Scientific Research Publishing

Open Journal of Statistics, 2021, 11, 713-719
https://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

Transmission Based Conditional Logistic Model for Testing Main and Interaction Effects

Caixia Li[1], Peixing Li[1,2*]

[1]School of Mathematics, Sun Yat-sen University, Guangzhou, China
[2]Guangdong Province Key Laboratory of Computational Science, Sun Yat-sen University, Guangzhou, China
Email: *lnslpx@mail.sysu.edu.cn

How to cite this paper: Li, C.X. and Li, P.X. (2021) Transmission Based Conditional Logistic Model for Testing Main and Interaction Effects. *Open Journal of Statistics*, **11**, 713-719.
https://doi.org/10.4236/ojs.2021.115042

Received: August 30, 2021
Accepted: October 8, 2021
Published: October 11, 2021

Abstract

Transmission disequilibrium test (TDT) is a popular family based genetic association method. Under multiplicative assumption, a conditional logistic regression for matched pair, affected offspring with allele transmitted from parents and pseudo-offspring (control) with allele non-transmitted from parents, was built to detect the main effects of genes and gene-covariate interactions. When there exist genotype uncertainties, expectation-maximization (EM) algorithm was adopted to estimate the coefficients. The transmission model was applied to detect the association between M235T polymorphism in AGT gene and essential hypertension (ESH). Most of parents are not available in the 126 families from HongKong Chinese population. The results showed M235T is associated with hypertension and there is interaction between M235T and the case's sex. The allele T is higher risk for male than female.

Keywords

Transmission Disequilibrium Test, Gene-Covariate Interaction, Conditional Logistic Model, Expectation-Maximization Algorithm

1. Introduction

To avoid false positive results because of confounding, some genetic association methods based on pedigree were proposed. Transmission disequilibrium test (TDT) introduced by Spielman *et al.* [1] is a family-based test. Only trios, including parents and one affected offspring, are needed in TDT. TDT were generalized for multi-allelic markers [2] [3].

For many diseases, especially those of late onset age, parental information is not available, and the classical TDT for triad data cannot implement. Some methods for incomplete families were proposed, e.g. S-TDT [4] and Sibass [5] for

siblings, and PDT (pedigree disequilibrium test) for general pedigrees [6].

Single nucleotide polymorphisms (SNPs) are highly abundant, stable genetic markers in humans. TDT methods for haplotype transmission using multiple tightly linked loci were proposed [7] [8]. In these approaches, the individuals' underlying haplotypes must be reconstructed using observed genotypes even if there are no missing genotype data.

Genotype relative risk may vary across levels of environmental exposure. That is to say, there maybe exist interactions between gene and covariates. Taub *et al.* [9] derive an extension of genotypic TDT to assess gene-environment interactions for binary environmental variables. However, joint effects of genotype and exposure or environmental covariates were not considered in classical allelic TDT. In the present paper, under conditional logistic regression structure, an allele or haplotype transmission based model is built to detect and assess main effects and gene-environment interactions.

2. Method

2.1. Transmission Based Model

Let H denote the number of allele or haplotypes for one or multiple tightly linked loci. The collection of all possible $H(H+1)/2$ genotypes is $G=\{1/1,1/2,\cdots,1/H,2/2,2/3,\cdots,2/H,\cdots,(H-1)/H,H/H\}$.

For an affected offspring with genotype g and covariate vector $X=(X_1,X_2,\cdots,X_k)'$, the joint effects of gene and covariates are considered in the genotype risk relative to reference genotype $g_0=(H/H)$, *i.e.*

$$R(g\mid X)=P(A\mid g,X)/P(A\mid g_0,X) \tag{1}$$

where A is being affected. Let g_f,g_m,g_c denote the genotypes for father, mother and the affected offspring respectively. Under some conditional independence, $P(g_c\mid g_f,g_m,X)=P(g_c\mid g_f,g_m)$ and $P(A\mid g_c,g_f,g_m,X)=P(A\mid g_c,X)$, then

