Implementation of a Combined Association-Linkage Model for Quantitative Traits in Linear Mixed Model Procedures of Statistical Packages

A. Leo Beem and Dorret I. Boomsma

Department of Biological Psychology, Vrije Universiteit, Amsterdam, the Netherlands

A transmission disequilibrium test for quantitative traits which combines association and linkage analyses is currently available in several dedicated software packages. We describe how to implement such models in linear mixed model procedures that are available in widely used statistical packages such as SPSS. We also briefly mention a few extensions of the model that become naturally available once the model is implemented in such procedures.

Genotyping of many microsatellite markers or single nucleotide polymorphisms (SNPs) over the entire genome is becoming increasingly common in human genetics. In those high-resolution maps the average distance between microsatellite markers may be as small as 5 cM and between SNPs one half cM or less. At those small distances it becomes fairly likely that some markers in the set are in linkage disequilibrium (LD) with a gene affecting the trait (a so-called quantitative trait locus or QTL if the trait or the vulnerability distribution is quantitative). Different alleles or combinations of alleles of the markers or SNPs can then be associated with different trait means. Association studies are conducted to discover such allelic effects.

Abecasis et al. (2000) generalized the model proposed by Fulker et al. (1999) for combined linkage and association tests, within and between families. The Fulker-Abecasis or F-A model is implemented in the program QTDT (http://www.sph.umich.edu/csg/abecasis/QTDT/index.html) and provides a transmission disequilibrium test (TDT) statistic for quantitative traits, with tests of linkage in the presence of association and tests of association that are robust to population stratification. Here we will describe a method to implement the model in the SPSS procedure for the analyses of data by linear mixed models. The method can also be used with other statistical packages such as SAS, STATA, S-Plus or R, which have similar or more extended capabilities for mixed model analyses.

There are several reasons for wanting to perform such analyses with a procedure available in widely used statistical packages. First, the data do not need to be written to a file for subsequent reading and analysis by for example QTDT. This step can be a source of error, both in the writing and reading step, but such external programs usually have only limited options for checking that the data analyzed have been transported correctly to the program. Second, repeating analyses on subsets of the data requires only some simple statements for selection of the subsets. Third, general statistical packages make it fairly easy to perform various checks, such as checks on model assumptions. Residuals, expected values under the model, and the so-called best linear unbiased predictors or BLUPs (see, for example, Robinson, 1991), among others, can be useful in this regard. SAS has extensive (though still experimental) capabilities for model diagnostics. Finally, the models can be extended in several ways as long as the model fits into the general linear mixed model framework implemented in the programs.

We will first briefly describe the F-A model. Next we give a short introduction to linear mixed models and in particular describe the way they are presently implemented in statistical packages using what has become a common representation of mixed models. We then describe how the F-A model can be implemented in mixed model procedures that rely on such representations of the mixed model. SPSS code for implementation of several models and documentation are available from http://www.psy.vu.nl/mxbib.

The Combined Linkage-Association Model

The basis of this model is a regression of a quantitative dependent variable y on (a function of) the number of

Received 11 November, 2005; accepted 23 January, 2006.

Address for correspondence: A. L. Beem, Department of Biological Psychology, Vrije Universiteit, van der Boechorstraat 1, 1081 BT Amsterdam, the Netherlands. E-mail: al.beem@psy.vu.nl

alleles of a particular type. Effects of stratification are accounted for by assuming that each family in the sample comes from a different subpopulation with different allele frequencies. A separate regression equation is defined for each family. The regression weights corresponding to the additive and dominance effects, if included, are assumed to be the same for each family, but the intercepts are possibly different across families. For subject j in family i, define g_{ij} as the number of alleles of a particular type minus 1. Including only the additive effect, the regression for this family then becomes $y_{ij} = \mu_i + \beta g_{ij} + \varepsilon_{ij}$. If the regression is now formulated for the entire population as $y_{ij} = \mu + \gamma g_i + \varepsilon_{ij}$, the regression weight γ for the entire population will usually be different from β. The model is therefore formulated as $y_{ij} = \mu + \beta_b b_i + \beta_w w_{ij} + \varepsilon_{ij}$ for $w_{ij} = g_{ij} - b_i$ and b_i the expected family mean of g_{ij}. The coefficients β_b and β_w are called the between and within-family regression coefficients and $\beta_w = \beta$, the additive effects in the subpopulations. If the father and mother genotypes g_{iF} and g_{iM} are available, then $b_i = (g_{iF} + g_{iM}) / 2$. Otherwise, b_i is estimated as the average of the offspring values g_{ij}. A dominance effect can be included by adding a quadratic term to the model, which then has the form $y_{ij} = \mu_i + \beta_a g_{ij} + \beta_d g_{ij}^2 + \varepsilon_{ij}$. If parental genotypes are available, the expected value of g_{ij}^2 can be computed as $(g_{iF} g_{iM} + 1) / 2$ (see Appendix A). Otherwise it can be estimated as the average of g_{ij}^2. The addition of a quadratic term to account for deviations from linearity (or additivity) is a standard method in (polynomial) regression. For a variable with only three values, as is the case for two allele markers (i.e., three genotypes), the linear and quadratic term account for all possible deviations from linearity. The biometrical model (e.g., Falconer & Mackay, 1996) uses $1 - g_{ij}^2$ instead of g_{ij}^2. The expected value for known parental genotypes is $1 - (g_{iF} g_{iM} + 1) / 2 = (1 - g_{iF} g_{iM}) / 2$. Families can only be informative for the within-family effect if at least one offspring has a genotype score different from the expected value or its estimate.

