Generalized Linear (Mixed) Models

© Reg Morrison / www.ardea.com

JANUARY 14

Echidna Mixed Models Software

Authored by: Arthur R Gilmour

Logo
Name

Generalized Linear (Mixed) Models

In ASReml and Echidna

ASReml and Echidna both fit linear mixed models and use iteratively reweighted least squares to fit linear models to discrete valued data assumed to follow a binomial (or other distribution from the exponential family) using several link functions.

This report seeks to explain the way these models are fitted. Two forms are discussed: the usual PQL or 'Schall' method where working variables and weights depend on both fixed and random effects, and a 'Maximization- Expectation' method of Gilmour et al (1985) where the weights just depend on the fixed effects.

An Example

First, consider some binary data collected from the lamb progeny of 34 sires. These data were collected by M Alwan for his masters project at Massey University as quoted by Gilmour (Ph D thesis, 1983). The sires represent 5 cohorts representing 2 years and 3 breed types:

Rams	Ewes	Year lambs born	
7 Perendale	Perendale	1980	
6 Booroola x Romney	Perendale	1980	
3 Booroola	Romney	1980	
6 Perendale	Perendale	1981	
12 Booroola x Romney	Perendale	1981	

While the 12 Booroola x Romney used in 1981 were progeny from group 3, the details are not available.

Booroola Merino sheep were reputed to be more susceptible to foot problems than Perendale sheep and the study was conducted to investigate this. The male and female lambs were scored for foot rot, foot scald and foot shape (the number of deformed feet, 0:4).

First analyzing L5, (all foot shape OK), as a normal variable,

L5 ~ mu SEX Breed Year !r SIRE

Echidna reports

```
Data File: alwan2.asd

 Summary of 2513 data records

 Variable   Levels Miss Zero      Min      Max    Distribution or Mn SD Sk Kt
 Year            2    0    0        1        2   1141 1372
 Breed           3    0    0        1        3   1071 1323 119
 SEX             1    0 1227  0.00000  1.00000  0.51174  0.49996  -0.05  -2.00
 SIRE           34    0    0        1       34
 FSscore         1    0    0  1.00000  3.00000  1.40708  0.59808   1.18   0.34
 L5              1    0  877  0.00000  1.00000  0.65101  0.47674  -0.63  -1.60
 L4              1    0 1782  0.00000  1.00000  0.29089  0.45426   0.92  -1.15
 LS              1    0 2341  0.00000  1.00000  0.06844  0.25256   3.42   9.67
 LR              1    0 2446  0.00000  1.00000  0.02666  0.16112   5.88  32.51
    1 LogL= 653.61  0.2104          2508 DF
    2 LogL= 654.90  0.2106          2508 DF
    3 LogL= 655.82  0.2123          2508 DF
    4 LogL= 656.28  0.2117          2508 DF
    5 LogL= 656.32  0.2116          2508 DF
    6 LogL= 656.32  0.2116          2508 DF

 Akaike Information Criterion   -1308.64 (assuming 2 parameters).
 Bayesian Information Criterion -1296.98

          Analysis of L5

                         Wald F statistics
Source of Variation           NumDF     DenDF     F-inc     F-con    P-inc
 mu                               1             1126.21   1126.21
 SEX                              1                4.80      4.53
 Breed                            2                6.20      3.38
 Year                             1                8.09      8.09

 Model_Term                     Order    Gamma        Sigma     Z_ratio  %C
 SIRE                              34 0.449257E-01 0.950481E-02    2.98   0 P
 Residual_units                  2513  1.00000     0.211567       35.21
  SIRE                                    34 effects fitted.
```

and

```
     Heritability = Genetic    4/Total      3 =         0.17198       0.05550
```

