Linkage mapping for complex traits : a regression-based approach

Lebrec, J.J.P.

Chapter 2

Score Test for Detecting Linkage to Complex Traits in Selected Samples

Abstract

We present a unified approach to selection and linkage analysis of selected samples, for both quantitative and dichotomous complex traits. It is based on the score test for the variance attributable to the trait locus and applies to general pedigrees. The method is equivalent to regressing excess IBD sharing on a function of the traits. It is shown that, when population parameters for the trait are known, such inversion does not entail any loss of information. For dichotomous traits, pairs of pedigree members of different phenotypic nature (e.g. affected sib pairs and discordant sib pairs) can easily be combined as well as populations with different trait prevalences.

This chapter has been published as: J. Lebrec, H. Putter and J.C. van Houwelingen (2004). Score Test for Detecting Linkage to Complex Traits in Selected Samples. *Genetic Epidemiology* **6** (2), 97–108.

2.1 Introduction

In complex traits where the effect of each contributing locus is very small, the sample sizes needed to carry out linkage analysis usually result in costs far beyond research budgets, even when using new high throughput genotyping technologies [Risch, 2000]. Geneticists have been aware of this fact for a while and many designs and selection strategies have been proposed [Risch and Zhang, 1995; Dolan and Boomsma, 1998a; Purcell et al., 2001]. In the search for genes, prior to any linkage study, researchers usually gather evidence of heritability for the trait of interest. This is often done in twin studies including both monozygotic and dizygotic twins from the general population. In addition to heritability of the trait, these studies provide precise population marginal means, variability and twin-twin correlation estimates for the trait of interest.

Complex traits have small locus effect and this is probably why the search for the corresponding susceptibility loci has proved so disappointing. However this is also the reason why a score test constitutes a promising testing strategy in this context since it has local optimality properties [Cox and Hinkley, 1974]. In this article, using the variance components framework we give a general formulation for a score test to detect linkage to a putative quantitative trait locus under selective sampling based on the trait values of the pedigree members. We give simple formulae for the test in a number of commonly used designs (sibships and nuclear families of arbitrary size). Using a liability threshold model, we extend our results to dichotomous traits. In particular, they apply to sib pair designs where different types of pairs (e.g. affected and discordant sib pairs) can be combined in an optimal way, and subpopulations with different disease prevalences can be incorporated in a straightforward manner. Our approach provides a unified framework in which both optimal selection and subsequent analysis are combined in a natural way, both for quantitative and dichotomous traits.

2.2 Score test for quantitative traits in selected samples

Model

Our starting point is the variance components model, where we assume that $\mathbf{x} = (x_1, \ldots, x_m)'$, the vector of phenotypes of the pedigree members, has been standardized so that it has mean vector 0 and variances equal to 1. The $m \times m$ matrix $\boldsymbol{\pi}$ contains the identity-by-descent (IBD) information at a marker, more precisely $[\boldsymbol{\pi}]_{jk} = \pi_{jk}$ is the proportion of alleles shared IBD by pedigree members j and k. For now, we assume that the marker map is fully informative, the consequences of relaxing this assumption will be examined in Section 2.6. The variance components model specifies that the conditional distribution of the standardized $\mathbf{x}$ given IBD information $\boldsymbol{\pi}$ follows a normal distribution with zero mean and variance-covariance matrix $\boldsymbol{\Sigma}$ given by

$$[\boldsymbol{\Sigma}]_{jk} = \begin{cases} a^2 + c^2 + e^2 \;=\; 1 \;, & \text{if } j = k \;, \\ (\pi_{jk} - \mathbf{E}\pi_{jk})q^2 + (\mathbf{E}\pi_{jk})a^2 + c^2 \;, & \text{if } j \neq k \;. \end{cases}$$

where a^2 denotes the total additive genetic variance, c^2, the common-environment variance and e^2, the residual variance. This parameterization of the problem was initially introduced by Tang and Siegmund [2001] and is crucial to the obtention of simple results. For the time being we will assume absence of any dominance component of variance. We show an extension incorporating dominance variance in section 2.4. Since the trait values are standardized to unit variance, these variance components can also be interpreted as proportions of variance explained by the appropriate components. The total additive genetic variance a^2 includes both additive polygenic variance and the (additive) variance q^2 attributable to the putative quantitative trait locus (QTL). The factor $\mathbf{E}\pi_{jk}$ denotes the expected proportion of alleles shared identical by descent between pedigree members j and k; it is determined solely by the family relationship between j and k and equals twice the kinship coefficient between j and k.

The key parameter in this model is the variance component q^2 determining the presence of linkage (no linkage is equivalent to $q^2 = 0$). It is the only unknown parameter in the model and we shall denote it by γ in the sequel. Two important

properties of the variance components model are: that $\mathbf{x}$ and $\boldsymbol{\pi}$ are independent under the hypothesis of no linkage ($\gamma = 0$) and that the marginal distribution of $\boldsymbol{\pi}$ does not depend on γ.

Score test for quantitative traits

A score test for detecting linkage to quantitative traits in random samples for general pedigrees was given by Putter et al. [2002] and by Wang [2002]. Here we extend those results to a sampling scheme where data are selected based on phenotypic values. We generalize results obtained by Tang and Siegmund [2001] for sibships to arbitrary pedigrees and use the continuous case as a building block to the dichotomous case as exposed in Section 2.5.

The following expression for the score function $\ell_\gamma^{\mathbf{x}}$ in the variance components model is obtained in the appendix:

$$\ell_\gamma^{\mathbf{x}} = \frac{1}{2} \operatorname{tr}\left(\boldsymbol{\Sigma}^{-1}(\boldsymbol{\pi} - \mathbf{E}\boldsymbol{\pi})(\boldsymbol{\Sigma}^{-1}\mathbf{x}\mathbf{x}' - \mathbf{I})\right) \ .$$

Here $\operatorname{tr}(A)$ stands for the trace (sum of the diagonal elements) of matrix A. Using elementary matrix theory, in particular $\operatorname{tr}(AB) = \operatorname{tr}(BA)$ and $\operatorname{tr}(AB) = \operatorname{vec}(A')'\operatorname{vec}(B)$ (here $\operatorname{vec}(A)$ places the n columns of the $m \times n$ matrix A into a vector of dimension $mn \times 1$), this score function can be rewritten as

$$\ell_\gamma^{\mathbf{x}} = \frac{1}{2} \operatorname{vec}(\mathbf{C})'\operatorname{vec}(\boldsymbol{\pi} - \mathbf{E}\boldsymbol{\pi}) \tag{2.1}$$

with $\mathbf{C} = \boldsymbol{\Sigma}^{-1}\mathbf{x}\left(\boldsymbol{\Sigma}^{-1}\mathbf{x}\right)' - \boldsymbol{\Sigma}^{-1}$. Note that the $\boldsymbol{\pi} - \mathbf{E}\boldsymbol{\pi}$ matrix has all diagonal elements equal to 0.

For selected samples, the conditional distribution of IBD sharing $\boldsymbol{\pi}$ given the trait values $\mathbf{x}$ gives a natural framework for testing linkage [Sham et al., 2000; Dudoit and Speed, 2000] and we shall refer to this setting as the *selection model*. It turns out that the score function for this *selection model*, and for the *joint model* of $\mathbf{x}$ and $\boldsymbol{\pi}$ remains the same. As we show below, this is true for any joint model of $\mathbf{x}$ and $\boldsymbol{\pi}$ under the following general conditions, which are satisfied for the variance components model:

1. $\mathbf{x}$ and $\boldsymbol{\pi}$ are independent at $\gamma = 0$ and
2. the marginal distribution of $\boldsymbol{\pi}$ does not depend on γ.

