Likelihood-Based Inference on Genetic Variance Component with a Hierarchical Poisson Generalized Linear Mixed Model[1]

C. Lee*
Laboratory of Statistical Genetics, Institute of Environment & Life Science, Hallym University
Chuncheon, Kangwon-do 200-702, Korea

ABSTRACT : This study developed a Poisson generalized linear mixed model and a procedure to estimate genetic parameters for count traits. The method derived from a frequentist perspective was based on hierarchical likelihood, and the maximum adjusted profile hierarchical likelihood was employed to estimate dispersion parameters of genetic random effects. Current approach is a generalization of Henderson's method to non-normal data, and was applied to simulated data. Underestimation was observed in the genetic variance component estimates for the data simulated with large heritability by using the Poisson generalized linear mixed model and the corresponding maximum adjusted profile hierarchical likelihood. However, the current method fitted the data generated with small heritability better than those generated with large heritability. (*Asian-Aus. J. Anim. Sci. 2000. Vol. 13, No. 8 : 1035-1039*)

Key Words : Count Variable, Dispersion Parameter, Maximum Adjusted Profile Hierarchical Likelihood

INTRODUCTION

Various quantitative genetic analyses of non-normal data have been extended from Hendersonian mixed model methodology. For instance, ordinary categorical traits such as dystocia were analyzed with threshold models (Gianola and Foulley, 1983; Harville and Mee, 1984; Zhao, 1987). A probit link function was recommended for analysis of a Bernoulli variable such as survival data (Foulley et al., 1987; Everett, 1996). Poisson models have been suggested for count variates, e.g. litter size (Foulley et al., 1987), prolificacy (Perez-Enciso et al., 1993), and embryo yield (Tempelman and Gianola, 1994). These models can be categorized as generalized linear mixed model (GLMM) which is the mixture of mixed model and generalized linear model. In GLMM, marginal maximum likelihood (MML) estimation was traditionally applied to estimating dispersion parameters in animal genetic analyses (Foulley et al., 1987; Foulley and Im, 1993; Tempelman and Gianola, 1993). However, it is computationally troublesome in GLMM because it requires the numerical evaluation of high dimensional integrals. Generally, they cannot be evaluated in closed forms, so approximation must be used. For instance, Stiratelli et al. (1984) introduced an expectation-maximization (EM) algorithm, Breslow and Clayton (1993) used penalized quasi marginal likelihood, and Tempelman and Gianola (1993) employed Laplace method to marginalize posterior densities in the field of animal breeding.

One of the major interests in the models which include genetic random effects is to develop a better method that estimates genetic variance components. Along with mixed models, Patterson and Thompson's (1971) restricted maximum likelihood (REML) has been employed as a standard method to estimate variance components. Breslow and Clayton (1993) extended this approach to GLMM by using the normal likelihood.

As a choice for genetic analysis of count traits, the current study attempted to derive a likelihood-based method with Poisson GLMM from a frequentist perspective. The methodology was developed based on Lee and Nelder's (1996) hierarchical generalized linear models (HGLM) where the random effects can have any kind of arbitrary density function. In HGLM, if the random effects follow a normal distribution as in the current study, it is reduced to GLMM.

Lee and Nelder (1996) introduced the maximum adjusted profile hierarchical likelihood estimator (MAPHLE) in HGLM. The MAPHLE was derived to estimate genetic variance component in the Poisson GLMM. While previous studies (Foulley et al., 1987; Tempelman and Gianola, 1994) on animal genetic analyses with Poisson GLMM dealt with approximations, the method presented in the current study did not.

MATERIALS AND METHODS

Hierarchical Poisson generalized linear mixed model

Genetic analyses for count data can be performed with a Poisson error model with random effects (Foulley et al., 1987; Tempelman and Gianola, 1993). The models with random effects can be represented as various two-stage hierarchical structures as shown by

* Address reprint request to C. Lee. Tel: +82-361-240-1794, Fax: +82-361-242-7534, E-mail: clee@sun.hallym.ac.kr.

