix in version 1 is based on true parameter values it is unachievable in practice for a real data set and is used here as a "gold standard" for evaluating the performance of the other methods. The other versions of $\dot{\mathbf{W}}_i^{-1}$ use the estimated values of the parameters obtained for each simulated data set at each iteration of the estimation procedure. Finally, we estimate parameters in the model by the use of ML provided by the SAGE package.

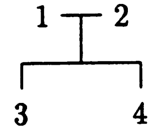

Figure 2 Nuclear family structure pedigree used in simulation of family data for illustration of the method.

Simulation

Data Set and Parameters

We performed simulations using the three-generation pedigree structure shown in figure 1 and the nuclear family pedigree structure in figure 2. In each case we simulated 100 data sets consisting of either 200 three-generation families or 400 nuclear families. A phenotype was generated under the regressive model equation (2). The population phenotype distribution conditional on the major genotype and shared environmental factor was assumed to be normal, $N(\mu_g, \sigma_\varepsilon^2 = 1.125)$, where $\mu_{AA} = \mu_{aA} = 6$, and $\mu_{aa} = 2$, respectively ($d = 1$, $\mu = 5$, and $\beta = 4$). The parent-offspring transmission of the major allele A with the allele frequency $p = .5$ was assumed to follow Mendelian inheritance. As additional correlations between members, not due to the major gene, a Markovian structure with first-degree relations having correlation of 0.5 was assumed and fixed, i.e., $\rho_{(j,k)} = 0.5^{R(j,k)}$, where $R(j,k)$ is the degree of relation between members j and k. Thus, $\sigma^2 = 1.5$ in equation (2). Since we assumed the complete dominance ($\mu_{AA} = \mu_{aA}$) of the major allele A to allele a with respect to the phenotype, the additive major genetic variance $\sigma_a^2 = 2pq^3 = 2 \times 0.5 \times 0.5^3 = 0.125$, and the dominant major genetic variance $\sigma_d^2 = p^2q^2 = 0.5^2 \times 0.5^2 = 0.0625$. The total phenotypic variance for an individual is $\beta^2(\sigma_a^2 + \sigma_d^2) + \sigma^2 = 4^2 \times (0.125 + 0.0625) + 1.5 = 4.5$; hence the heritability $h = 3/4.5 = 0.67$. Two binary dummy covariates, x_1 and x_2, representing the second and third generations, respectively, were created. The covariate effects on the phenotypes were chosen to reduce the mean by -1 and -2 from the top generation mean ($\alpha_1 = -1$; and $\alpha_2 = -2$).

Results

Table 2 displays all 19 possible third moments corresponding to the types of relationships given in the three-

Table 2

Third Moments Included in the GEE3: Eight-Member Three-Generation Family Structure Pedigree

Third Moments	GEE3-3	GEE3-11	GEE3-19
Self^3	In	In	In
Mother-father-offspring	In	In	In
Parent-sib1-sib2	In	In	In
Grandmother-grandfather-grandchild		In	In
Grandparent-parent-grandchild		In	In
Parent-uncle-nephew		In	In
Cousin-uncle-cousin		In	In
Grandparent-uncle-cousin		In	In
Grandparent-cousin1-cousin2		In	In
Grandparent-parent-uncle		In	In
Grandparent-parent-nephew		In	In
Parent^2-offspring			In
Parent-offspring^2			In
Sib1^2-sib2			In
Grandparent^2-grandchild			In
Grandparent-grandchild^2			In
Uncle^2-nephew			In
Uncle-nephew^2			In
Cousin1^2-cousin2			In

generational family pedigree shown in figure 1. Table 3 gives three results of estimating the parameters in the regressive model with ρ fixed by 0.5, using three different choices of the third moments to include in the working vector. In this table the estimates indicated by D for the SAGE estimation are indirectly derived estimates (SAGE provides mean estimates conditional on the genotypes, i.e., $\hat{\mu}_{aa}$, $\hat{\mu}_{aA}$, and μ_{AA}, instead of $\hat{\mu}$, $\hat{\beta}$, and $\hat{d}$). Thus, the $\underline{\text{SE}}^2$s for these estimates are the averages of the square of derived standard errors using a delta method approximation. The results labeled GEE3-3 correspond to using, in the working vector, the first three sample moments in table 2: self-self-self, mother-father-offspring, and parent-sib-sib. The results labeled GEE3-11 use the first 11 moments in table 2, and the GEE3-19 results use all 19 possible third sample moments. In table 3 the weight matrix used is the gold standard (method 1). Note that there is no gain in efficiency in estimating the nongenetic parameters, μ, a_1, a_2, or σ^2 as more higher moments are added to the working vector, which is not altogether surprising given that only GEE-2, i.e., equation (7), is used in updating these parameters estimates in the iterations. For the genetic parameters, β and d, a small gain in efficiency (as measured by the mean squared error [MSE] of estimation) is seen as the number of sample moments used increases from 3 to 19, but the gain appears to be relatively unimportant. For allele frequency, p, there is virtually no gain observed as the number of sample moments is increased. Bias in the parameter estimates using the GEE-3 methods does not appear to be very great for any of the model parameters. For all parameters except p, the information sandwich–based estimates of the variance of the parameter estimates and the 95% confidence intervals were reasonably accurate relative to their observed variances in the simulations.

