Biometrika (2005), **92,** 3, *pp.* 559–571

Printed in Great Britain

Locally-efficient robust estimation of haplotype-disease association in family-based studies

By ANDREW S. ALLEN

Department of Biostatistics and Bioinformatics and Duke Clinical Research Institute, Duke University, Durham, North Carolina 27710, U.S.A.

andrew.s.allen@duke.edu

GLEN A. SATTEN

Centers for Disease Control and Prevention, Atlanta, Georgia 30341, U.S.A.

gsatten@cdc.gov

and ANASTASIOS A. TSIATIS

Department of Statistics, North Carolina State University, Raleigh, North Carolina 27695, U.S.A.

tsiatis@stat.ncsu.edu

Summary

Modelling human genetic variation is critical to understanding the genetic basis of complex disease. The Human Genome Project has discovered millions of binary DNA sequence variants, called single nucleotide polymorphisms, and millions more may exist. As coding for proteins takes place along chromosomes, organisation of polymorphisms along each chromosome, the haplotype phase structure, may prove to be most important in discovering genetic variants associated with disease. As haplotype phase is often uncertain, procedures that model the distribution of parental haplotypes can, if this distribution is misspecified, lead to substantial bias in parameter estimates even when complete genotype information is available. Using a geometric approach to estimation in the presence of nuisance parameters, we address this problem and develop locally-efficient estimators of the effect of haplotypes on disease that are robust to incorrect estimates of haplotype frequencies. The methods are demonstrated with a simulation study of a case-parent design.

Some key words: Family-based association study; Haplotype; Nuisance parameter.

1. Introduction

In humans, 23 long strands of DNA called chromosomes contain the blueprint for the proteins and protein products that modify, regulate and, to a large extent, determine the physical processes that constitute health or lead to disease. In most cells, there are two versions of each chromosome, one inherited from the father of an individual and the other inherited from the mother. Hence, at each position, or locus, along the chromosome, there are two versions of the DNA sequence. These versions are termed alleles and it is the pair of alleles that make up an individual's single-locus genotype. When multiple loci

are considered, all the alleles that lie on the same chromosome, i.e. those alleles inherited from the same parent, make up a haplotype. The transcription of DNA that ultimately leads to protein synthesis takes place along each chromosome, making haplotypes important in identifying unique sequences of DNA that predispose individuals to disease. Unfortunately, haplotype-level data are rarely available and one is forced to attempt to reconstruct haplotypes based on multi-locus genotypes. The inability to reconstruct haplotype structure from multi-locus genotypes is termed haplotype phase uncertainty. To fix ideas consider two loci on the same chromosome consisting of 'biallelic' genotypes made up of binary alleles. Suppose that both genotypes are heterozygous, meaning that the maternal allele differs from the paternal allele at each locus; see Fig. 1. We will code each allele as a 0 or 1 and denote the genotype, G_l, at locus l by the sum of the coded alleles, so that, for two heterozygous loci, the multi-locus genotype is $G = (G_1, G_2) = (1, 1)$. Note that, when only multi-locus genotypes are available, the haplotype phase is uncertain; any of the possible haplotype phase structures seen in Fig. 1 are consistent with the observed genotype data.

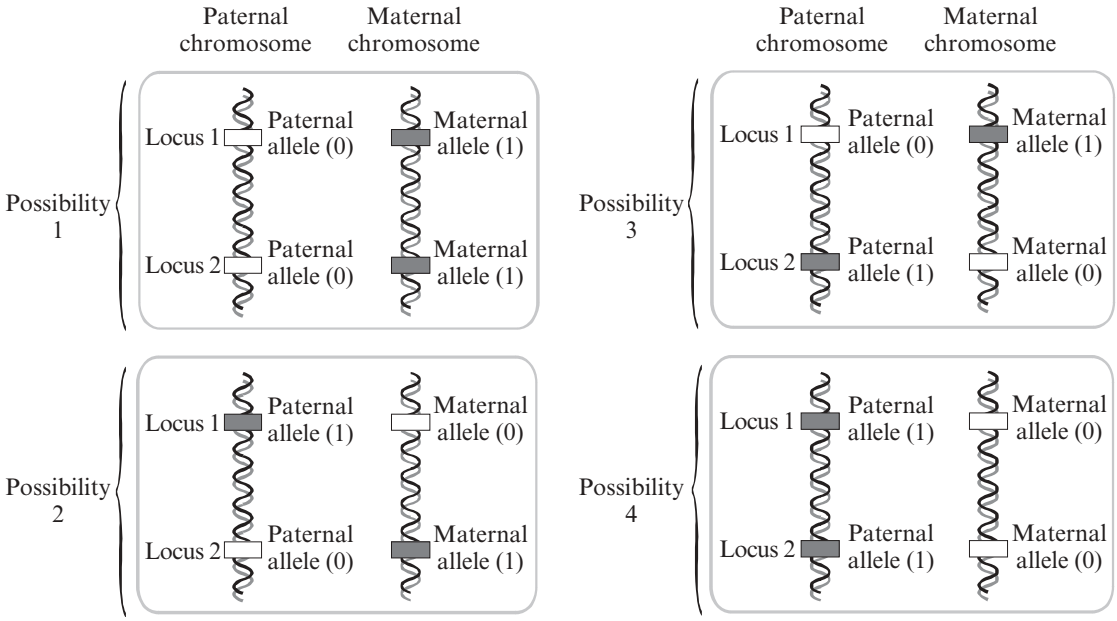

Fig. 1. Example of haplotype phase uncertainty when multi-locus genotype is $G = (G_1, G_2) = (1, 1)$.

If the loci considered are close enough that one can assume that there is no recombination between the loci, so that each parent passes one of their haplotypes intact to their offspring, parental multi-locus genotype data can reduce the amount of haplotype phase ambiguity. In addition, collecting this familial information has other benefits. Family-based designs, in which genotype data are collected on family members, often parents, of an affected individual, are often employed in genetic association studies because they avoid possible confounding due to ethnic stratification of the population (Self et al., 1991). As an example, consider the case in which a population consists of two subpopulations each having different allele frequencies at a locus as well as differing prevalences of disease. For simplicity assume that subpopulation 1 has both a higher allele frequency and prevalence of disease but that the locus is not causally related to disease. A random sample of cases from this population will tend to have more members of subpopulation 1 than a similar

sample of controls. Thus an unstratified case-control study will give inflated estimates of the effect of the locus on disease. A stratified analysis is complicated by the fact that it is notoriously difficult to measure ethnicity. Family-based studies avoid this complication by comparing how often a given parental allele is inherited by a diseased offspring with that expected if inheritance was random. A disease-haplotype association study can only be carried out along the same lines if the multi-locus genotype information resolves which parental haplotypes were inherited.