[1] This research was supported by the Korea Research Foundation, Project No. 1998-003-G00060.

Received July 13, 1999; Accepted January 5, 2000

Searle et al. (1992). According to them, mixed models were considered as normal-normal hierarchical models.

A Poisson-normal hierarchy was composed as follows. First, the conditional distribution of a count variate given fixed and random effects has the Poisson distribution:

$$f(y_{ijk}|\beta_i, u_{ij}) = \frac{e^{-\lambda_{ij}}\lambda_{ij}^{y_{ijk}}}{y_{ijk}!}$$

where y_{ijk} is observation, β_i is fixed effect, u_{ij} is random effect, and λ_{ij} is Poisson parameter. Secondly, the distribution of the vector v has the multivariate normal distribution with zero means and the covariances equal to $\mathbf{A}\sigma^2_a$ where v=log **u**, **A** is numerator relationship matrix, and σ^2_a is additive genetic variance. The linear predictor (η') takes the form η'=**X**β+**Zu** where β and **u** are vectors of unknown fixed and random effects, respectively, and **X** and **Z** are their corresponding known design matrices. For this Poisson error model, the canonical log link was used between linear predictor and the mean of the response, i.e. η'=ln μ' where μ' is the conditional mean of **y** given **u**. Thus, the distribution of **u** has a multivariate log normal distribution.

Then the hierarchical log likelihood was constructed by summing the logarithm of the density functions in two stages. The equation is presented below:

$$\begin{aligned} h &= l(\lambda; y \mid u) + l(\sigma^2_a; v) \\ &= \sum_{ijk} \ln\left(\frac{e^{-\lambda_{ij}}\lambda_{ij}^{y_{ijk}}}{y_{ijk}!}\right) + \ln\left(\frac{e^{-\frac{v' A^{-1} v}{2\sigma^2_a}}}{\sqrt{|2\pi A\sigma^2_a|}}\right) \\ &\propto \sum_{ijk}(y_{ijk}\ln\lambda_{ij} - \lambda_{ij}) - \frac{v' A^{-1} v}{2\sigma^2_a} - \frac{r\ln\sigma^2_a}{2} \end{aligned}$$

where r is the size of the matrix, **A**.

With known genetic variance component, the estimators and predictors can be derived by maximizing the hierarchical likelihood, i.e., the estimates are obtained by solving $\partial h/\partial\beta$=0 and $\partial h/\partial v$=0. The estimates are called maximum hierarchical likelihood estimates (MHLE). In field data, the genetic variance component estimation needs to be preceded obtaining the MHLE. The MHLE corresponds to the posterior mode of Tempelman and Gianola (1993) under the assumption of the uniform prior density for genetic variance component in Bayesian inference. But the assumption about the prior is theoretically improper (Hobert and Casella, 1996). In this study, solutions were obtained by Newton-Raphson method after constructing the expected Hessian matrix:

$$H = \begin{pmatrix} X^TWX & X^TWZ \\ Z^TWX & Z^TWZ + U \end{pmatrix}$$

where **W** is the GLM weight function, **W**=($\partial\mu$'/$\partial\eta$')^{2}V(μ')$^{-1}$, and U=-∂^2l(σ^2_a;v)/∂**v**∂**v**'. Let (β'$^{(k)}$ **v**'$^{(k)}$) be the kth solution vector. Then the solutions can be obtained by iteratively solving this equation:

$$\begin{pmatrix} \beta^{(k+1)} \\ v^{(k+1)} \end{pmatrix} = \begin{pmatrix} \beta^{(k)} \\ v^{(k)} \end{pmatrix} + (H^{(k)})^{-1} \left.\begin{pmatrix} \frac{\partial h}{\partial\beta} \\ \frac{\partial h}{\partial v} \end{pmatrix}\right|_{\substack{\beta=\beta^{(k)} \\ v=v^{(k)}}}$$

In every round, (k+1)th solutions satisfy the above equation. Iterations continue until solutions converge. Now the resulting solutions to β and **v** based on known variances become the best linear unbiased predictor and the marginal maximum likelihood estimator (Lee and Nelder, 1996).

