Analysis of Lamb Birth Weight Data Using Linear Mixed Effects Models and Maximum Likelihood Procedures

Michael Chen

June 13, 2019

1 Summary of Reference Article Data

In their 1985 paper, Harville and Fenech presented a method for constructing an exact confidence interval for the ratio of two variance components in a linear mixed model. To illustrate their results, the authors used data on the weights at birth of 62 single-birth male lambs. These lambs come from five population lines (two control lines and three selection lines). Each lamb was the offspring of one of 23 rams (male sheep), and each lamb had a different mother. The age of the mother is one of three categories: 1-2 years (category 1), 2-3 years (category 2), or 3+ years (category 3). The following linear mixed model was proposed:

$$y_{ijk} = l_i + a_1 x_{ijk,2} + a_2 x_{ijk,3} + s_{ij} + e_{ijk}$$

The line effect is denoted by l_i $(i = 1, ..., 5)$ and it is fixed. $x_{ijk,2} = 1$ if the age of the kth mother $(k = 1, ..., n_{ij})$ corresponding to line i and sire j is in category 2, and $x_{ijk,2} = 0$ otherwise; $x_{ijk,3} = 1$ if the age of the kth mother corresponding to line i and sire j is in category 3, and $x_{ijk,3} = 0$ otherwise. a_1 and a_2 are fixed effects corresponding to these indicator variables of mother's age. s_{ij} $(j = 1, ..., n_i; n_1 = n_2 = n_3 = 4, n_4 = 3, n_5 = 8)$ are the random sire (ram) effects nested within lines and are assumed to be i.i.d. $\mathcal{N}(0, \sigma_s^2)$. Finally, e_{ijk} are the random errors and are assumed to be i.i.d. $\mathcal{N}(0, \sigma_e^2)$.

This model can be written in the standard LMM form $y = X\beta + Zs + e$.

2 Data Entry

I entered the data into a .csv file using Table 1.2 on page 36 of the textbook as a reference. The first six rows of the data frame in R are presented below. Additionally, each of the covariates is a categorical variable, and therefore should be coded as factor.

```
> head(lamb)
  Weight Sire Line Age
1    6.2   11    1   1
2   13.0   12    1   1
3    9.5   13    1   1
4   10.1   13    1   1
5   11.4   13    1   1
6   11.8   13    1   2

> lamb$Sire = factor(lamb$Sire)
> lamb$Line = factor(lamb$Line)
> lamb$Age  = factor(lamb$Age)
```

3 Analysis Using Maximum Likelihood

Here is the model fit in R using maximum likelihood. Note that the model is fit without an intercept so that the fixed effects are identifiable.

```
> lamb_ml = lmer(Weight ~ Line + Age - 1 + (1|Sire), data = lamb, REML = FALSE)
```

The summary output is as follows:

```
> summary(lamb_ml)
Linear mixed model fit by maximum likelihood  ['lmerMod']
Formula: Weight ~ Line + Age - 1 + (1 | Sire)
Data: lamb

AIC      BIC   logLik deviance df.resid
261.5    280.6   -121.7    243.5       53

Scaled residuals:
     Min       1Q   Median       3Q      Max
-2.94579 -0.61274  0.00908  0.69072  1.84722

Random effects:
 Groups   Name        Variance Std.Dev.
 Sire     (Intercept) 0.000    0.000
 Residual             2.971    1.724
Number of obs: 62, groups:  Sire, 23

Fixed effects:
      Estimate Std. Error t value
Line1 10.69826    0.58274  18.359
Line2 12.29425    0.71501  17.195
Line3 10.87041    0.61835  17.580
Line4 10.19127    0.58366  17.461
Line5 10.95558    0.52080  21.036
Age2  -0.06316    0.67220  -0.094
Age3   0.02184    0.52187   0.042

Correlation of Fixed Effects:
       Line1  Line2  Line3  Line4  Line5  Age2
Line2  0.170
Line3  0.239  0.358
Line4  0.104  0.183  0.231
Line5  0.224  0.335  0.452  0.217
Age2  -0.317 -0.337 -0.534 -0.172 -0.501
Age3  -0.290 -0.513 -0.646 -0.358 -0.606  0.482
```

