病例-对照研究中遗传模型的估计*

李 娜[1)] 李正帮[2)] 朱家砚[3)†]

(1)北京信息科技大学理学院, 100192, 北京; 2)华中师范大学数学与统计学院, 430079, 湖北武汉;
3)湖北中医药大学信息工程学院, 430065, 湖北武汉)

摘要 在病例-对照研究中, 针对选择遗传模型的问题, 提出了对风险等位基因采用共显性编码, 利用 Logistic 回归模型选择遗传模型. 数值模拟结果表明: 新方法比哈迪-温伯格平衡检验选择模型效果更好. 将新方法应用于 6 个与乳腺癌有关、8 个与 II-型糖尿病有关的 SNPs 中, 进一步验证了新方法的可行性和有效性.

关键词 病例-对照研究; 遗传模型; 共显性编码

中图分类号 O212.1 **DOI**: 10.12202/j.0476-0301.2021215

0 引言

病例-对照设计是流行病学研究中探索病因的有效工具之一. 这虽是一个回顾性的设计, 但极大似然估计是弱相合的且具有渐近正态性, 因此将数据纳入 Logistic 回归模型来估计协变量的系数仍然是有效的[1]. 另外, 在体质量指数、年龄、血压等诸多协变量中, 遗传变异是其中最重要的一个. 单核苷酸多态性(single nucleotide polymorphisms, SNPs)是一种发生在基因组特定位置的遗传变异, 与插入、缺失和拷贝数变异等相比, 是一种更为常见的形式. 截至目前, 已经发现数百种人类疾病与 1 万多个 SNPs 有关.

全基因组关联分析[2−3]是应用基因组中数百万的 SNPs 作为遗传标记, 进行全基因组水平上的对照或相关性分析, 从而发现影响复杂性状的基因变异的一种新策略. 进行关联分析之前, 需要预先假定一个遗传模型, 指定一个遗传模型意味着指定一个替代假设. 实际上, 真正的遗传模型是未知的. 在许多遗传研究中使用的都是加性模型[4−6]; 也有一些 SNPs 是在其他模型下带来疾病风险. 例如: Moltke 等[7]在隐性模型下发现了一种与 II-型糖尿病相关的遗传变异; Nik-Zainal 等[8]报告了 5 个以隐性模型作用于乳腺癌的基因, 采用加性遗传模型会漏掉其他模型下与复杂疾病关联的遗传位点, 即错误地指定遗传模型将导致统计功效的损失[9−10]; Sladek 等[11]建议使用 MAX 统计量寻找与 II-型糖尿病相关的 SNPs; Li 等[12]、Zheng 等[6]建立了 MAX 统计量的理论框架, 并推导了 MAX 统计量的渐近分布. 已有文献只有基于哈迪-温伯格平衡检验(Hardy-Weinberg equilibrium test, HWET)选择遗传模型[13−15]的方法. 当遗传模型是加性模型时, 用 HWET 得到的遗传模型并不理想. 已有文献有关关联检验大多都要考虑遗传模型, 并且通常只考虑隐性模型、加性模型和显性模型 3 种常用的遗传模型, 其实还包含其他遗传模型, 例如共显性模型等, 而这些方法不能涵盖其他遗传模型. 如果能先找出遗传模型, 再构造检验统计量, 就能显著提高统计检验的功效.

在此提出一个通用的框架来估计遗传模型, 该框架有 3 个优点: 1) 用参数$\theta\in[0,1]$的不同取值表示遗传模型, 该模型不仅包括隐性模型、加性模型和显性模型, 还包含其他模型; 2) 提出的方法可以处理混杂因素, 而现有的方法无法处理; 3) 用一个二元变量代替原来的基因型得分, 它可以分离出基因型系数, 使参数估计变得可行.

1 符号说明与方法

1.1 变量符号 令 Y表示疾病状态, 则 $Y=1$表示患病状态, $Y=0$ 表示健康状态. 设$\boldsymbol{X}$和G分别为m维的协变量和基因型得分, 通过调整$\boldsymbol{X}$ 来检验Y和 G 之间的关系, 一个典型可用的模型是逻辑回归模型:

$$P(Y=1|X,G)=\frac{\exp(\alpha+\boldsymbol{X}^{\mathrm{T}}\gamma+G\beta)}{1+\exp(\alpha+\boldsymbol{X}^{\mathrm{T}}\gamma+G\beta)},$$

式中: α、γ和β是参数; T 表示矩阵或向量的转置. 考

* 北京市自然科学基金重点研究专题资助项目(Z180006); 华中师范大学“中央高校基本科研业务费”资助项目(CCNU20TS002)
† 通信作者: 朱家砚(1984—), 女, 博士, 讲师. 研究方向: 统计遗传、医学统计等. E-mail: zhujiayan999@163.com
收稿日期: 2021-08-30

虑一个双等位基因的 SNPs 位点, SNPs 基因座的 2 个等位基因分别为A和a, A一般称为次等位基因, 此时有 3 种基因型, 即aa、Aa和AA, 相应的基因型得分值分别为: 0、θ和 1, $0\leqslant\theta\leqslant1$, θ的不同取值表示不同的遗传模型. 例如, 对于常用的隐性、加性和显性模型, θ的取值则分别为 0、0.5 和 1.0. 假设从病例人群和对照人群中随机抽取r个病例和s个对照, $r+s=n$. 设$(y_i,x_i^{\mathrm{T}},g_i)^{\mathrm{T}}$为第$i$个个体对于$(Y,X^{\mathrm{T}},G)^{\mathrm{T}}$的观测值, $i=1,2,\cdots,n$. 假设前r个个体为病例, 后 s个个体为对照.

