First project of STA232B:Analyzing the Lamb's Weight Data

Yi Han

Department of Statistics, UC Davis

January 27th, 2020

Abstract

Lamb's weight data was first recorded by G.E.Bradford. And the data shows weights of different lambs and the feature that could influence weights such as lines' type, dams' age and etc. Harville and Fenech presented a method for constructing an exact confidence interval for the ratio of two variance components in a linear mixed model. In this project, I build a linear mixed model to depict such relationship because the data are not independent with each other. Maximum Likelihood(MLE) and Restricted Maximum Likelihood(RMLE) are used to evaluate parameters. I also use asymptotic covariance matrix and parametric bootstrap method to evaluate the standard error of variance components.

1 introduction

Linear mixed models are models that contains both fixed effects and one or more random effects. Each of the data sets, along with the set of residual effects, is auusmed to have a different common variance, known as a variance component[1]. Harville and Fenech[2] presented a method for constructing an exact confidence interval for the ratio of two variance components which may have very important biological and agricultural applications, for example, under certain conditions, the heritability can be expressed as a strictly increasing function of a variance components ratio, in a linear mixed model.

The data consists of 62-single birth male lambs' birth weight. Lambs came from five different lines(two control lines and three selection lines). Each Lamb was the offspring of one of 23 rams(father lambs), and each lamb has a different dam(mother lambs). The age of their dam(mother lamb) is one of the three age categories: 1-2 years (category 1), 2-3 years(category 2) and older than 3 years(category 3). The following linear mixed model was proposed:

$$y_{ijkd} = u + l_i + \pi_j + s_{ik} + e_{ijkd}$$

where the $l_i, i = 1, \ldots, 5$ represents the line effects and $\pi_j, j = 1, \ldots, 3$ represents the age effects, which are fixed effects. the sire(father lambs) effects(within line) $(s_{11}, s_{12}, \ldots, s_{58})$ are random effects that are distributed independently as $N(0, \sigma_s^2)$. The random errors $e_{1111}, e_{1121}, ..., e_{3582}$ are distributed as $N(0, \sigma_e^2)$ independently of each other and of the sire effects.

But for the later analyzing convenience, the model can be changed as the following one without intercept[2]:

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

The line effects is denoted by $l_i, i = 1, \ldots, 5$ and it is fixed. $x_{ijk,2} = 1$ if the age of the kth dam(mother lamb)$(k = 1, \ldots, n_{ij})$ 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 kth dam(mother lamb) corresponding to line i and sire j is in category 3, and $x_{ijk,2} = 0$ otherwise. a_1 and a_2 are fixed effects corresponding to these indicator variables of dams'(mother lambs) age. $s_{ij}(j = 1, \ldots, n; n_1 = n_2 = n_3 = 4, n_4 = 3, n_5 = 8)$ are the random sire(ram) effects nested within lines and are assumed to be $i.i.d \sim N(0, \sigma_s^2)$. And e_{ijk} are the random errors that distributed as $N(0, \sigma_e^2)$ independently of each other and of the sire effects.

These two formulas can all be written as

$$Y = X\beta + Z\alpha + \epsilon$$

where X, Z represents the design matrix of fixed effects and random effects. $\beta and \alpha$ represents the fixed effects and random effects and ϵ represents random errors. The difference between the two formulas is that they have different X and β, but their random effects α, Z and variance component remain the same. So, for convenience, we use the second formula in the following report.

2 Data Entry

The first 6 rows of the data is listed below for references:

Table 1: Part of Lambs' weight data

	line	sire	damage	weight
1	1	1	1	6.2
2	1	2	1	13.0
3	1	3	1	9.5
4	1	3	1	10.1
5	1	3	1	11.4
6	1	3	2	11.8

From the table above, we can observe that the "line", "sire" and "damage" variables are all categorical variable, which means they should be coded as factor.

Here we can going further regrading to the above model we choose, it can be written in the way:

$$Y = X\beta + Z\alpha + \epsilon$$

where $\alpha \sim N(0, \sigma_s^2 I_m)$, m=23 represents the number of sire. $\epsilon \sim N(0, \sigma_e^2 I_n)$, n=62 represents the number of lambs recorded. So we know that:

$$Y \sim N(X\beta, V)$$

$$V = Zvar(\alpha)Z' + R = \sigma_s^2 ZZ' + \sigma_e^2 In$$

So, we can use $\theta = (\sigma_s^2, \sigma_e^2)$ to denote the variance of components.