$$\begin{aligned}
&P(g_c\mid g_f,g_m,A,X)\\
&=\frac{P(g_c,g_f,g_m,A\mid X)}{\sum_{g\in G_c}P(g,g_f,g_m,A\mid X)}\\
&=\frac{P(g_f,g_m\mid X)P(g_c\mid g_f,g_m,X)P(A\mid g_c,g_f,g_m,X)}{\sum_{g\in G_c}P(g_f,g_m\mid X)P(g\mid g_f,g_m,X)P(A\mid g,g_f,g_m,X)}\\
&=\frac{P(g_c\mid g_f,g_m)P(A\mid g_c,g_f,g_m,X)}{\sum_{g\in G_c}P(g\mid g_f,g_m)P(A\mid g,g_f,g_m,X)}\\
&=\frac{P(A\mid g_c,X)}{\sum_{g\in G_c}P(A\mid g,X)}=\frac{R(g_c\mid X)}{\sum_{g\in G_c}R(g\mid X)},
\end{aligned} \tag{2}$$

where G_c is the collection of all possible genotypes of a child given both par-

ents' genotypes. Let $\left(g_f, g_m\right) = \left(i/j, k/l\right)$. Then

$$G_c = \begin{cases} \{i/k, i/l, j/k, j/l\}, & i \neq j, k \neq l, \\ \{i/k, i/l\}, & i = j, k \neq l, \\ \{i/k, j/k\}, & i \neq j, k = l, \\ \{i/k\}, & i = j, k = l. \end{cases} \tag{3}$$

Suppose that the genotype relative risk satisfies robust multiplicative model

$$R^2\left(i/j \mid X\right) = R\left(i/i \mid X\right) R\left(j/j \mid X\right) \tag{4}$$

and $R\left(i/i \mid X\right) = \exp\left(2\alpha_i + 2\beta_i' X\right)$ with $\beta_i = \left(\beta_{i1}, \beta_{i2}, \cdots, \beta_{ik}\right)'$, and then

$$R\left(i/j \mid X\right) = \exp\left(\alpha_i + \beta_i' X\right) \exp\left(\alpha_j + \beta_j' X\right). \tag{5}$$

Therefore, both parents' transmission probability

$$\begin{aligned} & P\left(\to i, \to k \mid i/j, k/l, A, X\right) \\ & = P\left(g_c = i/k \mid g_f = i/j, g_m = k/l, A, X\right) \\ & = \frac{R\left(i/k \mid X\right)}{R\left(i/k \mid X\right) + \delta_{kl} R\left(i/l \mid X\right) + \delta_{ij} R\left(j/k \mid X\right) + \delta_{ij}\delta_{kl} R\left(j/l \mid X\right)} \\ & = \frac{\exp\left(\alpha_i + \beta_i' X\right)}{\exp\left(\alpha_i + \beta_i' X\right) + \left(1 - \delta_{ij}\right) \exp\left(\alpha_j + \beta_j' X\right)} \\ & \quad \cdot \frac{\exp\left(\alpha_k + \beta_k' X\right)}{\exp\left(\alpha_k + \beta_k' X\right) + \left(1 - \delta_{kl}\right) \exp\left(\alpha_k + \beta_k' X\right)}, \end{aligned} \tag{6}$$

where

$$\delta_{ij} = \begin{cases} 1, & i \neq j, \\ 0, & i = j. \end{cases} \quad \delta_{kl} = \begin{cases} 1, & k \neq l, \\ 0, & k = l. \end{cases} \tag{7}$$

Equation (6) means paternal transmission and maternal transmission are independent, and transmission probability for a parent (father or mother) with genotype i/j is

$$P\left(\to i \mid i/j, A, X\right) = \frac{\exp\left(\alpha_i + \beta_i' X\right)}{\exp\left(\alpha_i + \beta_i' X\right) + \left(1 - \delta_{ij}\right) \exp\left(\alpha_j + \beta_j' X\right)}. \tag{8}$$

For a homozygous parent, transmission probability (8) is 1. For a heterozygous parent with genotype $i/j\left(i \neq j\right)$, we introduce dummy variables

$$Z_h^c = \begin{cases} 1, & \text{allele } h \text{ is tranmitted}, \\ 0, & \text{otherwise}. \end{cases} \quad Z_h^{pc} = \begin{cases} 1, & \text{allele } h \text{ is non-tranmitted}, \\ 0, & \text{otherwise}. \end{cases} \tag{9}$$

Then a heterozygous parent with genotype $i/j\left(i \neq j\right)$ is just

$$\begin{aligned} & P\left(\to i \mid i/j, A, X\right) \\ & = \frac{\exp\left[\sum_{h=1}^{H-1} \left(\alpha_h Z_h^c + \beta_h Z_h^c X'\right)\right]}{\exp\left[\sum_{h=1}^{H-1} \left(\alpha_h Z_h^c + \beta_h Z_h^c X'\right)\right] + \exp\left[\sum_{h=1}^{H-1} \left(\alpha_h Z_h^{pc} + \beta_h Z_h^{pc} X'\right)\right]} \end{aligned} \tag{10}$$