The maximum likelihood estimates of the model parameters and the standard errors of the fixed effects are shown in the above output (under “Random effects” and “Fixed effects”). Interestingly, the estimate of the variance component for the sire random effect is zero (as are all the random effects estimates). Since the output doesn’t provide the standard errors of the variance components’ estimates, those have to be obtained in a different way (asymptotic covariance matrix or bootstrap).

3.1 Asymptotic Covariance Matrix (ML)

For this method, I will use the equations from page 11 of the textbook. Denote θ_1 by σ_e^2 and θ_2 by σ_s^2. The ij^{th} element of the 2x2 Fisher information matrix is as follows:

$$-E\left(\frac{\partial^2 l}{\partial\theta_i\partial\theta_j}\right) = \frac{1}{2}tr\left(V^{-1}\frac{\partial V}{\partial\theta_i}V^{-1}\frac{\partial V}{\partial\theta_j}\right)$$

Note that $V = R + ZGZ^T$, where in this case $R = \sigma_e^2 I_n$ and $G = \sigma_s^2 I_m$ with the sample size $n = 62$ and the number of sires $m = 23$. Therefore,

$$\frac{\partial V}{\partial\theta_1} = I_n \qquad \frac{\partial V}{\partial\theta_2} = ZZ^T$$

Since the Fisher information matrix depends on the unknown variance components, those will be estimated using the maximumum likelihood estimates. The asympototic covariance matrix is the inverse of the Fisher information matrix. Calculations are below:

```
> sigma_e_sq_ml=2.971
> sigma_s_sq_ml=0
> R_ml=sigma_e_sq_ml*diag(1,62)
> G_ml=sigma_s_sq_ml*diag(1,23)
> Z_ml=getME(lamb_ml, "Z")
> V_ml=R_ml+Z_ml%*%G_ml%*%t(Z_ml)
> # elements of the fisher information matrix:
> fisher11=1/2*sum(diag(solve(V_ml)%*%diag(1,62)%*%solve(V_ml)%*%diag(1,62)))
> fisher12=1/2*sum(diag(solve(V_ml)%*%diag(1,62)%*%solve(V_ml)%*%Z_ml%*%t(Z_ml)))
> fisher21=1/2*sum(diag(solve(V_ml)%*%Z_ml%*%t(Z_ml)%*%solve(V_ml)%*%diag(1,62)))
> fisher22=1/2*sum(diag(solve(V_ml)%*%Z_ml%*%t(Z_ml)%*%solve(V_ml)%*%Z_ml%*%t(Z_ml)))
> fisher=matrix(c(fisher11,fisher12,fisher21,fisher22),2)
> solve(fisher)
            [,1]        [,2]
[1,]  0.37300522 -0.08826841
[2,] -0.08826841  0.08826841
```

From this result, we have that the standard errors of $\hat{\sigma}_e^2$ and $\hat{\sigma}_s^2$, respectively, are the square roots of the diagonal elements of the above matrix: 0.6107 and 0.2971.

3.2 Parametric Boostrap (ML)

For this method, I will use the R function `bootMer` that was showcased by the TA during the second lab. Note that in my version of the function `mySumm`, I am extracting the estimates of the variance components rather than their square roots.

```
> mySumm = function(.) {
+     c(sigma_e_sq = sigma(.)^2, sigma_s_sq = unlist(VarCorr(.)))
+ }
> booted_lamb_ml = bootMer(lamb_ml, mySumm, nsim = 100)
> summary(booted_lamb_ml)

Number of bootstrap replications R = 100
                original  bootBias  bootSE bootMed
sigma_e_sq        2.9709 -0.437637 0.53460  2.5622
sigma_s_sq.Sire   0.0000  0.052813 0.16286  0.0000
```

From the boostrap results, we have that the standard errors of $\hat{\sigma}_e^2$ and $\hat{\sigma}_s^2$, respectively, are 0.5346 and 0.1629.

