

Tilburg University

Testing for linkage disequilibrium, maternal effects, and imprinting with (in)complete case-parent triads using the computer program LEM

van den Oord, E.; Vermunt, J.K.

Published in:
American Journal of Human Genetics

Publication date:
2000

Document Version
Peer reviewed version

[Link to publication in Tilburg University Research Portal](#)

Citation for published version (APA):
van den Oord, E., & Vermunt, J. K. (2000). Testing for linkage disequilibrium, maternal effects, and imprinting with (in)complete case-parent triads using the computer program LEM. *American Journal of Human Genetics*, 66, 335-338.

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

Take down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Testing for Linkage Disequilibrium, Maternal Effects, and Imprinting

With (In)complete Case-Parent Triads Using the Computer Program LEM

To the Editor:

The traditional TDT (Transmission Disequilibrium Test) and related tests (see Thomson 1995) require complete triads of genotyped cases plus both parents to test for linkage disequilibrium in the presence of population admixture. A problem in empirical research is that some of the genotype measurements will usually be missing. These incomplete triads must be discarded to ensure the validity of the TDT (Curtis and Sham, 1995). Recently, Weinberg (1999a) developed likelihood ratio tests (LRTs) that employed the expectation-maximization algorithm (EM, Dempster et al. 1977) to use incomplete triads as well. Her tests capitalize on the fact that parent-child dyads may be informative about the genotype of the missing parent. For instance, if a child and parent are both homozygous for the variant allele, the genotype of the missing parent should comprise at least one copy. Simulations showed that the EM-LRTs were more powerful than the traditional tests that exclude incomplete triads and recaptured much of the loss in information caused by missing parental genotypes.

The widespread use this valuable approach seems however hampered by a lack of accessible software. Weinberg used for instance the commercial package GLIM. Although it is good and flexible software, GLIM is not very user-friendly (see remarks on their Internet site) and requires programming to perform the EM-LRTs. To suggest an alternative we discuss the script to perform Weinberg's tests (1999b) for linkage disequilibrium, maternal effects, or parent of origin effects in LEM which is a program for Log-linear analysis with missing data that employs the EM algorithm (Vermunt 1997a,b). An important advantage of using LEM is that with this script all the tests discussed by Weinberg (1999b) can readily be performed in the presence of all possible patterns of missing data without programming work or the need to learn more LEM syntax. Furthermore, the program is optimized for rapid convergence with EM algorithm, and standard errors of the estimates, fit indices, and a number of appropriate tests are automatically reported in the output so that they do not have to be programmed separately. A final

advantage is that the program (there is a DOS and Windows version) plus manual can be downloaded free of charge from Internet (http://cwis.kub.nl/~fsw_1/mto/mto3.htm#software).

Assume a biallelic locus so that the genotypes of the mother (M), father (P), and child (C) comprise no copy, one copy, or two copies of the variant allele. If the D s are dummy variables (e.g. $D_{(C=1)}$ means that the variable is one in all triads where $C = 1$, and zero otherwise), then the log of the expected cell counts $E(n_{MPC})$ of Weinberg's full model (1999b, see table 1) can be written as:

$$\ln[E(n_{MPC})] = \gamma_j + \beta_p D_{(C=1)} + \beta_2 D_{(C=2)} + \alpha_1 D_{(M=1)} + \alpha_2 D_{(M=2)} + \ln(w_{MPC})$$

where $e^{\gamma_j} = \mu_j$ are the mating-type-stratum effects (e is natural exponent), $e^{\beta_p} = R_p$ is the ratio of the risk of disease for genotypes with one copy versus genotypes with no copies of the variant allele, $e^{\beta_2} = R_2$ is the risk ratio when the genotype comprises two versus no copies of the variant allele, $e^{\alpha_1} = S_1$ is the risk ratio or maternal effect when the mother has one versus no copies of the variant allele, and $e^{\alpha_2} = S_2$ is the risk ratio when the mother has two versus no copies of the variant allele. The w_{MPC} are cell weights (this becomes clearer when the component is moved to the left hand side of the equation so that we obtain $\ln[E(n_{MPC})] - \ln[E(w_{MPC})] = \ln[E(n_{MPC}/w_{MPC})]$) or in GLIM terminology $\ln(w_{MPC})$ is called the offset. The weights can have four different values. Firstly, they can be zero. Because the expected counts in these cells have to be multiplied with $e^{\ln(0)} = 0$ the implication is that the cell frequencies are fixed to zero. This weight is therefore assigned to combinations such as $M=2, P=2, C<2$ that cannot occur for theoretical reasons. They are also useful in the context of recovering information from incomplete triads. For example, if in the above situation the genotype of the child would be missing the zero weights for $C<2$ imply that the missing genotype must comprise two copies of the variant allele. Secondly, the weights can be one so that the expected cell counts are multiplied with $e^{\ln(1)} = 1$ implying that the frequencies as predicted by R_p , R_2 , S_1 , and S_2 remain unaltered. Thirdly, in the triads $M=2, P=1, C=1$; $M=2, P=0, C=1$; and $M=1, P=0, C=1$ ($M > F$) where the child receives the copy of the variant allele from the mother, the weights equal the "parent of origin" or "imprinting" effect I_m . Because the models also specifies a "main" effect $e^{\beta_p} = R_p$, the total effect of $C=1$ on the expected count becomes $I_m R_p$. It is a bit unusual to use