3 Analysis using Maximum Likelihood

Here, I first use Maximum likelihood method to get the estimates of the model parameters as well as the standard errors of the variance component estimators. To estimate the standard errors of the variance component estimators, two method are used and the results are compared.
According to the result shown in R, the estimator of fixed effects parameters are:

Table 2: Estimates of fixed effects(ML)

Fixed Effects						
line1	line2	line3	line4	line5	damage2	damage3
10.69789	12.28864	10.89967	10.18697	10.95237	-0.07205	0.03258

The estimator of random effects' Intercepts are(Table 3):

The MLE estimates of the two variance components $\hat{\sigma_s^2}, \hat{\sigma_e^2}$ have standrad error:9.234×10^{-08} and 1.716 .Interestingly, the estimate of the random effects' intercepts and the variance component for the sire random effect are all very close to zero.

Table 3: Estimates of random effects(ML)

Ramdom Effects									
Sire's Intercept(part)									
-1.30e-14	6.66e-15	1.33e-14	-7.01e-15	4.11e-15	...	7.58e-15	-1.46e-14	1.58e-14	-4.13e-15

Since the output doesnt 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. ***Using the asymptotic covariance matrix***

According to Jiang(2007)[2], the asymptotic covariance matrix equals the inverse of Fisher Information. The following formula exists:

$$E(\frac{\partial^2 l}{\partial\beta\partial\beta'}) = -X'V^{-1}X, \tag{1}$$

$$E(\frac{\partial^2 l}{\partial\beta\partial\theta_r}) = 0, \quad 1 \le r \le q \tag{2}$$

$$E(\frac{\partial^2 l}{\partial\theta_r\partial\theta_s}) = -\frac{1}{2}tr(V^{-1}\frac{\partial V}{\partial\theta_r}V^{-1}\frac{\partial V}{\partial\theta_s}), \quad 1 \le r,s \le q. \tag{3}$$

where q is the elements number of variance components. In this model, $q = 2$. So, according to the equation(2), it is noted that the fisher information matrix denoted as $I(l)$ can be expressed as a block matrix with all the off-diagonal elements are 0. So the asymptotic covariance of θ can be obtained only by considering the inverse of the bottom of fisher matrix denoted as $I^*(l)$, i.e, $-E(\frac{\partial^2 l}{\partial\theta_r\partial\theta_s}), 1 \le r,s \le q$ part. And as shown in first chapter, there exists:

$$\frac{\partial V}{\partial\theta_1} = \frac{\partial V}{\partial\sigma_s^2} = ZZ'$$

$$\frac{\partial V}{\partial\theta_2} = \frac{\partial V}{\partial\sigma_e^2} = I_n$$

where $n = dim(Y) = 62$. Plug these two equation to the above equation(3), we can get the bottom fisher information matrix $I^*(l)$ we need. The diagonal elements of $I^*(l)^{-1}$ are the asymptotic variance of $\theta = (\sigma_s^2, \sigma_e^2)$, respectively. Calculate the lamb data in R, the result shows: the standard error of $\hat{\sigma_s}, \hat{\sigma_e}$ are 0.2944062 and 0.6052039.

3.2. ***Using the Bootstrap method***

The main idea of parametrical bootstrap method in estimating the standard errors of variance components estimators is replacing the distribution of Y, denoted as F by $\hat{F}$. $\hat{F}$ is the estimator of F, which is produced by substituting the parameters in F by their maximum likelihood estimation. Then bootstrap sample can be produced by $\hat{F}$. Let $\theta^{(b)}$ be the $b-th$ ML estimate of θ based on the $b-th$ bootstrap sample, then we can approximate the standard error of $\{\theta^{(b)}, b = 1, \ldots, B\}$.

By the function *bootMer* in R, we can get the result: the standard error of $\hat{\sigma_s}^2, \hat{\sigma_e}^2$ are 2.707×10^{-16} and 0.468 if we set $B = 100$.

4 Analysis using Restricted Maximum Likelihood

In this method, the above model is fitted using the *lmer* function by setting the option $REML = TRUE$ in *lme*4 package of R. In this method, we get the standard error of estimation of two variance components $\hat{\sigma_s^2}, \hat{\sigma_e^2}$ have value 0.7191 and 1.7209. The estimation of fixed effects are showing below:

The estimates of random effects is partly showing below:

Different from what we get when using maximum likelihood, the estimate of the random effects' intercepts and the variance component for the sire random effect are all relatively larger and not close to zero.