Inferring haplotype phase structure is an important problem in statistical genetics and a number of approaches have been advanced that address the reconstruction of ambiguous haplotypes. Most of these methods treat the haplotypes as missing data and maximise an observed-data likelihood to infer haplotype frequencies by assuming these frequencies are in Hardy–Weinberg equilibrium (Excoffier & Slatkin, 1995). This assumption results in estimates and tests that are sensitive to the genetic structure of the sampled population, potentially leading to a biased estimator and incorrect inferences. As one of the key reasons for collecting family-based data is to maintain an invariance to such genetic structure assumptions, sensitivity to departures from Hardy–Weinberg equilibrium is quite undesirable. Though some work has been done in the context of testing (Horvath et al., 2004), to the best of our knowledge, little is known about how to derive estimators of haplotype effects that remain robust to haplotype frequency misspecification due to unphased genotype data.

In this work we draw upon a geometric approach to estimation in the presence of nuisance parameters (Bickel et al., 1993, § 2.4) and derive locally efficient tests and estimators of haplotype effects that are robust to misspecification of the haplotype frequency distribution. Our approach is closely related to work done by Rabinowitz (2002) in the context of a genotype association study with missing parental genotype data. Rabinowitz developed a score function that, under the null hypothesis, has zero expectation regardless of the distribution of the parental genotypes. He also proved that, in a neighbourhood of the true parameter value, this score function had the smallest variance among all score functions with this robustness property. Recently, Whittemore (2004) has extended Rabinowitz's result so that the score can be used for parameter estimation in addition to testing. Unlike the work of Rabinowitz and Whittemore, which addresses the effect of missing parental genotype data, here we focus on estimating haplotype effects in family-based studies with unphased genotype data. Our methods draw upon a geometric approach to estimation in the presence of nuisance parameters, allowing proofs that are general while at the same time giving insight into how related problems may be approached. To be specific, we use a projection argument to derive the efficient score, $\tilde{S}_\gamma$, for the haplotype association parameter γ assuming a saturated multinomial model for the distribution of parental haplotypes. Though we give $\tilde{S}_\gamma$ in explicit form, evaluation of $\tilde{S}_\gamma$ is complicated by the fact that the nuisance parameter, η, in the saturated multinomial model is both high dimensional and nonidentifiable. Thus, a standard parametric theory approach, in which one finds a consistent estimator of the saturated multinomial nuisance parameter and then uses this estimator to evaluate $\tilde{S}_\gamma$, is impossible. However, we show that $\tilde{S}_\gamma$ has mean zero when evaluated at any η in the multinomial parameter space $\mathscr{E}$. Hence, our approach is to estimate parameters characterising a lower-dimensional, identifiable, but possibly misspecified parametric parental haplotype model. These estimators are then used to map to a corresponding multinomial parameter value η_*. Since $\tilde{S}_\gamma$ has zero mean for all $\eta \in \mathscr{E}$, constructing $\tilde{S}_\gamma$ using η_* will lead to an unbiased estimating function, yielding consistent estimators of γ. Such robustness to misspecification

has been seen in some semiparametric estimation contexts; see for example van der Laan & Robins (2003, § 1.5) and Tsiatis & Ma (2004). In addition, if the correct model for the distribution of parental haplotypes is used, then this estimator will yield the minimum asymptotic variance among all such regular robust estimators. This last property results in what is termed a locally-efficient estimator (van der Laan & Robins, 2003, p. 21).

2. Preliminaries

We assume a case-parent design in which individuals with disease or binary trait of interest T_o are sampled and their parents are also recruited. Denote offspring genotype data by G_o, parental genotype data by G_p, combined genotype data by $G=(G_o, G_p)$, offspring haplotype data by H_o and parental haplotype data by H_p; realisations are denoted by g_o, g_p, g, h_o and h_p.

If haplotypes were directly observed, inference on haplotype relative risk parameters could be based on the conditional likelihood of offspring haplotypes given the haplotypes of the parents and the trait of the offspring. In this case the distribution of parental haplotypes would not need to be specified and the biology of the problem used to motivate reasonable models for $\mathrm{pr}(H_o=h_o|H_p=h_p, T_o=1)$. To be specific, if we assume no recombination, Mendelian inheritance and that H_p and $T_o|H_o$ are independent, the conditional distribution of offspring haplotypes given parental haplotypes and offspring trait, $\mathrm{pr}(H_o=h_o|H_p=h_p, T_o=1)$, can be written as

$$\frac{\mathrm{pr}(T_o=1|H_o=h_o)}{\sum_{h_o^*\in\mathscr{C}(h_p)}\mathrm{pr}(T_o=1|H_o=h_o^*)}, \tag{1}$$

where $\mathscr{C}(h_p)$ denotes the set of all offspring haplotypes consistent with h_p; for details see for example Sham (1998, § 4.7.2). If we take h_o' to be the reference haplotype pair and divide numerator and denominator of (1) by $\mathrm{pr}(T_o=1|H_o=h_o')$, we obtain

$$\frac{r(h_o)}{\sum_{h_o^*\in\mathscr{C}(h_p)}r(h_o^*)}, \tag{2}$$

where $r(h_o)$ is the risk conferred by h_o relative to h_o'. Models for $r(h_o)$ can be obtained by taking $\log\{r(h_o)\}=x^{\mathrm{T}}\gamma$ where x is a coded version of h_o. For example, a popular model is to take the jth component, x_j, to be the number of copies of haplotype j in h_o ($x_j=0, 1, 2$). This coding corresponds to the so-called 'multiplicative' genetic model in which the number of variant haplotypes has a multiplicative effect on $r(h_o)$. Note that in order for this model to be identifiable we need to code x so that $\log\{r(h_o')\}=0$.

