Am. J. Hum. Genet. 49:773–785, 1991

Regressive Logistic Models for Familial Diseases: A Formulation Assuming an Underlying Liability Model

Florence M. Demenais

Division of Biostatistics and Epidemiology, Howard University Cancer Center, Washington, DC

Summary

Statistical models have been developed to delineate the major-gene and non-major-gene factors accounting for the familial aggregation of complex diseases. The mixed model assumes an underlying liability to the disease, to which a major gene, a multifactorial component, and random environment contribute independently. Affection is defined by a threshold on the liability scale. The regressive logistic models assume that the logarithm of the odds of being affected is a linear function of major genotype, phenotypes of antecedents and other covariates. An equivalence between these two approaches cannot be derived analytically. I propose a formulation of the regressive logistic models on the supposition of an underlying liability model of disease. Relatives are assumed to have correlated liabilities to the disease; affected persons have liabilities exceeding an estimable threshold. Under the assumption that the correlation structure of the relatives' liabilities follows a regressive model, the regression coefficients on antecedents are expressed in terms of the relevant familial correlations. A parsimonious parameterization is a consequence of the assumed liability model, and a one-to-one correspondence with the parameters of the mixed model can be established. The logits, derived under the class A regressive model and under the class D regressive model, can be extended to include a large variety of patterns of family dependence, as well as gene-environment interactions.

Introduction

Statistical models have been developed to delineate the major-gene and non-major-gene factors accounting for the observed familial transmission of complex diseases. The classical mixed model assumes an underlying variable, the liability to the disease, to which a major gene, a polygenic component, and random environment contribute independently (Morton and MacLean 1974; Lalouel and Morton 1981; Lalouel et al. 1983). This model explains familial aggregation essentially in terms of genetic causation, although variants of this model allow for environmental causes of dependence. Affection is defined by a threshold on this liability scale. Variation of the morbid risk according to various epidemiological factors is treated by a shift of this threshold on the liability scale. Specific liability classes are assigned to the family members on the basis of environmental and demographic factors, and the corresponding morbid risk is computed prior to segregation analysis and is held fixed.

On the other hand, the regressive models recently introduced by Bonney (1984, 1986) are constructed by conditioning each individual's observation on those of his antecedents, using logistic regression for binary traits. The log of the odds of being affected is assumed to be a linear function of major genotype, the phenotypes of antecedents, and other covariates. These models merge the goals of epidemiology and genetics by allowing simultaneous estimation of major-gene factors, residual covariation of unspecified origin, and measured environmental factors influencing the trait. The parameters of the regressive logistic models can be constrained or not to satisfy the observed morbid risk(s) in the population.

In the case of continuous traits, the mixed and regressive models have been compared theoretically and numerically, through computer simulations, in nu-

Received December 28, 1990; revision received June 6, 1991.

Address for correspondence and reprints: Dr. Florence Demenais, Division of Biostatistics, Howard University Cancer Center, 2041 Georgia Avenue, N.W., Washington, D.C. 20060.

0002-9297/91/4904-0010$02.00

clear families (Demenais and Bonney 1989). A one-to-one correspondence between the parameters of the two models has been established. However, for binary traits, there is no exact equivalence between the threshold-mixed models and the regressive logistic models. In the present paper, I propose a formulation of the regressive logistic models on the supposition of an underlying liability threshold model of disease. Relatives are assumed to have correlated liabilities to the disease, with a correlation structure following the regressive-model patterns of dependence. Affected persons have liabilities exceeding an estimable threshold. The probability that a person is affected, given the affection status of his or her antecedents, is a logistic regression function in which the regression coefficients are expressed in terms of the relevant familial correlations among liabilities. A correspondence with the parameters of the mixed model can thus be established. I will first summarize the main features of the mixed and regressive logistic models as originally described but with another coding scheme I define below. I will then present the liability formulation of the regressive logistic models and will compare these different approaches. For convenience, the original and liability formulations of the regressive logistic models will be denoted formulations I and II, respectively.

Likelihood of a Nuclear Family

Let $Y = (Y_F, Y_M, Y_{C1}, Y_{C2}, Y_{Ci}, \ldots Y_{Cn})$ be the vector of disease status (affected/unaffected) of father (F), mother (M), and a set of n children in a nuclear family. The joint likelihood of the observed phenotypes can be written, in general, as

$$P(Y_F,Y_M,Y_1,..,Y_n) = P(Y_F)P(Y_M|Y_F)$$
$$P(Y_1|Y_F,Y_M)..P(Y_n|Y_F,Y_M,..,Y_{n-1}) \quad (1)$$

If it is assumed that an unobservable discrete factor, g ($g = 1,2...k$), affects the variability in the trait, the likelihood becomes $P(Y) = \Sigma P(g)P(Y|g)$, where $P(g)$ is the probability of the vector of discrete factors (g) and the sum is over all possible g vectors. In segregation analysis, which is aimed at detecting major-gene effects, this discrete factor is a major genotype or, more generally, ousiotpe, and Mendelian transmission can be tested against discrete but non-Mendelian transmission (for a discussion, see Bonney 1986). Each model makes it possible to write the joint likelihood of the observed phenotypes (see eq. [1]) by specifying some type of dependence among the observations, as presented in the following sections.

Mixed Models

The mixed model, developed for discrete traits (Morton and MacLean 1974), assumes an underlying variable, the liability to the disease (l), resulting from the independent and additive contributions of a diallelic (A/a) major-gene component g, a polygenic or multifactorial transmissible component c, and random environment e, so that $l = g + c + e$. Affection is defined by being above a threshold (T) on this liability scale, determined from the morbid risk to affection. The major gene is characterized by the frequency of allele A in the population and by the three genotype-specific means, μ_{AA}, μ_{Aa}, and μ_{aa} (the overall mean, μ, is zero). Conditionally on the vector of major genotypes, multivariate normality of the liability is assumed. The original model assumes that the residual correlations among relatives' liabilities are due to additive polygenic inheritance. Therefore, the spouses' liabilities are uncorrelated, and the residual parent-offspring and sib-sib correlations are each equal to one-half the residual polygenic heritability (proportion of residual variance due to additive genetic variation). To allow the sib-sib correlation to differ from the parent-offspring correlation for any reason, different multifactorial transmissible components in adults (c_A with variance σ^2_{CA}) and children (c_K with variance σ^2_{CK}) have been introduced by Lalouel and Morton (1981). The likelihood of a nuclear family is written by integrating over all possible values of the multifactorial components of the father (c_F) and of the mother (c_M):

$$L(Y_F,Y_M,Y_1,\ldots,Y_n) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(c_F)f(c_M)$$
$$\Sigma P(g_F)P(Y_F|g_F,c_F)\Sigma P(g_M)P(Y_M|g_M,c_M)\prod_i \Sigma P(g_i|g_F,g_M)$$
$$P(Y_i|g_i,c_F,c_M)dc_F dc_M . \quad (2)$$

The summations are over the unobserved g's. The $P(g)$'s are genotypic frequencies, and the $P(g_i|g_F,g_M)$'s are transition probabilities expressed as functions of three transmission probabilities (as defined by Elston and Stewart [1971]). The $f(.)$'s are normal density functions, and the $P(Y|g,c)$'s are penetrances computed from cumulative normal distributions (Lalouel and Morton 1981).

The likelihood is a function of the major-gene parameters expressed in terms of the gene frequency q, the displacement t ($t = \mu_{AA} - \mu_{aa}$), the degree of dominance d [$d = (\mu_{Aa} - \mu_{aa})/(\mu_{AA} - \mu_{aa})$] and, generally, the three transmission probabilities plus the residual polygenic variances, σ^2_{CA} and σ^2_{CK}.