1.2　利用 HWET 选择模型　令(p_0,p_1,p_2)、(q_0,q_1,q_2)分别表示病例组和对照组(aa, Aa, AA)基因型频率, 设r个病例中基因型为(aa, Aa, AA)的个体数量为(r_0,r_1,r_2), s个对照中基因型为(aa, Aa, AA)的个体数量为(s_0,s_1,s_2). 令$(\hat{p}_0,\hat{p}_1,\hat{p}_2)=(r_0/r,r_1/r,r_2/r)$, $\hat{D}_1=\hat{p}_2-(\hat{p}_2+\hat{p}_1/2)^2$, $(\hat{q}_0,\hat{q}_1,\hat{q}_2)=(s_0/s,s_1/s,s_2/s)$, $\hat{D}_2=\hat{q}_2-(\hat{q}_2+\hat{q}_1/2)^2$. 在整个样本中与仅在对照组下得出的 HWET 分别表示为:

$$C_{\mathrm{HC}}=\frac{\sqrt{rs/n}(\hat{D}_1-\hat{D}_2)}{(1-n_2/n-n_1/(2n))(n_2/n+n_1/(2n))},$$

$$C_{\mathrm{H}}=\frac{\sqrt{r}\hat{D}_1}{(1-n_2/n-n_1/(2n))(n_2/n+n_1/(2n))},$$

式中: $n_1=r_1+s_1$、$n_2=r_2+s_2$. 利用 HWET[16−17] 选择遗传模型, 可总结为: 设置一个正阈值c, 如$c=1.654$; 遗传模型确定过程为: 若$Z>c$, 此时选择隐性模型; 若$Z<-c$, 确定为显性模型; 否则, 选择加性模型. Z可以是统计量C_{HC}或 C_{H} 的取值. C_{HC}和 C_{H}对 θ的估计值分别用θ_{CC}和θ_{C}表示, θ_{CC}和θ_{C}可能不是 θ 的一致估计.

1.3　分解基因型得分法　将基因型得分数据分解为

$$(G_1,G_2)=\begin{cases}(0,0), G=0,\\(1,0), G=\theta,\\(0,1), G=1.\end{cases}$$

此时 Logistic 回归模型为

$$P(Y=1|X,G)=\frac{\exp(\alpha+\boldsymbol{X}^{\mathrm{T}}\gamma+G_1\beta_1+G_2\beta_2)}{1+\exp(\alpha+\boldsymbol{X}^{\mathrm{T}}\gamma+G_1\beta_1+G_2\beta_2)}.$$

在罕见疾病假设下, 即$\alpha\ll0$时, 可得到如下 3 个关系式:

$$1+\exp(\alpha+\boldsymbol{X}^{\mathrm{T}}\gamma)\approx1,$$

$$1+\exp(\alpha+\boldsymbol{X}^{\mathrm{T}}\gamma+\theta\beta_1)\approx1,$$

$$1+\exp(\alpha+\boldsymbol{X}^{\mathrm{T}}\gamma+\beta_2)\approx1.$$

通过运算可得

$$P(Y=1|\boldsymbol{X},G=0)=\frac{\exp(\alpha+\boldsymbol{X}^{\mathrm{T}}\gamma)}{1+\exp(\alpha+\boldsymbol{X}^{\mathrm{T}}\gamma)}\approx\exp(\alpha+\boldsymbol{X}^{\mathrm{T}}\gamma),$$

$$P(Y=1|\boldsymbol{X},G=\theta)=\frac{\exp(\alpha+\boldsymbol{X}^{\mathrm{T}}\gamma+\theta\beta_1)}{1+\exp(\alpha+\boldsymbol{X}^{\mathrm{T}}\gamma+\theta\beta_1)}\approx\exp(\alpha+\boldsymbol{X}^{\mathrm{T}}\gamma+\theta\beta_1),$$

$$P(Y=1|\boldsymbol{X},G=1)=\frac{\exp(\alpha+\boldsymbol{X}^{\mathrm{T}}\gamma+\beta_2)}{1+\exp(\alpha+\boldsymbol{X}^{\mathrm{T}}\gamma+\beta_2)}\approx\exp(\alpha+\boldsymbol{X}^{\mathrm{T}}\gamma+\beta_2).$$

如果真实的遗传模型是加性的, 则满足:

$$\frac{P(Y=1|\boldsymbol{X},G=1)}{P(Y=1|\boldsymbol{X},G=0)}\approx\left(\frac{P(Y=1|\boldsymbol{X},G=\theta)}{P(Y=1|\boldsymbol{X},G=0)}\right)^2,$$

即

$$\frac{\exp(\alpha+\boldsymbol{X}^{\mathrm{T}}\gamma+\beta_2)}{\exp(\alpha+\boldsymbol{X}^{\mathrm{T}}\gamma)}\approx\left(\frac{\exp(\alpha+\boldsymbol{X}^{\mathrm{T}}\gamma+\theta\beta_1)}{\exp(\alpha+\boldsymbol{X}^{\mathrm{T}}\gamma)}\right)^2.$$

再通过运算可得到 $\theta=\beta_2/\beta_1\approx0.5$.