Unfortunately, haplotypes are rarely observed directly and need to be inferred from multi-locus genotype data. In the terminology of Heitjan & Rubin (1991), an individual's genotype data is a coarsened version of their haplotype data. Note that the deterministic nature of the coarsening ensures that the coarsening process can be safely ignored and inference based on the coarsened-data likelihood

$$\mathscr{L}=\prod_{i=1}^{n}\left\{\sum_{h_o,h_p\in\mathscr{H}(G_{oi},G_{pi})}\mathrm{pr}(H_o=h_o|H_p=h_p, T_{oi}=1;\gamma)\,\mathrm{pr}(H_p=h_p|T_{oi}=1;\eta)\right\}, \tag{3}$$

where γ, q-dimensional, are the parameters of interest, η, r-dimensional, are nuisance parameters, $\mathscr{H}(G_o, G_p)$ is the set of all haplotypes consistent with G_o and G_p, and i indexes the n families in the study. We assume there are r haplotype combinations and that η

represents the r-dimensional vector of multinomial parameters, constrained to sum to 1, used to specify a saturated model for $\mathrm{pr}(H_p = h_p | T_o = 1)$. Denote the set of all possible η by $\mathscr{E}$. When considering a parametric family of models for $\mathrm{pr}(H_p = h_p | T_o = 1)$, indexed by a parameter ξ belonging to a lower-dimensional parameter space Ξ, we make the relationship between the elements of $\mathscr{E}$ and Ξ explicit by writing $\eta(\xi)$. Denote the observed data from the ith family by O_i, so that $O_i = (G_{oi}, G_{pi}, T_i)$, and let $\mathscr{L}_i(O_i; \gamma, \eta)$ be the ith family's contribution to the likelihood function (3).

The likelihood (3) now involves the distribution of parental haplotypes. In contrast to the related missing parental genotype problem in which it is relatively straightforward to specify the distribution of parental genotypes nonparametrically (Weinberg, 1999), correctly specifying models for $\mathrm{pr}(H_p = h_p | T_o = 1)$ is problematic. First, a saturated model for $\mathrm{pr}(H_p = h_p | T_o = 1)$ would involve a multinomial distribution with a large, although not infinite, number of parameters. For example, when all loci have the same number of alleles, the number of parameters needed is a^{4l}, where a is the number of alleles at each locus and l is the number of loci considered. This number can grow astronomically as a and/or l increase. In the realistic situation where a is 2 and l is 5, the number of required parameters is 1048 576, the dimensionality of $\mathscr{E}$, making saturated models computationally infeasible. In addition, such models will often be nonidentifiable. In the missing parental genotype scenario considered by Weinberg (1999), at least some parents are assumed to be observed, yielding information on their genotype configuration and allowing a saturated parental genotype model to be identified. However, in the haplotype estimation problem considered here, certain haplotype combinations will always result in heterozygous genotypes, and hence will always have uncertain haplotype phase. These haplotype configurations will always be 'missing', resulting in the nonidentifiability of the saturated model. Another approach is to assume that the correct model is contained within a parametric family of models in which the parameters are identifiable and their number is manageable. A common identifiable parametric model for $\mathrm{pr}(H_p = h_p | T_o = 1)$ is to assume random mating and that the parental haplotype distribution is in Hardy–Weinberg equilibrium. This model corresponds to the assumption that all four parental haplotypes are independent. For the example above, this model reduces the number of parameters to a far more tractable 31, the dimensionality of Ξ. Unfortunately, there is no reason to believe that the parental haplotypes would be in Hardy–Weinberg equilibrium. In fact, even when the larger population is in Hardy–Weinberg equilibrium at a locus that modifies disease risk, families with affected offspring would not necessarily be in Hardy–Weinberg equilibrium. Therefore, our approach is to find estimators of γ that are insensitive to misspecification of the distribution of parental haplotypes.

We adopt the geometric perspective to estimation in the presence of nuisance parameters presented in Bickel et al. (1993). Define the Hilbert space $\mathbb{H}$ of all measurable functions u that map a single family's observed data, O_i, to $\mathbb{R}^q$ and have mean zero. We take the inner product on $\mathbb{H}$ to be $\langle u_1, u_2 \rangle = E\{u_1(O)^{\mathrm{T}} u_2(O)\}$, where $u_1, u_2 \in \mathbb{H}$. Note that we will often suppress the subscript i or even O_i when it is clear that we are operating on functions of a single observation. The Hilbert space will depend on the probability distributions used to compute the expectations needed above. Different probability distributions will generate different Hilbert spaces. We begin our exposition assuming that the expectations taken above are with respect to the true distribution that generated the data. We will later relax this assumption and investigate the impact of misspecifying the distribution of parental haplotypes. The geometry of $\mathbb{H}$ will provide the framework for deriving estimating functions that yield robust estimators of γ.

Define the observed-data score functions of γ and η by $S_\gamma(O_i) = \partial \log(\mathscr{L}_i)/\partial\gamma$ and $S_\eta(O_i) = \partial \log(\mathscr{L}_i)/\partial\eta$ respectively. Unless otherwise stated, we assume that the score vectors are evaluated at the true parameter values γ_0 and η_0. Note that, since $\mathscr{L}_i$ is correctly specified, $S_\gamma \in \mathbb{H}$. As a result of the nonidentifiability of the saturated multinomial model the elements of $S_\eta(O_i)$ may not be linearly independent. However, for any real $q \times r$ matrix B, it is still true that $BS_\eta \in \mathbb{H}$ and $\Lambda_\eta = \{BS_\eta | B \text{ is any real } q \times r \text{ matrix}\}$ forms a closed linear subspace of $\mathbb{H}$. We shall refer to the space Λ_η as the nuisance tangent space for the saturated multinomial model or, simply, the nuisance tangent space. The projection theorem for Hilbert spaces, see e.g. Ash (1972, Theorem 3.2.11, p. 121), states that for each $u \in \mathbb{H}$ there exists a unique element of Λ_η, denoted by $\mathscr{P}(u|\Lambda_\eta)$, that is 'closest' to u and yields a residual, $u\text{-}\mathscr{P}(u|\Lambda_\eta)$, that is orthogonal to all $\lambda \in \Lambda_\eta$. Thus, even though the saturated multinomial model is nonidentifiable, the projection on to its nuisance tangent space is guaranteed to exist and be unique. The efficient score for γ, $\tilde{S}_\gamma$, is the residual of the observed-data score vector S_γ after projecting it on to the nuisance tangent space, that is $\tilde{S}_\gamma = S_\gamma - \mathscr{P}(S_\gamma|\Lambda_\eta)$ (Bickel et al., 1993, § 2.4).