Regressive Logistic Models (Formulation I)

The regressive logistic models proposed by Bonney (1986) specify a regression relationship between the probability for a person being affected and a set of explanatory variables including major genotype, phenotypes of older relatives, and other covariates. Several classes of models have been described according to the patterns of dependence among sibs. I will discuss here the class A and class D models, which have common features with the mixed model (Demenais and Bonney 1989); they both assume (in different senses) equal sib-sib correlations or dependencies. We assume that spouses are independent and that father-child and mother-child dependencies are equal, as in the mixed model, but these assumptions can be readily relaxed.

The class D model assumes that, given the parental outcomes and all genotypes, the outcomes of offspring are equally predictive and uses a likelihood based on the following decomposition:

$$\begin{aligned} L = {} & \Sigma P(g_F)P(Y_F|g_F)\Sigma P(g_M)P(Y_M|g_M) \\ & \Sigma P(g_1|g_F,g_M)P(Y_1|g_1,g_F,g_M,Y_F,Y_M)\ldots \\ & \Sigma P(g_n|g_F,g_M)P(Y_n|g_F,\ldots,g_n,Y_F,\ldots,Y_{n-1}) . \end{aligned} \quad (3)$$

To reduce the number of parameters in the regression for binary traits, Bonney (1986) assumed that the major genotype of a relative affects the trait value of an individual only through that individual's own major genotype. Thus, the likelihood becomes

$$\begin{aligned} L = {} & \Sigma P(g_F)P(Y_F|g_F)\Sigma P(g_M)P(Y_M|g_M)\prod_i \\ & \Sigma P(g_i|g_F,g_M)P(Y_i|g_i,Y_F,Y_M,Y_1,\ldots,Y_{i-1}) . \end{aligned} \quad (4)$$

The penetrances $P(Y|g)$ and $P(Y|g,Y_F\ldots Y_{i-1})$ are logistic functions. The logistic function is defined as $\exp(\theta Y)/[1+\exp(\theta)]$, where a person's phenotype Y is coded 1 for affected and 0 for unaffected and where θ is the logit $= \ln[\Pr(Y=1)/\Pr(Y=0)]$. In logistic regression, discrete explanatory variables can be coded in different ways, each coding scheme having its own interpretation. Since the phenotypes of a person's antecedents, Y_A, can belong to either one of the three categories—i.e., affected, unaffected, or missing—I propose to use two dummy variables, Z_{A1} and Z_{A2}, to code the antecedents' phenotypes. These variables, Z_{A1} and Z_{A2}, are elements of a column vector Z_A, so that, when primes are used to denote transposes,

$$\begin{aligned} Z_A &= (1\ 0)' \text{ if A is affected} \\ Z_A &= (0\ 1)' \text{ if A is unaffected} \\ Z_A &= (0\ 0)' \text{ if A is missing} , \end{aligned}$$

and the corresponding vector of regression coefficients is $\Gamma'_A = (\gamma_{A1},\gamma_{A2})$, where A = F, M, or C represents father, mother, or child, respectively.

Thus, the logits for the parents, first child, and ith child are

$$\begin{aligned} \theta_F = \theta_M &= \alpha_g\ (g = AA,Aa,aa) \\ \theta_1|g_1,Y_F,Y_M &= \alpha_{g1} + \Gamma'_P Z_F + \Gamma'_P Z_M\ (\Gamma'_F = \Gamma'_M = \Gamma'_P) \\ \theta_i|g_i,Y_F,Y_M,Y_1\ldots Y_{i-1} &= \alpha_{gi} + \Gamma'_P Z_F + \Gamma'_P Z_M + \Gamma'_C \Sigma Z_j . \end{aligned}$$

The summation with respect to j is over the older sibs of i, and the Γ vectors are $\Gamma'_P = (\gamma_{P1},\gamma_{P2})$ and $\Gamma'_C = (\gamma_{C1},\gamma_{C2})$. The parameters are interpreted as follows: α_g is the genotype-specific baseline risk of the disease (on the logit scale); if the person's father is affected, the risk of the disease is modified by γ_{P1}; if the father is unaffected, it is modified by γ_{P2}; and if the father is unknown, the risk remains unchanged. The other γ parameters are defined in a similar manner. Usually, a person's risk is expected to increase or decrease, according to whether his or her antecedent is affected or unaffected, respectively, but negative phenotypic correlations are possible. If the ith child has an affected father, an unaffected mother, and three preceding sibs, two affected and one unaffected, then the logit is

$$\theta_i|g_i,Y_F,Y_M,Y_1,Y_2,Y_3 = \alpha_{gi} + \gamma_{P1} + \gamma_{P2} + 2\gamma_{C1} + \gamma_{C2} .$$

The parameters of the class D model are the major gene parameters: q (allele A frequency); the three genotype-specific baseline parameters, α_{AA}, α_{Aa}, α_{aa} (and generally the three transmission probabilities); plus the four γ's, which are used to express the residual dependency on parents (γ_{P1},γ_{P2}) and on previous sibs (γ_{C1},γ_{C2}).

Bonney (1986) used the transformation $Z = 2Y - 1$ to code the phenotypes of antecedents, so that $Z =$

+ 1 if the antecedent is affected, Z = − 1 if the antecedent is unaffected, and Z = 0 if the antecedent is missing. This can be obtained here by setting $\gamma_{P2} = -\gamma_{P1}$ and $\gamma_{C2} = -\gamma_{C1}$. The coding scheme proposed by Bonney (1986) implies symmetry on the logit scale: the risk of the disease is increased or decreased by the same quantity (γ), whether a person's relative is affected or unaffected, respectively. It may lead to detection of a spurious major gene, as shown in simulation studies (Demenais et al. 1990*b*), since a person's risk may be modified differently according to the relative's disease status. For example, among relatives of affected persons, polygenic inheritance leads to a risk increase that is higher than the decrease of risk among relatives of unaffected persons, except when the frequency of the disease in the population is $\geqslant .50$ (when the frequency equals .50, the increase and decrease of risks are equal with opposite sign). Note that other coding schemes have been discussed by Bonney (1987).

If we assume that, given the outcomes of the parents, the outcomes of sibs are independent, then the logit for the ith child is simply

$$\theta_i|g_i, Y_F, Y_M = \alpha_{gi} + \Gamma_P' Z_F + \Gamma_P' Z_M ,$$

and the likelihood of a nuclear family reduces to

$$L = \Sigma P(g_F)P(Y_F|g_F)\,\Sigma P(g_M)P(Y_M|g_M)\prod_i \Sigma P(g_i|g_F,g_M)P(Y_i|g_i,Y_F,Y_M) . \quad (5)$$

This is the class A regressive model. The parameters are only the major-gene parameters and the two γ's for the residual dependency on parents (γ_{P1},γ_{P2}).

A correspondence between the parameters of the mixed model and those of the regressive logistic model can be established analytically only for the major-gene component, by equating the penetrances. The major-gene component is expressed in terms of gene frequency and genotype-specific means of the liability for the mixed model and in terms of gene frequency and genotype-specific baseline parameters for the regressive models. If we let F be the standard cumulative normal distribution, then the probabilities to be affected, given the major genotype, are

$P(Y = 1|g) = \lambda_g = F[(\mu_g - T)/\sigma]$ under the mixed model; g = AA, Aa, aa, and $\sigma^2 = 1 - [\Sigma P(g)\mu_g^2]$, where the $P(g)$'s are population genotypic frequencies and where the sum is over the major genotypes;

$\lambda_g = \exp(\alpha_g)/[1 + \exp(\alpha_g)]$ under the regressive models, so that $\alpha_g = \ln\left(F[(\mu_g - T)/\sigma]/\{1 - F[(\mu_g - T)/\sigma]\}\right)$.