Association effects are often studied after linkage has been observed at one or more loci. An important purpose of association studies is to investigate whether an association effect can partly or wholly account for the observed linkage. If it does, the covariance among offspring due to linkage is reduced or disappears.

Linear Mixed Models

A standard regression or linear model contains an intercept, regressors and their associated regression parameters, and a residual to account for deviations from the model. A factor in an analysis of variance (ANOVA) is often modeled by a set of regressors with values zero or one, where the value one represents a particular factor level. The intercept and parameters associated with the regressors are called fixed effects, because they are assumed to describe properties of the regression in the population. For different random samples from the same population, the parameters associated with the regressors remain the same (their estimates will of course generally differ in different samples). This often means in practice, that the effects of the regressors or factor levels are of particular interest in the study. The residual is the difference between the value of the dependent variable of a given observation and the expected value (i.e., the mean) of the dependent variable in the population for a given combination of values of the regressors. Since the observations are different in different random samples, the residuals are called random effects. They are considered a random sample from a population of residuals. Individual residuals are usually not of interest. Instead, properties of their distribution in the population, such as their variance, are estimated. This distinction between fixed and random effects is one among several and need not cover all possible and sometimes conflicting meanings given to those terms (cf. Gelman, 2005).

A model that contains the residual as the only random effect is called a fixed effects model. A model with the intercept as the only fixed effect and with random effects beside the residual is called a random effects model. Such additional random effects often correspond to levels of factors, which are regarded as a random sample from a population of factor levels. The effect of the factor is then often associated with the variance of the level effects. Mixed models contain fixed effects in addition to the intercept and random effects in addition to the residual. The elementary mixed model is an ANOVA model with a fixed experimental factor β and a random sample of units (e.g., subjects), observed once for each level of the experimental factor. Such a model can be used for a repeated measures design where the units are measured at different levels of the factor. The model for unit i observed at level j of the fixed effect can be written as $y_{ij} = \mu + u_i + \beta_j + \varepsilon_{ij}$, where ε_{ij} is the residual. Each unit is associated with a random effect u_i. The expectation $E\, u_i$ (i.e., the mean in the population) of the random effects can always be defined as zero if μ is included in the model. The variance of the effects is then $E\, u_i^2 = \sigma_u^2$. Because the units are sampled independently, the effects u_i are independent and their covariances are zero. The effects u_i may be regarded as a general effect on the response level of unit i. All residuals are also assumed independent with zero mean and variance σ_e^2 and are independent of u_i. The random effects induce covariances among the observations of the same unit for different levels β_j: $\mathrm{cov}\,(y_{ij}, y_{ij}) = cov\,(u_i + \varepsilon_{ij}, u_i + \varepsilon_{ij}) = \sigma_u^2$. One purpose of such a model is that it can serve as a heuristic device to generate a covariance structure for the observations. This can be especially useful for complicated designs including several random effects, as it may then be quite difficult to specify the covariance structure without the benefit of a model. The above model implies the same variance $\sigma_u^2 + \sigma_e^2$ for

Table 1

Example of Data Set

Y	FAM	OFFS	G	B	W	F	P1	P2	P3	P4	C1	C2	C3	C4
8	1	1	0	.33	–.33	1	1	1	.69	.59	1	0	0	0
4	1	2	0	.33	–.33	1	1	1	.69	.59	1	0	0	0
9	1	4	1	.33	.67	2	.69	.69	1	.44	.69	0	.72	0
18	1	7	0	.33	–.33	3	.59	.59	.44	1	.59	0	.05	.81
3	2	1	–1	–.67	–.33	1	1	.50	0	—	1	0	0	0
2	2	2	–1	–.67	–.33	2	.50	1	.50	—	.50	.87	0	0
14	2	11	0	–.67	.67	3	0	.50	1	—	0	.58	.82	0

Note: Rows are subjects; columns are variables.

Y is the dependent variable; FAM indexes families; OFFSPRING indexes offspring within families; G is the genotype score; B is the average within-family genotype score; W = G – B; F codes factor levels; P1-P4: columns of Π_i matrices; C1-C4: columns of Cholesky decomposition of Π_i matrices.

each level *j* of the fixed effect and the same covariance σ_u^2 among levels of the fixed effect.

