中国农业大学学报 2005,10(6):33-36
Journal of China Agricultural University

利用 GLMM 方法估计家畜阈性状的遗传力

殷宗俊[1,2] 张 勤[2]
(1. 安徽农业大学 动物科技学院,合肥 230036; 2. 中国农业大学 动物科技学院,北京 100094)

摘 要 将广义线性混合模型(GLMM)技术引入家畜阈性状的遗传分析,对不同参数组合下阈性状的遗传力及准确度进行了估计,模拟研究的性状为单阈值和 3 阈值离散性状,试验设计为全同胞-半同胞混合家系。结果表明,GLMM 方法能有效地估计阈性状的遗传力,其准确性有较大的优势。不同参数组合下,单阈值与 3 阈值分类性状的遗传力估计均方误(MSE)均在 0.018 4 以下;遗传力参数的估计偏差(BIAS)均小于 0.429 0。同时,性状的遗传力和性状的表型发生率对遗传力估计效果都有直接影响,随着性状遗传力真值的提高,GLMM 方法出现高估现象;随着性状发生率的提高,GLMM 方法估计的遗传力结果有上升的趋势。说明利用 GLMM 方法估计家畜阈性状的遗传力是可行的。

关键词 广义线性混合模型(GLMM);阈性状;遗传力;遗传分析
中图分类号 Q 348 **文章编号** 1007-4333(2005)06-0033-04 **文献标识码** A

Estimation of heritability for threshold traits in livestock using GLMM method

Yin Zongjun[1,2], Zhang Qin[2]
(1. College of Animal Science and Technology, Anhui Agricultural University, Hefei 230036, China;
2. College of Animal Science and Technology, China Agricultural University, Beijing 100094, China)

Abstract The method of generalized linear mixed models (GLMM) was described. It allowed genetic analyses and heritability efficiency estimate for threshold traits in livestock under different parameters. The results showed that GLMM is advantageous in heritability estimation and accurate for discrete traits. The estimated maximal mean standard error (MSE) and heritability bias (BIAS) are respectively 0.018 4 and 0.429 0 for all different parameters. The increase of the true heritability and incidence of categorie induced the improvement of heritability accuracy, showing that the accuracy of estimation directly depended on these two parameters.

Key words generalized linear mixed model (GLMM); threshold traits; heritability; genetical analysis

目前,对阈性状的遗传分析有 2 种完全不同的思想,主要区别在于是否具有潜在连续分布的假设——一种是不考虑性状的实际离散性,在假设表型为连续分布的前提下,采用线性的方法对阈性状进行遗传分析,称之为线性分析思想。但对于阈性状而言,由于其观察值不能表达为遗传效应与环境效应的线性组合,从而限制了线性方法的有效应用;同时,在离散的表型分布下,常规线性模型的多个理论假设均无法成立。因此,线性思想下的阈性状遗传分析得不到理想效果。另一种思路就是阈值概念[1],认为在性状表型离散分布的背后存在着一个潜在的连续分布,而离散的表型值与潜在的连续值是通过一系列固定的阈相联系。阈性状非线性遗传分析方法的导出,主要是基于广义线性混合模型(generalized linear mixed model, GLMM)的发展应用。GLMM 是常规正态线性模型的直接推广[2],它适用于连续性状和离散性状的遗传分析。广义线性模型与典型线性模型的区别是其随机误差的分布没

收稿日期:2005-09-01
基金项目:国家重点基础研究发展规划资助项目(G2000016103);安徽省自然科学基金资助项目(050410204);安徽省教育厅资助项目(2002jq126,2004kj151)
作者简介:殷宗俊,博士,副教授,主要从事动物遗传育种方面的研究,E-mail:yinzongjun@yahoo.com.cn;张勤,教授,博士生导师,主要从事动物遗传育种方面的研究,E-mail:qzhang@cau.edu.cn

有正态性要求,可通过连接函数(link function)将观察值与遗传效应相联系,从而使转化后的响应变量期望值线性化,以实现参数的无偏估计;另外,广义线性模型的参数估计量具有大样本正态分布,因而具有良好的统计性质。

现代动物育种中,人们对离散性状的重视程度不断增大,一方面是因为许多重要的经济性状如记数繁殖性状、与疾病和生存有关的性状都属于阈性状;再者是由于许多选择试验[3-4]都证明了对阈性状的选择是有效的。特别是随着家畜育种技术的进一步发展,一些当前育种目标中的主选性状不断趋于选择极限,进一步选择的进展甚微,从而一些重要的次级性状受到育种学家们的极大重视,特别是与经济效益密切相关的阈值性状,如能将其纳入到家畜的育种目标中将会获得更大的育种效益。在此前提下,家畜阈性状的遗传分析方法显得日益迫切,本研究将 GLMM 方法引入家畜阈性状的遗传分析,并模拟分析了不同参数组合下 GLMM 方法的相对效率及精度,旨在探索合理的阈性状遗传分析方法。

