232B Project 1

Yan-Yu Chen

December 28, 2024

1 Linear Mixture Model

In this project, we demonstrate the liner mixture model (LLM) using a dataset of birth weights of lambs from Harville and Fenech (1985). The observations y consist of birth weight of $n = 62$ lambs from 23 rams with the ages of dam and the five distinct population lines together being recorded as covariates. The age of the dam is a categorical variable with three categories numbered 1 (1-2 years), 2 (2-3 years), and 3 (over 3 years). Let $x_{ijk,s} = 1$ if the age of the kth dam corresponding to line i and sire j is in category s, and $x_{ijk,s} = 0$ otherwise. Another fixed effects are the line effects, denoted by μ_i, $i = 1, \ldots, 5$. In the model, we consider the sire effects as random effects, denoted by α_{ij}'s for $i = 1, \ldots, 5$, $j = 1, \ldots, n_i$, with $n_1 = n_2 = n_3 = 4$, $n_4 = 3$, and $n_5 = 8$, where $\alpha_{ij} \overset{iid}{\sim} N(0, \sigma_\alpha^2)$. Finally, we denote the error terms as ϵ_{ijk}'s, $i = 1, \ldots, 5$, $j = 1, \ldots, n_i$, and $k = 1, \ldots, n_{ij}$, where $\epsilon_{ijk} \overset{iid}{\sim} N(0, \sigma_\epsilon^2)$, and n_{ij} is the number of measures in the (i,j) cell. In the meanwhile, we assume α_{ij}'s and ϵ_{ijk}'s are mutually independent.

Under the previous assumptions, we consider the model, $y_{ijk} = \mu_i + a_1 x_{ijk,1} + a_2 x_{ijk,2} + \alpha_{ij} + \epsilon_{ijk}$. In general, we have

$$\underbrace{\begin{bmatrix} y_{111} \\ y_{121} \\ \vdots \\ y_{585} \end{bmatrix}}_{y \in \mathbb{R}^{62}} = \underbrace{\begin{bmatrix} 1 & 0 & \cdots & 0 & 1 & 0 \\ \vdots & \ddots & & & & \vdots \\ 1 & 0 & \cdots & 0 & 0 & 1 \\ \vdots & & & & \ddots & \vdots \\ 0 & 0 & \cdots & 1 & 0 & 0 \end{bmatrix}}_{X \in \mathbb{R}^{62\times 7}} \underbrace{\begin{bmatrix} \mu_1 \\ \vdots \\ \mu_5 \\ a_1 \\ a_2 \end{bmatrix}}_{\beta \in \mathbb{R}^7} + \underbrace{\begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}}_{Z \in \mathbb{R}^{62\times 23}} \underbrace{\begin{bmatrix} \alpha_{11} \\ \alpha_{12} \\ \vdots \\ \alpha_{58} \end{bmatrix}}_{\alpha \in \mathbb{R}^{23}} + \underbrace{\begin{bmatrix} \epsilon_{111} \\ \epsilon_{121} \\ \vdots \\ \epsilon_{585} \end{bmatrix}}_{\epsilon \in \mathbb{R}^{62}},$$

where $\alpha \sim N_{23}(0, \sigma_\alpha^2 I_{23})$ and $\epsilon \sim N_{62}(0, \sigma_\epsilon^2 I_{62})$ are mutually independent. Lately, we denote the parameters of the components of the marginal covariance matrix as $\theta = (\sigma_\alpha^2, \sigma_\epsilon^2)^t$, and the parameter of interests in this model as $\psi = (\beta, \theta^t)^t$. In the next section, we presented our analysis of ψ using maximum likelihood estimation (MLE) methods and restricted maximum likelihood estimation (RMLE) methods.

2 Analysis

2.1 Maximum Likelihood Estimation

We implemented the model using package `lm4` in R. After doing the analysis of MLE, we estimated the coefficients β in Table 1. We also obtained the estimates $\hat{\sigma}_\alpha^2 = 0$, and $\hat{\sigma}_\epsilon^2 = 2.971$ from our analysis. To further investigate θ, we presented two fashions to analyze the dispersion of our estimate $\hat{\theta}$.