We consider regular asymptotically linear estimators, $\hat{\gamma}_n$, of the parameter of interest γ. We restrict ourselves to regular estimators in the sense of Newey (1990). Asymptotically linear estimators satisfy

$$n^{\frac{1}{2}}(\hat{\gamma}_n - \gamma) = n^{-\frac{1}{2}} \sum_{i=1}^{n} \phi(O_i) + o_p(1),$$

where $\phi(O_i)$, for $i = 1, \ldots, n$, are independent, identically distributed zero-mean q-dimensional random vectors and $o_p(1)$ is a term that converges in probability to zero. The random vector $\phi(O_i)$ is called the ith influence function of $\hat{\gamma}_n$ and the variance of this influence function is equal to the asymptotic variance of $\hat{\gamma}_n$. Note that we define the asymptotic variance of $\hat{\gamma}_n$ as the variance of the limiting distribution of $n^{\frac{1}{2}}(\hat{\gamma}_n - \gamma)$. The efficient influence function is given by $\phi_\gamma(O_i) = (E[\{\tilde{S}_\gamma(O)\tilde{S}_\gamma(O)^{\mathrm{T}}\}])^{-1}\tilde{S}_\gamma(O_i)$. Since we are assuming here that $\mathscr{L}_i$ is correctly specified, the variance of $\phi_\gamma(O_i)$, and hence the asymptotic variance of $\hat{\gamma}_n$, will attain the efficiency bound, given by $(E[\{\tilde{S}_\gamma(O)\tilde{S}_\gamma(O)^{\mathrm{T}}\}])^{-1}$, for all estimators that make no assumption about the distribution of parental haplotypes. However, even when $\mathrm{pr}(H_p = h_p | T_o = 1; \eta)$ is misspecified, we will show that the projection procedure outlined here gives score functions for γ that have zero mean under the truth, resulting in consistent estimators of γ.

3. The efficient score

We begin by assuming that both $\mathrm{pr}(H_o = h_o | H_p = h_p, T_o = 1; \gamma)$ and $\mathrm{pr}(H_p = h_p | T_o = 1; \eta)$ are correctly specified and evaluated at the truth (γ_0, η_0). Let n_h be the total number of parental/offspring haplotype combinations, let n_{g_p} be the number of parental genotypes, and let n_g be the total number of parent/offspring genotype combinations. Note that, since the offspring's trait value is always one by design, the combined offspring-parent genotype data G will completely determine the observed data. Hence the observed data can take on one of n_g possible values. As above we assume that there are r possible parental haplotypes. Let h_{pk} denote the kth parental haplotype in some arbitrary ordering of parental haplotypes. Let o_j denote the jth possible value of O in some arbitrary ordering of observed outcomes.

Define the following matrices: ϕ is an $n_g \times r$ matrix with (j,k)th element given by

$$\mathrm{pr}\{H_p = h_{pk} | O = o_j\};$$

η is the r-dimensional vector with jth element given by $\mathrm{pr}\{H_p = h_{pj} | T_o = 1\}$; θ is the n_g-dimensional vector with jth element given by $\mathrm{pr}\{G = g_j | T_o = 1\}$; and $\mathbb{S}_\gamma$ is an $n_g \times q$ matrix with rows made up of the vectors $S_\gamma(o_j)$, for $j = 1, \ldots, n_g$. Also define $[\theta]$ and $[\eta]$ to be the diagonal matrices with diagonal elements given by the elements of θ and η respectively.

THEOREM 1. *The efficient score, $\tilde{S}_\gamma(o_j)$, corresponds to the jth row of*

$$\mathbb{S}_\gamma - \phi(\phi^{\mathrm{T}}[\theta]\phi)^{-}\phi^{\mathrm{T}}[\theta]\mathbb{S}_\gamma,$$

where $(\phi^{\mathrm{T}}[\theta]\phi)^{-}$ denotes a generalised inverse of $\phi^{\mathrm{T}}[\theta]\phi$.

Proof. We characterise the nuisance tangent space and project the observed data score function for γ on to it. To simplify the algebra, and without lost of generality, we assume that $q = 1$. Since the offspring's trait is always 1 by design, we suppress $T_o = 1$ from our notation. The saturated multinomial model for parental haplotypes is given by

$$\mathrm{pr}(H_p; \eta) = \prod_{j=1}^{r} \eta_j^{I\{H_p = h_{pj}\}},$$

where $\sum_{j=1}^{r} \eta_j = 1$ and $I(.)$ is an indicator function. Letting $\eta_r = 1 - \sum_{j=1}^{r-1} \eta_j$, we see that the score with respect to η_j $(j = 1, \ldots, r-1)$ is

$$\mathrm{pr}\{H_p = h_{pj} | O = o\}/\eta_j - \mathrm{pr}\{H_p = h_{pr} | O = o\}/\eta_r.$$

The nuisance tangent space for this saturated multinomial model is given by

$$\Lambda_\eta(o) = \{c^{\mathrm{T}} S_\eta(o) : c \text{ is any } (r-1)\text{-dimensional real vector}\}.$$

An element of Λ_η can be written as

$$\sum_{j=1}^{r-1} c_j [\mathrm{pr}\{H_p = h_{pj} | O = o\}/\eta_j - \mathrm{pr}\{H_p = h_{pr} | O = o\}/\eta_r]$$

$$= \sum_{j=1}^{r-1} c_j \,\mathrm{pr}\{H_p = h_{pj} | O = o\}/\eta_j - \left(\sum_{j=1}^{r-1} c_j\right) \mathrm{pr}\{H_p = h_{pr} | O = o\}/\eta_r.$$

If we let $b_j = c_j/\eta_j$ $(j = 1, \ldots, r-1)$ and $b_r = -(\sum_{j=1}^{r-1} c_j)/\eta_r$, the nuisance tangent space can be written in matrix form as

$$\Lambda_\eta = \left\{ \phi b : b \text{ is any } r\text{-dimensional real vector such that } \sum_{j=1}^{r} b_j \eta_j = 0 \right\}.$$

The problem of projecting $\mathbb{S}_\gamma$ on to Λ_η can then be phrased in terms of the weighted regression of $\mathbb{S}_\gamma$ on ϕ; that is we need to find b such that $(\mathbb{S}_\gamma - \phi b)^{\mathrm{T}}[\theta](\mathbb{S}_\gamma - \phi b)$ is minimised and $\sum_{j=1}^{r} b_j \eta_j = 0$. An unconstrained solution is given for any solution, in b, to the normal equations

$$\phi^{\mathrm{T}}[\theta]\phi b = \phi^{\mathrm{T}}[\theta]\mathbb{S}_\gamma; \tag{4}$$

for example, one solution is

$$b = (\phi^{\mathrm{T}}[\theta]\phi)^{-}\phi^{\mathrm{T}}[\theta]\mathbb{S}_\gamma.$$

