OVERVIEW OF LEAST-SQUARES, MAXIMUM LIKELIHOOD AND BLUP PART 2

"ALL MODELS ARE WRONG, BUT SOME ARE USEFUL"

(Box, 1976)

Carl Friedrich Gauss (1777, Brunswick, Germany-1855, Göttingen, Germany)

LIKELIHOOD-BASED INFERENCE

1. The likelihood function

$\mathbf{y}$ = observed data vector

$\mathbf{u}$ = latent variables, unobserved random effects or "missing" data

$\boldsymbol{\theta}$ = unknown parameter vector

$$L(\boldsymbol{\theta}|\mathbf{y}) \propto p(\mathbf{y}|\theta) = \int p(\mathbf{y}|\mathbf{u},\boldsymbol{\theta})p(\mathbf{u}|\boldsymbol{\theta})d\mathbf{u}$$

Marginal density of observations viewed as function of $\boldsymbol{\theta}$

2. The Maximum Likelihood Estimator

$$\hat{\boldsymbol{\theta}} = Arg \max_{\boldsymbol{\theta}} L(\boldsymbol{\theta}|\mathbf{y})$$

$$\hat{\boldsymbol{\theta}} = Arg \max_{\boldsymbol{\theta}} l(\boldsymbol{\theta}|\mathbf{y})$$

Likelihood
Log-likelihood

Value of parameter that makes observed sample “most probable”

$$p\left(\mathbf{y}|\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}\right) \geqq p(\mathbf{y}|\boldsymbol{\theta}) \quad \vee \boldsymbol{\theta}$$

Some difficulties:

1. May not be easy to postulate a distribution for the data (more on this later)

2. $$L(\boldsymbol{\theta}|\mathbf{y}) \propto p(\mathbf{y}|\theta) = \int p(\mathbf{y}|\mathbf{u},\boldsymbol{\theta})p(\mathbf{u}|\boldsymbol{\theta})d\mathbf{u}$$

   After integration of random effects (if possible), latent variables or missing data, likelihood may not be in closed form

3. Difficulty in locating a maximum (flat or complicated likelihoods)

   $$\frac{\partial^2 l(\boldsymbol{\theta}|\mathbf{y})}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}'}$$

   Must be negative-definite (negative eigenvalues)

4. Does not provide guidance on prediction of random effects

   Infer $[\mathbf{u}|\mathbf{y},\boldsymbol{\theta}]$ via $[\mathbf{u}|\mathbf{y},\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}]$??

3. Information (Fisher's) contained in the likelihood

Multi-parameter situation: Expected information matrix

$$\mathbf{I}(\boldsymbol{\theta}) = E_{\mathbf{y}}\left[\left(\frac{\partial l}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial l}{\partial \boldsymbol{\theta}}\right)'\right] = -E_{\mathbf{y}}\left(\frac{\partial^2 l}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'}\right)$$

Multi-parameter situation: Observed information matrix

$$\mathbf{J}(\boldsymbol{\theta}, \mathbf{y}) = -\frac{\partial^2 l}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'}$$

Asymptotic variance-covariance matrix of estimates

$$Var(\hat{\theta}) = \boldsymbol{I}^{-1}(\theta)$$

4. Distribution of score vector (gradient)

$$S(\boldsymbol{\theta}|\mathbf{y}) = \frac{\partial l}{\partial \boldsymbol{\theta}}$$

$$S(\boldsymbol{\theta}|\mathbf{y}) \sim [\mathbf{0}, \mathbf{I}(\boldsymbol{\theta})]$$

- Corollary: If data are normal, and score is linear function of data ➔score is normal
- Asymptotic tests can be used for non-normal situations: score test

$$H_0 : \boldsymbol{\theta}_{p\times 1} = \boldsymbol{\theta}_0$$

$$U(\mathbf{y};\boldsymbol{\theta}_0) = \mathbf{S}(\boldsymbol{\theta}|\mathbf{y})'\mathbf{I}^{-1}(\boldsymbol{\theta}|\mathbf{y})\mathbf{S}(\boldsymbol{\theta}|\mathbf{y}) \sim \chi^2_p$$

Why?

$$E[U(\mathbf{y};\boldsymbol{\theta}_0)|H_0] = tr[\mathbf{I}^{-1}(\boldsymbol{\theta}|\mathbf{y})\mathbf{I}(\boldsymbol{\theta}|\mathbf{y})] = p$$

if information matrix non-singular

(identification)