Mixed Models Applied to Family Data

In the association model $y_{ij} = \mu + \beta_b\, b_i + \beta_w\, w_{ij} + \varepsilon_{ij}$ the residual ε_{ij} is the only random effect. The covariance structure conditional on fixed effects is usually formulated to include expected covariances among observations due to several sources, which are then represented by additional random effects. Those sources may include a family effect and an effect due to linkage, in order to test whether the association effect can account for an observed linkage effect at the locus. If monozygotic (MZ) twin pairs are available, the family effect can be separated into a polygenic effect and a common environment effect. In this covariance structure, the phenotypic variance conditional on the general mean and regressors (the fixed effects) is modeled as $\sigma_a^2 + \sigma_g^2 + \sigma_c^2 + \sigma_e^2$, where the subscript *a* refers to the additive effect of a gene linked to the marker but not necessarily in linkage disequilibrium, and the subscripts *g*, *c* and *e* refer to background polygenic, common environment and residual or unique environment effects. For a family with n_f offspring ($n_f \geq 2$), the within-family expected phenotypic covariances of any two offspring *j* and *k* in family *i* are modeled as $\pi_{ijk}\, \sigma_a^2 + \delta\, \sigma_g^2 + \sigma_c^2$, where π_{ijk} is the proportion of alleles shared identical by descent (IBD) by the pair and δ is 1 for MZ twins and .5 for other offspring pairs. Dominance effects due to background polygenic sources and linkage can also be included. The estimate of σ_a^2 in the model without b_i and w_{ij} included provides a baseline of the amount of linkage. After including b_i and w_{ij} the estimate of σ_a^2 should be close to zero or at least be substantially less than the baseline estimate.

In this situation, random effects are not needed as a heuristic device for generating covariances among observations. Instead, the variances and covariances are known from the underlying genetic model and the random effects must be specified so that they correspond to this covariance structure. In order to do that, we need to know how mixed models are represented in mixed model procedures.

Representation of Mixed Models

Software packages currently often use the following matrix representation of mixed models:

$$y = X\beta + Zb + \varepsilon.$$

The vector *y* and the columns of X and Z correspond to the variable columns in a dataset. Here *y* is a vector of observations on the dependent variable (the phenotype of interest). Its length *n* is the total number of offspring over all families. The number of rows of X and Z is *n*. The n_β columns of X contain regressors or zero-one dummy variables for encoding levels of factors; the vector β contains n_β fixed effect parameters associated with the columns of X, which usually also contains one column of 1s for the general mean or intercept. The matrix Z has a similar form as the matrix X, but its n_b columns are associated with the n_b random effects collected in the vector b. The vector ε of length *n* contains the residuals.

Table 1 gives an example of a dataset for two families with four offspring in the first family and three offspring in the second family. The variable Y is the dependent variable. The variables G, B, and W can be columns of X. G is the genotype score, B is the family average of G, and W is the difference G – B. Offspring 1 and 2 are two MZ twins; only one of the two genotypes of this twin pair is used to compute B. The levels of the factor F can become coded as zero-one dummy variables in the matrix Z. This factor and the other variables are discussed below.

In order to explicitly separate different sets of random effects, the matrix Z can be partitioned into matrices Z_s whose columns are associated with different sets of random effects collected in vectors b_s. The total number of columns over all the matrices Z_s is equal to the number of columns of Z. The model can then be written as

$$y = X\beta + \Sigma_s\, Z_s\, b_s + \varepsilon.$$

This representation of the mixed model is due to Hartley and Rao (1967). They formulated the model as a classical variance components model, for which

Y	S	R	Z_s	Z_r	Model Equations
9	1	1	1 0	1 0	$y_{11} = \mu + s_1 + r_1 + \varepsilon_{11}$
6	1	2	1 0	0 1	$y_{12} = \mu + s_1 + r_2 + \varepsilon_{12}$
1	2	1	0 1	1 0	$y_{21} = \mu + s_2 + r_1 + \varepsilon_{21}$
7	2	2	0 1	0 1	$y_{22} = \mu + s_2 + r_2 + \varepsilon_{22}$

Figure 1

Example of Z matrices and model equations for a two-way crossed random effects model.

all random effects are uncorrelated. In the present formulation in statistical packages, various types of covariance structures can be specified for the random effects b_s and for ε. Searle et al. (1992) and McCulloch and Searle (2001) provide mathematically quite detailed presentations of mixed models. Fitzmaurice et al. (2004), Verbeke and Molenberghs (2001) and Pinheiro and Bates (2000) discuss mixed models with various degrees of emphasis on applications.

As a small example, consider a model with two crossed random factors (i.e., a design in which each level of a factor S is combined with all levels of the other factor R). The levels of the factor S may correspond to subjects and the levels of the factor R may correspond to raters, who rate the subjects' behavior in a certain situation. The subjects and raters are randomly sampled from a population of subjects and raters. Let y_{ij} be the rating of subject i by rater j. Since each subject is observed only once by each rater, the subject by rater interaction cannot be separated from the residual and the linear model for the effect of subject s_i and rater r_j becomes $y_{ij} = \mu + s_i + r_j + \varepsilon_{ij}$. The variances of the random effects are σ_s^2 and σ_r^2. Assume that, unrealistically, each factor has only two levels in the sample. The effects are collected in the vector $b_1' = [s_1, s_2]$ and $b_2' = [r_1, r_2]$. For a dataset with four observations, Figure 1 contains the values of the dependent variable Y and the variables S and R, which encode the levels of the factors S and R. The first row contains the score for Subject 1 as obtained from Rater 1. For the four observations in the four cells of the design the matrices Z_1 and Z_2, which represent the levels of S and R, are presented in Figure 1. The zero-one columns in those matrices are called dummy variables. The model can now be written in matrix notation as $y = \mu + Z_1b_1 + Z_2b_2 + \varepsilon$. The products Z_1b_1 and Z_2b_2 multiply the elements in the first and second columns of these matrices by s_1 and s_2, and r_1 and r_2. This gives the model equations presented in Figure 1.