如果真实遗传模型是显性模型, 则满足:

$$\frac{P(Y=1|\boldsymbol{X},G=1)}{P(Y=1|\boldsymbol{X},G=0)}\approx\frac{P(Y=1|\boldsymbol{X},G=\theta)}{P(Y=1|\boldsymbol{X},G=0)},$$

即

$$\frac{\exp(\alpha+\boldsymbol{X}^{\mathrm{T}}\gamma+\beta_2)}{\exp(\alpha+\boldsymbol{X}^{\mathrm{T}}\gamma)}\approx\frac{\exp(\alpha+\boldsymbol{X}^{\mathrm{T}}\gamma+\theta\beta_1)}{\exp(\alpha+\boldsymbol{X}^{\mathrm{T}}\gamma)}.$$

通过运算可得到 $\theta=\beta_2/\beta_1\approx1$.

通过推导可知: 在常见的隐性、显性和加性遗传模型下, β_2/β_1表示病例-对照研究中的遗传模型, 因此在病例-对照研究中, β_2/β_1在一定程度上能度量基因遗传模型.

利用$\boldsymbol{X}^{\mathrm{T}}$与$G$的观测值$x_i^{\mathrm{T}}$与$(g_{i1},g_{i2})$, $i=1,2,\cdots,n$, 可得似然函数

$$L(\alpha,\beta_1,\beta_2)=\prod_{i=1}^{n}\left(\frac{\exp(\alpha+x_i^{\mathrm{T}}\gamma+g_{i1}\beta_1+g_{i2}\beta_2)}{1+\exp(\alpha+x_i^{\mathrm{T}}\gamma+g_{i1}\beta_1+g_{i2}\beta_2)}\right)^{y_i}\cdot\left(\frac{1}{1+\exp(\alpha+x_i^{\mathrm{T}}\gamma+g_{i1}\beta_1+g_{i2}\beta_2)}\right)^{1-y_i}.$$

对数似然函数为

$$l(\alpha,\beta_1,\beta_2)=\sum_{i=1}^{n}((y_i(\alpha+x_i^{\mathrm{T}}\gamma+g_{i1}\beta_1+g_{i2}\beta_2))-\ln(1+\exp(\alpha+x_i^{\mathrm{T}}\gamma+g_{i1}\beta_1+g_{i2}\beta_2))).$$

通过求解约束优化问题, 可得 β_1、β_2 的估计 $\hat{\beta}_1$、$\hat{\beta}_2$分别为$(\hat{\beta}_1,\hat{\beta}_2)=\underset{\beta_1,\beta_2\geqslant0,\beta_1\neq0}{\arg\max}\, l(\alpha,\beta_1,\beta_2)$; 用 $\hat{\theta}$ 表示θ 的估计, 则 $\hat{\theta}=\hat{\beta}_2/\hat{\beta}_1$, 式中$(\hat{\beta}_1,\hat{\beta}_2)^{\mathrm{T}}$是一般 Logistic 回归假定下$(\beta_1,\beta_2)^{\mathrm{T}}$的相合估计. 还可以根据对数似然函数得