We now turn to the proof of our previous statement regarding the equality of the scores for the selection model and the joint model. We denote the conditional distribution of $\mathbf{x} \,|\, \boldsymbol{\pi}$ and $\boldsymbol{\pi} \,|\, \mathbf{x}$ by $f_\gamma(\mathbf{x} \,|\, \boldsymbol{\pi})$ and $f_\gamma(\boldsymbol{\pi} \,|\, \mathbf{x})$ respectively, and the joint distribution of $\mathbf{x}$ and $\boldsymbol{\pi}$ by $f_\gamma(\mathbf{x}, \boldsymbol{\pi})$. The subscript γ expresses the dependence of those distributions on γ. The marginal distributions of $\mathbf{x}$ and $\boldsymbol{\pi}$ are denoted by $f_\gamma(\mathbf{x})$ and $f(\boldsymbol{\pi})$ respectively. With this notation, the score function for γ in the $\mathbf{x} \,|\, \boldsymbol{\pi}$ model is denoted by $\ell_\gamma^{\mathbf{x}}$, so $\ell_\gamma^{\mathbf{x}} = \frac{\partial}{\partial \gamma} \log f_\gamma(\mathbf{x} \,|\, \boldsymbol{\pi})$; and in the selection model by $\ell_\gamma^{\boldsymbol{\pi}}$, so $\ell_\gamma^{\boldsymbol{\pi}} = \frac{\partial}{\partial \gamma} \log f_\gamma(\boldsymbol{\pi} \,|\, \mathbf{x})$. By Bayes' rule, we have

$$f_\gamma(\boldsymbol{\pi} \,|\, \mathbf{x}) = \frac{f_\gamma(\mathbf{x}, \boldsymbol{\pi})}{f_\gamma(\mathbf{x})} = \frac{f_\gamma(\mathbf{x} \,|\, \boldsymbol{\pi})\ f(\boldsymbol{\pi})}{\int f_\gamma(\mathbf{x} \,|\, \boldsymbol{\pi})\ f(\boldsymbol{\pi})\ d\boldsymbol{\pi}}\ . \tag{2.2}$$

As a result,

$$\begin{aligned} \ell_\gamma^{\boldsymbol{\pi}} = &\ \frac{\partial}{\partial \gamma} \log f_\gamma(\mathbf{x} \,|\, \boldsymbol{\pi}) - \frac{\partial}{\partial \gamma} \log \left(\int f_\gamma(\mathbf{x} \,|\, \boldsymbol{\pi}) f(\boldsymbol{\pi})\ d\boldsymbol{\pi} \right) \\ = &\ \ell_\gamma^{\mathbf{x}} - \frac{\partial}{\partial \gamma} \log \left(\int f_\gamma(\mathbf{x} \,|\, \boldsymbol{\pi}) f(\boldsymbol{\pi})\ d\boldsymbol{\pi} \right)\ . \end{aligned} \tag{2.3}$$

For the score test for linkage in selected samples, we need this score function evaluated at $\gamma = 0$. Since score functions have mean 0, the second term $\frac{\partial}{\partial \gamma} \log \left(\int f_\gamma(\mathbf{x} \,|\, \boldsymbol{\pi}) f(\boldsymbol{\pi})\ d\boldsymbol{\pi} \right)$ equals the expectation of $\ell_\gamma^{\mathbf{x}}$ under $\boldsymbol{\pi} \,|\, \mathbf{x}$ evaluated at $\gamma = 0$. Since $\mathbf{x}$ and $\boldsymbol{\pi}$ are independent at $\gamma = 0$, this is just the distribution $\boldsymbol{\pi}$ (independent of γ). As a result we obtain,

$$\ell_\gamma^{\boldsymbol{\pi}} = \ell_\gamma^{\mathbf{x}} - \mathbf{E}_{\boldsymbol{\pi}} \ell_\gamma^{\mathbf{x}}\ .$$

Hence, in our case $\ell_\gamma^{\boldsymbol{\pi}} = \ell_\gamma^{\mathbf{x}}$, since $\ell_\gamma^{\mathbf{x}}$ is already, due to the parameterization used, centered with respect to the distribution of $\boldsymbol{\pi}$. The score $\ell_\gamma^{\mathbf{x}}$ is also centered with respect to the distribution of $\mathbf{x}$. Looking back at equation (2.2), we see that the score function for γ in the joint model of $\mathbf{x}$ and $\boldsymbol{\pi}$ also equals $\ell_\gamma^{\mathbf{x}} = \ell_\gamma^{\boldsymbol{\pi}}$. This has the important consequence that there is no loss of information by basing inference only on the conditional distribution of $\mathbf{x} \,|\, \boldsymbol{\pi}$ for random samples, or only on the distribution of $\boldsymbol{\pi} \,|\, \mathbf{x}$, the selection model for selected samples.

Fisher's information $\mathcal{I}_\gamma^{\boldsymbol{\pi}} = \mathbf{E}\left(-\frac{\partial^2}{\partial \gamma^2} \log f_\gamma(\boldsymbol{\pi} \,|\, \mathbf{x}) \right)$ for γ in the selection model is also the variance of the score function $\mathrm{var}_{\boldsymbol{\pi}}(\ell_\gamma^{\boldsymbol{\pi}})$ and is thus given by

$$\mathcal{I}_\gamma^{\boldsymbol{\pi}} = \frac{1}{4}\ \mathrm{vec}(\mathbf{C})'\ \mathrm{var}_{\boldsymbol{\pi}}\left(\mathrm{vec}(\boldsymbol{\pi})\right)\ \mathrm{vec}(\mathbf{C})\ . \tag{2.4}$$

The exact calculation of $\mathrm{var}_{\boldsymbol{\pi}}(\mathrm{vec}(\boldsymbol{\pi}))$ involves enumeration of all joint probabilities $\mathbf{P}(\pi_{ij}, \pi_{kl})$ for each possible inheritance vector in a pedigree. In practice, this is efficiently achieved through the use of the `--ibd` and `--matrices` options in the `MERLIN` software [Abecasis et al., 2002] with a pedigree file describing the appropriate pedigree structure and one marker with all values as missing. Note that under the assumption of complete IBD information, Fisher's information as given in Formula (2.4) can be directly used as a criterion for selection of the most informative individuals based on trait values.

The score test statistic z is formed by adding the scores from independent pedigrees and dividing by the square root of its variance under the null hypothesis:

$$z = \frac{\sum_i \ell^{\boldsymbol{\pi}}_{\gamma,i}}{\sqrt{\sum_i \mathcal{I}^{\boldsymbol{\pi}}_{\gamma,i}}} \ . \tag{2.5}$$

Under the null hypothesis of no linkage, z has asymptotically a standard normal distribution. The test is one-sided, only positive values of z being regarded as evidence for linkage. In other words, z_+^2 defined as being equal to 0 if $z \leq 0$ and to z^2 if $z > 0$ is asymptotically distributed as $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$.

Formulae (2.1) and (2.4) provide an interpretation of this score test in terms of regression. Similar to Sham et al. [2002], the numerator of the score test statistic z can be interpreted as an estimate of the slope of the regression through the origin of excess IBD sharing on a function of the trait values. The dependent variables are the observed excess IBD sharing between all $\frac{m(m-1)}{2}$ pairs of members in pedigree of size m while corresponding observations of the explanatory variable are quadratic functions of the original trait values as defined above. Those results are applicable to general pedigrees but take a very simple and appealing form in sib pairs and some other specialized cases as shown below. The slope estimate of the score test statistic is standardized by the square root of Fisher's information, but this standardization can also be interpreted as the standard error of the slope estimate of the numerator under the null hypothesis.