In order to implement a mixed model in statistical packages, the mixed model procedure needs information corresponding to the elements in the mixed model representation $y = X\beta + \Sigma_s Z_s b_s + \varepsilon$. The user must specify whether a variable is regarded as numerical or as a factor, in which case its values indicate the levels of the factor. Second, the user must specify whether the effects associated with the variable or with the factor levels are regarded as fixed or as random. In some models, both fixed and random effects may be associated with the same variable. Third, the covariance structure of the random effects must be specified.

The dummy variables corresponding to factor levels in the matrices Z_s (or X) can be coded explicitly as variables in the dataset, but it is more efficient to let the program (e.g., SPSS) generate them. If the user has specified that a variable encodes levels of a factor, the matrix corresponding to the factor is generated by the program as in Figure 1, without becoming a permanent part of the dataset. For a factor with k levels, k columns are generated for the matrix. The column k of the matrix is set to 1 if y is observed at level k of the factor and is zero elsewhere. If a variable is not defined as a factor, the original values are stored in a column of Z. The fixed effect parameters or (co)variances of the random effects are usually estimated by maximum likelihood or restricted maximum likelihood under multivariate normality.

For each factor, set of factors or set of variables, a different covariance structure can be specified for the random effects associated with the levels of the factors or with the variables. For each different level the program implicitly generates a random effect (implicitly, because programs usually work only with the variances and covariances of the effects). If for the example above, the effects are specified as independent with variances σ_s^2 and σ_r^2, the covariance matrices of the random effects become two by two diagonal matrices with σ_s^2 and σ_r^2 on the diagonal.

If the effects of a variable or of the levels of a factor are assumed to be family specific (i.e., the effect of level k of a factor is different for different families), those effects can then be modeled as random effects whose values vary independently over families. This can be modeled if in the model the random effect corresponding to a level k is different in different families. Mixed model procedures can be informed that different values of a variable define a partitioning of the data into different groups or units of observations. For example, the different values of the variable FAM in Table 1 define groups of offspring in the same family. For each such a group g, different sets of random effect parameters for variables or factor levels are generated. Then if b_s contains the effects of the levels of a given factor, it has the form $[b_{s1}' \mid b_{s2}' \mid \ldots \mid b_{sg}']'$. The covariance structure of the random effects is the same for each group. The matrix Z_s associated with the effect b_s can be partitioned accordingly as $[Z_{s1} \mid Z_{s2} \mid \ldots \mid Z_{sg}]$. Z_{sg} contains zeros only for observations in groups other than group g. For group g its rows encode the random effect part of the model for the observations in the group, as the Z matrices in Figure 1. If a factor level is not observed in a particular group g, the corresponding column of Z_{sg} is zero. Programs may actually store the model in another way.

If the expected covariance matrix for the random effects b_s is V_s, then the contribution to the expected covariance matrix for all random effects combined (i.e., for the dependent variable) becomes $\Sigma_s Z_s V_s Z_s'$. Suppose that the observations within each group are

stored consecutively in the vector y as in Table 1. The matrices V_s then have the simple form of a so-called block diagonal matrix. They contain along the diagonal the expected covariance matrices V_{si} for each family *i* and are zero elsewhere, since the families are sampled independently. Let now for convenience Z_{si} contain only factor level encodings or variables for a family *i* after deleting the zero rows for observations in the other families. The contribution to the expected covariance matrix can then be written as $\Sigma_i Z_{si} V_{si} Z_{si}'$ for each family separately. The number of offspring may vary over families and is equal to the number of rows of Z_{si}.

The specification of the fixed effects of the model is straightforward and similar to specifications of fixed effects in standard ANOVA and regression programs. We will therefore first describe the specification of the within-family covariance structure.