Tables 4–7 illustrate the impact on estimation (only for the β, d, and σ^2) of choosing among the three variants

Table 3

Third Moments Included in the GEE3: Four-Member Nuclear Family Structure Pedigree

Third Moments	GEE3-2	GEE3-3	GEE3-4	GEE3-5	GEE3-6
Self^3	In	In	In	In	In
Mother-father-offspring	In	In	In	In	In
Parent-sib1-sib2		In	In	In	In
Parent^2-offspring			In	In	In
Parent-offspring^2				In	In
Sib1^2-sib2					In

Table 4

Summary of Model Fitting in the Simulation: Analyzing Eight-Member Three-Generation Family Structure Pedigree

Parameter	True	Method	D/I[a]	Mean	Variance	$\overline{SE^2}$	MSE	95% Confidence Interval Coverage[b] (%)
μ	5.0	GEE3-3	I	4.996	.0098	.0087	.0099	94
		GEE3-11	I	4.996	.0096	.0086	.0096	94
		GEE3-19	I	4.997	.0098	.0087	.0098	94
		SAGE	D	4.989	.0065	.0116	.0066	92
α_1	−1.0	GEE3-3	I	−1.009	.0084	.0079	.0085	97
		GEE3-11	I	−1.008	.0082	.0079	.0083	97
		GEE3-19	I	−1.007	.0082	.0081	.0082	98
		SAGE	I	−.998	.0017	.0019	.0017	97
α_2	−2.0	GEE3-3	I	−2.015	.0140	.0132	.0143	95
		GEE3-11	I	−2.015	.0140	.0132	.0142	95
		GEE3-19	I	−2.017	.0147	.0134	.0150	94
		SAGE	I	−2.005	.0045	.0043	.0046	97
β	4.0	GEE3-3	I	4.062	.1161	.1759	.1199	99
		GEE3-11	I	4.059	.1215	.1392	.1250	99
		GEE3-19	I	3.893	.0716	.0657	.0831	89
		SAGE	D	4.010	.0233	.0214	.0234	95
d	1.0	GEE3-3	I	0.992	.0085	.0078	.0085	96
		GEE3-11	I	0.993	.0082	.0071	.0082	96
		GEE3-19	I	1.025	.0066	.0049	.0073	91
		SAGE	D	0.993	.0012	.0011	.0012	92
σ^2	1.5	GEE3-3	I	1.459	.0389	.0891	.0406	99
		GEE3-11	I	1.457	.0500	.0710	.0519	95
		GEE3-19	I	1.519	.0428	.0433	.0432	94
		SAGE	I	1.490	.0044	.0033	.0045	91
p	0.5	GEE3-3	I	.5015	.00055	.00213	.00055	96
		GEE3-11	I	.5014	.00055	.00235	.00055	95
		GEE3-19	I	.5005	.00056	.00182	.00056	99
		SAGE	I	.4991	.00028	.00023	.00028	93

[a] D = indirectly derived parameters in SAGE; and I = independent parameters.

[b] The percentage of 95% confidence intervals that included the true value.

of the weight matrices, $\dot{\mathbf{W}}_i$, used in the simulations during stage 1. In this elaboration, the GEE3-2 to GEE3-6 methods (tables 5–7) used third moments listed in table 4. As expected, both method 2 (assuming multivariate normality) and method 3 (empirical covariances) show some loss of efficiency relative to the unattainable gold standard (method 1). It is interesting that method 3 tended to introduce somewhat more bias in the genetic parameter estimates than did method 2, but the overall MSE of estimation with method 3 was generally better than method 2.