到θ的估计,即为约束最优化问题

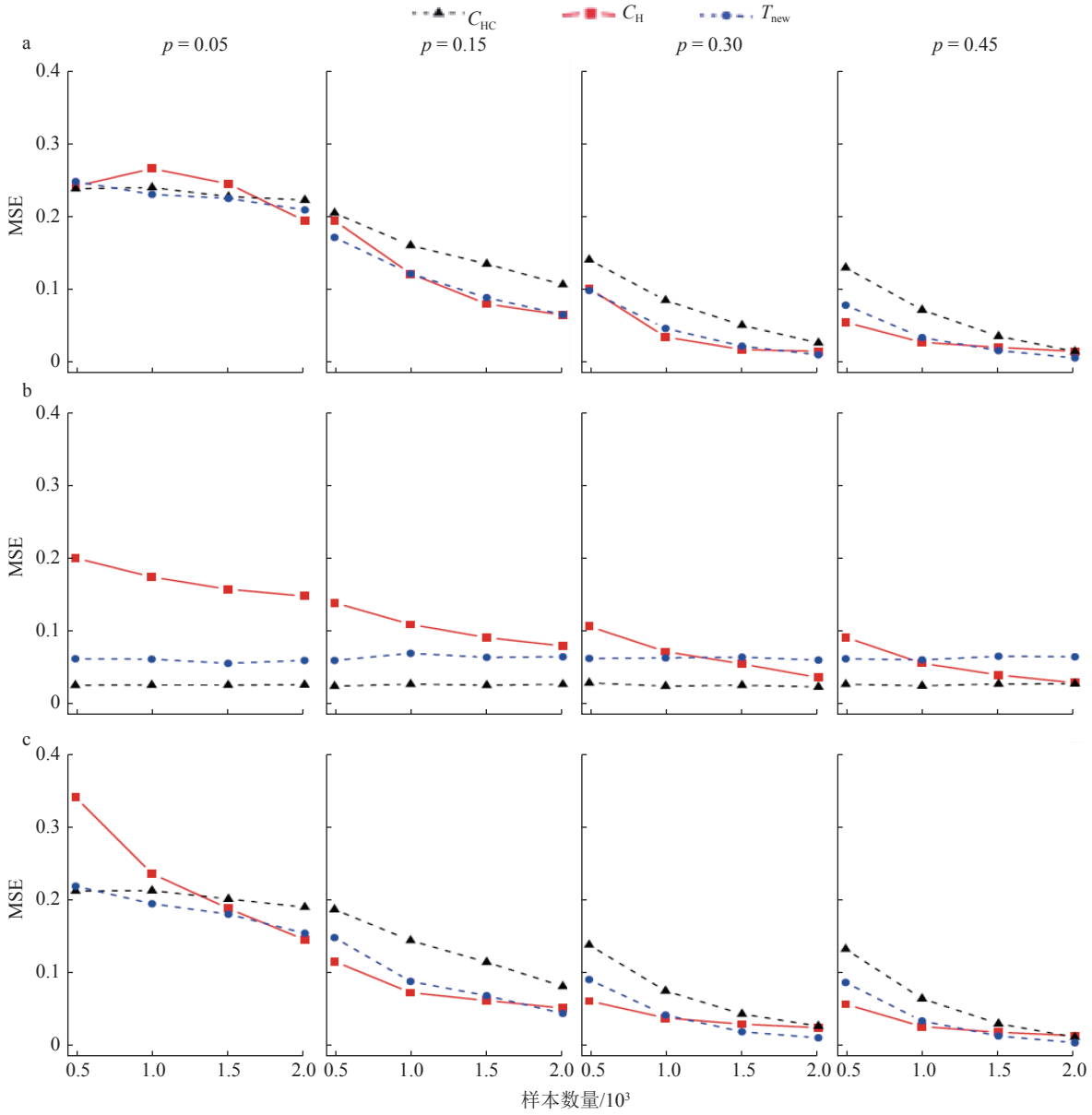
$$(\hat{\beta}_1,\hat{\theta})=\underset{\theta\geqslant 0,\beta_1\neq 0}{\arg\max}\, l(\alpha,\beta_1,\beta_2).$$

2 模拟结果

为说明本文所提方法(用T_{new}表示)的良好性能,本研究做了大量模拟,将T_{new}估计遗传模型的均方误差(mean square error, MSE)与统计量C_{HC}[13]、C_H[14]所选遗传模型的MSE进行比较.假设在一般群体中哈迪-温伯格平衡成立,即基因型频率满足$P(aa)=(1-p)^2$, $P(Aa)=2p(1-p)$, $P(AA)=p^2$, P = 0.05、0.15、0.30、0.45, $p=P(A)$.

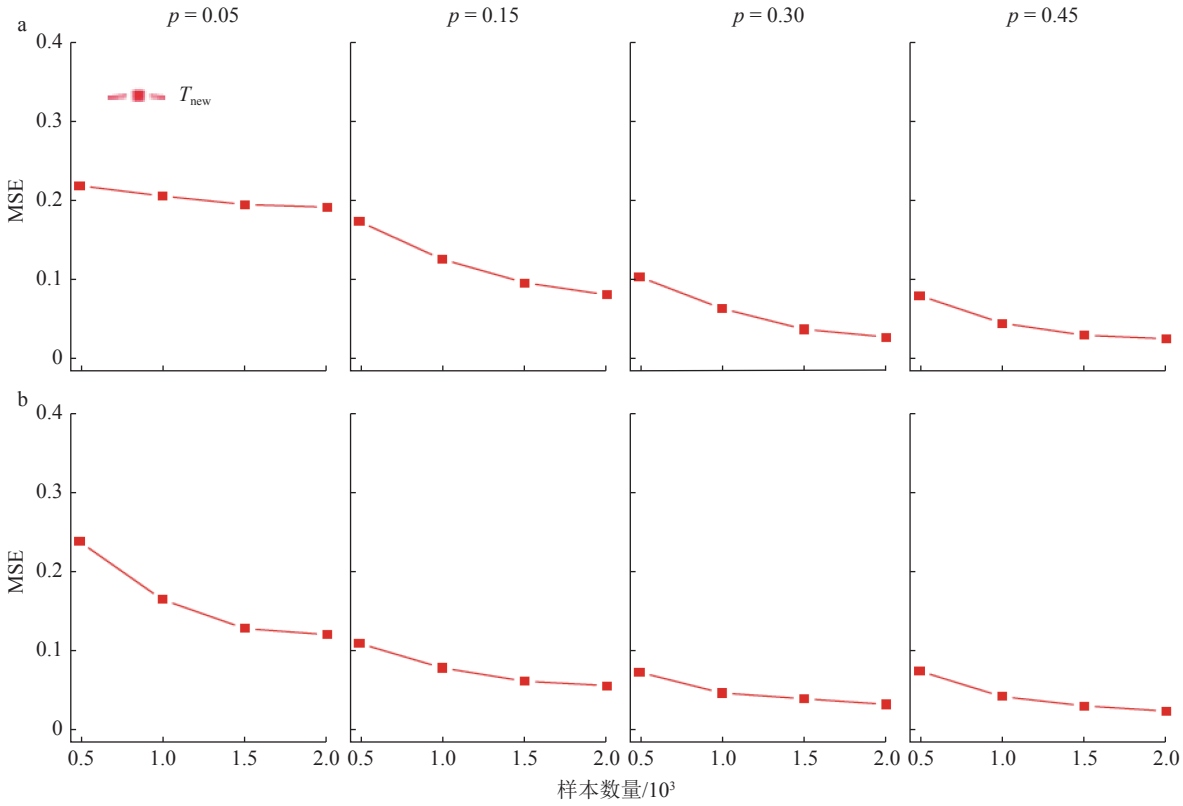
考虑$k=0.02$、0.05这2种疾病流行率.假设$\boldsymbol{X}$不受G的影响,且服从标准正态分布.令:$\gamma=\ln 1.1$、$\beta=\ln 1.5$.样本量$r=s=0.5\times10^3$、1.0×10^3、1.5×10^3、2.0×10^3, $\theta=0$、0.25、0.50、0.75、1.模拟重复1 000次.