parameters as weights. The cause are the triads consisting entirely of heterozygotes ($M=P=C=1$) for whom only the total cell count is observed, and it is unclear how many children receive the variant allele from the mother and how many from the father. As a result the effect of $C=1$ on the cell count involves the sum $R_p + I_m R_p$ which cannot be modeled in the usual way as products of effects. The effect of $C=1$ is therefore written as $(1 + I_m)R_p$ where $(1 + I_m)$ is the cell weight that can be modeled as a sum of effects.

A LEM script that estimates this model in the presence of all possible patterns of missing genotypes is shown in the Appendix. The data are analyzed as a 3 by 3 by 3 table (indicated in the script by the last three numbers after the statement *dim*) defined by the three manifest (*man 3*) or measured genotypes of the mother, father and child labeled M, P, and C (see *lab* statement). The cell indices correspond with the number of copies of the variant allele plus one. Thus, the count of the triads $M=0, P=2, C=1$ falls into cell 1,3,2. The cells are numbered in increasing order where the last indices change first (1,1,1; 1,1,2; 1,1,3; 1,2,1; 1,2,2 etc.). The statements *mod* and *des* are used to specify the model and parameters. The *mod* statement indicates the number of parameters and the margin of the table that is affected. For instance, *fac(C,2)* means that two parameters or main effects are estimated for the effects of the genotype of the child. The margin of C consists of three cells and the *des* statement specifies how the parameters affect these cells. In this case “0 1 2” means that the effect in all cells where $C=0$ is zero so that this category is used as the baseline, that the first parameter represents the effect in all cells where $C=1$ (β_p), and the second parameter the effect in all cells where $C=2$ (β_2). The mating-type-stratum effects are defined by the specific combination of the maternal and paternal genotype and therefore pertain to the margin MP. Although there are $3 \times 3 = 9$ possible combinations, due to the assumed symmetry across parents within each mating type (e.g. $M=1, P=2$ and $M=2, P=1$ have equal effects) only 6 effects are estimated. LEM knows such a symmetric margin as the pre-specified design 3a, so that by using the statement *spe(MP,3a)* there is no need for further specification in the *des* statement. The weights are combinations of constants and the imprinting parameter β_m and specified with the help of a latent variable X (statement *lat 1*) that has two discrete classes (the second number after the command *dim*). The effects of the first class are 0 implying an impact of $e^0 = 1$ on the cell counts, and the effects of the second class

are β_m corresponding with the imprinting parameter $e^{\beta_m} = I_m$. Because only one parameter is estimated and this parameter is modeled as an effect of the second latent class, $fac(X, I)$ is used in the model statement and “0 1” in the design statement. The command $wei(XMPC)$ means that the effects of the latent classes on the cell counts are mediated by the weight vector. The values for $X=1$ after the statement sta $wei(XMPC)$ specify which of the 27 cells are affected by the first latent class (zero is not affected and one means affected), and the values for $X=2$ the cells that are affected by the second latent class. For the combinations that cannot occur, two zeroes are specified so that the expected cell counts are multiplied with $e^{\ln(w_{MPC})} = e^{\ln(0 \times 1 + 0 \times I_m)} = 0 \times 1 + 0 \times I_m = 0$. For the triads where $M > F$ a value of zero is specified for the first latent class, and a value of one for the second latent class. This implies an effect of $0 \times 1 + 1 \times I_m = I_m$ on the cell count. For the triads $M=P=C=1$ ones are specified for both latent classes so that the total impact becomes $1 \times 1 + 1 \times I_m = (1 + I_m)$. Note that by fixing the effect of the second latent class to zero as well (no imprinting $\beta_m = 0$ and $e^{\beta_m} = 1$), the weight would become one for all combinations that can occur and 2 for triads consisting entirely of heterozygotes.