5. Asymptotic Properties of Maximum Likelihood Estimators

- **Asymptotic normality** $\hat{\boldsymbol{\theta}}_n \sim N[\boldsymbol{\theta}_0, \mathbf{I}^{-1}(\boldsymbol{\theta}_0)]$

- **Asymptotic unbiasedness** $E(\hat{\boldsymbol{\theta}}) \to \boldsymbol{\theta}$

- **Consistency**: converges to true parameter value when sample size is infinite

- **Efficiency:** reaches (asymptotically) Cramer-Rao lower bound for the variance of an unbiased estimator

$$\mathbf{I}^{-1}(\boldsymbol{\theta})$$

Additional finite sample properties →

- Functional invariance
- Sufficiency

Matrix representation of the mixed linear model

assumptions

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e} \\ \mathbf{u} &\sim N(\mathbf{0}, \mathbf{G}(\boldsymbol{\theta})) \\ \mathbf{e} &\sim N(\mathbf{0}, \mathbf{R}(\boldsymbol{\theta})) \\ Cov[\mathbf{u}, \mathbf{e}'] &= \mathbf{0} \\ \boldsymbol{\theta} &= \text{variance or covariance components} \end{aligned}$$

mean vector

$$E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$$

covariance matrix

$$\begin{aligned} Var(\mathbf{y}) &= Var(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}) = Var(\mathbf{Z}\mathbf{u} + \mathbf{e}) \\ &= \mathbf{Z}Var(\mathbf{u})\mathbf{Z}' + Var(\mathbf{e}) \\ &= \mathbf{Z}\mathbf{G}(\boldsymbol{\theta})\mathbf{Z}' + \mathbf{R}(\boldsymbol{\theta}) = \mathbf{V} \end{aligned}$$

Covariance between **u** and **y**

$$\begin{aligned} Cov(\mathbf{u}, \mathbf{y}') &= Cov\left(\mathbf{u}, (\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e})'\right) \\ &= Cov(\mathbf{u}, \mathbf{u}'\mathbf{Z}') = Var(\mathbf{u})\mathbf{Z}' \\ &= \mathbf{G}(\boldsymbol{\theta})\mathbf{Z}' = \mathbf{C} \end{aligned}$$

Conditional expectation of **u** given **y** (“best predictor” and “best linear predictor”)

$$\begin{aligned} E(\mathbf{u}|\mathbf{y}) &= E(\mathbf{u}) + Cov(\mathbf{u}, \mathbf{y}')\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \mathbf{C}\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \end{aligned}$$

Constructing the likelihood function in matrix form

$$\mathbf{y} \sim N\left(\mathbf{X}\boldsymbol{\beta}, \mathbf{V} = \mathbf{Z}\mathbf{G}(\boldsymbol{\theta})\mathbf{Z}' + \mathbf{R}(\boldsymbol{\theta})\right)$$

Multivariate normal density

$$p(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\theta}) = \frac{1}{(2\pi)^{\frac{n}{2}} |\mathbf{V}|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right]$$

$$l(\boldsymbol{\beta}, \boldsymbol{\theta}; \mathbf{y}) \propto |\mathbf{V}|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right]$$

Likelihood function

Maximize log-likelihood via some iterative algorithm
(solutions to first-order condition are not explicit)

Matrix representation of likelihood in a mixed linear model
(two uncorrelated sets of random effects assumed here)

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_s\mathbf{u}_s + \mathbf{Z}_d\mathbf{u}_d + \mathbf{e}$$

$$\begin{bmatrix} \mathbf{u}_s \\ \mathbf{u}_d \\ \mathbf{e} \\ \mathbf{y} \end{bmatrix} \sim N\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{X}\boldsymbol{\beta} \end{bmatrix}, \begin{bmatrix} \mathbf{I}_{N_S}\sigma_s^2 & 0 & 0 & \mathbf{Z}_s'\sigma_s^2 \\ 0 & \mathbf{I}_{N_D}\sigma_d^2 & 0 & \mathbf{Z}_d'\sigma_d^2 \\ 0 & 0 & \mathbf{I}_N\sigma_e^2 & \mathbf{I}_N\sigma_e^2 \\ \mathbf{Z}_s\sigma_s^2 & \mathbf{Z}_d\sigma_d^2 & \mathbf{I}_N\sigma_e^2 & \mathbf{Z}_s\mathbf{Z}_s' + \mathbf{Z}_\mathbf{d}\mathbf{Z}_\mathbf{d}' + \mathbf{I}_N\sigma_e^2 \end{bmatrix} \right)$$

$$p(\mathbf{u}_s, \mathbf{u}_d, \mathbf{y}) = p(\mathbf{u}_s|\sigma_s^2)\,p(\mathbf{u}_d|\sigma_d^2)\,p(\mathbf{y}|\boldsymbol{\beta}, \mathbf{u}_s, \mathbf{u}_d, \sigma_e^2)$$