Discussion

This paper has continued the exploration of the use of GEE-based methods for segregation analysis which was begun by Lee et al. (1993) and Whittemore and Gong (1994). The ultimate usefulness of these methods depends on how well they live up to their claims of robustness against violations of model assumptions while still providing reasonable statistical power, relative to such model-based approaches as the method of ML. We have not attempted to evaluate their robustness here but have focused our comparisons on their efficiency, compared to ML, when the correct likelihood model is known. In this situation the GEE-3 methods we have described are substantially outperformed by the MLE (table 3). For example the GEE-3–based MSEs for the genetic parameters, p, β and d, using all 19 third moments, ranged from two to six times as large as those obtained from SAGE. It appears that the main loss of efficiency in the GEE methods comes about because of poor estimation of the dominance parameter, d. Another possible explanation of the efficiency loss is that the founders-only method

Table 5

Summary of Simulation Analyzing Four-Member Nuclear Family Structure Pedigree Data Using GEE-3: Weight Option 1 (Optimal Weight)

Parameter	True	Method	Mean	Variance	$\overline{SE^2}$	MSE
β	4.0	GEE3-2	4.028	.1176	.1788	.1184
		GEE3-3	4.015	.0924	.1110	.0926
		GEE3-4	4.017	.0908	.0812	.0911
		GEE3-5	4.023	.0761	.0641	.0766
		GEE3-6	4.020	.0740	.0627	.0745
d	1.0	GEE3-2	1.000	.0078	.0091	.0078
		GEE3-3	1.001	.0062	.0057	.0062
		GEE3-4	1.001	.0060	.0058	.0060
		GEE3-5	.999	.0053	.0047	.0053
		GEE3-6	.999	.0049	.0043	.0049
σ^2	1.5	GEE3-2	1.441	.0518	.0800	.0552
		GEE3-3	1.455	.0556	.0641	.0577
		GEE3-4	1.451	.0542	.0495	.0567
		GEE3-5	1.452	.0516	.0420	.0539
		GEE3-6	1.454	.0546	.0391	.0566

of estimating the allele frequency (i.e., stage 2 of the estimation method), p, is suboptimal. However, in the simulations p was the best estimated of the three genetic parameters, relative to the MLE, and it appears possible that the founders-only method could work well, if the other parameters, particularly d, were better estimated in stage 1.

It has been remarked by a reviewer that the second-stage estimate of p may suffer greatly in the case when phenotype data are missing, since founders typically will have fewer phenotype data available. In order to see whether the two-stage method for estimating p was more sensitive to missing data than the MLE, an additional analysis, on just one of the simulated eight-member data sets, was performed. We randomly replaced 50% of the phenotype data for all the members of the data set with missing values and then, with all the parameters except the allele frequency fixed at their true values, we estimated p in two ways: first, using SAGE while using the full pedigrees; and second, using the data from the founders only (the second stage). We found that the lengths of likelihood-based 95% confidence in-

Table 6

Summary of Simulation Analyzing Four-Member Nuclear Family Structure Pedigree Data Using GEE-3: Weight Option 2 (Sample Moments Based on Multivariate Normal Assumption)

Parameter	True	Method	Mean	Variance	$\overline{SE^2}$	MSE
β	4	GEE3-2	4.038	.1129	.2077	.1143
		GEE3-3	4.024	.0974	.1283	.0980
		GEE3-4	4.023	.0951	.1009	.0957
		GEE3-5	4.025	.0863	.0847	.0869
		GEE3-6	4.018	.0806	.0815	.0810
d	1	GEE3-2	.998	.0081	.0103	.0081
		GEE3-3	.998	.0067	.0069	.0067
		GEE3-4	.998	.0066	.0066	.0066
		GEE3-5	.998	.0065	.0063	.0065
		GEE3-6	.999	.0056	.0059	.0056
σ^2	1.5	GEE3-2	1.436	.0637	.0973	.0678
		GEE3-3	1.461	.0914	.0833	.0930
		GEE3-4	1.456	.0912	.0793	.0931
		GEE3-5	1.455	.0784	.0711	.0804
		GEE3-6	1.460	.0787	.0654	.0830

Table 7

Summary of Simulation Analyzing Four-Member Nuclear Family Structure Pedigree Data Using GEE-3: Weight Option 3 (Empirical Covariance of the Working Vector)