Tests can be performed by merely changing the number of parameters in the *mod* statement plus the parameter specification in the *des* statement. For instance, to fit a model without imprinting we would use $fac(X, 0)$ instead of $fac(X, I)$, and “0 0” instead of “0 1”. The output of LEM reports the log-likelihoods plus a variety of other fit indices, parameter estimates, standard errors of the estimates, and comparisons between estimated and observed cell frequencies. To perform a likelihood ratio test one needs to take two times the difference between the log-likelihoods of the full model and the model without imprinting. Because one parameter is fixed to zero, this statistic will be chi-square distributed with one degree of freedom. A number of submodels are worth mentioning. If we assume that there is no imprinting and no maternal effects ($fac(M, 0)$ and *des* “0 0 0”), Schaid and Sommer’s Genotype Relative Risk method (1993) is obtained where $e^{\beta_p} = P_1$ and $e^{\beta_2} = P_2$. Recessive models $\beta_p = 0, \beta_2 > 0$ can be specified by $fac(C, I)$ and *des* [0 0 1], and dominance models $\beta_p = \beta_2$ by $fac(C, I)$ and *des* [0 1 1]. Although for polygenic traits it may be a somewhat coincidental situation (Van den Oord, in press), a gene-dose model is obtained by

imposing the constraint $\beta_2 = 2 \times \beta_p$ using *cov(C,1)* and *des [0 1 2]*. Note that the command *cov* instead of *fac* has to be used. The reason is that C is now treated as a “covariate” instead of a nominal factor because the expected cell frequencies are linear in C (if C=0 the effect is $0 \times \beta_p$, if C=1 the effects is $1 \times \beta_p$, and if C=2 the effect is $2 \times \beta_p$). This latter test is asymptotically equivalent to the traditional TDT (Spielman et al. 1993), so that LEM also enables one to perform a variant of the TDT with incomplete triads.

The name after the command *dat* in the LEM script means that the data are in the file TEST.DAT. The number after *rec* shows that there are 100 triads. The data are in free format with one record for each triad. The first two records are:

3 3 3

1 0 1

The numbers indicate the cell to which the triad belongs and zeroes are used for missing genotypes. Thus, 3 3 3 pertains to a triad where all three members have two copies of the variant allele (M=F=C=2), and 1 0 1 to a triad where the genotype of father is missing and the mother as well as the child have zero copies of the variant allele. There are seven possible data patterns. This is indicated by the first number after the command *dim*. To inform LEM about the nature of patterns the statement *sub* is used. For example, MPC pertains to triads with no missings, and MC to triads where the genotype of the father is missing.

Maximum likelihood estimates are obtained by means of the EM algorithm. The E step of this iterative method is of the form:

$$n_{MPC}^e = n_{MPC} + n_{MP0} \pi_{C|MP}^e + n_{M0C} \pi_{P|MC}^e + n_{0PC} \pi_{M|PC}^e + n_{M00} \pi_{PC|M}^e + n_{0P0} \pi_{MC|P}^e + n_{00C} \pi_{MP|C}^e$$

The zeroes indicate that the genotype is missing and superscript *e* reflects that the statistic is estimated and not observed. Thus, estimates of observed cell entries are computed using the observed data plus the current estimates of the predicted cell frequencies that are based on the information from incomplete triads as well. In the M step of the EM algorithm, the predicted cell counts n_{MPC}^e are treated as if they were really observed to obtain new estimates of the log-linear parameters and the cell frequencies. To speed up the estimation the program is instructed to switch to Newton-Raphson after 10 iterations (command *new*). Convergence is usually reached within a second on an ordinary computer.