2.3 Special designs

In this section we give explicit formulae for the score test in general sibships and nuclear families. The interpretation of the test in terms of regression for sib pairs provides interesting insight into the relation of our method with the so called Haseman-Elston regressions and helps us understand why these optimal methods for random samples turn out to be sub-optimal when data are subject to selection unless modified as in Sham and Purcell [2001]. We refer the reader to Skatkiewicz et al. [2003]; Cuenco et al. [2003] for a comprehensive review and numerical comparison of methods for selected sib pairs.

Sibships

In a **sibship of size** m consisting of m siblings, $\mathbf{\Sigma}$ is given by

$$[\mathbf{\Sigma}]_{jk} = \begin{cases} 1 & \text{if } j = k \\ (\pi_{jk} - \frac{1}{2})\gamma + \frac{1}{2}a^2 + c^2 & \text{if } j \neq k \ . \end{cases} \tag{2.6}$$

Hence, for $\gamma = 0$, with $\rho = \frac{1}{2}a^2 + c^2$,

$$\mathbf{\Sigma} = (1-\rho)\mathbf{I} + \rho\mathbf{J} \ \text{ so } \ \mathbf{\Sigma}^{-1} = \frac{1}{1-\rho}\left(\mathbf{I} - \omega_m\mathbf{J}\right) \ , \tag{2.7}$$

with $\omega_m = \frac{\rho}{1+(m-1)\rho}$ where $\mathbf{I}$ is the $m \times m$ identity matrix and $\mathbf{J}$ is the $m \times m$ matrix whose elements are all equal to 1. It can be shown mathematically that the elements of the matrix $\mathbf{C} = \mathbf{\Sigma}^{-1}\mathbf{x}\left(\mathbf{\Sigma}^{-1}\mathbf{x}\right)' - \mathbf{\Sigma}^{-1}$ are given by

$$C_{ij} = \frac{1}{(1-\rho)^2}\left(x_i x_j - m\omega_m\bar{\mathbf{x}}(x_i + x_j) + (m\omega_m\bar{\mathbf{x}})^2\right) + \frac{1}{1-\rho}\omega_m \ . \tag{2.8}$$

Under the assumption of perfect marker information, the IBD distributions are uncorrelated for sib pairs within a sibship and have mean $\frac{1}{2}$, the score function is thus given by

$$\ell_\gamma^{\boldsymbol{\pi}} = \sum_{1 \le i < j \le m} C_{ij} \ \left(\pi_{ij} - \frac{1}{2}\right)$$

and Fisher's information by

$$\mathcal{I}_\gamma^{\boldsymbol{\pi}} = \frac{1}{8} \sum_{1 \le i < j \le m} C_{ij}^2 \ .$$

In **sib pair** designs, the two by two covariance matrix $\mathbf{\Sigma}$ is given by

$$\begin{pmatrix} 1 & \gamma(\pi - \frac{1}{2}) + \rho \\ \gamma(\pi - \frac{1}{2}) + \rho & 1 \end{pmatrix} .$$

The score function and information in $\gamma = 0$ are

$$\begin{aligned} \ell^{\pi}_{\gamma}(x_1, x_2; \rho) &= (\pi - \frac{1}{2})\ C(x_1, x_2; \rho) \\ \mathcal{I}^{\pi}_{\gamma}(x_1, x_2; \rho) &= \frac{1}{8}\ C^2(x_1, x_2; \rho) \end{aligned}$$

where

$$C(x_1, x_2; \rho) = \frac{(1+\rho^2)x_1x_2 - \rho(x_1^2 + x_2^2) + \rho(1-\rho^2)}{(1-\rho^2)^2} .$$

The score test in a sample of n independent sib pairs with phenotypes $(x_{i1}, x_{i2})_{i=1,\ldots,n}$ is given by

$$\frac{\sum_{i=1}^{n} \left(\pi_i - \frac{1}{2}\right)\ C(x_{i1}, x_{i2}; \rho)}{\sqrt{\frac{1}{8}\sum_{i=1}^{n} C^2(x_{i1}, x_{i2}; \rho)}}$$

and its robust version by

$$\frac{\sum_{i=1}^{n} (\pi_i - \frac{1}{2})\ C(x_{i1}, x_{i2}; \rho)}{\sqrt{\sum_{i=1}^{n} \left(\pi_i - \frac{1}{2}\right)^2\ C^2(x_{i1}, x_{i2}; \rho)}} .$$

The score test in that instance simply is the regression of the excess IBD sharing $\pi - \frac{1}{2}$ on a function of the trait values $C(\mathbf{x}; \rho)$ through the origin. This method was already proposed by Tang and Siegmund [2001] and Sham and Purcell [2001]. In a recent numerical comparison of methods for selected samples, Skatkiewicz et al. [2003] and Cuenco et al. [2003] showed that it has good properties in finite samples for extreme proband ascertained sib pairs and discordant sib pairs designs. The same test was also motivated heuristically using an approximation for excess IBD sharing in Putter et al. [2003].

In selected samples, one crucial feature of this regression as far as power is concerned, is that it is constrained through the origin. Indeed, the variance of the slope estimate in an unconstrained regression, which is inversely proportional to $\sum_i (C_i - \bar{C})^2 = \sum_i C_i^2 - n\bar{C}^2$, will always be greater than its constrained version, whose variance is inversely proportional to $\sum_i C_i^2$. The contour plot of C is displayed in Figure 2.1 for $\rho = 0.2$ and $\rho = 0.5$, with the corresponding trait values density indicated in gray scale (the density plots were generated using the scatterplots function

of Eilers and Goeman [2004]). It clearly shows that extreme concordant sib pairs have moderately large positive C values whereas extremely discordant sib pairs have large negative C values. As long as sib pairs are selected so that $\bar{C}$ is close to 0, whether the regression is constrained through the origin or not is irrelevant. However, should one consider only extremely discordant pairs, then $\bar{C}$ is negative and the power can increase dramatically, when using methods for selected samples.

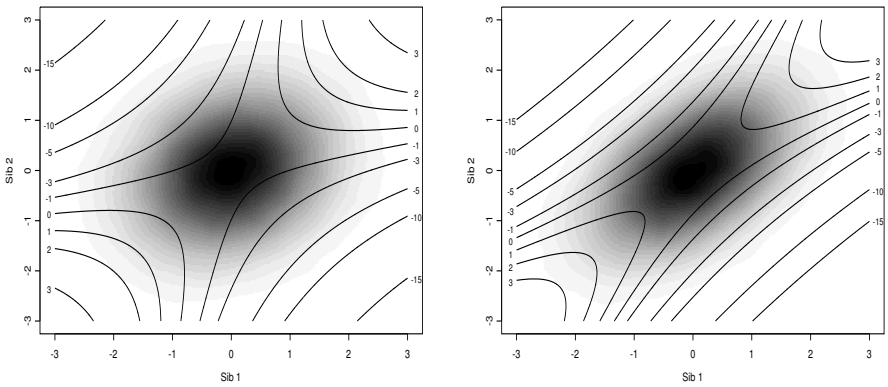

Figure 2.1: Joint distribution of sib trait values $\mathbf{x}$ (gray scale) and contour plot of $C(\mathbf{x}, \rho)$ ($\rho = 0.2$, left panel and $\rho = 0.5$, right panel)

