BMC Genetics

BioMed Central

Proceedings

Open Access

A new family-based association test via a least-squares method

Song Yang*[1], Jungnam Joo[1], Ziding Feng[2] and Jing-Ping Lin[1]

Address: [1]Office of Biostatistics Research, National Heart, Lung, and Blood Institute, Bethesda, Maryland 20892, USA and [2]Cancer Prevention and Research Program, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA

Email: Song Yang* - yangso@nhlbi.nih.gov; Jungnam Joo - jooj@nhlbi.nih.gov; Ziding Feng - zfeng@fhcrc.org; Jing-Ping Lin - linj@nhlbi.nih.gov

* Corresponding author

from Genetic Analysis Workshop 14: Microsatellite and single-nucleotide polymorphism
Noordwijkerhout, The Netherlands, 7-10 September 2004

Published: 30 December 2005

BMC Genetics 2005, **6**(Suppl 1):S110 doi:10.1186/1471-2156-6-S1-S110

Abstract

To test the association between a dichotomous phenotype and genetic marker based on family data, we propose a least-squares method using the vector of phenotypes and their cross products within each family. This new approach allows covariate adjustment and is numerically much simpler to implement compared to likelihood- based methods. The new approach is asymptotically equivalent to the generalized estimating equation approach with a diagonal working covariance matrix, thus avoiding some difficulties with the working covariance matrix reported previously in the literature. When applied to the data from Collaborative Study on the Genetics of Alcoholism, this new method shows a significant association between the marker rs1037475 and alcoholism.

Background

Case-control studies provide an important tool to test for the association between disease outcomes and genetic markers [1,2]. Family-based association studies take advantage of existing data, such as data from a previous linkage analysis [3,4].

Incorporation of covariates into the analysis should increase the power to detect associations. However, within-family correlations must be considered. For this purpose the generalized estimating equation (GEE) approach [5] is often used. In the case of dichotomous phenotypes, the GEE approach usually specifies that the mean response is related to a set of covariates via a link function. As for the correlation, usually a common correlation is assumed for each pair of relatives in the working correlation matrix, although more accurate correlation structures are possible and may be more efficient. However, a common problem of GEE is that the working correlation matrix may be singular. Recently Slager et al. [6] showed in various simulation studies that the failure rate for the GEE could be quite high in some cases, and should not be ignored. To remedy this problem they proposed a score test approach for tests of association.

In this article we propose a new association test and apply it to the data from Collaborative Study on the Genetics of Alcoholism (COGA). This new test is derived from a least-squares approach in which the dichotomous responses and their cross products are used, rather than the usual procedure in which the estimating equations only use the observed responses themselves. This approach is asymptotically equivalent to a GEE approach with a diagonal working correlation matrix, and therefore the estimating equation is always well defined.

Methods

Let y_{ij} be the dichotomous phenotype from the j^{th} individual of the i^{th} family, where there are k_i members from the i^{th} family and n families in the sample. Let x_{ij} be the covariate vector, decomposed as $x_{ij} = (x_{ijm}, x_{ije})$, where x_{ijm}, x_{ije} represent the marker allele effect and measured covariates respectively. Suppose that, for the i^{th} family, the phenotypes are conditionally independent given a common ran-

dom effect u_i, where the u_i values are independent and identically distributed with gamma distribution with mean 1 and variance $\theta > 0$, and that, given u_i and x_{ij},

$$P(\gamma_{ij} = 1 | x_{ij}, u_i) = \exp(-u_i \exp(x_{ij}\beta)), \; j = 1, \ldots, k_i, \; i = 1, \ldots, n, \quad (1)$$

where $x_{ij}\beta = x_{ijm}\beta_m + x_{ije}\beta_e$. From this, we obtain that the mean of γ_{ij} is $\{1 + \theta \exp(x_{ij}\beta)\}^{-1/\theta}$. Numerically, it is more stable to work with the reparametrization $\eta = log(\theta)$. In this reparametrization, the mean of γ_{ij} is

$$P(\gamma_{ij} = 1) = A_{ij}(\eta, \beta), \; j = 1, \ldots, k_i, \; i = 1, \ldots, n, \quad (2)$$

where $A_{ij} = \{1 + \exp(\eta + x_{ij}\beta)\}^{-\exp(-\eta)}$. The joint distribution of $Y_i = (\gamma_{i1}, \ldots, \gamma_{ik_i})$ can also be obtained by integrating out the random effect u_i and thus the likelihood function has a closed form. This is an appealing and important feature of the above modelling approach when using the log-log link function and log-gamma random effect. In comparison, for dealing with correlated dichotomous responses, a commonly used model specifies that, conditional on a normal random effect, the marginals of the conditional distribution are given by the logistic link. In that situation, the likelihood function does not have a closed form and extensive numerical methods are needed.