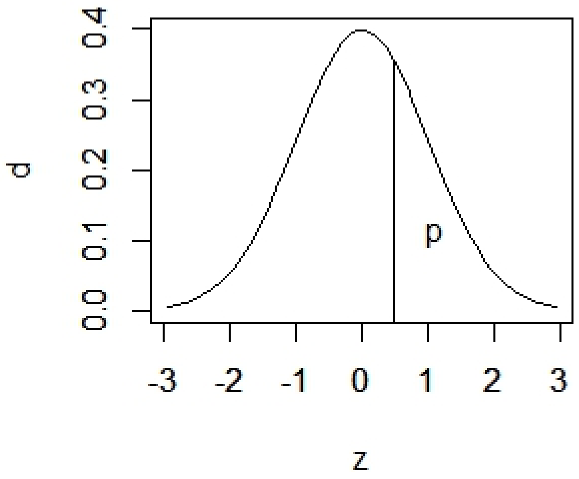
The liability scale

There are two issues with analysis on the observed scale.

1. For some models, predicted values may be outside of the range 0,1. This can be overcome by mapping the proportion (p: 0,1) to the real line (z) using a link

function p=f(z). Three link functions that do this (in R notation) are probits (z=qnorm(p); p=pnorm(z)), logits (z=log(p/(1-p)); p=exp(z)/(1+exp(z)) and complementary log log (z=log(-log(1-p)); p=1-exp(-exp(z))).

2. The variance of binary data is p(1-p), a function of the mean. For normal data, mean and variance are two separate parameters. Consequently, heritability is dependent on the mean also.



In this figure of a Normal (0,1) distribution, z measures standard deviations from the mean, d (=dnorm(z)) is the density and p (=1-pnorm(t)) is the area under the curve to the right of a threshold point (value of z) t.

Iteratively reweighted least squares

A typical formulation of the mixed model is y ~ N(XB, V=R+ZGZ') where X and Z are design matrices, V, R and G are variance matrices. B are fixed effects estimated as $B=(X'V^{-1}X)^{-1}X'V^{-1}y$ assuming X, V, and y are known. In ASReml and Echidna, given the form of R and G are specified, the parameters are estimated using the mixed model equations which also produce estimates of random effects (u).

$$\begin{array}{llllll} [X'R^{-1}X & X'R^{-1}Z &] & [\,B\,] & & [\,X'R^{-1}y] \\ [Z'R^{-1}X & Z'R^{-1}Z+G^{-1} &] & [\,u\,] & = & [\,Z'R^{-1}y] \end{array}$$

Now for binary, we can allow for the mean/variance issue by using a weighted analysis, calculating the weights from proportions predicted under the model W=1/(p(1-p)). The typical GL(M)M approach is to predict the proportions from the mixed model (p=XB+Zu) but an alternative available in Echidna is to use the marginal model (p=XB) to calculate the proportions from which the weights are calculated.

If we also use a link function, then we form a working response variable
w = XB + (y-p)/d for the marginal model, and
w = XB + Zu + (y-p)/d for the full mixed model,
where p = f(XB) or p=f(XB+Zu) depending on the model being used, and d is the derivative of the link function, a function of p, used to rescale the residual (y-p) onto the working variable scale. We can then analyse the working variable using weights d^2W.

The common GLMM method (using w = XB + Zu + (y-p)/d) is well known to be seriously biased when estimating variance components for binary traits under the animal model. The marginal model has therefore been implemented in Echidna as an alternative to be considered.

Under the GLMM method, the variance components for the random effects are on a scale determined by the link function, relative to a variance of 1 for probit, $\pi^2/3=3.28987$ for logit and $\pi^2/6=1.6449$ for complementary log log. So, for a sire model, using a logit link and allowing a dispersion parameter (Vd), the heritability is calculated as 4*Vg/(3.28987*Vd+Vg) and pertains to an underlying liability variable.

Under the marginal method, the variance components for the random effects are relative to a total variance of 1. So, for a sire model, using a logit link and allowing a dispersion parameter, the heritability is calculated as 4*Vg/(Vd+Vg) and pertains to the binary variable itself. It can be converted to a liability value by dividing by d^2W calculated at the average incidence.