Generating the Model Covariance Matrix

We will explain here how the covariance matrix for data from dizygotic (DZ) twins, sibs and possibly MZ twin pairs can be specified. We assume that only the phenotypes of one or more offspring from nuclear families are included in the analysis. Inclusion of MZ twin pairs allows for the possibility to also estimate a common environment effect. The expected covariance matrix for a family of size *n* (i.e., *n* offspring) is a matrix with *n* rows and *n* columns. The diagonal elements contain the expected variances $\sigma_a^2 + \sigma_g^2 + \sigma_c^2 + \sigma_e^2$, and the off-diagonal elements contain the expected covariances $\pi_{ijk}\,\sigma_a^2 + \delta\,\sigma_g^2 + \sigma_c^2$, for δ as specified previously. The residual or unique environmental variance σ_e^2 is automatically estimated by the program and need not be specified explicitly. The common environment effect contributes the variance component σ_c^2 to the variance and covariance. For a family of size *n*, the contribution to the expected family covariance matrix is therefore $\sigma_c^2\ 1_n 1_n'$, where 1_n is a vector with *n* 1s. We present two alternatives to generate this structure. For the first alternative, a variable should be created that is equal to one for all observations in the dataset. Then this variable is declared as a random effect with separate random effects for each family using the grouping option (e.g., using the variable FAM in Table 1 to group the observations). The effects are specified as having the same variance and zero covariance. Then for family *i* $Z_{si} = 1_n$, $V_{si} = \sigma_c^2$ and the expected covariance matrix is $\sigma_c^2\ Z_{si}\ Z_{si}' = \sigma_c^2\ 1_n 1_n'$ (i.e., a matrix with all elements equal to σ_c^2). For the second alternative, a variable should be available in the dataset which has the same value for each family and different values for different families, such as FAM in Table 1. If this variable is declared as a random effects factor, the model contains a separate random effect for each level of this factor (i.e., for each family). The matrix Z_c for this factor has a similar form as Z_1 in Figure 1 with offspring within families as subjects. For n_f families the matrix has n_f columns and column *k* contains 1s for the observations from offspring in family *k* and 0s for other families. The random effects are specified as having the same variance and zero covariance. The expected covariance matrix for the random effects is then an $n_f \times n_f$ diagonal matrix with σ_c^2 on the diagonal. The matrix $\sigma_c^2\ Z_c Z_c'$ contains matrices of the form $\sigma_c^2\ 1_n 1_n'$ for families of size *n* on the diagonal and is zero elsewhere.

The additive polygenic genetic effect contributes a variance component σ^2 to the variance and $\delta\,\sigma^2$ to the covariance, where δ is 1 for MZ twins and .5 for full first degree offspring pairs. Thus for siblings or DZ twins the variance is σ^2 and their covariance is $.5\ \sigma^2$. The variance for MZ twins is also σ^2, their covariance is σ^2, and the covariance of MZ twins with additional siblings is $.5\ \sigma^2$. We will demonstrate how to implement an equivalent but slightly different structure by specifying $2\ \sigma_g^2$ for the variance, σ_g^2 for the covariance of siblings including DZ twins and for the covariance of an MZ twin and siblings, and $2\ \sigma_g^2$ for the covariance of MZ twin pairs. This specification fits the same structure but with $\sigma_g^2 = .5\ \sigma^2$.

Suppose that for family *i* scores y_{ij} are available for a pair of MZ twins and for two additional siblings. For example, in Family 1 in Table 1 Offspring 1 and 2 are MZ twins and Offspring 3 and 4 are additional siblings. A fairly simple way to let the program generate the required covariance structure is as follows. First, specify a variable according to the first alternative for the common environment effect. Second, specify a factor which has the same value (i.e., level) for the MZ twin pair and different values for each of the other offspring, such as the factor F for the Family 1 in Table 1. For a family *i* with one MZ twin pair with observed scores y_{i1} and y_{i2} and two sibs with observed scores y_{i3} and y_{i4}, this specification generates the matrices Z_1 and Z_2 in Figure 2, where the first two rows are associated with the observations y_{i1} and y_{i2} for the MZ twins. The matrix Z_1 corresponds to the variable and Z_2 to the factor. The grouping option is used to assign different matrices and different random effects corresponding to the matrices' columns to different families. The variances of those effects are specified to be the same with no covariance among the effects.

Figure 2 also contains the contributions to the model equations for the effect g_1 corresponding to the column of Z_1 and g_{21}, g_{22}, g_{23} corresponding to the columns of Z_2. These equations are generated by $Z_1\ g_1 + Z_2\ g_2$ for $g_2 = [g_{21}, g_{22}, g_{23}]'$. Because the covariance of different effects is zero, the expected variance of each observation is the sum of the expected variances of the effects in the equation. Hence the expected variance is $2\ \sigma_g^2$. The expected covariance for the two siblings is $cov\ (g_1 + g_{22}, g_1 + g_{23}) = cov\ (g_1, g_1) = \sigma_g^2$ and similarly for the covariance of a MZ twin and a sibling. For the two MZ twins the

Z_1		Z_2		Contribution to Model Equations	
1	1	0	0	for y_{i1}:	$p_1 + p_{21}$
1	1	0	0	for y_{i2}:	$p_1 + p_{21}$
1	0	1	0	for y_{i3}:	$p_1 + p_{22}$
1	0	0	1	for y_{i4}:	$p_1 + p_{23}$

Figure 2

Example of Z matrices encoding for the additive polygenic effect.

expected covariance is $cov\ (g_1 + g_{21},\ g_1 + g_{21}) = cov\ (g_1, g_1) + cov\ (g_{21}, g_{21}) = 2\ \sigma_g^2$. It is easily verified that the expected covariance matrix $\sigma_g^2\ Z_1\ Z_1' + \sigma_g^2\ Z_2\ Z_2'$ (because $Z_2\ [\sigma_g^2\ I]\ Z_2' = \sigma_g^2\ Z_2\ Z_2'$) has the desired structure for the family in Figure 2.