Parameter	True	Method	Mean	Variance	$\overline{\text{SE}^2}$	MSE
β	4	GEE3-2	3.896	.1272	.1676	.1379
		GEE3-3	3.914	.1111	.1061	.1185
		GEE3-4	3.944	.1131	.0775	.1162
		GEE3-5	3.961	.0978	.0648	.0994
		GEE3-6	3.945	.0934	.0610	.0965
d	1	GEE3-2	1.011	.0090	.0093	.0090
		GEE3-3	1.000	.0072	.0058	.0072
		GEE3-4	.990	.0072	.0051	.0073
		GEE3-5	.980	.0068	.0047	.0072
		GEE3-6	.982	.0060	.0043	.0063
σ^2	1.5	GEE3-2	1.445	.0539	.0729	.0568
		GEE3-3	1.444	.0568	.0591	.0599
		GEE3-4	1.405	.0582	.0446	.0672
		GEE3-5	1.383	.0539	.0385	.0675
		GEE3-6	1.375	.0537	.0348	.0692

tervals for p were actually quite similar (0.074 and 0.0755 for SAGE and the founders-only method, respectively), using the two methods on this 50% missing data set. This analysis again appears to indicate that the efficiency problems of the GEE-3–based two-stage approach relate to poor estimation of the parameters estimated in stage 1 rather than inferior estimation of the allele frequency using only the founders data.

Further evaluation of these methods may be useful in the future. For example, the use of fourth, or even higher sample moments, could be considered and probably are necessary for good estimation of the dominance parameter d. In the present study we have taken a brute-force approach toward determining which of the third sample moments appear to provide information regarding the genetic parameters. In the construction of estimating equations for more complex pedigrees we need to better understand which higher-order sample moments are the most informative about the genetic parameters. In table 3 it appears that very little gain is made in efficiency of the estimation procedure once the first three third moments given in table 1 are included in the first stage: the generality of this observation, however, is as yet unknown. A GEE-based approach that used only a few very informative higher-sample moments might be an attractive addition, rather than alternative, to MLE methods, particularly since with such methods it would be possible to compare the moment estimates from the model with those observed in the sample data, thereby providing measures of the fit of the model as a whole.

Tables 5–7 address the issue of which of two proposed weight matrices is superior in terms of their performance relative to the supposed "gold standard." The results here indicate that, with the possible exception of σ^2, the weight matrix based on the higher moments of the multivariate normal distribution produced more accurate estimated variances of the parameter estimates than did the use of the empirical score covariances. Again, the generality of this observation requires further investigation.

Many segregation analyses involve data collection designs that may differ from the simple random sampling of families discussed in this paper. Families may be included in the sample on the basis of the phenotype of one or more members of the family or on the basis of the occurrence of a disease potentially related to the phenotype being examined. For example, studies of the genetics of blood pressure may ascertain only families for whom high blood pressure, or a disease related to blood pressure level, is detected in at least one pedigree member. Understanding the effects of these and other sampling schemes on the expectation of sample moments of the phenotype is key to extending GEE-based procedures for segregation analysis outside the random-sampling framework.

Acknowledgments

This work was supported by Department of Health and Human Services grant CA52862-02. Some of the results reported in the present study were obtained by using the program package SAGE, supported by U.S. Public Health Service grant RR03655. The authors would like to thank two reviewers for suggestions that led to improvements in the article.

References

Amos CI (1994) Robust variance-components approach for assessing genetic linkage in pedigrees. Am J Hum Genet 54:535–543

Bonney GE (1984) On the statistical determination of major gene mechanisms in continuous human traits: regressive models. Am J Med Genet 18:731–749

——— (1986) Regressive logistic models for familial disease and other binary traits. Biometrics 42:611–625

Elston RC (1992) Statistical Analysis of Genetic Epidemiology version 2.1. LSU Medical Center, New Orleans

Elston RC, Stewart J (1971) A general model for the genetic analysis of pedigree data. Hum Hered 21:523–542

——— (1978) Statistical modeling and analysis of in human genetics. Annu Rev Biophys Bioeng 7:253–286

Fisher RA (1918) The correlation between relatives on the supposition of Mendelian inheritance. Trans R Soc Edinb 52:399–433

Gong G, Samaniego FJ (1981) Pseudo maximum likelihood estimation: theory and applications. Ann Stat 9: 861–869

Huber P (1981) Robust statistics. John Wiley & Sons, New York

Lee H, Stram DO, Thomas DC (1993) A generalized estimating equation approach to fitting major gene models in segregation analysis of continuous phenotypes. Genet Epidemiol 10:61–74

Liang KY, Zeger SL (1986) Longitudinal data analysis using generalized linear models. Biometrics 73:13–22

Prentice RL, Zhao LP (1991) Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. Biometrics 47: 825–829

Stram DO, Lee H, Thomas DC (1993) Use of generalized estimating equations in segregation analysis of continuous outcomes. Genet Epidemiol 10:575–579

Whittemore AS, Gong G (1994) Segregation analysis of case-control data using generalized estimating equations. Biometrics 50:1073–1087

Zhao LP (1994) Segregation analysis of human pedigrees using estimating equations. Biometrika 81:197–209