ESTIMATION OF LINKAGE AND ASSOCIATION FROM ALLELE TRANSMISSION (TDT) DATA

James D. Malley[1], Richard A. Redner[2], Thomas A. Severini[3], Judith A. Badner[4], Sinisa Pajevic [1], and Joan E. Bailey-Wilson[5]

[1]Center for Information Technology, National Institutes of Health, [2]Department of Mathematical and Computer Sciences, University of Tulsa, OK, [3]Department of Statistics, Northwestern University, IL, [4]Department of Psychiatry, University of Chicago, IL, and [5]National Human Genome Research Institute, National Institutes of Health.

Data

2 x 2 table of allele transmission counts from both parents

x_{11} = allele 1 is transmitted, allele 1 is not

x_{12} = allele 1 not transmitted, allele 2 is

x_{21} = allele 2 is transmitted, allele 1 is not

x_{22} = allele 2 not transmitted, allele 2 is

$$\begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{bmatrix}$$

Probability Model

$$\pi_{11} = p^2(1 + aB)$$

$$\pi_{12} = p(1-p)[1 + aB - aB\theta/(1-p)]$$

$$\pi_{21} = p(1-p)[1 - aBp/(1-p) + aB\theta/(1-p)]$$

$$\pi_{22} = (1-p)^2[1 - aBp/(1-p)]$$

where a = linkage disequilibrium -1
θ = linkage recombination fraction
B = mode of inheritance parameter
p = marker allele frequency
(Sham and Curtis, 1995)

Statistical Issues

Model is not identifiable at any point of the null space: not when $a = 0$ (no association) or when $\theta = 0.5$ (no linkage)

Information matrix is singular when $a = 0$. In fact when $a = 0$, linkage term θ drops out of the model

Standard estimation methods using maximum likelihood do not apply, since regularity conditions aren't valid for usual large-sample theory.

Solutions

(1) Introduce first parameterization: $d = aB$ = generalized association $d = 0$ if and only if $a = 0$

(2) Use parametric bootstrap for estimation of association, linkage and marker allele frequency

(3) Use parametric bootstrap percentile method for confidence intervals and testing

(4) Introduce second parameterization to improve testing $\theta = 0.5$

$$\tau = (d^*, p^*, \theta^*) = (-d/(d+1), p(d+1), 1-\theta) \quad \text{if } d < 0$$
$$= (d, p, \theta) \quad \text{if } d > 0.$$

maps $d < 0$ to $d^* > 0$, and θ^* in [0, 1]. Now use likelihood ratio test (*LR*) for $\theta^* = 0.5$, as $\theta = 0.5$ if and only if $\theta^* = 0.5$.

(5) Using an integrated Bayesian marginal likelihood yields a test (*IL*) and confidence interval for association that is entirely free of identifiability and information matrix problems

Results

(1) Confidence intervals used for testing association and linkage are essentially as powerful as the TDT, but some power loss for association test when marker allele is rare (e.g., marker allele frequency $p = 0.05$)

(2) Confidence intervals have good coverage probabilities, and intervals for θ will usefully distinguish between weak, moderate and tight linkage.

(3) *LR* test for $\theta = 0.5$ is essentially equivalent to the *TDT* at all values of association and linkage (!)

(4) *IL* test for association has **significantly** greater power than *TDT*: for $d = 2.5$, $p = 0.25$, $\theta = 0.35$, power for *TDT* = **0.63**, but power of *IL* test = **0.94**

(5) C code for program **ELAAT** and pdf file for paper (under review at *Genetic Epidemiology*) available at:

http://mscl.cit.nih.gov/spaj/elaat

Robustness under Population Stratification

(1) admixture = two populations, both with zero association, but different marker allele frequencies.

Under admixture model the *d*-test for association is just as robust as the TDT

(2) linkage heterogeneity = two populations, one with zero association, the other with nonzero association and tight or moderate linkage ($\theta = 0.01$ or 0.20)

note: when $d = 0$ in a subpopulation θ drops out of the model, and is arbitrary.

Under linkage heterogeneity model the *d*-test has essentially the same robustness as the TDT, except for some power loss for rare marker allele.

Estimation with Realistic Models

Five models from Sham and Curtis (1995) were studied, having varying levels of association and linkage, with both parents' data included. The *d*-test generally agreed with the TDT, as did the θ^*-test and the *LR* test for linkage. Confidence intervals for d, θ and p were generated, all showing good coverage probabilities. When association is small (d near 0) then intervals for θ are large (see note above).

Discussion

Estimation and confidence intervals for linkage and association, is practical, robust and analytically valid, despite statistical problems of the allele transmission model. A Bayesian integrated marginal likelihood greatly increases power for testing, estimation and confidence intervals for association. It will be applied next to testing, estimation and intervals for linkage.

Work is being extended to multiple affecteds, missing parents, and multiple marker alleles.