If a family, as the second family in Table 1, contains no MZ twin pair but only other offspring, all levels of a factor such as F in Table 1 are different, also if one MZ twin is included. Then the matrix Z_2 becomes a matrix with 1s on the diagonal and 0s elsewhere. As g_1 is then the only effect shared by the offspring, the expected covariance among the offspring is σ_g^2.

If a dominance component is to be included, the same trick can be used, but with the factor corresponding to Z_2 replicated three times by creating three separate factors identical to F in Table 1. However, an alternative specification is more memory efficient. If the 1s of Z_2 in Figure 2 are replaced by $\sqrt{3}$, the covariance structure due to dominance is generated. This can be achieved by creating *variables*, which should not be specified as factors, with the same variance associated with the random effects and zero covariance among them (the desired matrix can be generated efficiently by nesting the variable within the factor used for generating Z_2). As $\sqrt{3}$ cannot be exactly represented digitally, this specification is slightly less accurate.

Finally, the covariance due to the IBD status among the offspring must be specified. The IBD status is estimated from IBD probabilities, which must be estimated using other software such as Merlin (Abecasis et al., 2002). For each pair of offspring in the family the estimated proportion of alleles shared IBD is computed. For family *i* the proportions are collected in the matrix Π_i, which has *n* rows and *n* columns for a family with *n* offspring. The elements *j, k* and *k, j* of this symmetric matrix equal π_{ijk}, the proportion IBD for offspring *j* and *k*. The diagonal elements of the matrix are equal to 1, which is the proportion IBD of an offspring with itself. Table 1 contains examples of those matrices, where the variables P1 to P4 are the matrix columns and the records of the subjects are the matrix rows. For Family 2, the variable P4 contains missing values, as this family consists of three subjects.

The contribution to the expected family covariance matrix is $\sigma_a^2\ \Pi_i$. Now we must specify random effects such that $\sigma_a^2\ \Pi_i = Z_{si}\ V_{si}\ Z_{si}'$. Set $V_{si} = \sigma_a^2\ I$, so that all random effects associated with the columns of Z_{si} are uncorrelated and have the same variance. Then $Z_{si}\ V_{si}\ Z_{si}' = \sigma_a^2\ Z_{si}\ Z_{si}'$. Therefore a matrix Z_{si} must be specified such that $Z_{si}\ Z_{si}' = \Pi_i$. There are various ways to do this. For example, the eigenvalue-eigenvector decomposition KDK' of Π_i can be calculated and Z_{si} can be set to $KD^{1/2}$. Here D is a diagonal matrix with the eigenvalues on the diagonal and $D^{1/2}$ contains the square roots of the diagonal elements. A simpler way is to compute the Cholesky decomposition of Π_i. The Cholesky decomposition of a symmetric nonnegative definite (i.e., positive definite or positive semidefinite) matrix A is defined as A = TT' for T a lower triangular matrix (the Cholesky decomposition is sometimes defined only for positive definite matrices; we follow Harville, 1997, and Rao & Rao, 1998). Statistical packages usually have a matrix algebra library with routines that compute eigenvalue-eigenvector and Cholesky decompositions. Some of these routines may be suitable only for positive definite matrices and then may return nonsensical results if the matrix Π_i is positive semidefinite (i.e., $\Pi_i \neq TT'$ for the T or T' that the routine returns). The matrix Π_i can become positive semidefinite, for example if MZ pairs are included or if π_{ijk} is equal to 1 for all sib pairs. We have implemented a routine in SPSS that is suitable for nonnegative definite matrices and is described for example by Harville (1997). The routine uses the *lag* function to compute the decomposition. No matrix algebra routine is needed. The routine computes T for each family separately and stores the columns of T as new SPSS variables. The maximum number of new variables is the maximum number of offspring over all families. For families that contain less than the maximum number of offspring, the columns corresponding to the superfluous variables are set to zero. Table 1 contains examples of those matrices encoded in the variables C1 to C4 for the Families 1 and 2. The variables C1 to C4 are the matrix columns, the records of the subjects are the matrix rows. In the mixed model procedure the new variables (i.e., the columns of T) have to be defined as *variables* (in contrast to factors) with associated uncorrelated random effects with the same variance. The grouping option is used to create separate random effects for each family. The covariance structure specification given here can of course also be used in other packages than SPSS. In SAS a covariance structure is available that can use Π_i directly.

It is not inconceivable that in rare cases the π_{ijk} that are computed from a program's estimated IBD probabilities may generate a Π_i matrix that is not nonnegative definite. Thus it seems good practice to check that, within some defined error level, TT' is indeed equal to Π_i. Finally, we note that the decomposition is not unique if Π_i is positive semidefinite. Therefore a substantive interpretation of the BLUPs of the random effects is probably quite useless.

Specification and Testing of the Fixed Part of the Model

The fixed part of the model as formulated earlier contains as regressors the mean b_i and the deviation $w_{ij} = g_{ij} - b_i$ as defined earlier. These variables must be included in the model as covariates. Other covariates or factors can also be included. The specification of these variables follows the same logic as in standard ANOVA or regression procedures.