However, as we show in the Appendix, this unconstrained solution satisfies the zero sum constraint. Therefore, the projection of $\mathbb{S}_\gamma$ on to Λ_η can be written as $\phi(\phi^{\mathrm{T}}[\theta]\phi)^{-}\phi^{\mathrm{T}}[\theta]\mathbb{S}_\gamma$. Since the efficient score $\tilde{S}_\gamma$ is the residual of $\mathbb{S}_\gamma$ after projecting on to Λ_η (Bickel et al., 1993, § 2.4, Proposition 1.A), the result follows. □

Theorem 1 assumed that both $\mathrm{pr}(H_o = h_o|H_p = h_p, T_o = 1; \gamma_0\}$ and $\mathrm{pr}(H_p = h_p|T_o = 1; \eta_0)$ were correctly specified. Next we explore the consequences of allowing the conditional distribution of parental haplotypes given the disease status of the child to be misspecified. To highlight the fact that this distribution is misspecified, we will write $\mathrm{pr}(H_p = h_p|T_o = 1; \eta_*)$, where η_* is a fixed value in $\mathscr{E}$. The matrices ϕ, $[\theta]$ and $\mathbb{S}_\gamma$, defined above, involve $\mathrm{pr}(H_p = h_p|T_o = 1; \eta_0)$ and so we denote their misspecified analogues by ϕ^*, $[\theta^*]$ and $\mathbb{S}^*_\gamma$. We will continue to assume that the conditional distribution of offspring haplotypes, given parental haplotypes and the affected status of the child, is generated at the truth with conditional probability $\mathrm{pr}(H_o = h_o|H_p = h_p, T_o = 1; \gamma_0)$. Taking the inner product with respect to this misspecified distribution will induce a Hilbert space $\mathbb{H}^*$ and nuisance tangent space $\Lambda_{\eta_*} \subset \mathbb{H}^*$. Our next result shows that, though Λ_{η_*} depends on the incorrectly specified probability $\mathrm{pr}(H_p = h_p|T_o = 1; \eta_*)$, the estimating function $\tilde{S}^*_\gamma$, obtained as the residual of the observed-data score vector $\mathbb{S}^*_\gamma$ after projecting it on to Λ_{η_*}, is robust to this misspecification.

Theorem 2. *The estimating function, $\tilde{S}^*_\gamma(o_j)$, corresponding to the jth row of $\mathbb{S}^*_\gamma - \phi^*(\phi^{*\mathrm{T}}[\theta^*]\phi^*)^{-}\phi^{*\mathrm{T}}[\theta^*]\mathbb{S}^*_\gamma$, has mean zero under the truth as long as the support of the misspecified distribution of parental haplotypes contains the support of the true distribution that generated the data.*

Proof. Again, without loss of generality assume that $q = 1$. In order to derive $\tilde{S}^*_\gamma(o)$ an argument entirely similar to the proof of Theorem 1 can be used. We make the dependence of $\tilde{S}^*_\gamma(o)$ on γ and η more explicit by writing $\tilde{S}^*_\gamma(o, \gamma, \eta)$. By construction, $\mathbb{S}^*_\gamma$ is orthogonal to the columns of ϕ^*; that is, for all $l = 1, \ldots, r$,

$$
\begin{aligned}
0 &= \sum_{j=1}^{n_g} \tilde{S}^*_\gamma(o_j, \gamma_0, \eta_*)\phi^*_{jl}\theta_j \\
&= \sum_{j=1}^{n_g} \tilde{S}^*_\gamma(o_j, \gamma_0, \eta_*)\,\mathrm{pr}\{H_p = h_{pl}|O = o_j; \gamma_0, \eta_*\}\,\mathrm{pr}(O = o_j; \gamma_0, \eta_*) \\
&= \sum_{j=1}^{n_g} \tilde{S}^*_\gamma(o_j, \gamma_0, \eta_*)\,\mathrm{pr}(O = o_j|H_p = h_{pl}; \gamma_0)\eta_{*l} \\
&= n_{*l} \sum_{j=1}^{n_g} \tilde{S}^*_\gamma(o_j, \gamma_0, \eta_*)\,\mathrm{pr}(O = o_j|H_p = h_{pl}; \gamma_0).
\end{aligned}
$$

Therefore, as long as the support of the misspecified distribution of parental haplotypes contains the support of the true distribution that generated the data, we have

$$
0 = \eta_{0l} \sum_{j=1}^{n_g} \tilde{S}^*_\gamma(o_j, \gamma_0, \eta_*)\,\mathrm{pr}(O = o_j|H_p = h_{pl}; \gamma_0),
$$

for all $l = 1, \ldots, r$. This implies that

$$
\sum_{j=1}^{n_g} \tilde{S}^*_\gamma(o_j, \gamma_0, \eta_*) \sum_{l=1}^{r} \mathrm{pr}(O = o_j|H_p = h_{pl}; \gamma_0)\eta_{0l} = \sum_{j=1}^{n_g} \tilde{S}^*_\gamma(o_j, \gamma_0, \eta_*)\,\mathrm{pr}(O = o_j; \gamma_0, \eta_0) = 0.
$$