Estimation of genetic variance component

MAPHLE was derived to estimate genetic variance component in the hierarchical Poisson generalized linear mixed model. Adjusted hierarchical likelihood is defined as follows:

$$h_A = h + .5\ln\{\det(2\Pi H^{-1})\}.$$

Then the adjusted profile hierarchical likelihood is

$$h_P = h_A \mid_{\beta=\hat\beta, v=\hat v},$$

where $\hat\beta$ and $\hat{v}$ are estimated values. The first and the second derivatives against genetic variance component were derived as:

$$\frac{\partial h_A}{\partial\sigma^2_a} = \sum_{ij}(v_{ij} - u_{ij} + \ln\sigma^2_a + 1 - \frac{d}{d\sigma^2_a}\ln\Gamma(\sigma^2_a)) - \frac{1}{2}tr(\mathbf{H}^{-1}\frac{\partial\mathbf{H}}{\partial\sigma^2_a})$$

$$\frac{\partial^2 h_A}{(\partial\sigma^2_a)^2} = \sum_{ij}(\frac{1}{\sigma^2_a} - \frac{d^2}{(d\sigma^2_a)^2}\ln\Gamma(\sigma^2_a)) + \frac{1}{2}tr(\mathbf{H}^{-1}\frac{\partial\mathbf{H}}{\partial\sigma^2_a}\mathbf{H}^{-1}\frac{\partial\mathbf{H}}{\partial\sigma^2_a})$$

Let the Hessian matrix be $\mathbf{H}_p$=-∂^2h$_A$/($\partial\sigma^2_a$)^2. Then the MAPHLE for genetic variance component can be obtained by iteratively solving the equation below:

$$(\sigma^2_a)^{(k+1)} = (\sigma^2_a)^{(k)} + (H_b^{(k)})^{-1}\left(\frac{\partial h_A}{\partial\sigma^2_a}\right)\Bigg|_{\sigma^2_a = (\sigma^2_a)^{(k)}}$$

Convergence criterion utilized in this study was:

$$\left|(\tilde\sigma^2_a)^{(k)} - (\tilde\sigma^2_a)^{(k-1)}\right| < 10^{-5}$$

In order to confirm that the value at convergence was not a local maximum, at least three runs with different initial values were performed for each analysis.

The MAPHLE becomes the REML estimator in mixed linear models, so it is the generalization of the REML estimator to non-normal mixed models. MAPHLE was justified by the method of moments (Lee and Nelder, 1996).

Simulation

Monte Carlo simulation was performed to examine whether the method introduced in this study was suitable for genetic evaluation of the count traits. Simulated were embryo yields within a nucleus breeding scheme combined with multiple ovulation and embryo transfer. An embryo yield was generated with a Poisson parameter whose logarithm could be additively explained by fixed and random effects. The fixed effects had five levels. The underlying means on the log scale were ln(3), ln(4), ln(5), ln(6), and ln(7) for the five levels.

Genetic merits for the base animals (25 sires and 50 dams) were generated from normal distribution with zero mean and variance equal to 0.05, 0.1, or 0.2. The data simulated with these three values were referred to Data 1, Data 2, and Data 3, respectively. The genetic merits for the subsequent generations were calculated as half of the genetic merits of their parents plus the Mendelian sampling. The Mendelian sampling is a random sampling of parental genes caused by segregation and independent assortment of genes during germ cell formation. The sampling was generated from $N(0, \sigma^2_a/2)$. Sex of progeny was randomly assigned. Among female progeny, fifty donor cows were also randomly selected per generation. They were superovulated and mated to randomly assigned sires. Random matings and random selections were applied all through the simulation. Finally the phenotypes were generated from Poisson distribution with the Poisson parameter equal to the exponent of the fixed effects multiplied by the exponent of the random effects. Two hundred fifty donor dams were produced and ten records were simulated per donor dam. Therefore, 2,500 embryo yields were generated per population. A total of 20 replicates were simulated. All the random deviates from Poisson and Normal distributions were generated based on the algorithms by Press et al. (1992).