Generalized Linear (Mixed) Models

The method of analysis for binomial (binary) and poisson data implemented as iteratively reweighted least squares is known as PQL (posterior quasi-likelihood). The process is invoked by specifying the distribution (!BINOMIAL or !POISSON) and the link function (!LOGIT, !PROBIT or !COMPLOGLOG for binomial, !LOG or !SQRT for Poisson) after the variable. For binomial data, a !TOTALS <totals> qualifier is required to set the count numbers for binomial data (else the response is assumed 0/1 (binary)). A dispersion parameter is estimated by default from the residual of the working variable), unless !DISP 1 is specified to set the dispersion to 1.

For binomial count data, the response is assumed to be a proportion unless the mean is greater than 1 in which case it is assumed to be counts.

See the LAMB example

```
L5 !BINOMIAL !LOGIT !TOTAL=TOT !DISP 1 ~ mu SEX GRP !r SIRE .16783
```

The PQL method is well known to give estimates of variance parameters which are biased down when class sizes are low (<10) for the random terms. This particularly applies to the so called ANIMAL model. An alternative intermediate method is available with the !MARGINAL qualifier. The concept is based on the 'Maximization- Expectation' method of Gilmour et al (1985) but the implementation is different. In this implementation, the weights are based only on the fixed part of the model, but the mixed model equations are used allow for a correlated correlation structure. It should only be used with binary (not grouped) data. The variance components are relative to a total variance of 1. See above.

GLMM Sire model for Lamb foot scoring data.

M Alwan in a masters project at Massey University score the feet of 2513 lambs and the results are summarized in the file lamb.dat. The lambs represented 34 sires in 5 groups representing 3 breed crosses and 2 years. The aim of this analysis is to estimate the heritability on the liability scale.

The analysis given above of the binary data produced a heritability estimate of 0.172. Scaling this up to the liability scale (mean proportion 0.65) gives 0.285

h2l=function(p,h2){x=qnorm(p)

z=dnorm(x)

h2l=h2*p*(1-p)/z^2}

```
!DEBUG !LOG !OUT  !REN !ARG  1
MOHAMMAD ALWAN lAMB DATA FOOT SCORES    !DOPART $1
Year 2 Breed 3 SEX SIRE !I FSscore L5 L4 LS LR
alwan2.asd !skip 1 !DDF -1   !SLN
!PART 2
L5 !bin  ~ mu SEX Breed Year !r SIRE
VPREDICT
F Total Residual*3.289 + SIRE
F Genetic SIRE*4
H Heritability Genetic Total

 Analysing L5 as Binomial with Logit link
    1 LogL= -3201.61  0.9811         2508 DF
    2 LogL= -3213.53  0.9830         2508 DF
    3 LogL= -3228.38  0.9853         2508 DF
    4 LogL= -3234.73  0.9864         2508 DF
    5 LogL= -3235.99  0.9866         2508 DF
    6 LogL= -3236.10  0.9867         2508 DF
    7 LogL= -3236.11  0.9867         2508 DF
    8 LogL= -3236.11  0.9867         2508 DF
          Analysis of L5
                         Wald F statistics
Source of Variation          NumDF     DenDF     F-inc     F-con    P-inc
 mu                              1               49.12     49.12
 SEX                             1                4.68      4.56
 Breed                           2                5.69      3.01
 Year                            1                8.70      8.70

 Model_Term                     Order     Gamma        Sigma     Z_ratio  %C
 SIRE                              34 0.210228     0.207425        2.95   0 P
 Residual_units                  2513  1.00000     0.986664       35.21
```

The heritability is reported in the .evp file.

```
   1 Residual                                         0.98666       0.28026E-01
   2 SIRE                                             0.20742       0.70234E-01
   3 Total                                               3.4534       0.11452
   4 Genetic                                            0.82970       0.28094
     Heritability = Genetic    4/Total      3 =         0.24025       0.07689
 Notice: The parameter estimates are followed by
          their approximate standard errors.
```