图1与2分别给出了$k=0.05$时,利用C_{HC}、C_H与T_{new}估计遗传模型的经验MSE.模拟结果表明T_{new}的经验MSE随样本量的增大而减小.例如:当$r=s=0.5\times10^3$, $p=0.15$,遗传模型为隐性模型时,T_{new}的经验MSE为0.18;当$r=s=1.0\times10^3$时,T_{new}的经验MSE为0.11.此外,当遗传模型为加性模型时,C_{HC}、C_H的经验MSE在不同样本量下几乎保持不变.例如:当$r=s=0.5\times10^3$, $p=0.30$时,C_{HC}、C_H的经验MSE分别为0.025、0.066;当$r=s=1.5\times10^3$时,C_{HC}、C_H的经验MSE分别为0.028、0.064.当遗传模型为加性模型,样



$p=0.05$、0.15、0.30、0.45;a.$\theta=0$时的4个MSE;b.$\theta=0.5$时的4个MSE;c.$\theta=1.0$时的4个MSE.

图1 $k=0.05$时C_{HC}、C_H、T_{new}的经验MSE



$p=0.05$、0.15、0.30、0.45；a．$\theta=0.25$时的 4 个 MSE，b．$\theta=0.75$时的 4 个 MSE.

图 2　$k=0.05$时C_{HC}、C_H、T_{new}的经验 MSE

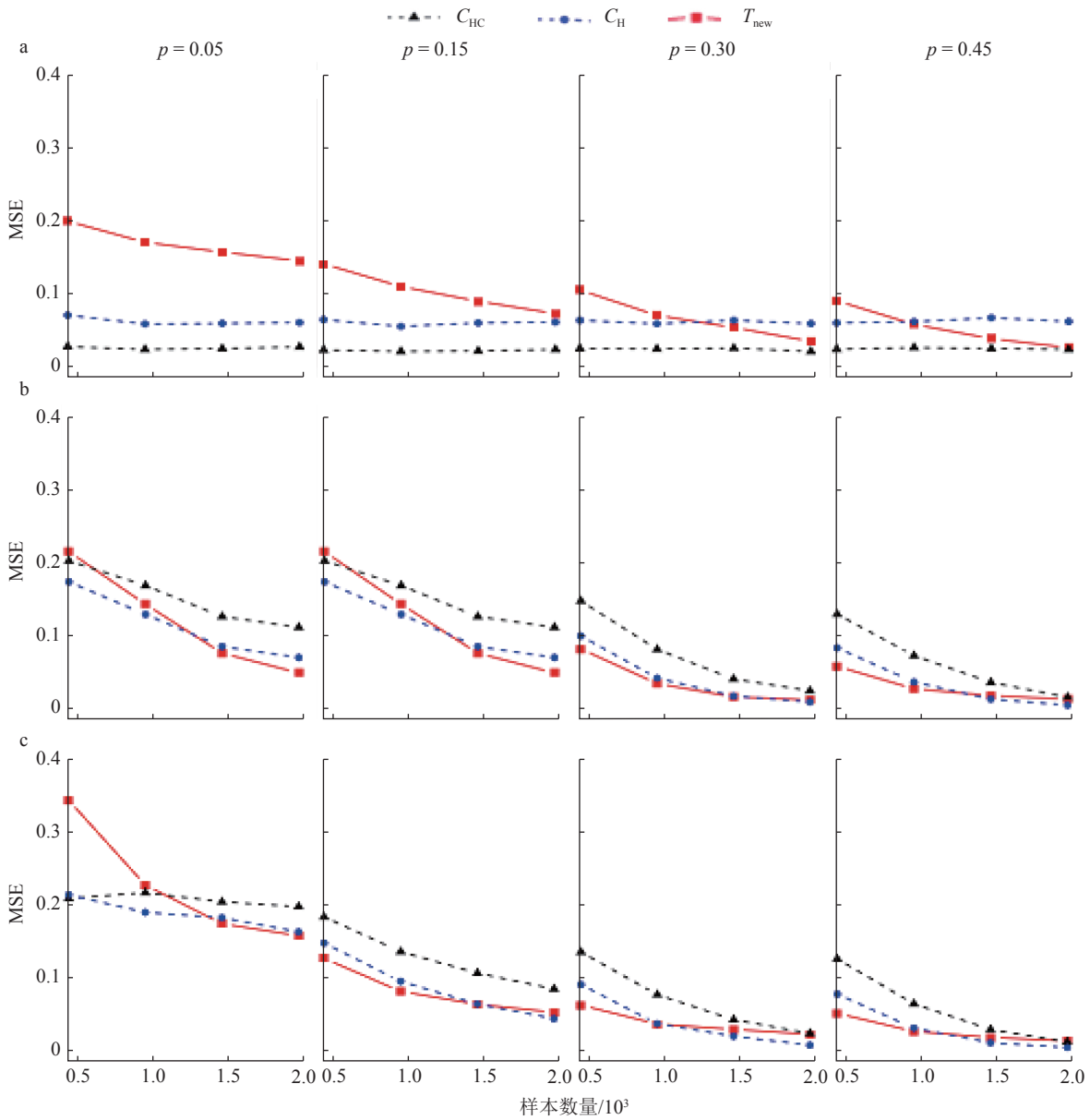
本量比较大时，T_{new} 的 MSE 小于C_{HC}和C_H的 MSE. 例如：当样本容量$r=s=0.8\times10^3$、$p=0.45$时，T_{new}、C_{HC}、C_H的经验 MSE 分别为 0.032、0.094、0.048.