Equation (10) can be regarded as conditional logistic model for n_{het} matched

pairs, where n_{het} is the number of heterozygous parents. The homozygous parents are excluded because of no contribution to likelihood. In such matched data, the affected offspring with predictors ($Z_1^c, Z_2^c, \cdots, Z_H^c, X$) was taken as case, the pseudo offspring with non-transmitted genotype with predictors ($Z_1^{pc}, Z_2^{pc}, \cdots, Z_H^{pc}, X$) was taken as matched controls. The parameters α_i and $\beta_i \left(i = 1, 2, H-1\right)$ measure the main effects of alleles and gene-covariate interaction effects. However, effects of covariates X cannot be included since there is no difference between the X values of the case and matched control.

The maximum likelihood estimates (MLEs) of the parameters can be given via standard conditional logistic model or stratified proportional hazard Cox model, such as PHREG (proportional hazard regression) procedure in statistical software SAS, or coxph in R package "survival".

2.2. EM Algorithm for Dealing Ambiguities in Allele Transmission

Haplotype phase is often uncertain for linked multi-locus genotype. There may be several haplotype pairs compatible with observed genotype. In addition, even when only one locus is considered, there might be missing parental genotypes, especially for late-onset diseases. Therefore, there are ambiguities to decide which allele or haplotype is transmitted from the parent.

Suppose there are N parents-case trios, and then there are $2N$ parents in all. The genotypes the r-th parent and his/her offspring are denoted by g_r, g_{rc}, covariates vector for the offspring is $X_r = \left(X_{r1}, X_{r2}, \cdots, X_{rk}\right)'$. The log-likelihood

$$\begin{aligned} \ln L\left(\alpha, \beta\right) &= \sum_{r=1}^{2N} \ln \left\{ \sum_{\left(i_r, j_r\right) \in \tilde{G}_r}^{H} P\left(\to i_r \mid i_r / j_r, A, X_r\right) \right\} \\ &= \sum_{r=1}^{2N} \ln \left[\sum_{\left(i_r, j_r\right) \in \tilde{G}_r} \frac{\exp\left(\alpha_{i_r} + \beta_{i_r}' X_r\right)}{\exp\left(\alpha_{i_r} + \beta_{i_r}' X_r\right) + \left(1 - \delta_{ij}\right) \exp\left(\alpha_{j_r} + \beta_{j_r}' X_r\right)} \right], \end{aligned} \tag{11}$$

where $\tilde{G}_r$ is the set of haplotype groups $\left\{i_r, j_r\right\}$ which haplotype pair $\left\{i_r, j_r\right\}$ is compatible with parent genotype g_r.

It is difficult to find the MLEs of parameter $\left(\alpha, \beta\right)$ directly. However, if we take underling haplotype pairs as "missing data" in Expectation-maximization (EM) algorithm, an iterative procedure can be provided to find the MLE. Given the current estimate, the expected complete-data log-likelihood in E (expectation) step is given by

$$Q\left(\alpha, \beta \mid \alpha^{(t)}, \beta^{(t)}\right) = \sum_{r=1}^{2N} \left\{ \sum_{i, j=1}^{H} \omega_r^{(t)}\left(i, j\right) \ln P\left(\to i \mid i / j, A, X_r\right) \right\}, \tag{12}$$

where

$$\omega_r^{(t)}\left(i, j\right) = \begin{cases} \dfrac{F_i F_j \exp\left(\alpha_i^{(t)} + \beta_i'^{(t)} X_r\right)}{\sum_{(k,l) \in \tilde{G}_r} F_k F_l \exp\left(\alpha_k^{(t)} + \beta_l'^{(t)} X_r\right)}, & \left(i, j\right) \in \tilde{G}_r, \\ 0, & \left(i, j\right) \notin \tilde{G}_r. \end{cases} \tag{13}$$

$Q\left(\alpha,\beta\,|\,\alpha^{(t)},\beta^{(t)}\right)$ can be regarded as the log-likelihood for a weighted conditional logistic model for matched case and controls. However, haplotype frequencies $\left(F_1,F_2,\cdots,F_H\right)$ are often unknown and must be estimated too. Therefore, starting with initial values $\left(\alpha^{(0)},\beta^{(0)}\right)$ and $\left(F_1^{(0)},F_2^{(0)},\cdots,F_H^{(0)}\right)$, the $(t+1)$-th iteration of EM algorithm consists of 2 steps.