GLMM Sire model for Lamb foot scoring data: Marginal model.

```
…
L5 !bin  !MARG ~ mu Breed Year !r SIRE
VPREDICT
F Total Residual + SIRE
F Genetic SIRE*4
H Heritability Genetic Total


Echidna 0.074 Aa  6 Jun 2019 Windows               Fri Jun  7 14:09:41 2019
 Licensed to Arthur(Arthur@cargovale.com.au)
 MOHAMMAD ALWAN lAMB DATA FOOT SCORES
  Folder: E:\MMX-II\Ex\GLMM

 Year 2 Breed 3 SEX SIRE !I FSscore L5 L4 LS LR

 Data File: alwan2.asd

 Summary of 2513 data records

 Variable   Levels Miss Zero      Min      Max    Distribution or Mn SD Sk Kt
 Year           2    0    0        1        2   1141 1372
 Breed          3    0    0        1        3   1071 1323 119
 SEX            1    0 1227  0.00000  1.00000  0.51174  0.49996  -0.05  -2.00
 SIRE          34    0    0        1       34
 FSscore        1    0    0  1.00000  3.00000  1.40708  0.59808   1.18   0.34
 L5             1    0  877  0.00000  1.00000  0.65101  0.47674  -0.63  -1.60
 L4             1    0 1782  0.00000  1.00000  0.29089  0.45426   0.92  -1.15
 LS             1    0 2341  0.00000  1.00000  0.06844  0.25256   3.42   9.67
 LR             1    0 2446  0.00000  1.00000  0.02666  0.16112   5.88  32.51
    1 LogL= 653.61  0.2104          2508 DF
    2 LogL= 654.90  0.2106          2508 DF
    3 LogL= 655.82  0.2123          2508 DF
    4 LogL= 656.28  0.2117          2508 DF
    5 LogL= 656.32  0.2116          2508 DF
    6 LogL= 656.32  0.2116          2508 DF

 Akaike Information Criterion   -1308.64 (assuming 2 parameters).
 Bayesian Information Criterion -1296.98

          Analysis of L5

                         Wald F statistics
Source of Variation           NumDF     DenDF     F-inc     F-con    P-inc
 mu                               1             1126.21   1126.21
 SEX                              1                4.80      4.53
 Breed                            2                6.20      3.38
 Year                             1                8.09      8.09

 Model_Term                     Order    Gamma        Sigma     Z_ratio  %C
 SIRE                              34 0.449257E-01 0.950481E-02    2.98   0 P
 Residual_units                  2513  1.00000     0.211567       35.21
  SIRE                                    34 effects fitted.
 Warning: If job runs slow, see if !EQN 2 or !EQN 3 is faster.
 Finished: Fri Jun  7 14:09:41 2019  LogL Converged      ALWN2/ALWN
```

```
   1 Residual                                           0.21157       0.60092E-02
   2 SIRE                                               0.95048E-02   0.31877E-02
   3 Total                                              0.22107       0.67364E-02
   4 Genetic                                            0.38019E-01   0.12751E-01
     Heritability = Genetic    4/Total      3 =         0.17198       0.05550
```

Note that although the analysis uses the logistic scale, this just pertains to the fixed effects. Therefore, 3.289 does not come into the calculation of heritability. Indeed the heritability estimate pertains to the binary scale with a mean incidence of 0.65 and using the $z^2/(pq)$ conversion, 0.172 (essentially heritability on the binary scale) maps to 0.285 on the liability scale. This value is higher than the 0.24 obtained with the GLMM approach.

The real test for this !MARGINAL method though will be for an animal model with low overall incidence but reasonable mean differences among fixed classes.

Multinomial data

ASReml has the ability to fit a GLMM model to multinomial data under the ordered threshold model but this is not available in Echidna. A simple alternative peoposed by Wilton (is to replace the class numbers with scores which are the average liability of those in the class.