Joint density of pertinent random variables

$$p(\mathbf{u}_s,\mathbf{u}_d,\mathbf{y}) = \frac{1}{(2\pi)^{\frac{N_S}{2}}|\mathbf{I}_{N_S}\sigma_s^2|^{\frac{1}{2}}}\exp\left(-\frac{1}{2\sigma_s^2}\mathbf{u}_s'\mathbf{u}_s\right)\times\frac{1}{(2\pi)^{\frac{N_D}{2}}|\mathbf{I}_{N_D}\sigma_d^2|^{\frac{1}{2}}}\exp\left(-\frac{1}{2\sigma_d^2}\mathbf{u}_d'\mathbf{u}_d\right)$$

$$\times\frac{1}{(2\pi)^{\frac{N}{2}}|\mathbf{I}_N\sigma_e^2|^{\frac{1}{2}}}\exp\left[-\frac{1}{2\sigma_e^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}_s\mathbf{u}_s-\mathbf{Z}_d\mathbf{u}_d)'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}_s\mathbf{u}_s-\mathbf{Z}_d\mathbf{u}_d)\right]$$

$$p(\mathbf{y}) = \int\int p(\mathbf{u}_s,\mathbf{u}_d,\mathbf{y})d\mathbf{u}_s d\mathbf{u}_d$$

Can show that

$$= \frac{1}{(2\pi)^{\frac{N}{2}}\left|\mathbf{Z}_s\mathbf{Z}_s'\sigma_s^2+\mathbf{Z}_d\mathbf{Z}_d'\sigma_d^2+\mathbf{I}_N\sigma_e^2\right|^{\frac{1}{2}}}\times$$

$$\exp\left[-\frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{Z}_s\mathbf{Z}_s'\sigma_s^2+\mathbf{Z}_d\mathbf{Z}_d'\sigma_d^2+\mathbf{I}_N\sigma_e^2)^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right]$$

Note

$$L(\boldsymbol{\beta},\sigma_s^2,\sigma_d^2,\sigma_e^2)$$

$$\propto \frac{1}{\left|\mathbf{Z}_s\mathbf{Z}_s'\sigma_s^2+\mathbf{Z}_d\mathbf{Z}_d'\sigma_d^2+\mathbf{I}_N\sigma_e^2\right|^{\frac{1}{2}}}\exp\left[-\frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{Z}_s\mathbf{Z}_s'\sigma_s^2+\mathbf{Z}_d\mathbf{Z}_d'\sigma_d^2+\mathbf{I}_N\sigma_e^2)^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right]$$

EXAMPLE

ESTIMATING 2 VARIANCE COMPONENTS AND HERITABILITY BY MAXIMUM LIKELIHOOD USING THE WHEAT DATA IN BGLR
(code uses an eigen-decomposition)

Model is a simple Gaussian random effects (additive) specification

$$\mathbf{y} = \mathbf{a} + \mathbf{e}$$

$$\mathbf{a} \sim \mathbf{N} \quad (0, \mathbf{A}\sigma_u^2)$$

$$\mathbf{e} \sim \mathbf{N} \quad (0, \mathbf{I}\sigma_e^2)$$

Pedigree based heritability : $h^2 = \dfrac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2}$

$$\mathbf{y} \sim N(\mathbf{0}, \mathbf{A}\sigma_g^2 + \mathbf{I}\sigma_e^2)$$

```
library(BGLR)
set.seed(1234567)
###LOAD DATA
data(wheat)
Y<-wheat.Y
X<-wheat.X
y<-Y[,1]
A<-wheat.A #pedigree based relationship matrix


n<-nrow(X)
p<-ncol(X)
```

CODE PROVIDED BY GUSTAVO DE LOS CAMPOS
USES EIGEN-DECOMPOSITION OF A, e.g.,

```
eiA<-eigen(A)
d<-eiA$values
U<-eiA$vectors
```