1 研究方法

1.1 研究设计

本研究模拟单阈值和多阈值 2 种离散表型的性状。基础群由 20 头公畜和 100 头母畜构成(无亲缘关系),其中公母畜随机交配,每头公畜与 5 头不同的母畜交配,每头母畜产 10 个后代,后代的性别比例为 1∶1,这样每头公畜的后代数为 50 头。一共模拟了 13 个世代的群体发展过程,在每个世代中,随机选择 20 头公畜和 100 头母畜作为下一世代的亲本,世代间不重叠。只利用最后 3 个世代的资料进行遗传分析。对上面设计的资源群体重复模拟产生 100 次,利用 GLMM 方法算出离散性状的方差组分及遗传力,之后进行 100 次模拟结果的平均。

1.2 性状的遗传模型及数据模拟

按照 Falconer[1]定义的阈值模型,对于阈性状,具有一个潜在的不可观察的连续分布,这个分布受多基因和环境的共同影响,潜在变量可用线性模型 $y = mu + u + e$ 来描述,其中 y 为潜在变量的取值,mu 为群体的均值,u 为加性遗传效应,e 为随机环境效应。而表观的不同类别是由若干个潜在的阈值决定的,高于或低于特定的阈值时会表现出不同的表型状态。模拟研究中,单阈值性状 2 种状态的表型发生率(π_1/π_2)设定为 3 个水平(10 %/90 %、20 %/80 %、40 %/60 %);3 阈值表型的发生率也设置为 3 个水平(5 %/15 %/60 %/20 %、10 %/40 %/40 %/10 %、25 %/25 %/25 %/25 %),并分别由此确定阈值 v。阈性状的遗传力也设定为 3 个水平(0.1、0.2、0.4)。

在基础群中,假定个体育种值效应 u 服从正态分布 $N(0, \sigma_u^2)$(σ_u^2 为加性遗传方差),e 服从 $N(0, \sigma_e^2)$(σ_e^2 为随机环境效应方差,设定为 1)。在给定性状遗传力 h^2 和 σ_e^2 的情况下,σ_u^2 的推算公式如下:$\sigma_u^2 = h^2\sigma_e^2/(1 - h^2)$。后代的加性遗传效应为 $u_i = 0.5u_i^s + 0.5u_i^d + m$,其中 u_i^s 和 u_i^d 分别为后代 i 的父亲和母亲加性遗传效应效应,m 为孟德尔抽样离差,服从分布 $N(0, \sigma_m^2)$,其中 $\sigma_m^2 = 0.25[(1 - f_s) + (1 - f_d)]\sigma_u^2$,$f_s$ 和 f_d 分别为父亲和母亲的近交系数。后代个体潜在表型值的产生同基础群,世代间 σ_e^2 保持不变。

1.3 遗传分析方法

广义线性模型定义为响应变量(Y)具有如下分布密度的指数分布族[2]

$$f_Y(y, \theta, \phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right\}$$

式中:$a(\phi)$、$b(\theta)$、$c(y, \phi)$均为已知函数,θ为典范参数,ϕ为离差参数。广义线性模型不同于一般线性模型,其响应变量的方差不是一个常数(σ^2),而是一个依赖于均值的函数。在广义线性模型中,通过反连接函数将条件均值 μ 与线性预测 η 相联系,并通过方差函数将协方差矩阵 R 与条件均值相联系,从而实现误差方差依赖于条件均值。本研究中,对于单阈值表型采用 logist 连接,对于 3 阈值表型采用 log 连接。观察值的方差构成主要有 2 个组分,一是随机效应方差,另一是误差分布方差。与混合线性模型方差组分估计类似,广义线性模型的方差组分估计可通过准似然函数(quasi likelihood)REML 方法实现,具体估计方程如下:

$$ql(\beta, \sigma) = -\frac{1}{2}\ln|V| - \frac{1}{2}\ln|X'HV^{-1}HX| - \frac{1}{2}(y^* - HX\beta)'V^{-1}(y^* - HX\beta)$$