1. The asymptotic covariance matrix (p. 11 of Jiang 2007);

   Let ℓ be the log-likelihood of y. The marginal variance of y in our model is $V = \sigma_\epsilon^2 I_{62} + \sigma_\alpha^2 ZZ^t$. By the theorem of MLE, we derived the asymptotic covariance matrix, $Var(\hat{\theta}) = -[(Var(\partial\ell/\partial\psi)^{-1}]_{8:9,8:9}$, where

$$Var\left(\frac{\partial \ell}{\partial \psi}\right) = \begin{bmatrix} X^t V^{-1} X & 0 & 0 \\ 0 & \frac{1}{2} tr(V^{-1} ZZ^t V^{-1} ZZ^t) & \frac{1}{2} tr(V^{-2} ZZ^t) \\ 0 & \frac{1}{2} tr(V^{-2} ZZ^t) & \frac{1}{2} tr(V^{-2}) \end{bmatrix}.$$

	Estimate	Std. Error	t value
Line1	10.72010	0.65981	16.247
Line2	12.31608	0.63314	19.452
Line3	10.89224	0.48753	22.342
Line4	10.21310	0.62860	16.247
Line5	10.97742	0.46298	23.710
Age1	-0.02184	0.52187	-0.042
Age2	-0.08499	0.62123	-0.137

Table 1: MLE of β

Let $\widehat{Var}_a(\hat{\theta})$ be the asymptotic covariance matrix of $\hat{\theta}$. Based on the expression, our implement showed that

$$\widehat{Var}_a(\hat{\theta}) = \begin{bmatrix} 0.088 & -0.088 \\ -0.088 & 0.372 \end{bmatrix}.$$

The standard errors of $\hat{\sigma}^2_\alpha$ and $\hat{\sigma}^2_\epsilon$ from the asymptotic covariance matrix are 0.297 and 0.610, respectively.

2. The bootstrap method with $B = 100$ being the number of bootstrap samples.

   Let $\widehat{Var}_b(\hat{\theta})$ be the bootstrapped covariance matrix of $\hat{\theta}$. For the reproducible purpose, we predetermined the random seed being `seed = 2022`. Using the parametric bootstrap method, we obtained

$$\widehat{Var}_b(\hat{\theta}) = \begin{bmatrix} 0.006 & -0.000 \\ -0.000 & 0.274 \end{bmatrix}.$$

   The standard errors of $\hat{\sigma}^2_\alpha$ and $\hat{\sigma}^2_\epsilon$ from the bootstrap method are 0.077 and 0.524, respectively.

2.2 Restricted Maximum Likelihood Estimation

Analogously, we conducted our analysis through `lm4`. The estimates $\tilde{\beta}$ are indicated in Table 2. We also obtained the estimates of θ, $\tilde{\sigma}^2_\alpha = 0.511$ and $\tilde{\sigma}^2_\epsilon = 2.995$. To complete our investigation of $\tilde{\theta}$, we performed the two methods again with a slight modification.

	Estimate	Std. Error	t value
Line1	10.500799	0.807022	13.012
Line2	12.299933	0.756868	16.251
Line3	11.042510	0.656169	16.829
Line4	10.286381	0.788240	13.050
Line5	10.962486	0.543773	20.160
Age1	-0.009646	0.548103	-0.018
Age2	-0.165080	0.643456	-0.257

Table 2: RMLE of β

1. The asymptotic covariance matrix;

   With a modification by replacing V^{-1} in $Var(\partial\ell/\partial\theta)$ by P, where

$$P = V^{-1} - V^{-1}X(X^tV^{-1}X)^{-1}X^tV^{-1},$$

   we obtained

$$\widehat{Var}_a(\tilde{\theta}) = \begin{bmatrix} 0.449 & -0.183 \\ -0.183 & 0.456 \end{bmatrix}.$$

   The standard errors of $\tilde{\sigma}^2_\alpha$ and $\tilde{\sigma}^2_\epsilon$ from the asymptotic covariance matrix are 0.670 and 0.675, respectively.

2. The bootstrap method with $B = 100$.

   Again, for the reproducible purpose, we predetermined the random seed being `seed = 2022`. The bootstrapped results suggested

$$\widehat{Var}_b(\tilde{\theta}) = \begin{bmatrix} 0.549 & -0.036 \\ -0.036 & 0.345 \end{bmatrix}.$$

   The standard errors of $\tilde{\sigma}_\alpha^2$ and $\tilde{\sigma}_\epsilon^2$ from the bootstrap method are 0.741 and 0.587, respectively.