The residual familial correlations are expressed differently in the two models: under the mixed model they are expressed by multifactorial transmissible components, specifying the correlations among the parent-offspring and sib-sib's liabilities (ρ_{PO},ρ_{SS}), and under the regressive models they are expressed in terms of regression coefficients on the phenotypes of parents (γ_{P1},γ_{P2}) and preceding sibs (γ_{C1},γ_{C2}). A correspondence between these parameters cannot be derived analytically. Comparisons can only be made numerically through computations of recurrence risks predicted by each model in various familial situations.

Regressive Logistic Models Based on an Underlying Liability Model (Formulation II)

This formulation assumes that an underlying liability to the disease is correlated among relatives. The correlation structure of liabilities follows the regressive model patterns of dependence. Affected persons have liabilities exceeding a threshold. A person's liability, given the affection status of antecedents, has mean and variance computed from regression theory applied to truncated distributions (Pearson 1903; Aitken 1934). Thus, the probability for that person to be affected is expressed in terms of correlation coefficients among relatives' liabilities. This approach was used by Mendell and Elston (1974) to compute recurrence risks under the polygenic model. The cumulative normal distribution is replaced here by the logistic function, so that the effects of measured environmental covariates can be more easily included in the penetrance function and estimated together with the other familial factors causing the disease.

To briefly summarize, in the general population a threshold T partitions the standardized normal distribution of liability, l, into two truncated distributions: affected (above the threshold) and unaffected (below the threshold). The probability for an individual (p) from the population to be affected is computed from the cumulative normal distribution:

$$P(l_p > T) = F(-T) = (1/\sqrt{2\pi}) \int_{-\infty}^{-T} \exp(-u^2/2)du .$$

Depending on his affection status, p will belong to

either one of the two truncated distributions of liability with the following means and variances:

$$\begin{aligned}
\text{Affected:}\quad & E(l|l_p>T) = d, \text{ and} \\
& V(l|l_p>T) = 1 - d(d-T) = 1 - k \\
\text{Unaffected:}\quad & E(l|l_p<T) = d', \text{ and} \\
& V(l|l_p<T) = 1 - d'(d'-T) = 1 - k',
\end{aligned}$$

where d and d' are as defined by Falconer (1965): $d = f(T)/F(-T)$, and $d' = -f(T)/F(T)$, with $f(T)$ being the standard normal density function at T. We let k, $= d(d-T)$ and $k' = d'(d'-T)$ be the coefficients of variance reduction in the selected populations. Note that Falconer (1965) assumed no reduction of the variance in the selected populations, whereas Mendell and Elston (1971, 1974) and Reich et al. (1972) relaxed that assumption, using Pearson's (1903) original results.

A relative, r, of individual p has a liability to the disease with mean and variance given by (see eqq. [A3] and [A4] in the Appendix):

$$\begin{aligned}
& E(l_r|l_p>T) = \rho_{pr}d, \text{ and} \\
& V(l_r|l_p>T) = 1 - \rho_{pr}^2 k \text{ if p is affected} \\
& E(l_r|l_p<T) = \rho_{pr}d', \text{ and} \\
& V(l_r|l_p<T) = 1 - \rho_{pr}^2 k' \text{ if p is unaffected},
\end{aligned}$$

where ρ_{pr} is the correlation between the liabilities of p and r. The probability that r is affected, given the affection status of p, is approximated from the cumulative normal distribution by using the adjusted threshold.

I replace the standard cumulative normal distribution by the logistic function $\exp(\theta)/[1+\exp(\theta)]$, θ being the logit. This logistic function is the cumulative distribution of the logistic density with variance $\pi^2/3$. Therefore,

$$F(x) = (1/\sqrt{2\pi}) \int_{-\infty}^{x} \exp(-u^2/2)du$$

is approximated by

$$G(x) = \exp(x\pi/\sqrt{3})/[1+\exp(x\pi/\sqrt{3})] .$$

It has been shown that for $-5<x<5$, the difference between $F(x)$ and $G(x)$ does not exceed .022 (Johnson and Kotz 1970). All previous quantities will be now defined on the logit scale. The threshold is replaced by the baseline parameter $\alpha \simeq -(\pi/\sqrt{3})T$. The means d and d' in the truncated distributions are replaced by $\delta \simeq (\pi/\sqrt{3})d$ and $\delta' \simeq (\pi/\sqrt{3})d' = -\exp(\alpha)\delta$, since $d' = -[F(-T)/F(T)]d$. The quantities k and k' become $\kappa = (3/\pi^2)\delta(\delta+\alpha)$ and $\kappa' = (3/\pi^2)\delta'(\delta'+\alpha)$. Although κ and κ' are equal to k and k', respectively, Greek letters will be used on the logit scale, since these quantities are now specified by α and δ. The logit for a random affected individual from the population is $\theta = \alpha$. I use the coding scheme proposed above, replacing the observed phenotype Y_P by the vector Z_P, in the regression

$$\begin{aligned}
Z_p &= (1\ 0)' \text{ if p is affected} \\
Z_p &= (0\ 1)' \text{ if p is unaffected} \\
Z_p &= (0\ 0)' \text{ if p is unknown} .
\end{aligned}$$

Thus, the logit for r, given the affection status of Y_p of p, is

$$\theta_r|Y_p = (\alpha + \rho_{pr}\boldsymbol{\Delta}'\mathbf{Z}_p)/(1-\rho_{pr}^2\mathbf{K}'\mathbf{Z}_p)^{1/2} ,$$

with

$$\boldsymbol{\Delta}' = (\delta,\delta') \text{ and } \mathbf{K}' = (\kappa,\kappa') .$$

Following the formulas given in the Appendix for the adjusted thresholds, under the class A regressive model and under the class D regressive model, the logits for each member of a nuclear family, conditional on antecedents, will be presented. Familial correlations will first be considered without a major-gene component, and then they will be considered together with a major-gene component.

Familial Correlations without a Major Gene

The class D model is characterized by equal sib-sib correlations of liabilities, and the likelihood of the observed phenotypes in a nuclear family is

$$\begin{aligned}
L = {} & P(Y_F)P(Y_M)P(Y_1|Y_F,Y_M)\ldots \\
& P(Y_n|Y_F,Y_M,Y_1,\ldots Y_{n-1}) , \qquad (6)
\end{aligned}$$

where all P's are logistic functions. For parents, the logits are simply $\theta_F = \theta_M = \alpha$, the baseline risk. The logit of the first child, given the observed phenotypes Y_F and Y_M, is regressed on the means of the parents' liabilities specified by their affection status; to be stan-

dardized, it is divided by the residual SD. Replacing d and d' by the vector $\Delta' = (\delta,\delta')$, k and k' by the vector $K' = (\kappa,\kappa')$, and the residual variance $v_{.i}$ by $w_{1(i)}$ in equations (A3) and (A4) of the Appendix, we find

$$\theta_1|Y_F,Y_M = (\alpha + \rho_{PO}\Delta_F'Z_F + \rho_{PO}\Delta_M'Z_M)/w_{1(1)}^{1/2},$$