Consider the following random effects model

$$\begin{cases} y = u + \varepsilon \\ \begin{pmatrix} u \\ \varepsilon \end{pmatrix} \sim MVN\left[\boldsymbol{0}, \begin{matrix} G\sigma_u^2 & \boldsymbol{0} \\ \boldsymbol{0} & I\sigma_\varepsilon^2 \end{matrix}\right] \end{cases} \quad [1]$$

The assumptions of the model imply the following marginal distribution of the data

$$y \sim MVN[0, P]$$

where $P = G\sigma_u^2 + I\sigma_\epsilon^2$And the following likelihood function

$$L(\sigma_\epsilon^2, \sigma_u^2|y) \propto \|P\|^{-\frac{1}{2}} Exp\{-\frac{y'P^{-1}y}{2}\}$$

$$P = G\sigma_u^2 + I\sigma_\epsilon^2 = \sigma_\epsilon^2\left[UDU'\frac{\sigma_u^2}{\sigma_\epsilon^2} + I\right]$$

$$= \sigma_\epsilon^2\left[UDU'\frac{\sigma_u^2}{\sigma_\epsilon^2} + UIU'\right] = \sigma_\epsilon^2[U\{D\lambda + I\}U']$$

$$= \sigma_\epsilon^2 U\widetilde{D}U'$$

where $\lambda = \frac{\sigma_u^2}{\sigma_\epsilon^2}$ and $\widetilde{D} = Diag\{d_i\lambda + 1\}$.

$$\|P\| = \left\|\sigma_\epsilon^2 U\widetilde{D}U'\right\| = \left\|I\sigma_\epsilon^2 U\widetilde{D}U'\right\| = \sigma_\epsilon^{2n}\prod_{i=1}(d_i\lambda + 1)$$

$$L(\sigma_\epsilon^2, \sigma_u^2|y) \propto \|P\|^{-\frac{1}{2}} Exp\left\{-\frac{y'P^{-1}y}{2}\right\}$$

$$\propto [\sigma_\epsilon^2]^{-\frac{n}{2}} \left[\prod_{i=1} (d_i\lambda + 1)\right]^{-\frac{1}{2}} Exp\left\{-\frac{y'P^{-1}y}{2}\right\}$$

$$\propto [\sigma_\epsilon^2]^{-\frac{n}{2}} \left[\prod_{i=1} (d_i\lambda + 1)\right]^{-\frac{1}{2}} Exp\left\{-\frac{y'U\widetilde{D}^{-1}Uy}{2\,\sigma_\epsilon^2}\right\}$$

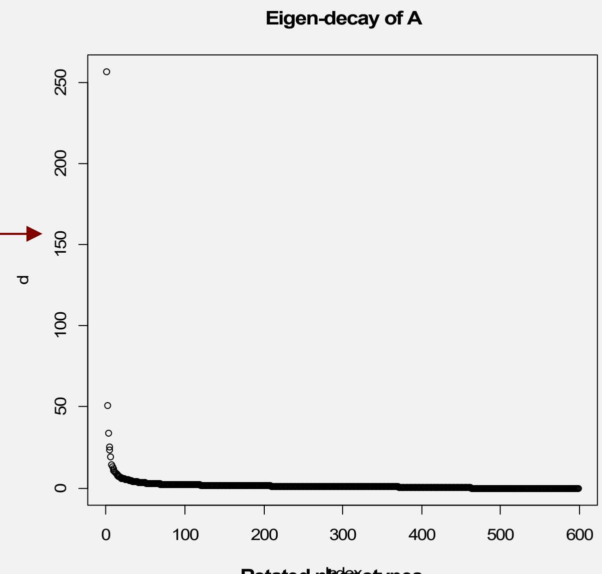
$$\propto [\sigma_\epsilon^2]^{-\frac{n}{2}} \left[\prod_{i=1} (d_i\lambda + 1)\right]^{-\frac{1}{2}} Exp\left\{-\frac{[U'y]'\widetilde{D}^{-1}[U'y]}{2\,\sigma_\epsilon^2}\right\}$$