Nuclear families

We now consider a general **nuclear family** with m sibs with trait value vector $\mathbf{x}_s$ and two parents with trait value vector $\mathbf{x}_p$, then the variance-covariance matrix $\mathbf{\Sigma}$ can be partitioned as

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{ss} & \mathbf{\Sigma}_{sp} \\ \mathbf{\Sigma}_{ps} & \mathbf{\Sigma}_{pp} \end{pmatrix} .$$

The sib-sib submatrix $\mathbf{\Sigma}_{ss}$ is the only submatrix to contain the linkage parameter γ. At $\gamma = 0$, $\mathbf{\Sigma}_{ss}$ is the same as (2.6) and (2.7) with ρ replaced by $\rho_{ss} = \frac{1}{2}a^2 + c^2$. The other submatrices are given by $\mathbf{\Sigma}_{sp} = \mathbf{\Sigma}'_{ps} = \rho_{sp}\mathbf{J}_{m2}$ and $\mathbf{\Sigma}_{pp} = (1 - \rho_{pp})\mathbf{I}_2 + \rho_{pp}\mathbf{J}_{22}$. Here, $\mathbf{I}_m$ is the identity matrix of dimension m and $\mathbf{J}_{ml}$ is the matrix of dimension $m \times l$ with all elements equal to 1. The parameter ρ_{sp} denotes the parent-sib trait

correlation and ρ_{pp} the father-mother trait correlation, both of which are assumed to be known. The correlations ρ_{ss}, ρ_{sp} and ρ_{pp} are given by 0.5, 0.5 and 0 times the additive genetic variance respectively, plus a scalar times the common environment variance. For ρ_{ss}, this multiplication factor will be 1 but we allow for smaller and mutually different factors for ρ_{sp} and ρ_{pp}. Matrices $\mathbf{\Sigma}_{sp}$ and $\mathbf{\Sigma}_{pp}$ do not involve the linkage parameter γ because there is no variation in IBD sharing between sibs and parents, nor between the two parents assuming they do not share alleles identical by descent. In practice however, parents are often genotyped because they are helpful in determining the IBD sharing of the siblings. With those conventions and using a similar reasoning as in (2.2) and (2.3), one can show that the score function for γ in the $\boldsymbol{\pi} \,|\, \mathbf{x}_p, \mathbf{x}_s$ model equals the score function for γ in the $\mathbf{x}_s \,|\, \boldsymbol{\pi}, \mathbf{x}_p$ model; in other words, the parents' phenotypes can simply be considered as 'covariates' in the analysis. Now, using standard results on conditional normal distributions, it turns out that

$$\mathbf{x}_s \,|\, \boldsymbol{\pi}, \mathbf{x}_p \;\sim\; N\left(\beta \bar{\mathbf{x}}_p,\; \mathbf{\Sigma}_{ss} - \rho_{sp}\beta \mathbf{J}_{mm}\right) \text{ with } \beta = \frac{2\rho_{sp}}{1+\rho_{pp}} \;,$$

thus

$$\left(\mathbf{x}_s - \beta \bar{\mathbf{x}}_p\right) \;/\; \left(1 - \rho_{sp}\beta\right)^{1/2} \,|\, \boldsymbol{\pi}, \mathbf{x}_p \;\sim\; N\left(0, \mathbf{\Sigma}_C\right) \;,$$

where $\mathbf{\Sigma}_C$ has diagonal elements equal to 1 and off-diagonal elements equal to

$$\left((\pi_{jk} - \frac{1}{2})\gamma + \rho_{ss} - \rho_{sp}\beta \right) \;/\; \left(1 - \rho_{sp}\beta\right) \;.$$

Finally, the score obtains as

$$\ell_\gamma^{\boldsymbol{\pi}} = \left(1 - \rho_{sp}\beta\right)^{-1} \sum_{1 \le i < j \le m} C_{ij} \left(\pi_{ij} - \frac{1}{2} \right)$$

and the information as

$$\mathcal{I}_\gamma^{\boldsymbol{\pi}} = \left(1 - \rho_{sp}\beta\right)^{-2} \frac{1}{8} \sum_{1 \le i < j \le m} C_{ij}^2 \;,$$

with C_{ij} given by formula (2.8) with $\mathbf{x} = \left(\mathbf{x}_s - \beta \bar{\mathbf{x}}_p\right) \;/\; \left(1 - \rho_{sp}\beta\right)^{1/2}$ and $\rho = \left(\rho_{ss} - \rho_{sp}\beta\right) \;/\; \left(1 - \rho_{sp}\beta\right)$. In most realistic situations ρ will be smaller than ρ_{ss}. The effect of including the parents on values of C is shown graphically in Figure 2.2. When the parent-sib trait correlation ρ_{sp} is small, whether parents are included or not

affects C mainly through the distortion of ρ. However when ρ_{sp} is substantial (e.g. high heritability or high household effect) and the parents' average trait values is high (or low), the effect is to shift the contour of C towards the north east quadrant (or south west quadrant) i.e. concordant siblings with non extreme values become valuable, whereas concordant siblings with extreme values become less attractive. For discordant pairs, the contour lines of C for average and extreme parents trait values cross, indicating that the inclusion of the extreme parents can affect C either way.

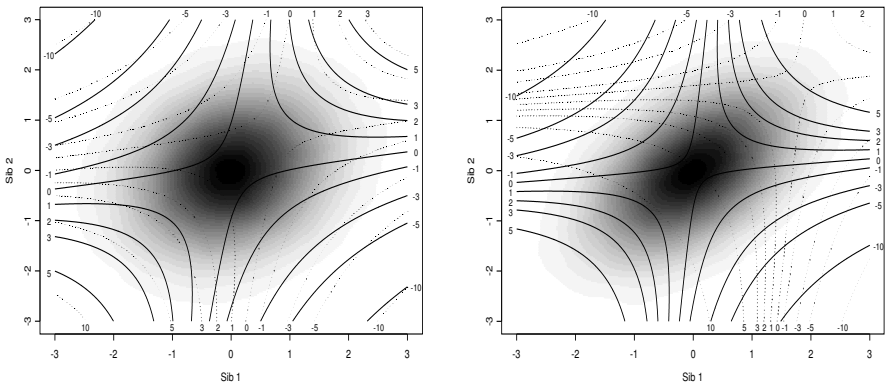

Figure 2.2: Joint distribution of sib trait values $\mathbf{x}$ (gray scale) and contour plot of $C(\mathbf{x}, \rho)$ (left panel: $\rho_{ss} = \rho_{sp} = 0.2$ and $\rho_{pp} = 0.1$, and right panel: $\rho_{ss} = \rho_{sp} = 0.5$ and $\rho_{pp} = 0.1$) for $\bar{\mathbf{x}}_p = 0$ (continuous lines, C values along vertical axis) and $\bar{\mathbf{x}}_p = 2$ (dotted lines, C values along horizontal axis)

Sibships and nuclear families of different sizes can easily be combined by weighting each family score according to its associated variance as suggested in Section 2.2.

2.4 Dominance

So far in our discussion we have neglected the effect of dominance. We show below what changes it involves in the score test compared to a fully additive model. We only consider here the most common design which allows evaluation of dominance variance component in non-inbred pedigrees: sibships consisting only of dizygotic twins or full

siblings. In presence of dominance, the conditional covariance $\mathbf{\Sigma}$ given the IBD status $\boldsymbol{\pi}$ becomes

$$[\mathbf{\Sigma}]_{jk} = \begin{cases} a^2 + d^2 + c^2 + e^2 \ = \ 1 \ , & \text{if } j = k \ , \\ (\pi_{jk} - \frac{1}{2})q^2 + (\mathbf{1}_{\{\pi_{jk}=1.0\}} - \frac{1}{4})t^2 & \text{if } j \neq k \ . \\ \quad +\frac{1}{2}a^2 + \frac{1}{4}d^2 + c^2 \ , \end{cases}$$

where d^2 denotes total dominance variance and t^2 represents the proportion of total variance attributable to the dominance component at the locus of interest.