式中:σ 为方差组分向量,$V = R + HZGZ'H'$。GLMM 中参数的估计方法及各方差组分迭代求解可参见文献[2]和[5]。

1.4 估计准确度的度量[6]

遗传力估计的准确度主要通过估计遗传力的偏

差(BIAS)和均方误(MSE)来度量。$MSE = \frac{\sum_{i=1}^{n}(\hat{h}_i^2 - h^2)^2/n}{h^2}$, $BIAS = \frac{\sum_{i=1}^{n}(\hat{h}_i^2 - h^2)/n}{h^2}$, n 是每一参数组合重复的次数, $\hat{h}_i^2$ 为第 i 次重复的估计遗传力, h^2 是遗传力真值。二者反映了遗传力估计的精度和准确性。

2　结果与分析

2.1　遗传力估计

各参数组合下, GLMM 方法的估计遗传力结果(表 1)可以看出, GLMM 方法的估计遗传力结果在多数参数组合下均处在真值附近,反映出广义线性模型在阈性状遗传分析上具有良好的统计特性。但是,不同参数设置下遗传力估计结果差异较大,性状遗传力较低时,遗传力估计结果出现偏低的趋势,随着性状遗传力真值的提高($h^2 > 0.2$), GLMM 方法出现高估现象,特别是在性状表型发生率相互接近的情况下, GLMM 方法的高估现象最为明显。

随着性状表型发生率的提高, GLMM 方法的估计遗传力结果有上升的趋势。从本研究的结果看,随着性状表型发生率的相互接近, GLMM 的遗传力估计出现偏高的趋势,总的来看, GLMM 方法对性状表型发生率的敏感性不如性状遗传力强。

潜在变量的阈值数决定着阈性状的表型分类情况,阈值的大小直接影响着表型发生率的高低。随着表型分类数的增加,其表型分布就愈接近正态分布,遗传信息的丢失程度也就愈少。因此,类似于常规的线性模型,随着阈值数的增加, GLMM 方法的估计遗传力效果也会进一步提高,但是在实际应用中,当性状的表型分类数较多时,可以直接利用常规的线性模型进行遗传分析。

表 1　GLMM 方法在不同参数组合下的估计遗传力及准确度

Table 1　Heritability estimates and accuracy computed by generalized linear mixed model (GLMM)

性状表型	表型发生率	遗传力真值	遗传力估计		
			估计遗传力	MSE	BIAS
单阈值表型	10 %/90 %	0.1	0.070 6	0.008 6	- 0.301 4
		0.2	0.140 0	0.018 0	- 0.323 1
		0.4	0.330 1	0.012 2	- 0.174 7
	20 %/80 %	0.1	0.123 5	0.005 5	0.235 0
		0.2	0.181 2	0.001 8	- 0.095 2
		0.4	0.381 2	0.000 9	- 0.047 0
	40 %/60 %	0.1	0.126 2	0.006 9	0.251 2
		0.2	0.230 5	0.004 7	0.152 5
		0.4	0.410 1	0.002 5	0.025 2
3 阈值表型	5 %/15 %/60 %/20 %	0.1	0.112 7	0.001 6	0.127 3
		0.2	0.210 4	0.005 4	0.065 6
		0.4	0.386 5	0.004 5	- 0.034 6
	10 %/40 %/40 %/10 %	0.1	0.115 8	0.002 5	0.152 8
		0.2	0.216 6	0.001 4	0.085 3
		0.4	0.407 0	0.000 5	0.017 5
	25 %/25 %/25 %/25 %	0.1	0.142 9	0.018 4	0.429 0
		0.2	0.227 0	0.003 6	0.142 5
		0.4	0.410 0	0.002 5	0.025 8

注:遗传力估计值为 100 次重复的均值。

2.2　遗传力估计的准确度

GLMM 方法遗传力估计的 MSE 与 BIAS(表 1)由 100 次重复模拟获得,二者分别反映了遗传力估计的准确性和精度。总的来看, GLMM 方法的估计均方误和偏差较小,特别是在性状的表型发生率较低的情况下, GLMM 方法的估计准确性最高。随着表型发生率的提高, GLMM 方法的遗传力估计偏差和估计均方误有增大的趋势,这可能是由于 GLMM 方法对遗传力的高估引起。

在相同的表型发生率下,随着性状遗传力的增

大(0.1→0.4),GLMM 方法的遗传力估计均方误和偏差有明显增大的趋势。从表 1 还可看出,在其他参数相同的条件下,随着性状表型分类数的增加,GLMM 方法的估计偏差和均方误下降。这主要是在多阈值情况下,性状的表型分布更趋向正态分布,遗传信息的丢失程度相应较少。