RESULTS

The estimates of genetic variance and fixed effects using the likelihood-based method derived in this study were obtained from the simulated data (tables 1 and 2). Based on the empirical standard errors from 20 replicates, t tests were performed to examine whether the estimates were different from the input values. Overall, the MAPHLEs of genetic variance components obtained from the three data sets tended to be underestimated (table 1). This concurred with Tempelman and Gianola (1993, 1994). However, the estimates of genetic variances obtained from Data 1 and from Data 2 corresponded to their input values ($p>0.05$), but not from Data 3 ($p<0.05$).

The fixed effect estimates of the underlying means on the log scale from Data 1 are shown in table 2. Fixed effect estimates for the first to the fifth levels did not differ from their corresponding input values ($p>0.05$). However, some estimates obtained from Data 2 were different from their corresponding input values ($p<0.05$). Furthermore, most estimates from Data 3 were not corresponding to their input values ($p<0.05$). These differences showed the evidence of bias produced in this estimation method. Such bias had been consistently identified in previous GLMM studies (Breslow and Clayton, 1993; Tempelman and Gianola, 1994).

Table 1. Maximum adjusted profile hierarchical likelihood estimates and their empirical standard errors of genetic variance components obtained in simulated data

Data	Input value	Estimate ± S.E.
1	0.05	0.047^{ns} ± 0.002
2	0.1	0.094^{ns} ± 0.003
3	0.2	0.186* ± 0.005

ns $p>0.05$; * $p<0.05$.

Table 2. Estimates and their empirical standard errors of fixed effects obtained in simulated data

Input value	Estimate ± S.E.		
	Data 1 (σ^2_a=.05)	Data 2 (σ^2_a=.1)	Data 3 (σ^2_a=.2)
1.10	1.11^{ns} ± 0.010	1.12^{ns} ± 0.012	1.14* ± 0.013
1.39	1.39^{ns} ± 0.009	1.41^{ns} ± 0.011	1.42^{ns} ± 0.016
1.61	1.62^{ns} ± 0.008	1.64* ± 0.012	1.65* ± 0.014
1.79	1.81^{ns} ± 0.007	1.82* ± 0.009	1.84* ± 0.016
1.95	1.96^{ns} ± 0.008	1.97^{ns} ± 0.012	1.98^{ns} ± 0.018

ns $p>0.05$; * $p<0.05$.

DISCUSSION

While animal breeders (e.g., Gianola and Foulley, 1983; Foulley and Gianola, 1984) had suggested Bayesian approach for estimating parameters of non-normal data, a likelihood-based method derived from a frequentist perspective was applied in this study. It should be noted that no assumptions on the distribution of genetic variance component was made here as opposed to subjective priors in the Bayesian analyses. The inferences using hierarchical likelihood can avoid the high dimensional integration of fixed and random effects, which is required for the use of marginal likelihood.

The method employed in this study is an expansion of Henderson's mixed model approach. If observations are assumed to have normal errors, the current model is equivalent to mixed model. Hierarchical likelihood is an expansion of Henderson's (1975) joint likelihood to non-normal mixed models, and MAPHLE is an expansion of Patterson and Thompson's (1971) REML.

It is theoretically more reasonable to analyze count traits with Poisson error model (or Poisson GLMM) than with Normal error model (or mixed model). As mentioned in the introduction, GLMM is a mixture of mixed model and generalized linear model. Therefore, GLMM has many advantages over mixed models. First, GLMM can accommodate non-normal or non-linear data because observations can have errors from certain exponential families. For instance, the current study dealt with count data employing Poisson distribution. Second, GLMM allows a link between the mean response and the predictor. A log link was utilized in the current study while the identity link is assumed in classical mixed models. Third, GLMM can explain overdispersion. Since Poisson error was assumed in this study, variance of response equaled its mean. However, dispersion parameters can be included even in the Poisson error model, i.e., $\mathrm{var}(y \mid \boldsymbol{u})= \varphi V(\mu')$ where φ is dispersion parameter. This modified model would be applied to field data to explain heterogeneity of dispersion parameters.