We re-parameterize the model as in Tang and Siegmund [2001] so as to make the terms involving π_{jk} uncorrelated, with mean 0 and same variance: let $\gamma = q^2 + t^2$ and $\delta = \frac{t^2}{\sqrt{2}}$. The covariance matrix $\mathbf{\Sigma}$ then writes

$$[\mathbf{\Sigma}]_{jk} = \begin{cases} 1 \ , & \text{if } j = k \ , \\ (\pi_{jk} - \frac{1}{2})\gamma - \frac{1}{\sqrt{2}}(\mathbf{1}_{\{\pi_{jk}=0.5\}} - \frac{1}{2})\delta & \text{if } j \neq k \ . \\ \quad +\frac{1}{2}a^2 + \frac{1}{4}d^2 + c^2 \ , \end{cases}$$

The score for γ is as in formula (2.1) (however γ is now the sum of the additive and the dominant QTL variances) and the score with respect to δ is given by

$$\ell_\delta^{\boldsymbol{\pi}} = -\frac{1}{2\sqrt{2}} \ \mathrm{vec}(\mathbf{C})' \mathrm{vec}(\mathbf{1}_{\{\boldsymbol{\pi}=0.5\}} - \frac{1}{2}) \ .$$

Due to the new parameterization, $\ell_\gamma^{\boldsymbol{\pi}}$ and $\ell_\delta^{\boldsymbol{\pi}}$ are orthogonal under complete information (this is because π_{jk} and $\mathbf{1}_{\{\pi_{jk}=0.5\}}$ are uncorrelated in sib pairs [Amos et al., 1989]), and Fisher's information in $(\gamma, \delta) = (0, 0)$ is given by

$$\mathcal{I}_{\gamma,\delta}^{\boldsymbol{\pi}} = \begin{pmatrix} \mathcal{I}_\gamma^{\boldsymbol{\pi}} & 0 \\ 0 & \mathcal{I}_\delta^{\boldsymbol{\pi}} \end{pmatrix}$$

where $\mathcal{I}_\delta^{\boldsymbol{\pi}} = \frac{1}{8}\mathrm{vec}(\mathbf{C})' \ \mathrm{var}_{\boldsymbol{\pi}} \left(\mathrm{vec}(\mathbf{1}_{\{\boldsymbol{\pi}=0.5\}})\right) \mathrm{vec}(\mathbf{C})$ and $\mathcal{I}_\gamma^{\boldsymbol{\pi}}$ is given by formula (2.4). Under the assumption of a fully informative marker map $\mathcal{I}_\gamma^{\boldsymbol{\pi}} = \mathcal{I}_\delta^{\boldsymbol{\pi}} = \frac{1}{8}\sum_{1 \leq i < j \leq m} C_{ij}^2$, $\ell_\gamma^{\boldsymbol{\pi}} = \sum_{1 \leq i < j \leq m} C_{ij} \ \left(\pi_{ij} - \frac{1}{2}\right)$ and
$\ell_\gamma^{\boldsymbol{\pi}} = -\frac{1}{\sqrt{2}}\sum_{1 \leq i < j \leq m} C_{ij} \ \left(\mathbf{1}_{\{\pi_{ij}=0.5\}} - \frac{1}{2}\right)$ with C_{ij} as in formula (2.8), and the one-sided score test of the joint null hypothesis $(\gamma, \delta) = (0, 0)$ under the constraint $0 \leq$

$\sqrt{2}\ \delta \leq \gamma$ is then given by

$$z_+^2 = \begin{cases} \frac{\ell_\gamma^{\boldsymbol{\pi}\ 2}}{\mathcal{I}_\gamma^{\boldsymbol{\pi}}} + \frac{\ell_\delta^{\boldsymbol{\pi}\ 2}}{\mathcal{I}_\delta^{\boldsymbol{\pi}}}\ , & \text{if } 0 \leq \sqrt{2}\ \ell_\delta^{\boldsymbol{\pi}} \leq \ell_\gamma^{\boldsymbol{\pi}}\ , \\ \frac{\ell_\gamma^{\boldsymbol{\pi}\ 2}}{\mathcal{I}_\gamma^{\boldsymbol{\pi}}}\ , & \text{if } 0 < \ell_\gamma^{\boldsymbol{\pi}} \text{ and } 0 < \ell_\delta^{\boldsymbol{\pi}}\ , \\ \frac{1}{3}\ \left(\sqrt{2}\ \ell_\gamma^{\boldsymbol{\pi}} + \ell_\delta^{\boldsymbol{\pi}}\right)^2\ , & \text{if } -\frac{1}{\sqrt{2}}\ \ell_\delta^{\boldsymbol{\pi}} < \ell_\gamma^{\boldsymbol{\pi}} < \sqrt{2}\ \ell_\delta^{\boldsymbol{\pi}} \text{ and } \ell_\delta^{\boldsymbol{\pi}} > 0\ , \\ 0\ , & \text{otherwise .} \end{cases}$$

The local optimality properties of the univariate score test are preserved by this statistic since it is asymptotically equivalent to the likelihood ratio test [Verbeke and Molenberghs, 2003]. Under the null hypothesis of no locus effect, z_+^2 is distributed as $(1-\kappa)\chi_0^2 + \frac{1}{2}\chi_1^2 + \kappa\chi_2^2$ with $\kappa = 0.098$ [Shapiro, 1988]. Note that this test is the same as the one proposed by Wang and Huang [2002b] (see Section 2.6 for a closer comparison).

2.5 Dichotomous traits

Zeegers et al. [2003] have developed a modified Haseman-Elston regression for binary traits and have shown that it is approximately equivalent in power to the liability-threshold variance components model. In order to apply similar ideas to those developed in previous sections to dichotomous traits we use this so-called liability threshold model. Under such setting, a continuous variable arbitrarily scaled to have mean 0 and variance 1 underlies the trait of interest. In pedigrees involving only one type of family members relationship like sibships, the model is fully characterized by two parameters: the overall prevalence of the trait K (or equivalently the liability threshold t where $K = 1 - \Phi(t)$, Φ denotes here the cumulative density function of a standard normal) and the correlation ρ between the scaled liabilities of two sibs, also known as the tetrachoric correlation for the trait of interest. Different types of family members relationship may correspond to different tetrachoric correlations. Provided population data are available, the maximum likelihood method can be used to obtain estimates of the tetrachoric correlation between different relative pairs. Approximate formulae due to Pearson [1901] appear in Sham [1998, Section 5.5.5].