□

4. Locally-efficient robust estimation

In the previous section we assumed that the efficient score was computed using known parameter values, γ and η. Here we consider the estimation of these parameters and propose an estimating function for γ motivated by the results of § 3. We assume that the model for the conditional probability $\mathrm{pr}(H_o = h_o | H_p = h_p, T_o = 1; \gamma)$ is correctly specified. We will consider a parametric family of distributions for $\mathrm{pr}(H_p = h_p | T_o = 1)$ indexed by the parameter vector $\xi \in \Xi$. Note that the induced multinomial parameter space $\mathscr{E}^* = \{\eta(\xi) : \xi \in \Xi\} \subset \mathscr{E}$ may not contain the truth η_0. We assume the weak condition that an estimator $\hat{\xi}_n$ is available such that $n^{\frac{1}{2}}(\hat{\xi}_n - \xi_*)$ is bounded in probability for some $\xi_* \in \Xi$ and take $\hat{\eta}_n = \eta(\hat{\xi}_n)$. Based on the results of § 3 we propose to estimate γ by solving

$$\sum_{i=1}^{n} \tilde{S}^*_\gamma(O_i, \gamma, \hat{\eta}_n) = 0, \tag{5}$$

and denote the resulting estimator by $\hat{\gamma}_n$. In the Appendix we argue that the estimator $\hat{\gamma}_n$ is consistent and that $n^{\frac{1}{2}}(\hat{\gamma}_n - \gamma_0)$ converges in distribution to a multivariate normal with mean zero and covariance matrix

$$[E\{D_\gamma(O_i, \gamma_0, \eta_*)\}]^{-1} E\{V_\gamma(O_i, \gamma_0, \eta_*)\} [E\{D_\gamma(O_i, \gamma_0, \eta_*)^{\mathrm{T}}\}]^{-1}, \tag{6}$$

where

$$D_\gamma(O_i, \gamma, \eta) = \frac{\partial}{\partial \gamma} \tilde{S}^*_\gamma(O_i, \gamma, \eta), \quad V_\gamma(O_i, \gamma, \eta) = \tilde{S}^*_\gamma(O_i, \gamma, \eta) \tilde{S}^*_\gamma(O_i, \gamma, \eta)^{\mathrm{T}}.$$

This covariance matrix (6) can be consistently estimated by

$$\{\bar{D}_{\gamma,n}(\hat{\gamma}_n, \hat{\eta}_n)\}^{-1} \bar{V}_{\gamma,n}(\hat{\gamma}_n, \hat{\eta}_n) \{\bar{D}_{\gamma,n}(\hat{\gamma}_n, \hat{\eta}_n)^{\mathrm{T}}\}^{-1}, \tag{7}$$

where

$$\bar{D}_{\gamma,n}(\gamma, \eta) = n^{-1} \sum_{i=1}^{n} \frac{\partial}{\partial \gamma} \tilde{S}^*_\gamma(O_i, \gamma, \eta), \quad \bar{V}_{\gamma,n}(\gamma, \eta) = n^{-1} \sum_{i=1}^{n} \tilde{S}^*_\gamma(O_i, \gamma, \eta) \tilde{S}^*_\gamma(O_i, \gamma, \eta)^{\mathrm{T}}.$$

The analytical derivation of the matrix of partial derivatives $\bar{D}_{\gamma,n}$ is straightforward though algebraically complicated. The analytic form of $\bar{D}_{\gamma,n}$ is available from A. S. Allen upon request. Alternatively, $\bar{D}_{\gamma,n}$ can be computed using numerical derivatives.

If the model for the parental haplotypes given the affected status of the offspring is correctly specified, so that $\eta_0 \in \mathscr{E}^*$, and $\hat{\eta}_n$ is a consistent estimator of the true value η_0 then $\tilde{S}^*_\gamma(O, \gamma_0, \eta_*)$ is the efficient score and $\hat{\gamma}_n$ has minimum variance among all estimators that make no assumption about the distribution of parental haplotypes. Thus $\hat{\gamma}_n$ is a locally-efficient, i.e. for $\eta_0 \in \mathscr{E}^*$, robust estimator of γ.

5. A multiplicative genetic model example

To illustrate the proposed methods, we consider an example based on the multiplicative genetic model presented in § 2. To be specific, the model for $\mathrm{pr}(H_o = h_o | H_p = h_p, T_o = 1; \gamma)$ is given by

$$\frac{r(h_o)}{\sum_{h_o^* \in \mathscr{C}(h_p)} r(h_o^*)},$$

where the relative risk, $r(h_o)$, is parameterised by $\log\{r(h_o)\} = x^{\mathrm{T}}\gamma$ with the jth component of x denoting the number of copies of haplotype j in h_o ($x_j = 0, 1, 2$). We report here a simulation designed to illustrate the robustness of the estimator $\hat{\gamma}_n$ to misspecification of the parental haplotype distribution. We assumed that the population was made up of two equal subpopulations and denote membership in a given subpopulation by Z. We generated biallelic genotypes for an affected offspring and their parents prospectively according to

$$\mathrm{pr}(T_o|H_o)\,\mathrm{pr}(H_o|H_p)\,\mathrm{pr}(H_p|Z)\,\mathrm{pr}(Z).$$

Thus, we generated each dataset by sampling the subpopulation, then sampling haplotypes given the subpopulation, then forming genotypes from the generated haplotypes, then randomly selecting one haplotype from each parent to be transmitted, and then randomly sampling the offspring trait value given the offspring's haplotype. Families in which the offspring was not affected were dropped from the dataset. The underlying haplotypes and subpopulation membership are considered unobserved and were not kept for analysis. Within each subpopulation, we assumed random mating and Hardy–Weinberg equilibrium, so that each parental haplotype could be sampled from a multinomial with cell probabilities given in Table 1. Offspring haplotypes were generated from parental haplotypes by randomly sampling one haplotype from each parent. Finally, the trait of the offspring was determined by assuming that those individuals with two copies of the reference haplotype (11) had a disease prevalence of 0·20 in the first subpopulation and 0·10 in the second. Other haplotype combinations had relative risks given by the multiplicative genetic model with $\gamma_0 = (\gamma_{00}, \gamma_{01}, \gamma_{10}, \gamma_{11}) = (-1, 1, 0, 0)$. Note that the parameter indexing the reference haplotype, that is γ_{11}, was set to zero for identifiability. A sample of 500 families was generated in each of 10 000 simulated datasets.