The QTDT program employs a likelihood ratio test for the null hypothesis $\beta_w = 0$ against the alternative $\beta_w \neq 0$. Rejection of the null hypothesis is interpreted as evidence for an association effect. The same test can be performed in SPSS by running the model once with and once without the variable w_{ij} included in the model. The output of each analysis provides a value of $-2\ Log\ L$, where L is the likelihood. The difference between those values for the two models can then be used to perform a chi-squared test on 1 degree of freedom. For likelihood ratio tests of fixed effects the maximum likelihood estimation option should be used instead of the restricted maximum likelihood estimation option. Alternatively, the approximate F test for the estimate of β_w in the output can be used. Such tests are recommended by Pinheiro and Bates (2000).

A contrast $\beta_b - \beta_w$ can be specified to obtain an approximate F test for the hypothesis $\beta_b = \beta_w$. If the hypothesis $\beta_b = \beta_w$ is not rejected, a more powerful test for association can be obtained using only g_{ij} as predictor. Alternatively, the model can be fitted using only the original variable g_{ij} and the value of $2\ Log\ L$ for this model can be subtracted from $2\ Log\ L$ for the model containing w_{ij} and b_i to perform a 1 degree of freedom chi-squared test.

Some Alternative Modeling Options

If a marker or gene contains multiple alleles, the effects of the alleles can be tested simultaneously. The variables g_{ij}, b_i and w_{ij} can be computed for each allele. The variables corresponding to each allele can then be specified as covariates. Using all alleles creates linear dependence among the columns of X. The dependence is automatically detected by the program, which then sets the parameters corresponding to the redundant columns to zero. The effect of the alleles can be tested simultaneously by the likelihood ratio test or by specifying contrast among the parameters.

Multiple SNPs or markers can be included in the model as multiple fixed effects covariates or factors. One possible advantage of the inclusion of multiple SNPs is that their contribution can be tested simultaneously, which may lead to a more powerful test. The inclusion of multiple SNPs may also entail little loss of information as compared to the use of haplotypes (Chapman et al., 2003; Clayton et al., 2004).

If genotypes of both parents are available b_i can be computed as the average of the parental genotypes. In the absence of parental genotypes from both parents b_i is estimated as the average of the offspring genotypes. When all offspring are included or offspring within families can be regarded as a random sample the estimates of β_b and β_w should be similar whether or not both parental genotypes are available. If the dataset contains families with and without both parental genotypes, this may be tested by dividing the sample into groups of families with and without parental genotypes, fitting the model with the regressors as group specific covariates and testing the equality of the regression weights of the corresponding covariates in the groups. Of course, families can be divided into other groups based on other criteria.

Finally, the original family specific model can be fitted as a so-called random-coefficient regression model (see, for example, Longford, 1993). In those models the intercepts and regressions weights vary randomly over families. Thus the assumption that the within-family regression weights are the same for each family is relaxed in those models. Since such models are a special case of linear mixed models, they can be implemented with the same procedures. The average and variance of the regression weights can then be estimated and tested for deviation from zero. A zero variance implies that the within-family regression weights are the same for each family. Such models are often motivated by the argument that it is unlikely that the model can be sufficiently completely specified for each family (e.g., interaction effects can produce varying coefficients).

Discussion

We have shown here how the F-A model can be implemented in the procedure for linear mixed models in SPSS or other statistical packages. The model component of primary interest, which describes the regression of the phenotype on the genotype, is specified in the fixed part of the model in a way familiar from standard regression or ANOVA models. The specification of the within-family covariance structure is somewhat more involved, because the model is not used as a basis for deriving the covariance structure, but must be formulated to satisfy a structure that derives from the underlying genetics. We have presented some ways to specify this structure, but others are also possible. Except for the covariance structure due to the IBD status, which does require some extra calculations, the required structure is fairly easy to specify using standard options available in the procedure.

Mx (Neale, 1999) has become a very popular program for performing statistical analyses that involve the specification of covariance structures. The program is more flexible regarding the covariance structures that can be specified than programs for linear mixed models. This is partly a programmer's choice and partly due to the traditions out of which the programs developed (the experimental design tradition for linear mixed models and a merger of the psychometric factor

analytic tradition with the econometric tradition of simultaneous or structural equations modeling for Mx). Programs for linear mixed models offer a set of predefined covariance structures for the random effects. Specification of other structures is either not possible or requires programming by the users themselves. Mx allows users to choose their own structure in a fairly simple way, but it handles families of unequal size less elegantly. Unequally sized families can be handled in Mx in the same way as in linear mixed models using Mx's *definition variables*, but then the columns of Z matrices and the covariance structure of the random effects are not automatically generated.

The F-A model is one way to investigate association. Advantages of the model are that it belongs to a quite general and familiar class of models with well understood properties and that it can be implemented in widely available software. The model is less appropriate for selected samples, although the selection will often be such that the estimate of β_w is not or hardly affected. However, for selected samples other models (e.g., Lange et al., 2004) can be more appropriate. Other approaches are discussed in a quite general context by Terwilliger and Göring (2000).

Acknowledgments

We thank the reviewers for their comments. This work was supported by the Centre for Neurogenomics and Cognition Research (CNCR) of the Vrije Universiteit, Amsterdam.