3 讨 论

广义线性混合模型是一般正态线性模型的直接推广[2],它适用于连续性状和离散性状的遗传分析,它与典型线性模型的主要区别是其随机误差的分布无需正态性假设。本研究对单阈值和 3 阈值离散性状的遗传分析结果显示,GLMM 方法能有效地估计阈性状的遗传力,在遗传力估计的准确性方面具有较大的优势。同时,性状的遗传力和性状的表型发生率对遗传力估计效果均有直接的影响,随着性状遗传力真值的提高,GLMM 方法出现高估现象;随着性状发生率的提高,GLMM 方法的估计遗传力结果有上升的趋势。说明 GLMM 方法在阈性状遗传分析上的效率很大程度上受到性状性质的影响,因此,在实际应用中应根据所研究性状的性质(包括性状的遗传力、阈性状的阈值数和表型发生率等)合理选择所要采用的遗传分析方法。

虽然笔者仅研究了单阈值和 3 阈值 2 种类型的阈值性状,但对于其他分布类型的阈性状而言,GLMM 方法的分析也是极为有效的[4,7-9],只是在连接函数及方差函数的选择上有所差异,这主要取决于阈性状的表型分布。动物育种中大多的经济性状表型都是连续的,且服从或近似服从正态分布,而对于阈性状,其表型值呈现出离散分布,在动物育种中有几种重要的离散分布,它们分别是二项分布(binomial)、多项分布(multinomial)、泊松分布(poisson)和负二项分布(negative binomial)。动物的生育能力、抗病或抗应激能力、成活或死亡等都服从二项分布或 0 - 1 分布,而有关母畜排卵数、产仔数、乳头数等记数性状则服从多项分布或者泊松分布。

目前,广义线性混合模型主要用于离散性状的遗传分析,包括育种值估计(效应预测)、方差组分及遗传参数的估计,而且分析的性状大多服从二项分布或者泊松分布。除此之处,GLMM 是否能渗入到动物遗传育种中的其他方面?答案应该是肯定的。近年来,随着统计基因组学的迅猛发展,对多基因控制的连续性状实现 QTL(数量性状位点)定位成为数量性状研究的热点,而对于阈性状由于其表型的不连续性,且不服从正态分布的假设,因此要实现阈性状 QTL 的准确有效定位,仍需要借助于非线性方法。目前,用于分类性状 QTL 定位的非线性方法主要有 3 种:Bayesian 法、GLMM 和非参数方法。这些方法在人类及植物分类性状的 QTL 定位中已发挥了重要的作用[10-12],但这些非线性方法在动物分类性状的 QTL 定位中则少见报道,这也可能是今后一段时间内动物分类性状的一个研究重点。

参 考 文 献

[1] Falconer D S. Introduction to Quantitative Genetics (4rd eds.)[M]. England: Longman, 1996

[2] McCullagh P, Nelder J A. Generalized Linear Models (2nd ed)[M]. London: Chapman & Hall, 1989

[3] Rekaya R, Gianola D. Application of a structural model for genetic covariance in international dairy sire evaluations[J]. J Dairy Sci, 2001, 84: 1525-1530

[4] Tempelman R J. A mixed effects model for overdispersed count data in animal breeding[J]. Biometrics, 1996, 52:265-279

[5] 殷宗俊,张勤等. 多基因抗性性状育种值预测的广义线性方法[A]. 李辉. 动物遗传育种研究进展[C]. 哈尔滨:中国农业科学技术出版社,2005. 21-24

[6] Gamal A, Berger P J. Properties of threshold model predictions[J]. J Anim Sci, 1999, 77: 582-590

[7] Gianola D. Sire evaluation for ordered categorical data with a threshold model[J]. Genet Sel Evol, 1983, 15: 210-224

[8] Foulley J L. Prediction of selection response for threshold dichotomous traits[J]. Genetics, 1992, 132:1187-1194

[9] Thomson P C. A generalized estimating equations approach to quantitative trait locus detection of non-normal traits[J]. Genet Sel Evol, 2003, 35: 257-280

[10] Christoph L. Mapping quantitative trait loci using generalized estimating equations[J]. Genetics, 2001, 159: 1325-1337

[11] Heather J. Statistical modeling of interlocus interactions in a complex disease[J]. Genetics, 2001, 158: 357-367

[12] Nengjun Y, Shizhong X. Bayesian mapping of quantitative trait loci for complex binary traits[J]. Genetics, 2000, 155: 1391-1403