Table 1: *Simulation study. Haplotype frequencies by subpopulation*

Haplotype	Subpopulation 1	Subpopulation 2
00	0·05	0·45
01	0·45	0·05
10	0·45	0·05
11	0·05	0·45

In applying the method to each dataset, we correctly specified the model for $\mathrm{pr}(H_o = h_o|H_p = h_p, T_o = 1; \gamma)$ but incorrectly specified $\mathrm{pr}(H_p = h_p|T_o = 1; \eta)$ by assuming random mating and Hardy–Weinberg equilibrium for the study population as a whole. This model assumes that the parental haplotype distribution can be expressed as the product of marginal haplotype frequencies:

$$\mathrm{pr}[H_p = \{(h_{m1i}, h_{m2j}), (h_{f1k}, h_{f2l})\}|T_o = 1] = \xi_i\xi_j\xi_k\xi_l,$$

where (h_{m1i}, h_{m2j}) and (h_{f1k}, h_{f2l}) are the maternal and paternal haplotype pairs respectively. For each dataset we obtained the maximum likelihood estimate $\hat{\xi}_n$ under this misspecified model via an EM algorithm detailed in Sham (1998). The estimate, $\hat{\eta}_n = \eta(\hat{\xi}_n)$, was then used to find the solution, $\hat{\gamma}_n$, to equation (5). Although not necessary in the scenarios we considered, one should constrain the haplotype frequencies to be nonzero to ensure the conditions for Theorem 2. The covariance matrix of the estimator was estimated using the sandwich variance estimate (7), and approximate 95% confidence

intervals were constructed of the form estimate $\pm 1{\cdot}96$ times estimated standard error. In order to gauge the possible bias inherent in the misspecification of $\mathrm{pr}(H_p = h_p | T_o = 1; \eta)$, we also present an alternative estimator $\hat{\gamma}_n^o$ obtained as the root of the observed data score function. The results of this simulation experiment are summarised in Table 2.

Table 2. *Summary of simulation statistics for the estimation of log relative risk parameters by the efficient score based estimator, $\hat{\gamma}_n$, and the observed data score based estimator, $\hat{\gamma}_n^o$. True values of log relative risk parameters are given in parentheses*

		$\gamma_{00}(-1{\cdot}0)$	$\gamma_{01}(1{\cdot}0)$	$\gamma_{10}(0{\cdot}0)$
$\hat{\gamma}_n$	Mean	$-1{\cdot}024$	$1{\cdot}004$	$-0{\cdot}002$
	Empirical variance	$0{\cdot}045$	$0{\cdot}059$	$0{\cdot}062$
	Estimated variance	$0{\cdot}043$	$0{\cdot}058$	$0{\cdot}060$
	Empirical coverage	$0{\cdot}946$	$0{\cdot}948$	$0{\cdot}946$
	Mean squared error	$0{\cdot}045$	$0{\cdot}059$	$0{\cdot}062$
$\hat{\gamma}_n^o$	Mean	$-1{\cdot}231$	$0{\cdot}731$	$-0{\cdot}203$
	Empirical variance	$0{\cdot}060$	$0{\cdot}043$	$0{\cdot}048$
	Mean squared error	$0{\cdot}114$	$0{\cdot}116$	$0{\cdot}089$

The estimator based on the efficient score exhibits very little bias, good coverage characteristics and close agreement between the estimated and empirical variances. These results are particularly noteworthy given that the parental haplotype distribution used to compute $\hat{\gamma}_n$ was misspecified. In contrast, γ_n^o demonstrated sensitivity to this misspecification resulting in bias. It is of particular interest to note the amount of bias exhibited by $\hat{\gamma}_n^o$ given the relatively small amount of haplotype phase ambiguity present in the two-loci case.

6. Discussion

Though we developed the efficient score based estimator in the context of complete genotype data, the approach here can be generalised to include missing genotype data. If the missing genotype data can be assumed to be missing at random, one simply needs to expand the set of haplotypes that are consistent with the observed genotype data. We plan to incorporate this feature into our current program and make the software publicly available.

In forming the efficient score based estimator we projected the observed data score function $\mathbb{S}_\gamma^*$ on to the nuisance tangent space Λ_{η_*}. However, any q-dimensional function of the observed data and γ can be projected on to Λ_{η_*}. The residual of this function can then be used as an unbiased estimating function for γ and will be robust to misspecification of the distribution of parental haplotypes. This approach to generating unbiased estimating functions may be useful where it is difficult to specify the observed data score, in complex pedigrees for example, or when the score would involve nuisance parameters in addition to η. Of course, the resulting estimator would no longer be locally efficient even when the model for the parental haplotypes given the affected status of the offspring is correctly specified.

Robustness properties have been previously noted for projection-based estimators. For example, Jørgensen & Knudsen (2004) note that the asymptotic variance of the efficient score is the same regardless of whether it is evaluated at the true nuisance parameter or

at a consistent estimate. The estimator highlighted here has the additional robustness property of being consistent even when the estimator of the nuisance parameter is not.

It should also be possible, using the theoretical framework presented here, to develop a score function that incorporates environmental covariates into the risk model (2), allowing one to estimate haplotype-environment interaction effects. Such an extension would be a valuable tool in dissecting the aetiology of complex diseases.

Acknowledgement

The authors would like to thank the editor and the referees, whose valuable remarks led to significant improvements in the presentation. Andrew S. Allen's research was supported by a research career award from the National Heart, Lung, and Blood Institute of the U.S. National Institutes of Health.

Appendix

Discussion of equation (4)

To see that the unconstrained solution to (4) satisfies the zero sum constraint, let ψ be an $r \times n_g$ matrix with (j, k)th element given by $\mathrm{pr}\{G = g_k | H_p = h_{pj}, T_o = 1\}$, and note the following matrix forms of standard probability relationships:

$$[\theta]\phi = ([\eta]\psi)^{\mathrm{T}}, \tag{A1}$$

$$\eta^{\mathrm{T}}\psi = \theta^{\mathrm{T}}, \tag{A2}$$

$$\theta^{\mathrm{T}}\phi = \eta^{\mathrm{T}}. \tag{A3}$$

We can write the constraint in matrix form as

$$\begin{aligned}
\eta^{\mathrm{T}} b &= \theta^{\mathrm{T}}\phi b && \text{(from (A3))}\\
&= \eta^{\mathrm{T}}\psi\phi b && \text{(from (A2))}\\
&= \eta^{\mathrm{T}}[\eta]^{-1}[\eta]\psi\phi b\\
&= \eta^{\mathrm{T}}[\eta]^{-1}\phi^{\mathrm{T}}[\theta]\phi b && \text{(from (A1))}\\
&= \eta^{\mathrm{T}}[\eta]^{-1}\phi^{\mathrm{T}}[\theta]\mathbb{S}_\gamma && \text{(from (4))}\\
&= \eta^{\mathrm{T}}[\eta]^{-1}[\eta]\psi\mathbb{S}_\gamma && \text{(from (A1))}\\
&= \theta^{\mathrm{T}}\mathbb{S}_\gamma && \text{(from (A2))}\\
&= 0,
\end{aligned}$$

where the last equality results from the standard zero-mean property of the score function.