References

Abecasis, G. R., Cardon, L. R., & Cookson, W. O. (2000). A general test of association for quantitative traits in nuclear families. *American Journal of Human Genetics, 66,* 279–292.

Abecasis, G. R., Cherny, S. S., Cookson, W. O., & Cardon, L. R. (2002). Merlin: Rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genetics, 30,* 97–101.

Chapman, J. M., Cooper, J. D., Todd, J. A., & Clayton, D. G. (2003). Detecting disease associations due to linkage disequilibrium using haplotype tags: A class of tests and the determinants of statistical power. *Human Heredity*, *56,* 18–31.

Clayton, D., Chapman, J., & Cooper, J. (2004). Use of unphased multilocus genotype data in indirect association studies. *Genetic Epidemiology*, 27, 415–428.

Falconer, D. S., & Mackay, T. F. C. (1996). *Introduction to quantitative genetics* (4th ed.). Harlow: Prentice Hall.

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2004). *Applied longitudinal analysis*. New York: Wiley.

Fulker D. W., Cherny S. S., Sham P. C., & Hewitt, J. K. (1999). Combined linkage and association analysis for quantitative traits. *American Journal of Human Genetics*, *64,* 259–267.

Gelman, A. (2005). Analysis of variance: Why it is more important than ever (with discussion). *The Annals of Statistics*, *33,* 1–53.

Hartley, H. O., & Rao, J. N. K. (1967). Maximum likelihood estimation for the mixed analysis of variance model. *Biometrika*, *54*, 93–108.

Harville, D. A. (1997). *Matrix algebra from a statistician's perspective*. New York: Springer.

Lange, C., DeMeo, D., Silverman, E. K., Weiss, S. T., & Laird, N. M. (2004). PBAT: Tools for family-based association studies. *American Journal of Human Genetics, 74,* 367–369.

Longford, N. T. (1993). *Random coefficient models*. Oxford: Clarendon Press.

McCulloch, C. E., & Searle, S. R. (2001). *Generalized, linear, and mixed models*. New York: Wiley.

Neale, M. C. (1999). *Mx: Statistical modeling* (5th ed.). Richmond, VA: Department of Psychiatry.

Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-Effects Models in S and S-Plus*. New York: Springer-Verlag.

Rao, C. R., & Rao, M. B. (1998). *Matrix algebra and its applications to statistics and econometrics*. Singapore: World Scientific.

Robinson, G. K. (1991). That BLUP is a good thing: The estimation of random effects (with discussion). *Statistical Science*, *6,* 15–51.

Searle, S. R., Casella, G., & McCulloch, C. E. (1992). *Variance components*. New York: Wiley.

Terwilliger, J. D., & Göring, H. H. H. (2000). Gene mapping in the 20th and 21st centuries: Methods, data analysis, and experimental design. *Human Biology*, *72,* 63–132.

Verbeke, G., & Molenberghs, G. (2001). *Linear mixed models for longitudinal data*. New York: Springer-Verlag.

Appendix A

We demonstrate here the calculation of the expected value of g_{ij}^2 if parental genotypes are available (this calculation is no doubt available somewhere).

Let the parental alleles m_{iF} and m_{jM} for $i, j = 1, 2$ be coded as zero or one for a given allele absent or present.

Then the expected value of g_{ij}^2 becomes
$\Sigma_i \Sigma_j (m_{iF} + m_{jM} - 1)^2 / 4 = \Sigma_i \Sigma_j [(m_{iF} + m_{jM})^2 + 1 - 2 (m_{iF} + m_{jM})] / 4.$

Because m_{iF} and m_{jM} are zero or one, $m_{iF}^2 = m_{iF}$ and $m_{jM}^2 = m_{jM}$.

Therefore this expression can be written as
$\Sigma_i \Sigma_j [(m_{iF} + m_{jM} + 2 m_{iF} m_{jM} + 1 - 2 (m_{iF} + m_{jM})] / 4 = \Sigma_i \Sigma_j [(2 m_{iF} m_{jM} + 1 - (m_{iF} + m_{jM})] / 4.$

Now $\Sigma_i \Sigma_j m_{iF} m_{jM} = (m_{1F} + m_{2F}) (m_{1M} + m_{2M})$ and $\Sigma_i \Sigma_j (m_{iF} + m_{jM}) = 2 (m_{1F} + m_{2F}) + 2 (m_{1M} + m_{2M})$.

Therefore the expectation becomes
$[2(m_{1F} + m_{2F}) (m_{1M} + m_{2M}) - 2 (m_{1F} + m_{2F}) - 2 (m_{1M} + m_{2M}) + 4] / 4$
$= [(m_{1F} + m_{2F}) (m_{1M} + m_{2M}) - (m_{1F} + m_{2F}) - (m_{1M} + m_{2M}) + 2] / 2$
$= [(g_F + 1) (g_M + 1) - (g_F + 1) - (g_M + 1) + 2] / 2$
$= (g_F g_M + g_F + g_M + 1 - g_F - g_M] / 2$
$= (g_F g_M + 1) / 2.$