To examine whether the script worked properly we first computed expected cell frequencies using the full model. Fitting the script to these frequencies gave a perfect fit and the correct parameters were recovered. Next we simulated 1,000 samples of 100 triads in 6 different conditions that assumed percentages of missing paternal genotypes of 0, 10, 20, 30, 40, and 50 per cent. The data were simulated assuming two completely segregated strata that were mixed so that the sample comprised about equal proportions of triads from each stratum. Within the first stratum the frequency of the disease allele was .10 and the risk of disease .01, and within the second stratum the frequency of the disease allele was .9 and the disease risk of .1 was 10 times larger. When data were simulated assuming no genetic effects, the null-hypothesis $\beta_p = \beta_2 = 0$ was rejected in 4.3, 6.1, 4.4, 5.5, 4.8, and 4.9 per cent of the 1,000 samples. Z-tests showed that *none* of the rejection rates differed significantly from the expected Type I or alpha error of 5 per cent. This showed that the tests for genetic effects were accurate even in conditions where the number of missing paternal genotypes was substantial. The whole simulation was repeated by generating data assuming $\beta_2 > 0$. The rejection rates of the null-hypothesis or the power in the 6 conditions was 52.1, 53.4, 48.0, 49.9, 42.2, and 43.8 per cent respectively. This confirmed results reported by Weinberg (1999a) showing that even with many incomplete triads the EM LRT recaptures much of the loss in information.

The scripts for all the tests discussed in this article, sample data, and output can be downloaded from Internet site “<http://www.fss.uu.nl/ped/welcome.html>”. In addition, Weinberg (1999b) proposed an alternative test for parent-of-origin effects that is also valid in situations where the locus is a marker rather than a candidate gene. A script plus documentation for this PO-LRT can be found on that site as well.

Edwin J. C. G. van den Oord,¹ and Jeroen K. Vermunt²

¹Department of Child and Adolescent Psychology, Utrecht University, Utrecht and ²Department of Methodology, Tilburg University, Tilburg

Appendix

The following LEM script estimates the full model reported Weinberg (1999b, table 1). The bold numbers and text indicate the only instructions that need to be changed to perform significance tests and adjust the data format to ones own data.

```

* variable and table definition
man 3          * # manifest variables
lat 1          * # latent variables
res 1          * # response variables
dim 7 2 3 3 3 * dimension table: R x X x M x P x C
lab R X M P C  * labels R=patterns, X=lat var, M=moth, P=fath, C=child
sub MPC MP MC PC M P C * possible data patterns or subgroups

* model
mod XMPC {spe(MP,3a) * mating-type-stratum effects
          fac(C,2)    * child effects
          fac(M,2)    * maternal effects
          fac(X,1)    * imprinting effect
          wei(XMPC)}  * weight vector

* data format
rec 100        * # records or triads
dat TEST.DAT    * data file

* design matrix/parameter specification
des [0 1 2     * child effects
     0 1 2     * maternal effects
     0 1 ]     * imprinting effect

* values weight vector
sta wei(XMPC)
*M=0,P=0;M=0,P=1;M=0,P=2;M=1,P=0;M=1,P=1;M=1,P=2;M=2,P=0;M=2,P=1;M=2,P=2
[ 1 0 0  1 1 0  0 1 0  1 0 0  1 1 1  0 1 1  0 0 0  0 0 1  0 0 1  * X=1
  0 0 0  0 0 0  0 0 0  0 1 0  0 1 0  0 0 0  0 1 0  0 1 0  0 0 0] * X=2

* optimization
new 10 1      * switch to Newton-Raphson after 10 EM iterations

```


References

- Curtis D, Sham PC (1995) A note on the application of the transmission disequilibrium test when a parent is missing. *Am J Hum Genet* 56:811-812
- Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc Ser B* 39:1-22
- Schaid DJ, Sommer SS (1993) Genotype relative risks: methods for design and analysis of candidate-gene association studies. *Am J Hum Genet* 53:1114-1126
- Spielman RS, McGinnis RE, Ewens WJ (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet* 52:506-516
- Thomson G (1995) Mapping disease genes: family-based association studies. *Am J Hum Genet* 57: 487-498
- Van den Oord EJCG. A comparison between different designs and tests to detect QTLs in association studies. *Behav Genet* (in press)
- Vermunt JK (1997a) LEM: a general program for the analysis of categorical data. Tilburg University, Tilburg
- Vermunt JK (1997b) Log-linear models for event histories. *Adv Quant Techn Social Sciences* 8, Sage, Thousand Oakes
- Weinberg CR (1999a) Allowing for missing parents in genetic studies of case-parent triads. *Am J Hum Genet* 64:1186-1193
- Weinberg CR (1999b) Methods for detection of parent-of-origin effects in genetic studies of case-parent triads. *Am J Hum Genet* 65:229-235

Address for correspondence and reprints: Dr. Edwin van den Oord, Department of Child and Adolescent Psychology, Universiteit Utrecht, Heidelberglaan 1, Centrumgebouw Zuid, Postbus 80140, 3508 TC, Utrecht, The Netherlands, E-mail: E.vandenOord@fss.uu.nl