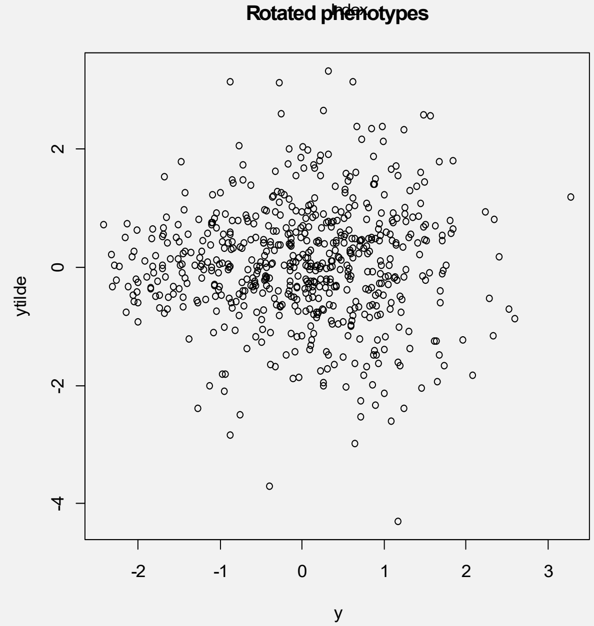
$$\propto [\sigma_\epsilon^2]^{-\frac{n}{2}} \left[\prod_{i=1} (d_i\lambda + 1)\right]^{-\frac{1}{2}} Exp\left\{-\frac{\tilde{y}'\widetilde{D}^{-1}\tilde{y}}{2\,\sigma_\epsilon^2}\right\} \quad where\ \tilde{y} = U'y$$

$$\propto [\sigma_\epsilon^2]^{-\frac{n}{2}} \left[\prod_{i=1} (d_i\lambda + 1)\right]^{-\frac{1}{2}} Exp\left\{-\frac{\sum_i\ \tilde{y}_i^2/(d_i\lambda + 1)}{2\,\sigma_\epsilon^2}\right\} \quad where\ \tilde{y} = U'y$$

$$log\{L(\sigma_\epsilon^2, \sigma_u^2|y)\} = l(\sigma_\epsilon^2, \sigma_u^2|y) \propto -\frac{1}{2}\left\{\mathrm{n} \times \log(\sigma_\epsilon^2) + \sum_i\ log(d_i\lambda + 1) + \frac{\sum_i \frac{\tilde{y}_i^2}{d_i\lambda+1}}{\sigma_\epsilon^2}\right\} \quad [5]$$

```
eiA<-eigen(A)
d<-eiA$values
U<-eiA$vectors


plot(d,main="Eigen-decay of A")

########################################################
################TRANSFORM PHENOTYPES

ytilde<-t(U)%*%y
plot(y,ytilde,main="Rotated phenotypes",xlab="y",ylab="ytilde")
################################################################
####Faster way:
ytilde2<-as.vector(crossprod(U,y)^2)
```





```
## R-function to evaluate -2*logLik=out in function below

neg2LogLik<-function(logVar,ytilde2,d,n=length(ytilde2)){
  varE<-exp(logVar[1])
  varA<-exp(logVar[2])
  lambda<-varA/varE
  dStar<-(d*lambda+1)
  sumLogD<-sum(log(dStar))
  logLik_1<- -0.5*( n*log(varE) + sumLogD )
  logLik_2<- (-0.5*sum(ytilde2/dStar))/varE
  out<--sum(logLik_1,logLik_2)*2
  return(out)
}
##End of function


## Optimize neg2LogLik

 fm<-optim(fn=neg2LogLik,ytilde2=ytilde2,par=log(c(.2,.8)),
         d=d,n=nrow(A),hessian=TRUE)

####Variances are varE=variances[1], varA=variances[2]

variances<-exp(fm$par)
Variances
[1] 0.5758496 0.2860183
```

→ additive=0.286
residual=0.576

```
h2<-variances[2]/sum(variances)
h2
0.3318586
###HESSIAN=NEGATIVE OF MATRIX OF SECOND DERIVATIVES
###OF LOG-LIKELIHOOD WITH RESPECT TO VARIANCES
###INVERSE(HESSIAN)=ESTIMATE OF ASYMPTOTIC VARIANCE-
###COVARIANCE OF ESTIMATES
optimHess(fn=neg2LogLik,ytilde2=ytilde2,par=log(c(.2,.8)),
        d=d,n=nrow(A))
       [,1]       [,2]
[1,] 198.1166 141.5810
[2,] 141.5810 225.8437
H<-matrix(nrow=2,ncol=2)
H[1,]<-c(198.1166, 141.5810)
H[2,]<-c(141.5810, 225.8437)
```

```
##Invert Hessian
Asycov<-solve(H)
Asycov
            [,1]         [,2]
[1,]  0.009144134 -0.005732441
[2,] -0.005732441  0.008021497

#Asympotic SE and correlation
AsySEvE<-sqrt(0.009144134 )
AsySEvE
[1] 0.09562497
AsySEvA<-sqrt(0.008021497)
 AsySEvA
[1] 0.08956281
Asycor<--0.005732441/( 0.09562497*0.08956281)
Asycor
[1] -0.6693304
```

```
> #####CHANGE STARTING VALUES TO (0.8,0.2)
>  fm<-optim(fn=neg2LogLik,ytilde2=ytilde2,par=log(c(.8,.2)),
+           d=d,n=nrow(A),hessian=TRUE)
> variances<-exp(fm$par)
> variances
[1] 0.5760111 0.2860044
>
> h2=variances[2]/sum(variances)
> h2
[1] 0.3317857
```