The probability $p_\gamma(\mathbf{y} \,|\, \boldsymbol{\pi})$ of the affection states of the pedigree members being $\mathbf{y}$, given $\boldsymbol{\pi}$, where $\mathbf{y}$ is one of the possible phenotypes, is obtained by integration of the density $f_\gamma(\mathbf{x} \,|\, \boldsymbol{\pi})$ for the underlying liability as expressed in the variance components

setting of Section 2.2 over $R_{\mathbf{y}}$, the region corresponding to phenotype $\mathbf{y}$ on the liability scale

$$p_\gamma(\mathbf{y} \,|\, \boldsymbol{\pi}) = \int_{\mathbf{x} \in R_{\mathbf{y}}} f_\gamma(\mathbf{x} \,|\, \pi) d\mathbf{x} \ .$$

The score $\ell_\gamma^{\mathbf{y}}$ for $p_\gamma(\mathbf{y} \,|\, \boldsymbol{\pi})$ at $\gamma = 0$ equals

$$\ell_\gamma^{\mathbf{y}} = \frac{\partial}{\partial \gamma} \log p_\gamma(\mathbf{y} \,|\, \boldsymbol{\pi}) = \frac{\int_{R_{\mathbf{y}}} \frac{\partial}{\partial \gamma} f_\gamma(\mathbf{x} \,|\, \pi) d\mathbf{x}}{\int_{R_{\mathbf{y}}} f_\gamma(\mathbf{x} \,|\, \pi) d\mathbf{x}} = \frac{\int_{R_{\mathbf{y}}} \ell_\gamma^{\mathbf{x}} f_\gamma(\mathbf{x} \,|\, \pi) d\mathbf{x}}{\int_{R_{\mathbf{y}}} f_\gamma(\mathbf{x} \,|\, \pi) d\mathbf{x}} = \mathbf{E}_{\mathbf{x}} \left(\ell_\gamma^{\mathbf{x}} \,|\, \mathbf{x} \in R_{\mathbf{y}} \right) \ .$$

As for the continuous case, the score $\ell_\gamma^{\boldsymbol{\pi}}$ for γ of the selection model $\boldsymbol{\pi} \,|\, \mathbf{y}$ is equal to the score $\ell_\gamma^{\mathbf{y}}$ for the $\mathbf{y} \,|\, \boldsymbol{\pi}$ model. Using formula (2.1) and by linearity of the expectation $\mathbf{E}$,

$$\ell_\gamma^{\boldsymbol{\pi}} = \ell_\gamma^{\mathbf{y}} = \frac{1}{2} \ \mathrm{vec}(\mathbf{C_y})' \mathrm{vec}(\boldsymbol{\pi} - \mathbf{E}\boldsymbol{\pi}) \ ,$$

and

$$\mathcal{I}_\gamma^{\boldsymbol{\pi}} = \frac{1}{4} \ \mathrm{vec}(\mathbf{C_y})' \ \mathrm{var}_{\boldsymbol{\pi}} \left(\mathrm{vec}(\boldsymbol{\pi}) \right) \ \mathrm{vec}(\mathbf{C_y})$$

with $C_{\mathbf{y}} = \mathbf{E}_{\mathbf{x}}(\mathbf{C}(\mathbf{x}, \rho) \,|\, \mathbf{x} \in R_{\mathbf{y}})$.

In the case of **sib pair** designs, there are only three possible unordered phenotypes: Affected/Affected (AA), Affected/Unaffected (AU) and Unaffected/Unaffected (UU). This implies that there are only three possible values of $\mathbf{C_y}$: C_{AA}, C_{AU}, C_{UU}, each corresponding to the conditional expectation of $C(\mathbf{x}, \rho)$, given $\mathbf{x}$ in the appropriate region on the liability scale. For a data set consisting of n_{AA} affected sib pairs, n_{AU} discordant sib pairs and n_{UU} unaffected sib pairs, the score test then equals

$$z = \frac{C_{AA} \sum_{i \in AA} \left(\pi_i - \frac{1}{2} \right) + C_{AU} \sum_{i \in AU} \left(\pi_i - \frac{1}{2} \right) + C_{UU} \sum_{i \in UU} \left(\pi_i - \frac{1}{2} \right)}{\sqrt{\frac{1}{8} \left(n_{AA} C_{AA}^2 + n_{AU} C_{AU}^2 + n_{UU} C_{UU}^2 \right)}} \ ,$$

and a robust score test is given by

$$z^* = \frac{C_{AA} \sum_{i \in AA} \left(\pi_i - \frac{1}{2} \right) + C_{AU} \sum_{i \in AU} \left(\pi_i - \frac{1}{2} \right) + C_{UU} \sum_{i \in UU} \left(\pi_i - \frac{1}{2} \right)}{\sqrt{C_{AA}^2 \sum_{i \in AA} \left(\pi_i - \frac{1}{2} \right)^2 + C_{AU}^2 \sum_{i \in AU} \left(\pi_i - \frac{1}{2} \right)^2 + C_{UU}^2 \sum_{i \in UU} \left(\pi_i - \frac{1}{2} \right)^2}} \ .$$

Nowadays, the $\mathbf{C_y}$ quantities can be approximated to a sufficient degree of precision using Monte Carlo simulation techniques.

Values of C_{AA}, C_{AU} and C_{UU} are provided in Table 2.1 for typical values of the tetrachoric correlation ρ and trait prevalence K. Under this liability threshold model, the main characteristics of the sib pair designs are that UU sib pairs provide

		AA		AU		UU				AA		AU		UU	
K	ρ	Prob.	$\bar{C}$	Prob.	$\bar{C}$	Prob.	$\bar{C}$	K	ρ	Prob.	$\bar{C}$	Prob.	$\bar{C}$	Prob.	$\bar{C}$
0.001	0.1	0.0000	9.63	0.0020	-0.04	0.9980	0.00	0.05	0.1	0.0037	3.68	0.0926	-0.29	0.9037	0.02
	0.2	0.0000	8.26	0.0020	-0.06	0.9980	0.00		0.2	0.0053	3.25	0.0895	-0.38	0.9053	0.02
	0.3	0.0000	7.24	0.0020	-0.10	0.9980	0.00		0.3	0.0071	2.92	0.0857	-0.49	0.9071	0.02
	0.4	0.0000	6.43	0.0019	-0.20	0.9980	0.00		0.4	0.0094	2.67	0.0812	-0.62	0.9094	0.03
	0.5	0.0001	5.82	0.0019	-0.34	0.9981	0.00		0.5	0.0122	2.48	0.0756	-0.80	0.9122	0.03
	0.6	0.0001	5.37	0.0018	-0.55	0.9981	0.00		0.6	0.0155	2.36	0.0690	-1.06	0.9155	0.04
0.01	0.1	0.0002	6.06	0.0196	-0.12	0.9802	0.00	0.1	0.1	0.0133	2.69	0.1733	-0.41	0.8133	0.04
	0.2	0.0003	5.27	0.0193	-0.18	0.9803	0.00		0.2	0.0172	2.40	0.1656	-0.50	0.8172	0.05
	0.3	0.0006	4.66	0.0189	-0.27	0.9806	0.00		0.3	0.0216	2.18	0.1567	-0.60	0.8216	0.06
	0.4	0.0009	4.20	0.0183	-0.40	0.9809	0.00		0.4	0.0267	2.02	0.1468	-0.73	0.8266	0.06
	0.5	0.0013	3.85	0.0174	-0.57	0.9813	0.01		0.5	0.0324	1.90	0.1352	-0.91	0.8324	0.07
	0.6	0.0019	3.60	0.0163	-0.83	0.9819	0.01		0.6	0.0390	1.83	0.1220	-1.17	0.8390	0.08
0.02	0.1	0.0007	5.02	0.0386	-0.18	0.9607	0.00	0.2	0.1	0.0481	1.75	0.3038	-0.55	0.6481	0.13
	0.2	0.0011	4.39	0.0378	-0.26	0.9611	0.01		0.2	0.0568	1.58	0.2864	-0.63	0.6568	0.14
	0.3	0.0017	3.91	0.0366	-0.35	0.9617	0.01		0.3	0.0662	1.46	0.2676	-0.72	0.6661	0.15
	0.4	0.0024	3.54	0.0352	-0.49	0.9624	0.01		0.4	0.0762	1.37	0.2476	-0.85	0.6762	0.15
	0.5	0.0034	3.26	0.0332	-0.67	0.9634	0.01		0.5	0.0871	1.31	0.2256	-1.02	0.6872	0.17
	0.6	0.0046	3.07	0.0307	-0.93	0.9646	0.01		0.6	0.0992	1.29	0.2015	-1.27	0.6992	0.18