3 Discussion

We found out that the estimates of β and σ_ϵ^2 are pretty similar no matter we used MLE or RMLE. However, the estimate of σ_α^2 is significantly different. The results from MLE suggested that α should degenerate to a point mass rather than be normally distributed. This meant there is no random effect from the sire effect to the weight of lambs. By contrast, the results from RMLE allowed the presence of the random effects in the model.

Another departure in our analyzes is their asymptotic behavior. The bootstrapped covariance matrix in MLE did not seem to converge to the asymptotic covariance matrix in MLE, whereas the bootstrapped covariance matrix in RMLE was closed to the asymptotic covariance matrix in RMLE. In fact, the RMLE remained consistent and asymptotically normal, while the MLE may fail to be consistent or asymptotically normal (Jiang, 1996). The huge difference between the standard errors of σ_α^2 from asymptotic covariance matrix and the bootstrapped method suggested that $\hat{\theta}$ failed to be asymptotically normally distributed in the weight of lambs dataset.

In short, because RMLE gave a similar result of MLE with much more flexibility, and its estimators behaved asymptotically normal, we suggested using RMLE for LLM to investigate this dataset.

Appendix

```
library(lme4)
lamb <- read.csv("lamb.csv")
cols <- c("Sire", "Line", "Age")
lamb[cols] <- lapply(lamb[cols], as.factor)
lamb <- within(lamb, Age <- relevel(Age, ref = 3))

# MLE
fit.mle <- lmer(Weight ~ Line + Age - 1 + (1|Sire), data = lamb, REML = F)
summary(fit.mle)

# asymptotic covariance
Z <- model.matrix(lm(Weight ~ Sire - 1, data = lamb))
asym.var <- function(obj, Z, REML = T){
  V <- sigma(obj)^2 * diag(nrow(Z)) + unlist(VarCorr(obj)) * Z %*% t(Z)
  V.m <- solve(V)
  X <- model.matrix(obj)
  S1 <- t(X) %*% V.m %*% X
  if (REML){
    P <- V.m - V.m %*% X %*% solve(t(X) %*% V.m %*% X) %*% t(X) %*% V.m
    S2 <- 0.5 * matrix(c(sum(diag(P %*% Z %*% t(Z) %*% P %*% Z %*% t(Z))),
                         rep(sum(diag(P %*% P %*% Z %*% t(Z))), 2),
                         sum(diag(P %*% P))), ncol = 2)
  } else{
    S2 <- 0.5 * matrix(c(sum(diag(V.m %*% Z %*% t(Z) %*% V.m %*% Z %*% t(Z))),
                         rep(sum(diag(V.m %*% V.m %*% Z %*% t(Z))), 2),
                         sum(diag(V.m %*% V.m))), ncol = 2)
  }
```

```
  O <- matrix(rep(0, ncol(S1)*2), ncol = 2)
  colnames(O) <- c("sigma_a^2", "sigma_e^2")
  S <- rbind(cbind(S1, O), cbind(t(O), S2))
  return(solve(S))
}
asym.var(fit.mle, Z, REML = F)[8:9,8:9]
sqrt(diag(asym.var(fit.mle, Z, REML = F)[8:9,8:9]))

# bootstrap
library(boot)
mySumm <- function(mod) {
  c("sigma_a^2" = unlist(VarCorr(mod)), "sigma_e^2" = sigma(mod)^2)
}
fit.boot <- bootMer(fit.mle, mySumm, nsim = 100, seed = 2022)
cov(fit.boot$t)
sqrt(diag(cov(fit.boot$t)))

# RMLE
fit.rmle <- lmer(Weight ~ Line + Age - 1 + (1|Sire), data = lamb)
summary(fit.rmle)

# asymptotic covariance
asym.var(fit.rmle, Z)[8:9,8:9]
sqrt(diag(asym.var(fit.rmle, Z)[8:9,8:9]))

# bootstrap
fit.boot2 <- bootMer(fit.rmle, mySumm, nsim = 100, seed = 2022)
cov(fit.boot2$t)
sqrt(diag(cov(fit.boot2$t)))
```