图 3 与 4 分别给出了$k=0.02$时，利用 C_{HC}、C_H、T_{new} 估计遗传模型的经验 MSE. 通过模拟，得到与$k=0.05$类似的结论. 例如：当遗传模型为加性模型，$p=0.3$, $r=s=1.0\times10^3$、1.5×10^3时，T_{new}的经验 MSE 值分别为 0.081、0.04. 可以看出，T_{new}的经验 MSE 数值随着样本量的增加而减小. 在加性模型下C_{HC}、C_H的经验 MSE 值在不同样本量下几乎保持不变. 例如：当$r=s=0.5\times10^3$、$p=0.15$时，C_{HC} 、C_H的经验 MSE 值分别为 0.023、0.06; 当$r=s=2.0\times10^3$时，C_{HC}、C_H的经验 MSE 值分别为 0.022、0.059. 同样，在隐性模型下，当样本量较大时，T_{new}的 MSE 值小于C_{HC}、C_H的 MSE 值. 例如当样本量$r=s=1.0\times10^3$、$p=0.45$时，T_{new}、C_{HC}、C_H的经验 MSE 值分别为 0.027、0.065、0.035.

3　实例分析

乳腺癌是女性中常见的癌症. 据 2018 年国际癌症研究机构(IARC)调查的最新数据显示，乳腺癌在全球患癌症女性中的发病率为 24.2%, 位居患癌症女性的首位，其中 52.9% 发生在发展中国家. 乳腺癌的病因尚不清楚，截至目前科学家还未找到乳腺癌的确切致癌原因，但已经发现诸多与乳腺癌发病有关的高危因素. 随着乳腺癌高危因素不断积聚，其患病风险就会增大. 几乎 15% 的女性乳腺癌患者的家庭成员都被诊断出患有乳腺癌，这意味着遗传基因变异可能会带来患乳腺癌的风险. Hunter 等[5] 进行了全基因组关联研究，鉴定出 rs10510126、rs12505080、rs17157903、rs1219648、rs7696175、rs2420946 这 6 个与乳腺癌有关的 SNPs, 表 1 给出了基因型数值(参阅文献 [12] 中表 2 的数据)以及遗传模型估计. II-型糖尿病原名为成人发病型糖尿病，多在 35~40 岁之后发病，占糖尿病患者 90% 以上. II-型糖尿病是一种终生疾病，常有家族史; 可发生于任何年龄，成人多见; 多数起病隐匿，症状相对较轻，仅有轻度乏力、口渴，半数以上无任何症状; 有些病人因慢性并发症、伴发病或体检时发现. 通常，遗传因素会增加患 II-型糖尿病的风险. Sladek 等[11] 进行了全基因组关联研究，确定了 8 个与 II-型糖尿病相关的 SNPs, 汇总数据见表 2.

用C_{HC}、C_H、T_{new}来判断这 14 个 SNPs 的基因模型，结果如表 1 与 2 所示. 对于乳腺癌而言，如果使用C_{HC}或 C_H, 一半的 SNPs 为显性模型，其余为加性模型. 但是利用T_{new}可以给出遗传模型的具体数值. 例