Table 2.1: Probabilities of affection states and average C values for sib pairs

very little information whereas AA sib pairs provide the most information especially as the trait becomes rare. However, it must be stressed that as the prevalence of the trait increases, AU sib pairs become more informative. If only one type of phenotype is used (say only affected sib pairs) the score test is equivalent to $z = \frac{(\bar{\pi}-\frac{1}{2})}{\sqrt{1/(8n)}}$ and the robust score test equal $z^* = \frac{(\bar{\pi}-\frac{1}{2})}{\widehat{\text{se}}(\bar{\pi})}$ which are two standardized versions of the mean IBD sharing test. These tests are well established [Blackwelder and Elston, 1985] and have been in popular use for decades. As for the continuous case the test can be seen as a regression through the origin of the excess IBD sharing on a function C of the trait, however the function C only takes a limited number of distinct values. To illustrate this regression, we generated the affection states for 10000 sib pairs using the liability threshold model with $K = 0.05$, $\rho = 0.4$ and $\gamma = 0.15$. The 150 most informative pairs were selected using the corresponding $\bar{C}^2$ obtained from table 2.1; this resulted in all 97 affected pairs and 53 random discordant pairs being selected. Figure 2.3 illustrates the regression for this simulated data set.

One attractive feature of our approach is that it naturally allows combination of sib pairs of different nature (more generally, pedigree pairs of different nature and familial relationships). Each type of pairs contributes to the deviation from average IBD sharing with a weight proportional to the average value of the C function in the corresponding region. Note that in practice, table I can also be used with pedigrees consisting of other types of relative pairs. For example, if n^c_{AA} pedigrees consisting of affected cousins also are available then their contribution to the numerator of the previous z will simply be $C_{AA} \sum_{i=1}^{n^c_{AA}} (\pi^c_i - \frac{1}{8})$ where C_{AA} is drawn from table I with K as the population prevalence of the trait and ρ equal to the trait tetrachoric correlation between cousins. Our approach also offers an elegant solution to the problem of prevalence heterogeneity in the population: if a data set consists of groups with different disease prevalence, the contribution of each group to the overall test is weighted accordingly (see Table I).

2.6 Discussion

In the context of the variance components model, we have given an expression of the score test for linkage under sample selection based on phenotype values. It is

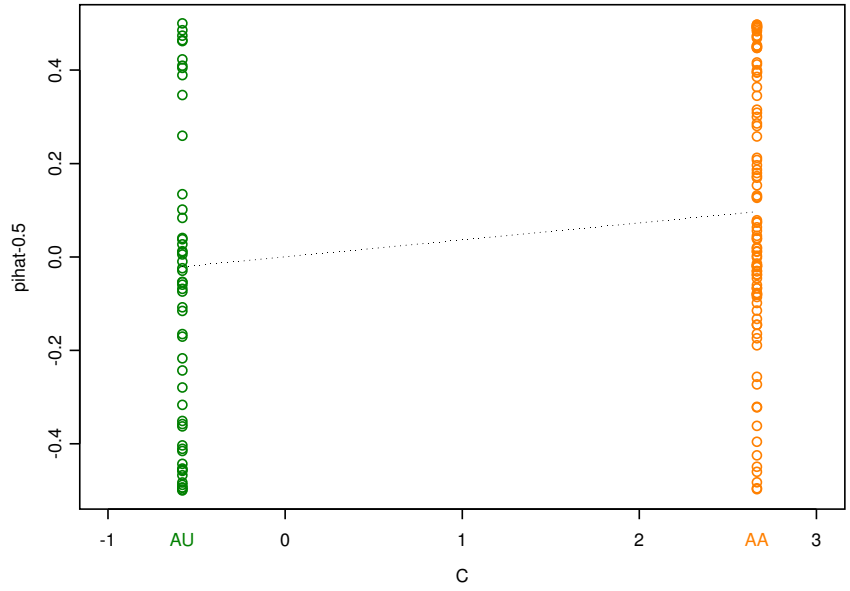

Figure 2.3: Regression of $\pi - \frac{1}{2}$ on $C(\mathbf{x}, \rho)$ for 150 selected sib pairs ($K = 0.05$, $\rho = 0.4$ and $\gamma = 0.15$)

a general expression for arbitrary pedigrees which takes a very simple form in some widely used designs. Commenges [1994] first introduced score tests in the context of linkage, however his approach is not conditional on trait values and therefore leads to reduced power in selected samples. In a recent article, Tritchler et al. [2003] give a general score test in nuclear families conditional on the trait values under the assumption that the trait distribution depends on different genetic models through the exponential family. Our results give a very similar expression to theirs. In their software implementation, they allow the population mean to be specified by the user but not the population sib-sib correlation and our understanding is that the authors attempt to estimate this correlation from the selected data, which potentially results in power loss (unless the ascertainment mechanism is known). Our approach is to fully acknowledge the fact that selected samples do not provide unbiased estimates of the population trait distribution characteristics and to assume that unbiased estimates

of the first and second moments of the population trait are available a priori. In the context of the GenomEUtwin project, where twin registries provide us with precise population mean and twin-twin correlation, this seems a realistic assumption.

The score test that we derive also has a simple interpretation in terms of regression of IBD sharing on a function of the phenotypes. Sham et al. [2002] have recently proposed a general method of analysis for quantitative linkage data which explicitly regresses IBD sharing on all possible squared sums and differences of trait values within a family. As shown in Section 2.2, the score test essentially is a regression of the excess IBD sharing on a quadratic function of the trait values whose shape depends on the normality assumption. When the data truly are normal, it seems reasonable to expect that the score test results in similar regressor as in the method of Sham et al. [2002]. We have compared the information content provided by the two methods in sibships and nuclear families of different sizes and they happen to exactly coincide. In fact, as demonstrated in a recently published paper [Chen et al., 2004], the two methods are the same for quantitative traits under an additive model (with trait correlations assumed to be the same over all pairs of relatives). The IBD covariance matrix is determined solely by family relations; no marker information is needed to compute it, which is a prerequisite to make it useful for selection prior to genotyping. Note that calculation of the information index in [Sham et al., 2002] does not require marker information either.

One possible criticism of the variance components model is that departure from the normality assumption might invalidate its results. However, the analogy of the test with regression methods, very much as the score test in unselected data coincides with the optimally weighted Haseman-Elston regression [Putter et al., 2002], pleads in favor of its robustness. In fact, as the regression interpretation of the score reveals, the test depends on the distribution of the trait values only through its second order moments. So as long as the shape of the distribution does not show any great departure from normality for those moments (e.g. heavy tail) then the test should remain valid. When the model clearly is wrong, the robust version of the test should preclude over-optimistic inference.

We showed in Section 2.2 that in the current variance components setting under which population marginal characteristics are known and the only unknown parameter

is the linkage parameter γ, there is no loss of information when conditioning on trait values. This is a direct consequence of the fact that scores for the selection model $\boldsymbol{\pi} \,|\, \mathbf{x}$, the $\mathbf{x} \,|\, \boldsymbol{\pi}$ model and the joint $(\mathbf{x}, \boldsymbol{\pi})$ model are identical. The situation becomes more complicated when population parameters are unknown and need to be conjunctly estimated.