Step 1: Calculate the weights $\omega_r^{(t)}\left(i,j\right)\left(i,j=1,2,\cdots,H\right)$, and obtain MLE of $\left(\alpha^{(t+1)},\beta^{(t+1)}\right)$ via weighted conditional logistic model for matched case-pseudo-controls.

Step 2: Update hapotype frequencies

$$F_i^{(t+1)}=\frac{\sum_{r=1}^{2N}\sum_{u=1}^{H}\omega_r^{(t)}\left(u,i\right)}{2N},\ i=1,2,\cdots,H. \tag{14}$$

Likelihood ratio test (LRT) can be used to detect gene effect and gene-covariates interactions. Likelihood ratio tests can be used to select model or to test gene effect and gene-covariates interaction. For example, if we consider only one SNP and one covariate, we can construct three models, the null model in which $\alpha=\beta=0$, the model without interaction in which $\beta=0$, and the full model with interaction. Then we can use $\Lambda_1=2\ln L\left(\hat{\alpha},0\right)-2\ln L\left(0,0\right)$ to test gene effect and $\Lambda_2=2\ln L\left(\hat{\alpha},\hat{\beta}_k\right)-2\ln L\left(\hat{\alpha},0\right)$ to test gene-covariates interaction.

3. Application

Essential hypertension is a multi-factional disorder that is influenced by genetic and environmental factors. The angiotensinogen (AGT) gene of the renin-angiotensin system (RAS) has been considered important elements in blood pressure regulation. Some studies show the M/T polymorphism in exon 2 of the AGT gene at position 235 (M235T) has been related to essential hypertension with controversy in white Europeans [10] [11].

In our study, 126 families with at least one hypertensive sibling, a total of 434 siblings from Hong Kong Chinese population are included in the analysis. As shown in **Table 1**, 59.5% of the families had two or three siblings with a further 33.4% having four or five siblings, and parents are not available in most of the families (86.5%). The information of siblings is very useful to reduce the uncertainty of the transmission from parent with unknown genotype.

The AGE gene M235T and covariate gender are introduced into the proposed model. Give initial value $\left(\alpha^{(0)},\beta^{(0)}\right)=\left(0,0\right)$ and $F^{(0)}=\left(0.5,0.5\right)$ and precision $\varepsilon=10^{-5}$. After the EM iterative procedure (shown in Section 2.2) stops, the MLEs for the parameters are $\hat{\alpha}=1.2876$ and $\hat{\beta}=-0.7378$, where allele M is

Table 1. Nuclear families in the analyse.

With parents			With siblings				
0	1	2	2	3	4	5	≥6
109 (86.5%)	11 (8.7%)	6 (4.8%)	40 (31.7%)	35 (27.8%)	22 (17.5%)	20 (15.9%)	9 (7.1%)

reference allele. To detect the effect of M235T and interaction effect of M235T*sex, we perform likelihood ratio test (LRT) with statistic $\Lambda_1 = 2\ln L(\hat{\alpha},0) - 2\ln L(0,0)$ and $\Lambda_2 = 2\ln L(\hat{\alpha},\hat{\beta}) - 2\ln L(\hat{\alpha},0)$, respectively. The log-likelihoods with $\ln L(\hat{\alpha},\hat{\beta}) = -85.477$, $\ln L(\hat{\alpha},0) = -87.471$, and $\ln L(0,0) = -100.937$ yield $\Lambda_1 = 26.932$ ($p < 0.001$) and $\Lambda_2 = 3.987$ ($p = 0.046$).

The results show that M235T is association with hypertension and there is interaction between M235T and gender. The relative risk for allele T is $\exp(\hat{\alpha}) = 3.624$ for male and $\exp(\hat{\alpha}+\hat{\beta}) = 1.732$ for female. This finding overlaps with several other association reports about gene-by-sex interaction of insulin-related traits and demonstrates the importance of considering interactions in the search for related genes [12] [13].