Table 4: Estimates of fixed Effects(REML)

Fixed Effects						
line1	line2	line3	line4	line5	damage2	damage3
10.48907	12.28554	11.07547	10.27414	10.95082	-0.16967	0.01959

Table 5: Estimates of random effects(REML)

Ramdom Effects									
Sire's Intercept(part)									
-0.63753	0.37322	0.51127	-0.24697	0.17761	...	0.40413	-0.64076	0.58563	-0.12988

Here we also use the same methods as chapter 3 to estimate the standard errors of variance components.

4.1. ***Using the asymptotic covariance matrix***

Under some suitable conditions, REML is consistent and asymptotically normal with the asymptotic covariance matrix, which equals to the inverse of the restricted Fisher Information Matrix, calculated as:

$$I(\theta) = -E(\frac{\partial^2 l_R}{\partial\theta\partial\theta'}) \tag{4}$$

According to Jiang(2007)[2], formula(4) can be later expressed in a clearer formula:

$$-E(\frac{\partial^2 l_R}{\partial\theta_r\partial\theta_s}) = \frac{1}{2}tr(P\frac{\partial V}{\partial\theta_r}P\frac{\partial V}{\partial\theta_s}), \qquad 1 \le r, s \le q. \tag{5}$$

where

$$P = V^{-1} - V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1} \tag{6}$$

And like ML method, we also have:

$$\frac{\partial V}{\partial\theta_1} = \frac{\partial V}{\partial\sigma_s^2} = ZZ', \qquad \frac{\partial V}{\partial\theta_2} = \frac{\partial V}{\partial\sigma_e^2} = I_n \tag{7}$$

where $n = dim(Y) = 62$

So, after implementing the code in R, the result shows that the standard error of $\hat{\sigma_s}$ and $\hat{\sigma_e}$ are 0.6670731 and 0.6678501

4.2. ***Using the bootstarp method(REML)***

The bootstrap method for REML is the same as ML, the only different thing is that we replace θ with $\hat{\theta}$, which is REML estimator of θ instead of ML estimator. Using R function $bootMer$, we can conveniently get the standard error of $\hat{\sigma_s}^2$ and $\hat{\sigma_e}^2$ are 0.727 and 0.605.

5 Discussion

The comparison between the results of these two methods are listing below(table 6):

From the table below, we can find some interesting conclusions. Firstly, when using ML method the standard error of σ_s is very close to zero while using REML method, the standard error of σ_s is not close zero. The reason for such phenomenon may be when applying MLE method, if the computed value of an MLE is negetive, then R will put it as zero automatically. When applying bootstrap method for the MLE method, it may due to the same reason that the estimator of standard error of σ_s is also close to zero. But when using REML, this problem can be avoided.

Secondly, from both results of MLE and REML, the estimators of dam age effects are very close to zero which suggests that this fixed effect may be not significant. In fact, when we build another model

Table 6: Comparison the results of ML and REML

Estimation Methods	Estimator	Point Estimation	Asymptotic Covariance Matrix	Bootstrap
MLE	$\text{sd}(\hat{\sigma_s^2})$	9.234×10^8	0.294	2.037488×10^{-08}
	$\text{sd}(\hat{\sigma_e^2})$	1.716	0.605	0.1432732
REML	$\text{sd}(\hat{\sigma_s^2})$	0.719	0.667	0.482
	$\text{sd}(\hat{\sigma_e^2})$	1.721	0.668	0.177

without this fixed effect, and do a comparison between new model and the original one. The results shows below:

Figure 1: The Comparison between two models

```
Data: lamb
Models:
lamb_reml_noage: weight ~ line - 1 + (1 | sire)
lamb_mle: weight ~ line + damage - 1 + (1 | sire)
                Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
lamb_reml_noage  7 256.92 271.81 -121.46   242.92
lamb_mle         9 260.89 280.04 -121.45   242.90 0.0287      2     0.9857
```

From the above table, we can easily observe that the value of AIC and BIC are both smaller in new model which deletes the age effects, suggesting deleting this effect can give us an easier and better model.

Thirdly, the asymptotic standard error of variance components estimators are very different between MLE and REML method. The reason may be that when applying Asymptomatic normality, it is required that $\frac{p}{\sqrt{n}} \to 0$. But in our data, $\frac{p}{\sqrt{n}} = 0.89$ which doesn't satisfied the prerequisite. So maybe bootstrap is a better option here. For bootstrap method, we can increase the number of B to 1000 or even 10000 in order to get more accurate results.

References

[1] Harville D A, Fenech A P. Confidence intervals for a variance ratio, or for heritability, in an unbalanced mixed linear model[J]. *Biometrics*, 1985, 41(1): 137-152.

[2] Jiang, Jiming. *Linear and generalized linear mixed models and their applications*. Springer Science Business Media, 2007.