Note that Equation (1) imposes the same correlation structure regardless of family relations. More accurate descriptions are possible by assuming different random effects for different family relations, but this increases the number of parameters to estimate. Petersen [7] discussed some random effect models for correlated life times. Similar structures can also be adapted for the dichotomous phenotypes. For ease of presentation and due to space limitation, we work with the simplified but illustrative Equation (1) here.

There have been some results on analysis of Equation (1). Conaway [8] proposed the log-log link and log-gamma random effect for correlated binary data. He focused on the case in which there are no covariates, but this model can be easily extended to accommodate covariates. Pulkstenis et al. [9] used the log-log link and log-gamma random effect in a case study of longitudinal binary data for pain relievers. Both of these papers focused on the maximum likelihood estimators (MLE) based on the marginal likelihood function.

For families of larger size, the likelihood function becomes increasingly more complicated. Contribution to the likelihood function from the i^{th} cluster has 2^{k_i} terms. Also MLE may be sensitive to model misspecifications. Here we propose a new least-squares approach for testing $H_0: \beta_m = 0$. Note that the parameters β and η can be identified from the marginal mean response function, thus a natural and simple approach is to use the GEE based on Equation (2). We further observe that, for the cross products $\gamma_{ij}\gamma_{il}$, $j \neq l$, in the i^{th} family, we have

$$E(\gamma_{ij}\gamma_{il}) = P(\gamma_{ij} = 1, \gamma_{il} = 1) = B_{ijl}(\eta, \beta),$$

where

$$B_{ijl}(\eta, \beta) = \{1 + \exp(\eta + x_{ij}\beta) + \exp(\eta + x_{il}\beta)\}^{-\exp(-\eta)}.$$

Considering that η is involved in the random effect induced correlation among family members, it may be more efficient to work with Y_i as well as cross products $\gamma_{ij}\gamma_{il}$. For the i^{th} family let Z_i be the $k_i(k_i + 1)/2 \times 1$ vector consisting of Y_i and the $k_i(k_i - 1)/2$ cross products $\gamma_{ij}\gamma_{il}$, $j \neq l$, $j, l = 1, \ldots, k_i$. Let $m_i = E(Z_i)$, and V_i be the diagonal matrix with variance of the components of Z_i on the diagonal. Then we define (β, η) as the minimizer of

$$\sum_{i=1}^{n} (Z_i - m_i)' V_i^{-1} (Z_i - m_i) \quad (3)$$

For obtaining V_i, m_i, we have

$$E(\gamma_{ij}) = P(\gamma_{ij} = 1) = A_{ij}(\eta, \beta), \quad (4)$$

$$Var(\gamma_{ij}) = A_{ij}(\eta, \beta) - A_{ij}^2(\eta, \beta), \quad (5)$$

$$Var(\gamma_{ij}\gamma_{il}) = B_{ijl}(\eta, \beta) - B_{ijl}^2(\eta, \beta), \quad (6)$$

with $A_{ij}(\eta, \beta)$, $B_{ijl}(\eta, \beta)$ defined previously. Note that (β, η) is asymptotically equivalent to the root of the estimating equation

$$\sum_{i=1}^{n} (\partial m_i)' V_i^{-1} (Z_i - m_i) = 0, \quad (7)$$

where ∂m_i is the vector of partial derivatives of m_i with respect to (β, η). In the above estimating equation the working covariance matrix is diagonal, and thus the estimating equation is always well defined. However, numerically it is more stable to use the least squares approach.

Once the estimators (β, η) are obtained, due to the asymptotic equivalence to the GEE approach, the covariance matrix of (β, η) can be estimated by the robust estimator

$$V_\beta = A^{-1}BA^{-1} \quad (8)$$

with

$$A = \sum_{i=1}^{n} \partial m_i V_i^{-1} (\partial m_i)', B = \sum_{i=1}^{n} \partial m_i V_i^{-1} (Z_i - m_i)(Z_i - m_i)' V_i^{-1} (\partial m_i)', \quad (9)$$

where (β, η) are replaced by (β, η). A more stable but numerically more intensive alternative for estimating the covariance matrix is to use a bootstrapping method to resample family units a large number of times. Decompose $\hat{\beta} = (\hat{\beta}_m, \hat{\beta}_e)$ where $\hat{\beta}_m$ is the estimator for β_m. Now the hypothesis H_0:β_m = 0 can be tested using the asymptotic normality of the z score based on $\hat{\beta}_m$.

Note that in the least-squares approach above, β can be interpreted as a regression parameter in the mean response function EZ_i, which includes the cross product terms. We can similarly define a least squares estimator of (β, η) by working with Y_i and its mean response function EY_i, without the addition of the cross product terms. In that case, β would be interpreted as a regression parameter in the mean response function EY_i. In various numerical studies, the addition of the cross product terms improves the efficiency for small and moderate samples sizes.