As announced in Section 2.2, we now turn to the case of imperfect IBD information. In practice, $\boldsymbol{\pi}$ is not known with certainty. In fact, the only available data are marker information which we denote M and the phenotypes $\mathbf{x}$. Strictly speaking, the likelihood to be considered should be expressed in terms of those data, i.e. we should write $f_\gamma(M, \mathbf{x})$ for the joint distribution of M and $\mathbf{x}$ and $f_\gamma(M \,|\, \mathbf{x})$ for the conditional distribution of $M \,|\, \mathbf{x}$. It turns out that the score ℓ_γ^M for the $M \,|\, \mathbf{x}$ distribution simply becomes the weighted average of the score $\ell_\gamma^{\boldsymbol{\pi}}$ for the idealized fully informative model $\ell_\gamma^M = \sum_{\boldsymbol{\pi}} P(\boldsymbol{\pi} \,|\, M)\ \ell_\gamma^{\boldsymbol{\pi}}$ and thus, with $\hat{\boldsymbol{\pi}} = \mathbf{E}(\boldsymbol{\pi} \,|\, M)$,

$$\ell_\gamma^M = \frac{1}{2} \mathrm{vec}(\mathbf{C})' \mathrm{vec}(\hat{\boldsymbol{\pi}} - \mathbf{E}\hat{\boldsymbol{\pi}}) \ .$$

Since $\mathbf{E}\hat{\boldsymbol{\pi}} = \mathbf{E}\pi$, this result means that Formula (2.1) still holds true with imperfect data but $\boldsymbol{\pi}$ values have to be replaced by estimates given marker data available $\hat{\boldsymbol{\pi}}$. Values of $P(\boldsymbol{\pi} \,|\, M)$ and $\hat{\boldsymbol{\pi}}$ are calculated using for example the Lander-Green or Elston-Stewart algorithms [Lander and Botstein, 1989] as implemented in publicly available softwares like `GENEHUNTER` [Kruglyak et al., 1996] or `MERLIN` [Abecasis et al., 2002]. Note that this result theoretically justifies (as mentioned by Commenges [1994] and Tang and Siegmund [2001]) the use of the so-called $\hat{\pi}$ approach in variance components linkage modelling for arbitrary pedigrees. The corresponding Fisher's information is given by

$$\mathcal{I}_\gamma^M = \frac{1}{4}\ \mathrm{vec}(\mathbf{C})'\ \mathrm{var}_M\left(\mathrm{vec}(\hat{\boldsymbol{\pi}})\right)\ \mathrm{vec}(\mathbf{C}) \ .$$

Given a marker map and a certain pedigree structure, Monte Carlo simulations can be used to approximate $\mathrm{var}_M\left(\mathrm{vec}(\hat{\boldsymbol{\pi}})\right)$. A conservative alternative is to use $\mathcal{I}_\gamma^{\boldsymbol{\pi}}$ as given by Formula (2.4) instead of $\mathcal{I}_\gamma^M$ in the standardization of ℓ_γ^M. One consequence of imperfect information in the case of sibships for example is that negative terms appear on the off-diagonal components of the $\mathrm{var}_M\left(\mathrm{vec}(\hat{\boldsymbol{\pi}})\right)$ matrix. When considering both additive and dominance variance components, the scores $\ell_\gamma^{\boldsymbol{\pi}}$ and $\ell_\delta^{\boldsymbol{\pi}}$ derived

in Section 2.4 are no longer orthogonal and the use of the test as defined in that section is not optimal. It is possible to obtain the expression of a multivariate score test that is asymptotically optimal [Verbeke and Molenberghs, 2003] and whose null distribution ($(1-\kappa)\chi^2_0 + \frac{1}{2}\chi^2_1 + \kappa\chi^2_2$, where κ depends on informativeness) can be obtained using results in Shapiro [1988]. The details are beyond the scope of this article, however the results appear in Wang and Huang [2002b] who consider only random samples and therefore suggest to estimate the sib-sib correlation as well as $\mathbf{P}(\boldsymbol{\pi} = 0.5 \,|\, M)$, $\mathbf{E}(\hat{\boldsymbol{\pi}})$ and $\mathrm{var}(\hat{\boldsymbol{\pi}})$ from the data. Interestingly, our derivation shows that their approach is perfectly valid in selected samples too, provided the population sib-sib correlation is known and unbiased values for $\mathbf{P}(\boldsymbol{\pi} = 0.5 \,|\, M)$, $\mathbf{E}(\hat{\boldsymbol{\pi}})$ and $\mathrm{var}(\hat{\boldsymbol{\pi}})$ are calculated (e.g. using Monte Carlo simulation technique described above). Note that in selected samples, the use of population estimates for those 'nuisance' parameters amounts to constraining the regression through the origin and is critical in order to maintain maximum power. In practice, the asymptotic results might fail to hold in finite samples and it seems wise to use re-sampling methods (bootstrap) in order to obtain a robust assessment of significance.

By use of the liability threshold model, the continuous case extends to the case of dichotomous traits. Because of the well-known optimality properties of the score test (which is asymptotically equivalent to the likelihood-ratio test), it provides an efficient means to test for linkage in affected sib pairs and in discordant sib pairs as well as a way to combine the two types of data when needs arise. More complicated pedigrees can also be handled in a very flexible manner. In this selection framework where IBD sharing $\boldsymbol{\pi}$ is considered conditional on the trait values $\mathbf{x}$, the extension to multiple traits, in analogy with multiple regression, should be fairly straightforward.

This score test approach has been implemented into a C program calling upon the publicly available software `MERLIN` [Abecasis et al., 2002] and is available at `http://www.msbi.nl/Genetics` .

2.7 Appendix

Score test

The score function for γ in the $\mathbf{x} \,|\, \boldsymbol{\pi}$ model is denoted by $\ell_\gamma^{\mathbf{x}}$ and by definition equals $\ell_\gamma^{\mathbf{x}} = \frac{\partial}{\partial \gamma} \log f_\gamma(\mathbf{x} \,|\, \boldsymbol{\pi})$ with

$$\log f_\gamma(\mathbf{x} \,|\, \boldsymbol{\pi}) = -\frac{m}{2} \log(2\pi) - \frac{1}{2} \log(|\boldsymbol{\Sigma}|) - \frac{1}{2} \mathbf{x}' \boldsymbol{\Sigma}^{-1} \mathbf{x}$$

Standard results on matrix algebra (see, e.g. [Searle et al., 1992, Appendix M.7]) show that

$$\ell_\gamma^{\mathbf{x}} = \frac{1}{2} \left\{ \mathbf{x}' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\pi} - \mathbf{E}\boldsymbol{\pi}) \boldsymbol{\Sigma}^{-1} \mathbf{x} - \operatorname{tr}(\boldsymbol{\Sigma}^{-1} (\boldsymbol{\pi} - \mathbf{E}\boldsymbol{\pi})) \right\}$$

Because of the relation $a'b = \operatorname{tr}(ba')$, the previous equation can be rewritten

$$\ell_\gamma^{\mathbf{x}} = \frac{1}{2} \operatorname{tr} \left(\boldsymbol{\Sigma}^{-1} (\boldsymbol{\pi} - \mathbf{E}\boldsymbol{\pi}) (\boldsymbol{\Sigma}^{-1} \mathbf{x}\mathbf{x}' - \mathbf{I}) \right) \ .$$