and

$$w_{1(1)} = 1 - \rho_{PO}^2K_F'Z_F - \rho_{PO}^2K_M'Z_M, \qquad (6a)$$

with $0 \leqslant \rho_{PO}^2K_F'Z_F + \rho_{PO}^2K_M'Z_M<1$, where ρ_{PO} is the parent-offspring correlation of liabilities (the father-child and mother-child correlations are assumed to be equal: $\rho_{FO} = \rho_{MO} = \rho_{PO}$). Note that $\Delta_F' = \Delta_M' = \Delta'$ and $K_F' = K_M' = K'$. Also, note that $w_{1(1)}$ denotes the variance of the first child, given the phenotypes of antecedents to the first child (i.e., the parents). The logit of the ith child, given the disease status of antecedents, is similarly obtained by regressing recursively on the means of the liabilities of the parents and of each preceding sib. It is also standardized by dividing by the corresponding residual SD. We let Δ_j', K_j', $w_{j(i)}$, $w_{i(i)}$ ($j<i$), and υ_{ij} denote quantities which are conditional on the phenotypes of the respective antecedents, then we have, from equations (A6) and (A7) of the Appendix,

$$\theta_i|Y_F,Y_M,Y_1\ldots Y_{i-1} = (\alpha + \rho_{PO}\Delta_F'Z_F + \rho_{PO}\Delta_M'Z_M + \Sigma w_{j(i)}^{1/2}\upsilon_{ij}\Delta_j'Z_j)/w_{i(i)}^{1/2}, \qquad (6b)$$

where the summation with respect to j is over the older sibs of i, and

$$w_{i(i)} = (1 - \rho_{PO}^2K_F'Z_F - \rho_{PO}^2K_M'Z_M)\prod_j(1 - \upsilon_{ij}^2K_j'Z_j), \qquad (6c)$$

where the product is over j older sibs of i. The variance $w_{j(i)}$ in equation (6b) is the variance of the ith child, given the phenotypes of antecedents to the jth child; it is similar to $w_{i(i)}$, with the product being over the older sibs of j. Note that $w_{j(i)} = w_{j(j)}$, if the variances of the liability in all family members are equal, as assumed here. The vectors $\Delta_j' = (\delta_j,\delta_j')$ and $K_j' = (\kappa_j,\kappa_j')$ are completely specified by the logit θ_j, as Δ and K depend on α. The partial correlations, υ_{ij}'s, between the liabilities of i and j, given the phenotypes of antecedents, are computed recursively (see eqq. [A5], [A8], and [A9] in the Appendix). For example, for $j = 1$, $i = 2,\ldots,n$,

$$\upsilon_{i1} = (\rho_{SS} - \rho_{PO}^2K_F'Z_F - \rho_{PO}^2K_M'Z_M)/(1 - \rho_{PO}^2K_F'Z_F - \rho_{PO}^2K_M'Z_M); \qquad (6d)$$

and for $j = 2$, $i = 3,\ldots,n$,

$$\upsilon_{i2} = \upsilon_{i1}(1 - \upsilon_{i1}K_1'Z_1)/(1 - \upsilon_{i1}^2K_1'Z_1), \qquad (6e)$$

where ρ_{SS} is the sib-sib correlation of liabilities. Each υ_{ij}, when $j>2$, is obtained similarly. Thus, the parameters of the class D model with no major gene are the baseline parameter α, the parent-offspring correlation ρ_{PO}, and the sib-sib correlation ρ_{SS}.

If we assume that, given the parental affection status, the childen's liabilities are independent, then all partial correlations υ_{ij}'s = 0. The regression on preceding sibs reduces to the regression on parents only, so that, from equation (6a),

$$\theta_i|Y_F,Y_M = (\alpha + \rho_{PO}\Delta_F'Z_F + \rho_{PO}\Delta_M'Z_M)/w_{1(i)}^{1/2}$$

and

$$w_{1(i)} = w_{1(1)}.$$

This is the class A regressive model. The likelihood function takes the simpler form:

$$L = P(Y_F)P(Y_M)\prod_i P(Y_i|Y_F,Y_M). \qquad (7)$$

The parameters of the class A regressive model with no major gene are only α and ρ_{PO}. The sib-sib correlation ρ_{SS} is constrained to be equal to $\rho_{PO}^2K_F'Z_F + \rho_{PO}^2K_M'Z_M$. Note that, in the case of continuous traits (Bonney 1984; Demenais and Bonney 1989), ρ_{SS} was equal to $2\rho_{PO}^2$. It can be pointed out that the correlation pattern specified by the class A model has the following meaning: sibs are correlated only because they have common parents with certain characteristics, measured by the phenotype, the underlying mechanisms (genetic and/or environmental) causing this correlation being unknown. The sib-sib correlation changes according to the parental affection status, because of the truncation of the liability distribution.

Major-Gene Component and Residual Familial Correlations

For an autosomal diallelic major locus, the overall liability is a mixture of three distributions, each characterized by its own mean μ_g and residual variance σ^2 which is assumed to be equal in each genotypic distribution. Each distribution itself can be parti-

tioned into two truncated distributions (affected/unaffected) by its specific threshold: $T_g = (T-\mu_g)/\sigma$. T_g will correspond to the genotype-specific baseline parameter α_g on the logit scale ($\alpha_g \simeq -(\pi/\sqrt{3})T_g$).

Under the class D model, the likelihood of the observed phenotypes is

$$\begin{aligned} L = {} & \Sigma P(g_F)P(Y_F|g_F)\Sigma P(g_M)P(Y_M|g_M) \\ & \Sigma P(g_1|g_F,g_M)P(Y_1|g_1,g_F,g_M,Y_F,Y_M)\ldots \\ & \Sigma P(g_n|g_F,g_M)P(Y_n|g_F,\ldots g_n,Y_F,\ldots,Y_{n-1}), \end{aligned} \quad (8)$$

which, under the class A model, simplifies to

$$\begin{aligned} L = {} & \Sigma P(g_F)P(Y_F|g_F)\Sigma P(g_M)P(Y_M|g_M) \\ & \prod_i \Sigma P(g_i|g_F,g_M)P(Y_i|g_i,g_F,g_M,Y_F,Y_M). \end{aligned} \quad (9)$$

The penetrances depend now on both phenotypes and genotypes of antecedents. For parents, the logits are $\theta_{gF} = \theta_{gM} = \alpha_g$, the genotype-specific baseline parameters. The logit of the first child under class D (or of the ith child under class A), is

$$\begin{aligned} & \theta_1|g_1,g_F,g_M,Y_F,Y_M \\ & = (\alpha_{g1} + \rho_{PO}\Delta'_{gF}Z_F + \rho_{PO}\Delta'_{gM}Z_M)/w_{1(1)}^{1/2}, \end{aligned}$$

and

$$w_{1(1)} = w_{1(i)} = 1 - \rho_{PO}^2 K'_{gF}Z_F - \rho_{PO}^2 K'_{gM}Z_M$$

with $0 \leqslant \rho_{PO}^2 K'_{gF}Z_F + \rho_{PO}^2 K'_{gM}Z_M < 1$; the variance conditional on major genotype (σ^2) cancels out, as shown in the Appendix. The logits of subsequent children become

$$\begin{aligned} \theta_i|g_F..g_i,Y_F..Y_{i-1} = {} & (\alpha_{gi} + \rho_{PO}\Delta'_{gF}Z_F \\ & + \rho_{PO}\Delta'_{gM}Z_M + \Sigma w_{j(i)}^{1/2}\upsilon_{ij}\Delta'_{gj}Z_j)/w_{i(i)}^{1/2}, \end{aligned} \quad (10)$$

where the summation with respect to j is over the older sibs of i. The variances $w_{j(i)}$ and $w_{i(i)}$ and the partial correlations υ_{ij} are similar to those shown above (eqq. [6c–6e]) but depend now on the genotypes and phenotypes of antecedents. Again, the vectors Δ'_g and K'_g for parents depend on α_g, and Δ'_{gj} and K'_{gj} for jth child are specified by the logit θ_{gj}.