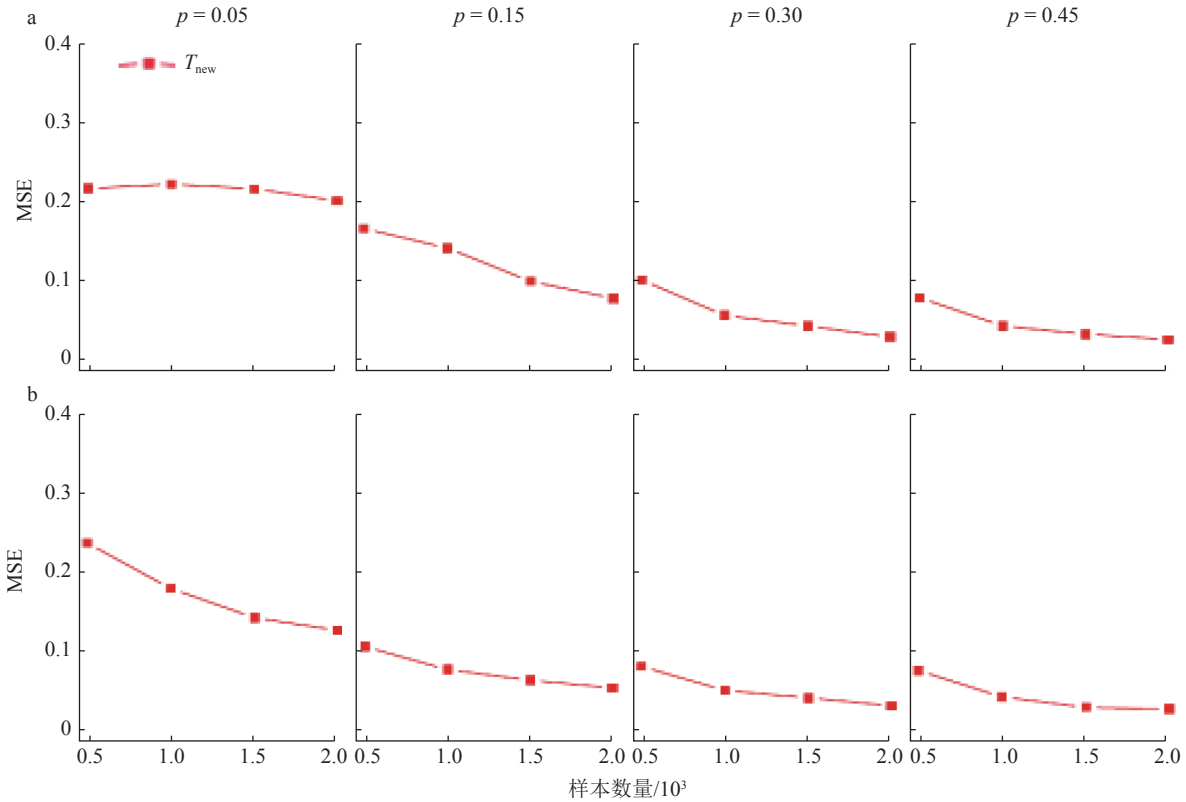
$p = 0.05$、0.15、0.30、0.45；a. $\theta = 0$时的 4 个 MSE；b. $\theta = 0.5$时的 4 个 MSE；c. $\theta = 1$时的 4 个 MSE.

图 3 $k = 0.02$时C_{HC}、C_H、T_{new}的经验 **MSE**

如：对于 SNPs rs10510126, T_{new}估计的遗传模型$\theta = 2.36\times10^{-11}$, 为隐性模型；但如果使用$C_H$, 给出的是加性模型. 除此之外, T_{new}还可以给出其他类型的遗传模型. 例如对于 SNPs rs2420946, 使用T_{new}得到的遗传模型数值为 0.38.

4 结论

病例-对照遗传关联研究已被证明是通过扫描人类基因组来识别有害变异的有效工具. 遗传变异有插入、缺失、拷贝数变异、单核苷酸多态性等, 其中以 SNPs 最为常见. 人类基因组中有 29.6 亿个碱基对, 单核苷酸多态性约3.0×10^7个. 截至目前, 已有超过 1 万个 SNPs 被发现与数百种疾病或性状有关. 为了评估 SNPs 的重要性, 必须指定一个遗传模型. 在隐性、加性和显性 3 种遗传模型中, 关联性研究常假设遗传模型为加性模型. 然而, 在实践中, 真正的遗传模型是未知的. 因果性 SNPs 和替代者之间的遗传模型可能不同[18]. 错误地指定遗传模型可能会导致统计功效的损失, 新方法使用θ来表示遗传模型, 且$\theta \in [0,1]$. 已有方法仅对$\theta = 0$、0.5、1.0进行推断, 不能估计θ的其他值. 新方法通过对基因型得分进行分解, 提出了一种新的估计θ的方法, 并得出比已有方法更理想的结果. 遗传模型的选择与关联检验之间存在一定的相关性, 以后的研究希望在所选模型的基础上构造关联检验.



$p = 0.05$、0.15、0.30、0.45；a．$\theta = 0.25$时的 4 个 MSE；b．$\theta = 0.75$时的 4 个 MSE.