4 Analysis Using Restricted Maximum Likelihood

Here is the model fit in R using REML. Again, the intercept term is removed to allow for the fixed effects to be identifiable.

```
> lamb_reml <- lmer(Weight ~ Line + Age - 1 + (1|Sire), data = lamb)
```

The summary output is as follows:

```
> summary(lamb_reml)
Linear mixed model fit by REML ['lmerMod']
Formula: Weight ~ Line + Age - 1 + (1 | Sire)
Data: lamb

REML criterion at convergence: 238.9

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.5602 -0.6572  0.1012  0.6616  1.7770

Random effects:
 Groups   Name        Variance Std.Dev.
 Sire     (Intercept) 0.5114   0.7151
 Residual             2.9959   1.7309
Number of obs: 62, groups:  Sire, 23

Fixed effects:
       Estimate Std. Error t value
Line1 10.491153   0.726188  14.447
Line2 12.290287   0.824231  14.911
Line3 11.032864   0.764787  14.426
Line4 10.276735   0.738914  13.908
Line5 10.952840   0.608411  18.002
Age2  -0.155435   0.715703  -0.217
Age3   0.009646   0.548103   0.018

Correlation of Fixed Effects:
       Line1  Line2  Line3  Line4  Line5  Age2
Line2  0.113
Line3  0.161  0.263
Line4  0.062  0.125  0.151
Line5  0.162  0.271  0.357  0.156
Age2  -0.260 -0.297 -0.479 -0.141 -0.476
Age3  -0.222 -0.450 -0.542 -0.278 -0.562  0.508
```

The REML estimates of the model parameters and the standard errors of the fixed effects are shown in the above output under "Random effects" and "Fixed effects". Note that this time the estimate of the variance component for the sire random effect is not zero. Overall, the estimates are similar to the ones from the maximum likelihood method but they are not the same. Once again, the asympototic covariance matrix and the boostrap method will be used to estimate the standard errors of the variance components' estimates.

4.1 Asymptotic Covariance Matrix (REML)

For this method, I will use the equations from pages 10 and 15 of the textbook. Again, denote θ_1 by σ_e^2 and θ_2 by σ_s^2. The ij^{th} element of the 2x2 Fisher information matrix is as follows:

$$-E\left(\frac{\partial^2 l_R}{\partial\theta_i\partial\theta_j}\right) = \frac{1}{2}tr\left(P\frac{\partial V}{\partial\theta_i}P\frac{\partial V}{\partial\theta_j}\right)$$

As before, $V = R + ZGZ^T$, where $R = \sigma_e^2 I_n$ and $G = \sigma_s^2 I_m$ with the sample size $n = 62$ and the number of sires $m = 23$. Therefore,

$$\frac{\partial V}{\partial\theta_1} = I_n \qquad\qquad \frac{\partial V}{\partial\theta_2} = ZZ^T$$

Additionally, $P = V^{-1} - V^{-1}X(X^TV^{-1}X)^{-1}X^TV^{-1}$. Now for the calculations:

```
> sigma_e_sq_reml=2.9959
> sigma_s_sq_reml=0.5114
> R_reml=sigma_e_sq_reml*diag(1,62)
> G_reml=sigma_s_sq_reml*diag(1,23)
> Z_reml=getME(lamb_reml,"Z")
> X_reml=getME(lamb_reml,"X")
> V_reml=R_reml+Z_reml%*%G_reml%*%t(Z_reml)
> P=solve(V_reml)-solve(V_reml)%*%X_reml%*%solve(t(X_reml)%*%solve(V_reml)%*%X_reml)%*%t(
  X_reml)%*%solve(V_reml)
# elements of the fisher information matrix:
> fisher11reml=1/2*sum(diag(P%*%diag(1,62)%*%P%*%diag(1,62)))
> fisher12reml=1/2*sum(diag(P%*%diag(1,62)%*%P%*%Z_reml%*%t(Z_reml)))
> fisher21reml=1/2*sum(diag(P%*%Z_reml%*%t(Z_reml)%*%P%*%diag(1,62)))
> all.equal(fisher12reml,fisher21reml)
[1] TRUE
> fisher22reml=1/2*sum(diag(P%*%Z_reml%*%t(Z_reml)%*%P%*%Z_reml%*%t(Z_reml)))
> fisherreml=matrix(c(fisher11reml,fisher12reml,fisher21reml,fisher22reml),2)
> solve(fisherreml)
           [,1]       [,2]
[1,]  0.4561320 -0.1831565
[2,] -0.1831565  0.4492569
```