Results

We applied the proposed approach to the data from COGA. The data provide alcoholism diagnosis on 1,614 individuals from 143 families. We focus on two distinct categories for the alcoholism diagnosis, "affected" as case (609 individuals) and "purely unaffected" as control (261 individuals). The preliminary genome scan carried out for linkage analysis using the microsatellite data identified a gene *ADH3* on chromosome 4 as a candidate gene. We found 4 single-nucleotide polymorphisms (SNPs) (rs1036475, rs1491233, rs749407, rs980972), which are located in the physical map location of *ADH3* genes from the Illumina SNPs data. Without correcting the correlated structure between family members, a logistic regression on these 4 SNPs suggested that rs1037475 and rs980972 were significant predictors (p-values of 0.0032 and 0.0284, respectively). Also, a quick look at the 2 × 2 table stratified by sex showed some differences. This led us to the consideration of using sex as a covariate. Assuming a recessive genetic model, the new least-squares method showed a significant association between rs1037475 and alcoholism, with a p-value of 0.013. Further, the analysis showed a significant sex effect with p-value < 0.001. Without using the cross product terms, the corresponding least squares method also showed a significant association between rs1037475 and alcoholism with a p-value of 0.002, and a significant sex effect with p-value < 0.001. The smaller p-value for the association between rs1037475 and alcoholism might be due to the fact that a common correlation was assumed among all family members for the 870 individuals and 143 families. Violation of this assumption does not affect the mean response function $E\gamma_{ij}$ but would introduce some bias in the mean response function $E\gamma_{ij}\gamma_{il}$ for the cross product terms. This in turn might reduce the power of the corresponding association test. When we restricted our analysis to the 499 siblings in 141 families, we still found a significant sex effect, with or without using the cross product terms. However, with the cross product terms, a significant association between rs1037475 and alcoholism was found; and without the cross product terms, no such association was established. In all cases with the reduced dataset, the p-value was smaller with the cross product terms than without them.

Conclusion

In this article we have proposed a new test of association between dichotomous disease outcomes and genetic markers for family data. When applied to the data from COGA, this new approach indicated an association between SNP marker rs1037475 and alcoholism. This new approach has the flexibility of adjusting for covariates, and sex was a significant covariate in this analysis. The use of complementary log-log link function and the conjugate log-gamma random effect, rather than the more common combination of logistic link function and normal random effect, allowed us to obtain closed forms for the means and variances for the responses and their cross products. Using these quantities enables us to derive parametric estimators via the least squares approach that avoids the difficulty in the GEE approach created by singularity of the working correlation matrix. The least squares approach is more robust and computationally much simpler to implement than the likelihood approach.

Simulation studies also yielded evidence that the efficiency of the new approach is high and often its behavior on small samples is better than the more complicated likelihood-based approach.

Abbreviations

COGA: Collaborative Study on the Genetics of Alcoholism

GEE: Generalized estimating equation

MLE: Maximum likelihood estimators

SNP: Single-nucleotide polymorphism

Authors' contributions

SY was involved in the design of the study and statistical analysis, and drafted the manuscript. JJ and J-PL performed the statistical analysis and participated in revising

the manuscript. ZF was involved in the design of the study and participated in revising the manuscript.

References

1. Risch N: **Searching for genetic determinants in the new millennium.** *Nature* 2000, **405:**847-856.
2. Risch N, Merikangas K: **The future of genetic studies of complex human diseases.** *Science* 1996, **273:**1516-1517.
3. Whittaker JC, Morris A: **Family-based tests of association and/or linkage.** *Ann Hum Genet* 2001, **65:**407-419.
4. Witte JS, Gauderman W, Thomas D: **Asymptotic bias and efficiency in case-control studies of candidate genes and gene-environment interactions: basic family designs.** *Am J Epidemiol* 1999, **149:**693-705.
5. Liang KY, Zeger S: **Longitudinal data analysis using generalized linear models.** *Biometrika* 1986, **73:**13-33.
6. Slager SL, Schaid DJ, Wang L, Thibodeau SN: **Candidate-gene association studies with pedigree data: controlling for environmental covariates.** *Genet Epidemiol* 2003, **24:**273-283.
7. Petersen JH: **An additive frailty model for correlated life times.** *Biometrics* 1998, **54:**646-661.
8. Conaway MR: **A random effects model for binary data.** *Biometrics* 1990, **46:**317-328.
9. Pulkstenis EP, Ten Have TR, Landis JR: **Model for the analysis of binary longitudinal pain data subject to informative dropout through remediation.** *J Am Stat Assoc* 1998, **93:**438-450.