To argue the consistency of $\hat{\gamma}_n$, first note that $E\{\tilde{S}^*_\gamma(O, \gamma, \eta_*)\} = 0$ follows directly from Theorem 2. This, along with the invertibility of $E\{D_\gamma(O, \gamma, \eta)\}$ and technical smoothness conditions, in γ and η, on the estimating function $n^{-1}\sum_{i=1}^n \tilde{S}^*_\gamma(O_i, \gamma, \eta)$, its expectation $E\{\tilde{S}^*_\gamma(O, \gamma, \eta)\}$, and the derivatives $\bar{D}_{\gamma,n}(\gamma, \eta)$ and $E\{D_\gamma(O, \gamma, \eta)\}$, are sufficient conditions for invoking Foutz's theorem (Foutz, 1977). The smoothness and invertibility conditions depend on the particular models being considered. We assume that the models considered are sufficiently smooth that these technical conditions hold.

To show asymptotic normality, begin by expanding, with respect to γ, the estimating function (6) about the truth γ_0, keeping $\hat{\eta}_n$ fixed. This gives

$$n^{1/2}(\hat{\gamma}_n - \gamma_0) = \{-\bar{D}_{\gamma,n}(a_n, \hat{\eta}_n)\}^{-1} n^{-1/2} \sum_{i=1}^n \tilde{S}^*_\gamma(O_i, \gamma, \hat{\eta}_n), \tag{A4}$$

where a_n is a value between $\hat{\gamma}_n$ and γ_0. Expanding the $\tilde{S}^*_\gamma(O_i, \gamma, \hat{\eta}_n)$ on the left-hand side of (A4) with respect to $\hat{\eta}_n$, about η_*, shows that $n^{1/2}(\hat{\gamma}_n - \gamma_0)$ is equal to

$$\{-\bar{D}_{\gamma,n}(a_n, \hat{\eta}_n)\}^{-1} n^{-1/2} \sum_{i=1}^{n} \{\tilde{S}^*_\gamma(O_i, \gamma, \eta_*) + \bar{D}_{\eta,n}(\gamma_0, b_n) n^{1/2}(\hat{\eta}_n - \eta_*)\},$$

where $\bar{D}_{\eta,n}(\gamma, \eta)$ is the $q \times r$ matrix of partial derivatives

$$\bar{D}_{\eta,n}(\gamma, \eta) = n^{-1} \sum_{i=1}^{n} \frac{\partial}{\partial \eta} \tilde{S}^*_\gamma(O_i, \gamma, \eta),$$

and b_n is a value between $\hat{\eta}_n$ and η_*. We will assume that $\tilde{S}^*_\gamma(O_i, \gamma_0, \eta)$ is smooth enough in η so that $\bar{D}_{\eta,n}(\gamma_0, \eta)$ converges to

$$E\left\{\frac{\partial}{\partial \eta} \tilde{S}^*_\gamma(O_i, \gamma_0, \eta)\right\} \tag{A5}$$

uniformly for η in a neighbourhood of η_*. As long as the support of the true distribution of parental haplotypes that generated the data is contained in the support of the misspecified distribution, Theorem 2 implies that $E\{\tilde{S}^*_\gamma(O_i, \gamma_0, \eta)\} = 0$ for all η. Differentiating both sides with respect to η and interchanging expectation and differentiation implies that (A5) is equal to zero. Therefore, assuming that $n^{1/2}(\hat{\eta}_n - \eta_*)$ is bounded in probability, we have that

$$n^{1/2}(\hat{\gamma}_n - \gamma_0) = [E\{-D_\gamma(\gamma_0, \eta_*)\}]^{-1} n^{-1/2} \sum_{i=1}^{n} \tilde{S}^*_\gamma(O_i, \gamma_0, \eta_*) + o_p(1).$$

Thus $\phi(O_i) = [E\{-D_\gamma(\gamma_0, \eta_*)\}]^{-1} \tilde{S}^*_\gamma(O_i, \gamma_0, \eta_*)$ is the ith influence function of $\hat{\gamma}_n$. It follows that $\hat{\gamma}_n$ is asymptotically normal with asymptotic variance given by $E\{\phi(O)\phi(O)^{\mathrm{T}}\}$.

References

Ash, R. B. (1972). *Real Analysis and Probability*. Boston, MA: Academic Press.

Bickel, P. J., Klassen, C. A. J., Ritov, Y. & Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Baltimore, MD: The Johns Hopkins University Press.

Excoffier, L. & Slatkin, M. (1995). Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Molec. Biol. Evol.* **12**, 921–7.

Foutz, R. V. (1977). On the unique solution to the likelihood equations. *J. Am. Statist. Assoc.* **72**, 147–8.

Heitjan, D. F. & Rubin, D. B. (1991). Ignorability and coarse data. *Ann. Statist.* **19**, 2244–53.

Horvath, S., Xu, X., Lake, S. L., Silverman, E. K., Weiss, S. T. & Laird, N. M. (2004). Family-based tests for associating haplotypes with general phenotype data: Application to asthma genetics. *Genet. Epidemiol.* **26**, 61–9.

Jørgensen, B. & Knudsen, S. J. (2004). Parameter orthogonality and bias adjustment for estimating functions. *Scand. J. Statist.* **31**, 93–114.

Newey, W. K. (1990). Semiparametric efficiency bounds. *J. Appl. Economet.* **5**, 99–135.

Rabinowitz, D. (2002). Adjusting for population heterogeneity and misspecified haplotype frequencies when testing nonparametric null hypotheses in statistical genetics. *J. Am. Statist. Assoc.* **97**, 742–51.

Self, S. G., Longton, G., Kopecky, K. J. & Liang, K. Y. (1991). On estimating HLA/disease association with application to a study of aplastic-anemia. *Biometrics* **47**, 53–61.

Sham, P. (1998). *Statistics in Human Genetics*. London: Arnold.

Tsiatis, A. A. & Ma, Y. (2004). Locally efficient semiparametric estimators for functional measurement error models. *Biometrika* **91**, 835–48.

Van der Laan, M. J. & Robins, J. M. (2003). *Unified Methods for Censored Longitudinal Data and Causality*. New York: Springer-Verlag.

Weinberg, C. R. (1999). Allowing for missing parents in genetic studies of case-parent triads. *Am. J. Hum. Genet.* **64**, 1186–93.

Whittemore, A. (2004). Estimating genetic association parameters from family data. *Biometrika* **91**, 219–25.

[*Received March* 2004. *Revised March* 2005]