Therefore, the parameters of the class D model are the major-gene parameters—the gene frequency q and the three baselines α_{AA}, α_{Aa}, and α_{aa} (and, generally, the three transmission probabilities)—plus the residual parent-offspring (ρ_{PO}) and sib-sib (ρ_{SS}) correlations. Under the class A model, these parameters reduce to the major-gene parameters and to ρ_{PO}.

As can be seen from the above formula (eq. [10]) of the logit θ_i, where all quantities depend on the genotypes of antecedents (parents and preceding sibs), the Elston-Stewart algorithm (Elston and Stewart 1971), which replaces the sum of products by products of sums, cannot be applied to compute the likelihood (eq. [8]) under the class D model. Computation of this likelihood becomes extremely time consuming as the number of sibs increases. The performance of different approximations to allow use of the Elston-Stewart algorithm has been recently explored in the case of continuous traits through simulations (Demenais et al. 1990*a*). Approximation 6 was found to work appropriately in terms of estimation of parameters and hypothesis testing and will be used here. The genotype-specific mean for each preceding sib in the regression was replaced by a weighted mean, the weights being the probabilities of the possible genotypes, given the parental gentoypes and the sib's own phenotype. Thus, the genotype-specific vectors Δ'_{gj} and K'_{gj} will be replaced, respectively, by a vector of weighted means, $\overline{\Delta}'_j$, and a vector of weighted-coefficients-of-variance reduction, $\overline{K}'_j$:

$$\overline{\Delta}'_j = (\overline{\delta}_j, \overline{\delta}'_j) = (\Sigma P_{gj}\delta_{gj}, \Sigma P_{gj}\delta'_{gj})$$

with

$$P_{gj} = P(g_j|g_F,g_M,Y_j)$$

and

$$\begin{aligned} P_{gj} = {} & \{P(g|g_F,g_M)\exp(\alpha_g Y_j)/ \\ & [1+\exp(\alpha_g)]\}/ \\ & \{\Sigma P(g|g_F,g_M)\exp(\alpha_g Y_j)/[1+\exp(\alpha_g)]\}. \end{aligned}$$

$Y_j = 1$ or 0, depending on whether j is affected or not. Similarly,

$$\overline{K}'_j = (\overline{\kappa}_j, \overline{\kappa}'_j) = (\Sigma P_{gj}\kappa_{gj}, \Sigma P_{gj}\kappa'_{gj}).$$

The logit becomes

$$\begin{aligned} \theta_i|g_F,\ldots g_i,Y_F,\ldots Y_{i-1} = {} & (\alpha_{gi} + \rho_{PO}\Delta'_{gF}Z_F + \rho_{PO}\Delta'_{gM}Z_M \\ & + \Sigma w_{j(i)}^{1/2}\upsilon_{ij}\overline{\Delta}'_j Z_j)/w_{i(i)}^{1/2} \end{aligned}$$

with $\overline{K}'_j$ replacing $\overline{K}'_{gj}$ in the formulas for $w_{i(i)}$, $w_{j(i)}$, and υ_{ij}. The likelihood, under the class D model, can therefore be written as

$$L = \Sigma P(g_F)P(Y_F|g_F)\Sigma P(g_M)P(Y_M|g_M)\prod_i \Sigma P(g_i|g_F,g_M)P(Y_i|g_i,g_F,g_M,\overline{g_1},\ldots \overline{g_{i-1}},Y_F,Y_M,Y_1\ldots Y_{i-1}), \quad (11)$$

by using the Elston-Stewart algorithm.

A correspondence between the parameters of the mixed model and this liability formulation can now be established. For the major-gene component, the two models are comparable in terms of the gene frequency and penetrances, as shown above with the original formulation I of the regressive logistic models. For the residual familial correlations, there is a one-to-one correspondence between the heritabilities, H_A and H_K, of the mixed model and the residual correlations, ρ_{PO} and ρ_{SS}, of this liability formulation, as in the case of continuous traits (Demenais and Bonney 1989). The pure polygenic model ($\sigma^2_{CA} = \sigma^2_{CK} = \sigma^2_C = H\sigma^2_T$ with the total variance $\sigma^2_T = 1$ and with H being the usual heritability) and the class D model are equivalent if, in the class D model, we set $\rho_{PO} = \rho_{SS} = H/2$, with no major-gene component, and set $\rho_{PO} = \rho_{SS} = H/2\sigma^2$, with a major-gene component; σ^2 is the variance conditional on major genotype, as defined above. If we consider two different multifactorial components in adults and children (with respective heritabilities H_A and H_K), we have $\rho_{PO} = \sqrt{H_A}\sqrt{H_K}/2$ and $\rho_{SS} = H_K/2$, with no major gene, and have $\rho_{PO} = \sqrt{H_A}\sqrt{H_K}/2\sigma^2$ and $\rho_{SS} = H_K/2\sigma^2$, with a major gene. The polygenic and class A regressive models are equivalent in the parent-offspring correlation (ρ_{PO}). Note that, in the polygenic model, ρ_{PO} is interpreted as one-half of the heritability. The two models differ in terms of the sib-sib correlation, the difference being the portion of ρ_{SS} not explained by common parentage. The correspondence between the parameters of the different models is presented in table 1.

Table 1

Correspondence between Parameters of Mixed Model and Formulations I and II of Regressive Logistic Models

Parameter	Mixed Model	Regressive Model	
		I[a]	II
Major gene:			
Gene frequency	q	q	q
Penetrance of AA[b].....	λ_{AA}	λ_{AA}	λ_{AA}
Penetrance of Aa[b]	λ_{AA}	λ_{Aa}	λ_{Aa}
Penetrance of aa[b].....	λ_{aa}	λ_{aa}	λ_{aa}
Residual correlations:			
Parent-offspring.....	$\sqrt{H_A}\sqrt{H_K}/2\sigma^2$	γ_{P1},γ_{P2}	ρ_{PO}
Sib-sib	$H_K/2\sigma^2$	γ_{C1},γ_{C2}	ρ_{SS}

[a] Note that there is no direct mathematical correspondence between the γ parameters and the residual correlations of the other models.

[b] Computed from cumulative normal distributions and logistic functions under the mixed and regressive models, respectively (see text).

Missing Values

The problem of missing values can be handled in a manner similar to that which Bonney (in press) proposed for the case of continuous traits. If the affection status of an individual is not observed, the penetrance function for that individual will be equal to unity. Under the class D model, children with one unobserved parent will be regressed on the observed parent only. If both parents are missing, the logit for the first child will be equal to the baseline risk, and the correlation between the liability of the first child and that of subsequent sibs (υ_{i1} in eq. [6d]) will be simply ρ_{SS}. The logits for the other children will be obtained by regressing recursively on preceding sibs as above. For the class A model, when the parents' quantitative traits are missing, Bonney (in press) uses the class D likelihood formulation to account for the correlation among sibs. The regression coefficients on preceding sibs are specified to satisfy the constraint $\rho_{SS} = 2\rho^2_{PO}$. A problem arises here, since under the class A model, ρ_{SS} varies according to the parental affection status. Thus, in nuclear families where both parents are missing, we propose to use the class D formulation with the correlation υ_{i1} being set equal to a weighted average of all possible ρ_{SS} values specified by ρ_{PO} and the parents' affection status, the weight being the probability of a given parental mating type. Thus, when there is no major gene,

$$\upsilon_{i1} = \overline{\rho_{SS}} = \sum_{Y_F}\sum_{Y_M} P(Y_F)P(Y_M)[\rho^2_{PO}K'_F Z_F + \rho^2_{PO}K'_M Z_M],$$

and the sum for each parent is over the two possible types of affection status (unaffected [Y = 0] and affected [Y = 1]). When there is a major gene, a similar formula will be used for each parental genotypic combination:

$$\upsilon_{i1} = \overline{\rho_{SS}}$$

$$= \sum_{Y_F} \sum_{Y_M} P(Y_F|g_F)P(Y_M|g_M)[\rho^2_{FO}K'_{gF}Z_F + \rho^2_{FO}K'_{gM}Z_M] .$$

If only one parent is observed (e.g., the father), ρ_{SS} in the formula of υ_{i1} (eq. [6d] will be replaced by $\rho^2_{FO}K'_FZ_F + \Sigma P(Y_M)\rho^2_{FO}K'_MZ_M$, the sum being over the two possible types of affection status for the mother. Thus, after conditioning on the father is completed, the class D formulation for subsequent children will be applied. The proposed formulations for missing data will be evaluated by future simulation studies.