Single variance component for random effects was employed in this study. However, the analyses of field data often require more than one component. The vector of the random effects can be partitioned as u' $=[u_1' \ u_2' \ \dots \ u_s']$ where u_i is vector of the i^{th} random effects. Then the generalization of the procedure for the model with multiple variance components is straightforward. Furthermore, distribution of random effects can be generalized. Lee and Lee (1998) introduced Gamma distribution for random effects to sire evaluation.

The applicability of the MAPHLE in genetic parameter estimation was investigated in this study. The use of Poisson GLMM and the corresponding MAPHLE produced some biases as other methods did (Breslow and Clayton, 1993; Tempelman and Gianola, 1994). The proposed method fitted the data generated with small heritability (Data 1) better than those with large heritability (Data 3).

REFERENCES

Breslow, N. E. and D. G. Clayton. 1993. Approximate inference in generalized linear mixed models. J. Am. Stat. Assoc. 88:9-25.

Everett, R. W. 1996. Personal communication.

Foulley, J. L. and D. Gianola. 1984. Estimation of genetic merit from bivariate all-or-none responses. Genet. Sel. Evol. 16:285-306.

Foulley, J. L., D. Gianola and S. Im. 1987. Ginetic evaluation for discrete polygenic traits in animal breeding. In: Advances in Statistical Methods for Genetic Improvement of Livestock (Ed. D. Gianola, K. Hammond). Springer Verlag, Heidelberg. pp. 361-409.

Foulley, J. L. and S. Im. 1993. A marginal quasi-likelihood approach to the analysis of Poisson variables with generalized linear mixed models. Genet. Sel. Evol. 25:101-107.

Gianola, D. and J. L. Foulley. 1983. Sire evaluation for ordered categorical data with a threshold model. Genet. Sel. Evol. 15:201-224.

Harville, D. A. and R. W. Mee. 1984. A mixed model procedure for analyzing ordered categorical data. Biometrics. 40:393-408.

Henderson, C. R. 1975. Best linear unbiased estimation and prediction under a selection model. Biometrics. 31: 423-447.

Hobert, J. and G. Casella. 1996. Effect of improper priors on Gibbs sampling in Hierarchical linear mixed model. J. Am. Stat. Assoc. 91:1461-1473.

Lee, C. and Y. Lee. 1998. Sire evaluation of count traits with Poisson-Gamma hierarchical generalized linear model. Asian-Aus. J. Anim. Sci. 11:642-647.

Lee, Y. and J. A. Nelder. 1996. Hierarchical generalized linear models. J. R. Stat. Soc. B58:619-678.

Patterson, H. D. and R. Thompson. 1971. Recovery of interblock information when block sizes are unequal. Biometrika. 58:545-554.

Perez-Enciso, M., R. J. Tempelman and D. Gianola. 1993. A comparison between linear and Poisson mixed models for litter size in Iberian pigs. Livest. Prod. Sci. 35:303-316.

Press, W. H., S. A. Teukolsky, W. T. Vetterling and B. P. Flannery. 1992. Numerical Recipes in FORTRAN (2nd. ed.). Cambridge University Press, New York.

Searle, S. R., G. Casella and C. E. McCulloch. 1992. Variance Components. Wiley, New York.

Stiratelli, R., N. M. Laird and J. H. Ware. 1984. Random-effects models for serial observations with binary response. Biometrics. 40:961-971.

Tempelman, R. J. and D. Gianola. 1993. Marginal maximum likelihood estimation of variance components in Poisson mixed models using Laplacian integration. Genet. Sel.

Evol. 25:305-319.

Tempelman, R. J. and D. Gianola. 1994. Assessment of a Poisson animal model for embryo yield in a simulated multiple ovulation-embryo transfer scheme. Genet. Sel. Evol. 26:263-290.

Zhao, Y. 1987. Estimation of parameters in a mixed threshold model: its application to dystocia and birth weight in Simmental cattle. Ph.D. Dissertation. Cornell University, Ithaca, New York.