For example, for the trait Lambing Ease (LE) with 3 classes with proportions 0.80, 0.18. and 0.2, replace the class codes 1, 2 and 3 with scores -.350, 1.286 and 2.421 which were calculated (in R) as follows:

```
> qnorm(c(0.8,0.98))
[1] 0.8416212 2.0537489
> dnorm(qnorm(c(0.8,0.98)) )
[1] 0.27996192 0.04841814
> -.279962/0.8
[1] -0.3499525
> (0.279962-0.04841814)/0.18
[1] 1.286355
> 0.04841814/0.02
[1] 2.420907
```

Explanation: Category 2 has 18% of the values with 80% on the left and 2% on the right. These proportions (80% and 98%) correspond to threshold values for the normal distribution of 0.8416212 and 2.0537489 where the ordinates are dnorm(qnorm(c(0.8,0.98)) = 0.27996192 0.04841814

The average liability for a class is given by
(the change in ordinate)/probability = (0.279962-0.04841814)/0.18

```
[1] 1.286355
```

If you difference these 3 new score values (-.350, 1.286 , 2.421) you get 1.636 and 1.135 so the main effect (relative to the 1,2,3 scale) is to move the 3 class relatively closer to the 2 class.

For binary (2 classes) data, there is no gain because it is just a change of origin and scale.

A full GLMM analysis involves slightly different substitution because the model is more complex and it involves weights and iteration but the gain is usually small and this simple use of scores is easily extended to multivariate data. To assess the gain in your data, compare analyses of

1) Univariate of LE
2) Univariate of ZLE (LE transformed as above)
3) LE !MULTINOMIAL 3 (if you have access to ASReml)

Discussion

The !MARGINAL option as implemented here has not been published or implemented before in this form to my knowledge. It has also not been explored with respect to all the model options it opens up.

As currently implemented, the random effects are on what may be considered an relative scale, relative to a total variance of 1. As there are several scales on which residuals can be reported (see ASReml: working, deviance, etc) so the same options probably apply here.

While the random effects can be used directly say as breeding values, they are not on the same scale as the fixed effects so any prediction that would directly combine fixed and random effects will be invalid.

All random effects are currently treated as marginal. However there may be cases where random effects are well estimated and the user may like to have some estimated on the underlying scale. For example, if the random terms include a spl() term which is just one way of fitting a curvilinear response which might otherwise be fitted as a fixed polynomial.

The initial example described above related to a trait with fixed effect means ranging 0.45 to 0.75. As such, the GLM weights and derivatives have little effect. However, if the means ranged so 0.01 to 0.15, the weight effects would have a substantial impact.

Model Comparison

The likelihood reported by Echidna is for the working variable. This is not appropriate for comparing GL(M)M models because the working variable keeps changing. Ari Verbyla has proposed an adjustment which has been partially implemented in Echidna and is reported like

```
   6 LogL= -3221.65  0.9975        2507 DF
 Adjusted LogLikelihood suitable for comparing GLMM models: -1573.195
```

which for a model L5 !bin   ~ mu SEX GRP !r SIRE

can be compared with

```
   3 LogL= -3199.96   1.003        2507 DF
 Adjusted LogLikelihood suitable for comparing GLMM models: -1583.017
```

for a model L5 !bin   ~ mu SEX GRP

Notice that the LogL based on the working variable decreased 22.69 but the adjusted LogL increased 9.82.

The implementation is partial in that it has not been extended to more general forms of the G structure. (ex/glm/alwan.es 2 3)

List of References

Gilmour, A.R., Anderson, R.D. and Rae, A.L. (1985). The analysis of binomial data by a generalized linear mixed model. Biometrika, 72: 593 599. http://dx.doi.org/10.1093/biomet/72.3.593

Schaeffer, L,R. and Wilton, J. W. (1977) Evaluation of beef sires across breeds for calving ease. Canadian Journal of Animal Science, 57: 635-645.