An Example

As an illustration of the proposed liability formulation of the regressive models, we consider the analysis of a simulated sample of 500 six-member nuclear families selected at random. Random sampling was chosen instead of sampling families through at least one affected, since our aim was to compare models and not to compare methods of ascertainment correction. The disease status was generated under a mixed model including a dominant major gene (q = .05, t = 2, d = 1) and polygenic variance σ^2_C = .32, corresponding to residual familial correlations $\rho_{PO} = \rho_{SS}$ = .25. A morbid risk of .10 was assumed. The analysis was done under the general class D model and its different subhypotheses, including the class A model (table 2). When there is no major-gene effect, a model specifying that sib-sib correlations depend only on common parentage (class A) is highly rejected vis-à-vis a model where the parent-offspring (ρ_{PO}) and sib-sib correlation (ρ_{SS}) are freely estimated (7 vs. 5; χ^2_1 = 31.67, $P < 10^{-9}$). A model with equal parent-offspring and sib-sib correlations, as specified by a pure polygenic model, fits the data as well as when then they are both estimated (6 vs. 5; χ^2_1 = 1.74, $P > .10$). Equality of these correlations will be assumed subsequently. Under a model including a major gene and residual familial correlations, the presence of a major gene is highly significant (6 vs. 3; χ^2_2 = 16.07, $P < .001$). I verified that, in the presence of a major gene, the hypothesis $\rho_{PO} = \rho_{SS}$ fitted the data well. The Mendelian transmission of the dominant major effect (3 vs. 1) is compatible with the data (χ^2_3 = 0.51, $P > .30$), and the absence of parent-offspring transmission of this effect is rejected (2 vs. 1; χ^2_2 = 12.03, $P < .01$). We should also note that, when a major gene is present, the residual familial correlations (ρ_{PO} and ρ_{SS}) are not significant (4 vs. 3; χ^2_1 = 0.77, $P > .30$), although the estimates of these parameters, $\rho_{PO} = \rho_{SS}$ = .22, are close to the generated values of .25. This may be due to a lack of power, especially when a discrete trait is considered. The parameter estimates of the major-gene component (model 3) are close to the generated values, q being estimated at .04 and penetrances λ_{AA} (= λ_{Aa}) and λ_{aa} being estimated at .79 and .04, respectively, when the true values are q = .05, $\lambda_{AA} = \lambda_{Aa}$ = .74, and λ_{aa} = .03. The estimate of the morbid risk in the population is .11, which is also close to the expected value of .10. Therefore, in this particular example, the liability formulation of the class D regressive model fits well the mixed model, as it is confirmed by our current simulations (Demenais et al. 1990*b*).

Discussion

In conclusion, the regressive logistic models can be formulated by assuming a liability threshold model of

Table 2

Segregation Analysis of Sample of 500 Six-Member Nuclear Families Simulated under Mixed Model, Including Dominant Major Gene (q = .05, $t = \mu_{AA}.\mu_{aa}$ = 2) and Residual Correlations ($\rho_{PO} = \rho_{SS}$ = .25)

Model	q	α_{AA}	α_{aa}	τ_{AAA}	τ_{AaA}	τ_{aaA}	ρ_{PO}	ρ_{SS}	-2ln$L$ + C
1. General transmission of major effect.....	.04	1.69	−2.95	1.00	.45	.00	.27	.27	1,825.71
2. No transmission of major effect.....	.14	−1.14	−2.63	.09	.09	.09	.49	.49	1,837.74
3. Mendelian with $\rho_{PO} = \rho_{SS}$.....	.04	1.31	−3.04	(1)	(.5)	(0)	.22	.22	1,826.22
4. Mendelian ($\rho_{PO} = \rho_{SS}$ = 0).....	.05	1.31	−3.28	(1)	(.5)	(0)	(0)	(0)	1,826.99
5. Familial correlation ($\rho_{PO} \neq \rho_{SS}$; q = 0)	(0)	−2.17	−2.17	...	...	...	.39	.44	1,840.55
6. Familial correlation ($\rho_{PO} = \rho_{SS}$; q = 0)	(0)	−2.17	−2.17	...	...	...	.39	.39	1,842.29
7. Familial correlation (class A; q = 0).....	(0)	−2.17	−2.17	...	...	...	.41	...	1,872.22

NOTE.—Values in parentheses are fixed under a given hypothesis. The τ values are transmission probabilities.

disease. Parsimony is a consequence of the assumed liability model. If the correlation structure of the relatives' liabilities follows a class D regressive model, the residual familial aggregation of the disease is expressed in terms of two correlations among the liabilities of parent-offspring (ρ_{PO}) and sib-sib (ρ_{SS}). Formulation I of the regressive logistic models, which considered the disease status only, used four regression coefficients to express the dependency on parents (γ_{P1},γ_{P2}) and on preceding sibs (γ_{C1},γ_{C2}). Although these γ parameters are not independent, a general relationship among them is difficult to derive analytically. However, parsimony could be achieved by specifying some types of relationships corresponding to particular situations. When the liability model is used, the probability of being affected is conditioned on both the phenotypes and genotypes of antecedents, as required by the multivariate normal assumption of the liability model. In order to keep the parameters at a reasonable number, formulation I assumed that the major genotype of a relative affects the trait value of an individual through that individual's own major genotype (Bonney 1986), so that conditioning was made on the phenotypes of antecedents only. However, this assumption can be relaxed by increasing the number of parameters. We should also note that, under the class D regressive model, the likelihood of a nuclear family may differ according to the ordering of sibs in the sibship, since the penetrance function will vary depending on the phenotypes of preceding sibs, affected or unaffected. Although the likelihood, under the liability model, is not affected by the order of sibs if computed exactly, the approximation used here takes into account an order for the sibs. However, this liability formulation (formulation II) appears to be numerically less sensitive to a given order than is formulation I. The regression on preceding sibs, under formulation II, is partly expressed in terms of the parent-offspring correlation, which is independent of the sibs' order, whereas the dependency on sibs, under formulation I, is a function of the γ_C parameters. The likelihood of a nuclear family with two affected children among four was computed for all possible positions of the affected in the sibship, under different genetic models. When the familial correlations are due to a generating polygenic component (heritablity .7), the relative change in log likelihood between the two extreme situations —i.e., two first sibs affected and two last sibs affected—is 1% and 0.3% under formulations I and II, respectively. This relative change is higher (1.7%) when the γ_C's are constrained to be equal to the γ_P's (i.e., when $\gamma_{C1} = \gamma_{P1}$ and when $\gamma_{C2} = \gamma_{P2}$). When the generating model includes a major gene and residual correlations, the relative change is also 1% under formulation I and is negligible (0.1%) under formulation II.