图 4　$k = 0.02$时C_{HC}、C_H、T_{new}的经验 MSE

表 1　6 个与乳腺癌相关的 SNPs 的遗传模型估计

rs	r_0	r_1	r_2	s_0	s_1	s_2	$\hat{\theta}$	$\hat{\theta}_C$	$\hat{\theta}_{CC}$
10510126	955	180	10	854	272	14	2.36×10^{-11}	0.5	0.5
12505080	608	477	50	628	408	99	0.99	1.0	1.0
17157903	777	316	18	862	220	26	0.99	1.0	1.0
1219648	352	543	250	433	538	170	0.36	0.5	0.5
7696175	353	605	187	396	496	249	0.99	1.0	1.0
2420946	357	546	242	440	537	165	0.38	0.5	0.5

表 2　8 个与 II-型糖尿病相关的 SNPs 的遗传模型估计

rs	r_0	r_1	r_2	s_0	s_1	s_2	$\hat{\theta}$	$\hat{\theta}_C$	$\hat{\theta}_{CC}$
7903146	197	348	149	335	254	65	0.62	0.5	0.5
13266634	54	229	411	53	293	307	1.62×10^{-8}	0	0
1111875	77	302	315	119	308	227	0.54	0.5	0.5
7923837	66	300	328	116	296	242	0.66	0.5	0.5
7480010	301	327	66	363	246	353	0.82	1.0	1.0
3740878	25	273	386	65	249	353	1.00	1.0	1.0
11037909	25	274	387	65	251	353	0.99	1.0	1.0
1113132	25	271	390	63	251	355	0.98	1.0	1.0

5　参考文献

[1] PRENTICE R L, PYKE R. Logistic disease incidence models and case-control studies[J]. Biometrika, 1979, 66(3): 403

[2] HAYES B. Overview of statistical methods for genome-wide association studies(GWAS)[J]. Methods in Molecular

Biology, 2013, 1019: 149
[3] WANG M H, CORDELL H J, STEEN K V. Statistical methods for genome-wide association studies[J]. Seminars in Cancer Biology, 2019, 55: 53
[4] KLEIN R J, ZEISS C, CHEW E Y, et al. Complement factor *H* polymorphism in aged-related macular degeneration[J]. Science, 2005, 308(5720): 385
[5] HUNTER D J, KRAFT P, JACOBS K B, et al. A genome-wide association study identifies alleles in FGFR2 associated with risk of spo radic postmenopausal breast cancer[J]. Nature Genetics, 2007, 39(7): 870
[6] ZHENG G, LI Q Z, YUAN A. Some statistical properties of efficiency robust tests for genetic studies[J]. Scandinavian Journal of Statistics, 2014, 41(3): 762
[7] MOLTKE I, GRARUP N, TREEBAK J T, et al. A common Greenlandic TBC1D4 variant confers muscle insulin resistance and type 2 diabetes[J]. Nature, 2014, 512(7513): 190
[8] NIK-ZAINAL S, DAVIES H, STAAF J, et al. Landscape of somatic mutations in 560 breast cancer whole genome sequences[J]. Nature, 2016, 534(7605): 47
[9] GAYE A, DAVIS S K. Genetic model misspecification in genetic association studies[J]. BMC Research Notes, 2017, 10(1): 569
[10] GLOAGUEN E, DIZIER M, BOISSEL M, et al. General regression model: a "model- free" association test for quantitative traits allowing to test for the underlying genetic model[J]. Annals of Human Genetics, 2019, 84(3): 280
[11] SLADEK R, ROCHELEAU G, RUNG J, et al. A genome-wide association study identifies novel risk loci for type2 diabetes[J]. Nature, 2007, 445(7130): 881
[12] LI Q Z, ZHENG G, LI Z H, et al. Efficient approximation of *p*-value of the maximum of correlated tests with applications to genome-wide association studies[J]. Annals of Human Genetics, 2008, 72(3): 397
[13] ZHENG G, NG H K T. Genetic model selection in two-phase analysis for case-control association studies[J]. Biostatistics, 2008, 9(3): 391
[14] ZHENG G, ZHANG W, XU J F, et al. Genetic risks and genetic model specification[J]. Journal of Theoretical Biology, 2016, 403: 68
[15] HU X N, DUAN X G, PAN D D, et al. A model-embedded trend test with incorporating Hardy-Weinberg equilibrium information[J]. Journal of Systems Science and Complexity, 2017, 30(1): 101
[16] CHEN J B, CHATERJEE N. Exploiting Hardy-Weinberg equilibrium for efficient screening of single SNP associations from case-control studies[J]. Human Heredity, 2007, 63(3/4): 196
[17] SLAGER S L, SCHAID D J. Case-control studies of genetic markers: power and sample size approximations for Armitage's test for trend[J]. Human Heredity, 2001, 52(3): 149
[18] HORMOZDIAR F, KICHAEV G, YANG W Y, et al. Identification of causal genes for complex traits[J]. Bioinformatics, 2015, 31(12): i206

Estimation of genetic inheritance in case-control studies

LI Na[1] LI Zhengbang[2] ZHU Jiayan[3]

(1)School of Applied Science, Beijing Information Science and Technology University, 100192, Beijing, China; 2)School of Mathematics and Statistics, Central China Normal University, 430079, Wuhan, Hubei, China; 3)School of Information Engineering, Hubei University of Chinese Medicine, 430065, Wuhan, Hubei, China)

Abstract For selective genetic models, risk alleles were co-dominant coded. Logistic regression was used to select genetic model in case-control study. Numerical simulations show that our new method is more effective than Hardy-Weinberg equilibrium test to choose the genetic model. Applications to six single nucleotide polymorphisms (SNPs) for breast cancer and eight SNPs for Type 2 Diabetes further show the feasibility and effectiveness of our proposed method.

Keywords case-control study; genetic model; co-dominant code

【责任编辑: 陆有忠】