4. Discussions

Gene-covariates interactions are considered in allele/haplotype relative risk, and furthermore, in transmission probability in this transmission model. The missing parental genotypes and multiple tightly linked loci are allowed. For missing genotype or multi-locus genotype data, the underlying haplotypes or alleles are looked as missing data; the weighted conditional logistic models are given via EM algorithm.

As an application, 126 nuclear family data from Hong Kong Chinese population are used in haplotype-based model to detect the association between M235T in angiotensinogen gene and essential hypertension. The results suggest that the 235T is a risk allele with essential hypertension (ESH) for HongKong Chinese people, and contributes to higher risk in ESH men than in women. The 235T allele was more preferentially transmitted from heterozygous parents to ESH male patients than to female patients.

Acknowledgements

This work was supported by Guangdong Basic and Applied Basic Research Foundation (2020B1515310007), and Guangdong Province Key Laboratory of Computational Science, Sun Yat-sen University (2020B1212060032).

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Spielman, R.S., Mcginnis, R.E. and Ewens, W.J. (1993) Transmission Test for Linkage Disequilibrium: The Insulin Gene Region and Insulin-Dependent Diabetes Mellitus (IDDM). *The American Journal of Human Genetics*, **52**, 506-516.

[2] Spielman, R.S. and Ewens, W.J. (1996) The TDT and Other Family-Based Tests for Linkage Disequilibrium and Association. *American Journal of Human Genetics*, **59**, 983-989.

[3] Sham, P. and Curtis, D. (2012) An Extended Transmission/Disequilibrium Test (TDT) for Multi-Allele Marker Loci. *Annals of Human Genetics*, **59**, 323-336. https://doi.org/10.1111/j.1469-1809.1995.tb00751.x

[4] Spielman, R.S. and Ewens, W.J. (1998) A Sibship Test for Linkage in the Presence of Association: The Sib Transmission/Disequilibrium Test. *The American Journal of Human Genetics*, **62**, 450-458. https://doi.org/10.1086/301714

[5] Curtis, D. (1998) Use of Siblings as Controls in Case-Control Association Studies. *Annals of Human Genetics*, **61**, 319-333. https://doi.org/10.1017/S000348009700626X

[6] Martin, E.R., Monks, S.A., Warren, L.L. and Kaplan, N.L. (2000) A Test for Linkage and Association in General Pedigrees: The Pedigree Disequilibrium Test. *GeneScreen*, **1**, 65-67. https://doi.org/10.1086/302957

[7] Clayton, D. (1999) A Generalization of the Transmission/Disequilibrium Test for Uncertain-Haplotype Transmission. *The American Journal of Human Genetics*, **65**, 1170-1177. https://doi.org/10.1086/302577

[8] Zhao, H.Y., Zhang, S.L., *et al.* (2000) Transmission/Disequilibrium Tests Using Multiple Tightly Linked Markers. *The American Journal of Human Genetics*, **67**, 936-946. https://doi.org/10.1086/303073

[9] Taub, M.A., Schwender H., Beaty T.H., Louis T.A. and Ruczinski, I. (2012) Incorporating Genotype Uncertainties into the Genotypic TDT for Main Effects and Gene-Environment Interactions. *Genetic Epidemiology*, **36**, 225-234. https://doi.org/10.1002/gepi.21615

[10] Jeunemaitre, X., Inoue, I., Williams, C., Charru, A., Tichet, J., Powers, M., *et al.* (1997) Haplotypes of Angiotensinogen in Essential Hypertension. *American Journal of Human Genetics*, **60**, 1448–1460. https://doi.org/10.1086/515452

[11] Barley, J., Blackwood, A., Sagnella, G., Markandu, N. and Carter, N. (1994) Angiotensinogen met235-->thr Polymorphism in a London Normotensive and Hypertensive Black and White Population. *Journal of Human Hypertension*, **8**, 639-640.

[12] Freire, M., Ji, L., Onuma, T., Orban, T., Warram, J.H. and Krolewski, A.S. (1998) Gender-Specific Association of m235t Polymorphism in Angiotensinogen Gene and Diabetic Nephropathy in NIDDM. *Hypertension*, **31**, 896-899. https://doi.org/10.1161/01.HYP.31.4.896

[13] North, K.E., Franceschini, N., Borecki, I.B., Gu, C.C., Heiss, G., Province, M.A., *et al.* (2007) Genotype-by-Sex Interaction on Fasting Insulin Concentration: The Hypergen Study. *Diabetes*, **56**, 137-142. https://doi.org/10.2337/db06-0624