On the other hand, the liability formulation of the regressive models makes possible a one-to-one correspondence with the parameters of the mixed model. Furthermore, spouse correlation and unequal mother-child and father-child correlations can easily be accommodated in the class A regressive model and in the class D regressive model (see Appendix). The effects of measured environmental factors can be simultaneously estimated together with both the major-gene effect and residual familial covariation by adding regression coefficients for covariates in the logits. Other patterns of dependences, such as that of the class B model and that of the class C model, can also be specified (Bonney 1984), as can gene-environment interactions. Furthermore, the problems of variable age at onset and time-dependent covariates can be handled in a manner similar to that proposed by Abel and Bonney (1990).

The different penetrance functions have been derived here for nuclear families. However, the regressive models assume a Markov correlation structure across generations (Bonney 1984). Thus, given major genotypes, the liability of a person depends on those of ancestors only through the liabilities of the parents. In this case, the penetrance functions are applicable, without modification, to pedigrees. However, more complex patterns of dependence among liabilities in a pedigree could be considered.

In conclusion, the regressive models provide, in a computationally practical manner, a framework for understanding how genetic and environmental factors interact in the determination of complex diseases. The statistical properties of their different formulations are currently being assessed through simulation studies.

Acknowledgments

This work was supported in part by U.S. Public Service Research grant GM41885, by the Howard University Faculty Research Support Grant program, and by a grant from C.N.A.M.T.S.—I.N.S.E.R.M. We wish to thank George Bonney, Maria Martinez, Chris Amos, and two anonymous reviewers for their helpful comments on the draft.

Appendix

Calculation of the Adjusted Thresholds under the Regressive Models

I will first consider a model with no major gene and let L be a vector of normally distributed liabilities in parents (P) and a set of n offspring (O):

$$\mathbf{L} = \begin{bmatrix} L_P \\ L_O \end{bmatrix}$$

with

$$\text{mean } \mu = \begin{bmatrix} \mu_P\,(2\times 1) \\ \mu_O(n\times 1) \end{bmatrix}$$

and variance matrix

$$V = \left[\begin{array}{c|c} V_P(2\times 2) & V_{PO}\,(2\times n) \\ \hline V_{OP}\,(n\times 2) & V_O\,(n\times n) \end{array}\right].$$

Without loss of generality, each individual's liability in the population has mean 0 and variance 1. I consider that the correlation structure among the liabilities of relatives follows a class D regressive model which is characterized by equal sib-sib correlations (ρ_{SS}). I will also assume that the parents' liabilities are uncorrelated ($\rho_{FM} = 0$). The father-child and mother-child correlations are denoted by ρ_{FO} and ρ_{MO}, respectively. Thus,

$$\mu = \begin{bmatrix} 0 \\ 0 \\ \cdot \\ \cdot \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ 0 \end{bmatrix}$$

and

$$V = \left[\begin{array}{cc|ccc} 1 & 0 & \rho_{FO} & \rho_{FO} & \rho_{FO} \\ 0 & 1 & \rho_{MO} & \rho_{MO} & \rho_{MO} \\ \hline \rho_{FO} & \rho_{MO} & 1 & \rho_{SS} & \rho_{SS} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \rho_{FO} & \rho_{MO} & \rho_{SS} & \rho_{SS} & 1 \end{array}\right]$$

If we select the parents on the basis of their affection status, so that the L_P vector belongs to a subset of values Ω_P, we define:

$$E(L_P|L_P \varepsilon \Omega_P) = \mu_P^{\ddagger} \text{ and } \mathrm{Var}(L_P|L_P \varepsilon \Omega_P) = V_P^{\ddagger}\,.$$

As an example, let us assume that the father is affected and that the mother is unaffected ($\Omega = \{l_F > T, l_M < T\}$, T being the threshold); we have

$$\mu_P^{\ddagger} = \begin{bmatrix} d_F \\ d'_M \end{bmatrix}$$

and

$$V_P^{\ddagger} = \begin{bmatrix} 1-k_F & 0 \\ 0 & 1-k'_M \end{bmatrix}$$

with $d = E(l|l>T) = f(T)/F(-T)$ and $d' = E(l|l<T) = -f(T)/F(T)$, where f and F are the standard normal density function and cumulative normal distribution, respectively. $k = d(d-T)$ and $k' = d'(d'-T)$. The subscripts F and M are simply used to distinguish the father from the mother.

Then letting $\mu_{.O} = E(L_O|L_P \varepsilon \Omega_P)$ and $V_{.O} = \mathrm{Var}(L_O|L_P \varepsilon \Omega_P)$, Pearson (1903) and Aitken (1934) have shown that, provided that L_O is normally distributed in the selected populations,

$$\mu_{.O} = \mu_O + V_{OP}V_P^{-1}(\mu_P^{\ddagger} - \mu_P) \qquad \text{(A1)}$$

and

$$V_{.O} = V_O - V_{OP}(V_P^{-1} - V_P^{-1}V_P^{\ddagger}V_P^{-1})V_{PO}\,. \qquad \text{(A2)}$$

The rows of the matrix $V_{OP}V_P^{-1}$ give the multiple regression coefficients of the offspring's liabilities on the parents' liabilities (ρ_{FO},ρ_{MO}). The ith-row jth-column element of $V_{OP}(V_P^{-1} - V_P^{-1}V_P^{\dagger}V_P^{-1})V_{PO}$ is found to be $\rho_{FO}^2 k_F + \rho_{MO}^2 k'_M$, which is the same for any pair of sibs (i,j), since the father-child and mother-child correlations are assumed to be the same for all sibs.

From equations (A1) and (A2), the conditional mean and variance of the ith child's liability, given $l_F > T$ and $l_M < T$, are

$$\mu_{.i} = \rho_{FO} d_F + \rho_{MO} d'_M \qquad \text{(A3)}$$

and

$$v_{.i} = 1 - \rho^2_{FO} k_F - \rho^2_{MO} k'_M \text{ with}$$
$$0 \leqslant \rho_{FO}^2 k_F + \rho_{MO}^2 k'_M < 1 . \qquad \text{(A4)}$$

The conditional covariance of l_i and l_j, given $L_P \in \Omega$, is

$$v_{.ij} = \rho_{SS} - \rho_{FO}^2 k_F - \rho_{MO}^2 k'_M .$$

Therefore, the partial correlation of l_i and l_j, given $L_P \in \Omega$, is

$$\rho_{.ij} = \frac{\rho_{SS} - \rho_{FO}^2 k_F - \rho_{MO}^2 k'_M}{1 - \rho_{FO}^2 k_F - \rho_{MO}^2 k'_M} . \qquad \text{(A5)}$$

Class A models imply that $\rho_{.ij} = 0$, so that $\rho_{SS} = \rho_{FO}^2 k_F + \rho_{MO}^2 k'_M$. The adjusted threshold for the ith child is simply $T_{.i} = (T - \mu_{.i})/v_{.i}^{1/2}$ $(i = 1,2,\ldots,n)$.