From this result, we have that the standard errors of $\hat{\sigma}_e^2$ and $\hat{\sigma}_s^2$, respectively, are the square roots of the diagonal elements of the above matrix: 0.6754 and 0.6703.

4.2 Parametric Bootstrap (REML)

For the bootstrap estimation, I use the same method as in Section 3.2 except with the REML model in place of the ML model. Here is the code:

```
> booted_lamb_reml = bootMer(lamb_reml, mySumm, nsim = 100)
> summary(booted_lamb_reml)

Number of bootstrap replications R = 100
                original bootBias  bootSE bootMed
sigma_e_sq       2.99593 -0.10504 0.64573 2.77839
sigma_s_sq.Sire  0.51136  0.47274 0.82670 0.81907
```

From the boostrap results, we have that the standard errors of $\hat{\sigma}_e^2$ and $\hat{\sigma}_s^2$, respectively, are 0.6457 and 0.8267.

5 Discussion

The ML and REML methods produce similar results for the estimation of the fixed effects. In both cases, all line effects are significantly different from zero (with line 2 having the highest estimate) while the age effects are not significant. These results indicate that average lamb birth weight may differ across lines, while age category of the mother is not a relevant predictor. In the model fit using ML, the estimate for the variance component of the sire random effect is exactly zero. This indicates that ML may not be the best choice for model-fitting in this situation. In the model fit using REML, there is no such problem.

The estimates of the standard errors of the variance components estimates using the asympototic covariance matrix method and the parametric bootstrap method are quite similar, especially for $\hat{\sigma}_e^2$. It is understandable that there is some different in the results from the two methods since the accuracy of the asymptotic covariance matrix relies on certain assumptions (such as the sample size) and the bootstrap method is inherently random. The sample size of 62 in the data is decent, though not extremely large. Additionally, the number of simulations used in the bootstrap method ($B = 100$) is not particularly large. Running the bootstrap simulations several times produces different results every time; for better stability, one could increase B to 1000 or even 10000.

Since the age effects were not significant, I thought would be interesting to fit the model without the age variables and compare them. The results are as follows:

```
> lamb_reml_reduced=lmer(Weight ~ Line - 1 + (1 | Sire),lamb)
> AIC(lamb_reml_reduced)
[1] 254.4261
> AIC(lamb_reml)
[1] 256.8849
> BIC(lamb_reml_reduced)
[1] 269.3161
> BIC(lamb_reml)
[1] 276.0291
```

Using either the AIC or BIC criterion for model selection, the reduced model (without the age variables) appears to be a better fit. Out of curiosity, I also fit a fixed effects model without the sire effect:

```
> lamb_reml_reallyReduced=lm(Weight~Line-1,lamb)
> AIC(lamb_reml_reallyReduced)
[1] 255.4761
> BIC(lamb_reml_reallyReduced)
[1] 268.239
```

By the BIC criterion, this model is actually even better. Furthermore, fitting a model with only line as a predictor shows that none of the line variables are significant in the presence of the others (i.e. all p-values above 0.05, although the line 2 variable has p-value 0.0641). It seems that none of the variables sire, line, or mother's age are extremely significant in predicting lamb's birth weight, though sire and line are borderline significant based on these analyses.