For class D models, $\rho_{.ij} \neq 0$. The adjusted threshold for the first child, $T_{.1}$, is the same as that given above for the class A model. This threshold allows us to define a mean $d_1 = f(T_{.1})/F(-T_{.1})$ and a coefficient of variance reduction $k_1 = d_1(d_1 - T_{.1})$ in the distribution of affected children (and, alternatively, d'_1 and k'_1 in unaffected children). Given that the first child is affected, the conditional mean and variance of liabilities for the remaining sibs, when equations (A1) and (A2) are used, are

$$\mu_{..i} = \mu_{.i} + \rho_{.i1}(\mu_{.1}^{\dagger} - \mu_{.1})$$
$$= \rho_{FO} d_F + \rho_{MO} d'_M + v_{.i}^{1/2} \rho_{.i1} d_1$$
$$i = 2,3,\ldots,n \qquad \text{(A6)}$$

and

$$v_{..i} = v_{.i}(1 - \rho_{.i1}^2 k_1)$$
$$= (1 - \rho_{FO}^2 k_F - \rho_{MO}^2 k'_M)(1 - \rho_{.i1}^2 k_1) . \qquad \text{(A7)}$$

The adjusted threshold becomes $T_{..i} = (T - \mu_{..i})/v_{..i}^{1/2}$. Similarly, the subsequent children $(i = 3,4,\ldots n)$ will have mean and variance of liability adjusted on parents and on preceding sibs by following equations (A6) and (A7).

The partial correlations between susbsequent children are computed recursively as follows: given the phenotypes of a set of antecedents, the partial correlations between any pair of subsequent sibs are equal. The partial correlation between i and j ($j \geqslant 2$ and $i > j$), given the affection status of parents and of the first child ($l_F > T$, $l_M < T$, $l_1 > T_{.1}$), is

$$\rho_{..ij} = \frac{\rho_{.ij}(1 - \rho_{.ij} k_1)}{1 - \rho_{.ij}^2 k_1} \quad i = 3,4,\ldots,n;$$
$$j = 2,3,\ldots,n . \qquad \text{(A8)}$$

The partial correlation between i and j ($j \geqslant 3$ and $i > j$), given the affection status of parents and of two preceding children ($l_F > T$, $l_M < T$, $l_1 > T_{.1}$, $l_2 > T_{..2}$), is

$$\rho_{...ij} = \frac{\rho_{..ij}(1 - \rho_{..ij} k_2)}{1 - \rho_{..ij}^2 k_2} \quad i = 4,5,\ldots,n;$$
$$j = 3,4,\ldots,n . \qquad \text{(A9)}$$

Major-Gene and Residual Familial Correlations

When a major gene is included in the model, the threshold in each genotypic distribution is $T_g = (T - \mu_g)/\sigma$, where μ_g is the genotype-specific mean of liability and σ^2 is the variance conditional on major genotype (assumed equal in each genotypic distribution). From equations (A3) and (A4), the mean and variance of the liability of the ith child, given his parents' phenotypes and genotypes and his own genotype (g), become

$$\mu_{.i} = \mu_{gi} + \sigma \rho_{FO} d_{gF} + \sigma \rho_{MO} d'_{gM}$$

and

$$v_{.i} = \sigma^2(1 - \rho_{FO}^2 k_{gF} - \rho_{MO}^2 k'_{gM}) .$$

Therefore, the adjusted threshold for ith child is

$$T_{.i} = \frac{T - \mu_{.i}}{v_{.i}^{1/2}}$$
$$= \frac{1}{(1 - \rho_{FO}^2 k_{gF} - \rho_{MO}^2 k'_{gM})^{1/2}} [T_{gi} - \rho_{FO} d_{gF} - \rho_{MO} d'_{gM}] .$$

Since T_g corresponds to the baseline parameter α_g on the logit scale, the residual variance σ^2 cancels out and

will not be a parameter of the model. The adjusted logits for subsequent children are computed as above.

Familial Correlations Including a Spouse Correlation

When a spouse correlation (ρ_{FM}) is included in the model, the mean and variance of the liability of one spouse, e.g., the mother, given that the father is affected, become

$$\mu_{.M} = \rho_{FM} d_F$$

and

$$v_{.M} = 1 - \rho_{FM}^2 k_F .$$

The corresponding adjusted threshold for the mother is $T_{.M} = (T - \mu_{.M})/v_{.M}^{1/2}$. This threshold permits us to define a mean $d'_{.M} = -f(T_{.M})/F(T_{.M})$ and a coefficient of variance reduction $k'_{.M} = d'_{.M}(d'_{.M} - T_{.M})$ in the distribution of unaffected mothers (and, alternatively, $d_{.M}$ and $k_{.M}$ in affected mothers). Given that the father is affected, the children will have their mean and variance of liability equal to $\rho_{FO} d_F$ and $(1 - \rho_{FO}^2 k_F)$, respectively. Thus, given the affection status of both father and mother, the conditional mean and variance of the ith child's liability become

$$\mu_{.i} = \rho_{FO} d_F + \frac{\rho_{MO} - \rho_{FM}\rho_{FO} k_F}{(1 - \rho_{FM}^2 k_F)^{1/2}} d'_{.M}$$

and

$$v_{.i} = 1 - \rho_{FO}^2 k_F - [(\rho_{MO} - \rho_{FM}\rho_{FO} k_F)^2 / (1 - \rho_{FM}^2 k_F)] k'_{.M} ,$$

and the partial correlation between the liabilities of the ith and jth children, given the parents' affection status, is

$$\rho_{.ij} = \frac{\rho_{SS} - \rho_{FO}^2 k_F - [(\rho_{MO} - \rho_{FM}\rho_{FO} k_F)^2/(1 - \rho_{FM}^2 k_F)] k'_{.M}}{1 - \rho_{FO}^2 k_F - [(\rho_{MO} - \rho_{FM}\rho_{FO} k_F)^2/(1 - \rho_{FM}^2 k_F)] k'_{.M}} .$$

The class A model will imply $\rho_{.ij} = 0$, imposing the corresponding constraint on ρ_{SS}. Under the class D model, the same formulas as given above can be used to compute the regression on preceding siblings.

References

Abel L, Bonney GE (1990) A time-dependent logistic hazard function for modeling variable age of onset in analysis of familial diseases. Genet Epidemiol 7:391–407

Aitken (1934) Notes on selection from a multivariate normal population. Proc Edinb Math Soc [B] 4:106–110

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

——— (1987) Logistic regression for dependent binary observations. Biometrics 43:951–973

———. Compound regressive models for family data. Hum Hered (in press)

Demenais FM, Bonney GE (1989) Equivalence of the mixed and regressive models for genetic analysis. I. Continuous traits. Genet Epidemiol 6:597–617

Demenais F, Murigande C, Bonney GE (1990*a*) Search for faster methods of fitting the regressive models to quantitative traits. Genet Epidemiol 7:319–334

Demenais FM, Laing AE, Bonney GE (1990*b*) The fit of the logistic regressive models to the mixed model in segregation analysis of discrete traits. Am J Hum Genet 47 [Suppl]: A132

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

Falconer DS (1965) The inheritance of liability to certain diseases, estimated from the incidence among relatives. Ann Hum Genet 29:51–76

Johnson NL, Kotz S (eds) (1970) Distribution in statistics, vol 2: Continuous univariate distributions. Houghton Mifflin, Boston

Lalouel JM, Morton NE (1981) Complex segregation analysis with pointers. Hum Hered 31:312–321

Lalouel JM, Rao DC, Morton NE, Elston RC (1983) A unified model for complex segregation analysis. Am J Hum Genet 35:816–826

Mendell N, Elston RC (1971) Use of the tetrachoric correlation coefficient in the estimation of heritability of quasi-continuous traits. Biometrics 27:483–484

——— (1974) Multifactorial qualitative traits: genetic analysis and prediction of recurrence risks. Biometrics 30:41–57

Morton NE, MacLean CJ (1974) Analysis of family resemblance. III. Complex segregation analysis of quantitative traits. Am J Hum Genet 26:489–503

Pearson K (1903) On the influence of natural selection on the variability and correlation of organs. Philos Trans R Soc Lond [A] 200:1–66

Reich T, James JW, Morris A (1972) The use of multiple thresholds in determining the mode of transmission of semi-continuous traits. Ann Hum Genet 36:163–183