SIMILAR RESULTS!!

```
> #####CHANGE STARTING VALUES TO (0.01,0.40)
>  fm<-optim(fn=neg2LogLik,ytilde2=ytilde2,par=log(c(.01,.40)),
+           d=d,n=nrow(A),hessian=TRUE)
> variances<-exp(fm$par)
> variances
[1] 0.5759338 0.2859071
>
> h2=variances[2]/sum(variances)
> h2
[1] 0.33174
```

SIMILAR RESULTS!!

Asymptotic standard error of heritability estimates

$$\left(\frac{x}{x+y}\right) \approx \left(\frac{x_0}{x_0+y_0}\right) + \left[\begin{array}{cc} \frac{\partial}{\partial x}\left(\frac{x}{x+y}\right) & \frac{\partial}{\partial y}\left(\frac{x}{x+y}\right) \end{array}\right]\Bigg|_{\substack{x=x_0\\y=y_0}} \left[\begin{array}{c} x-x_0 \\ y-y_0 \end{array}\right]$$

$$\Rightarrow \quad Var\left(\frac{x}{x+y}\right) \approx \left[\begin{array}{cc} \frac{y_0}{(x_0+y_0)^2} & -\frac{x_0}{(x_0+y_0)^2} \end{array}\right] \left[\begin{array}{cc} V_X & C_{XY} \\ C_{XY} & V_Y \end{array}\right] \left[\begin{array}{c} \frac{y_0}{(x_0+y_0)^2} \\ -\frac{x_0}{(x_0+y_0)^2} \end{array}\right]$$

$$\Rightarrow \quad SE = \sqrt{Var\left(\frac{x}{x+y}\right)} \approx \sqrt{\frac{x_0^2 V_Y + y_0^2 V_X - 2x_0 y_0 C_{XY}}{(x_0+y_0)^4}}$$

$$\left[\begin{array}{c} x_0 \\ y_0 \end{array}\right] = \left[\begin{array}{c} 0.2860 \\ 0.5760 \end{array}\right]$$

$$\hat{h}^2 = \frac{0.2860}{0.2860+0.5760} = 0.332$$

$$\left[\begin{array}{cc} V_X & C_{XY} \\ C_{XY} & V_Y \end{array}\right] = \left[\begin{array}{cc} 0.0080 & -0.0057 \\ -0.0057 & 0.0091 \end{array}\right]$$

$$SE\left(\hat{h}^2\right) = 9.776 \times 10^{-2}$$

Asy 99% CI=$[0.03872, 0.62528]$

BEST LINEAR UNBIASED ESTIMATION IN THE MIXED EFFECTS MODEL

A. Model and requirements

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}$$

$$\mathbf{u} \sim (\mathbf{0}, \mathbf{G})$$

$$\mathbf{e} \sim (\mathbf{0}, \mathbf{R})$$

Assume X has full column rank

Normality is NOT assumed, but dispersion parameters assumed KNOWN

Wish to estimate linear function of fixed effects

$$\mathbf{t}'\boldsymbol{\beta}$$

Using linear function of the data

$$\boldsymbol{\lambda}'\mathbf{y}$$

Such that

$$a)\ E(\boldsymbol{\lambda}'\mathbf{y}) = \mathbf{t}'\boldsymbol{\beta}$$

$$b)\ Var(\boldsymbol{\lambda}'\mathbf{y}) < Var(\boldsymbol{\alpha}'\mathbf{y}) \quad for\ any\ E(\boldsymbol{\alpha}'\mathbf{y}) = \mathbf{t}'\boldsymbol{\beta}$$

Requirement (a)= "**unbiasedness**" of estimator of parametric function
Requirement (b)= "**best**" in the sense of minimum variance in the class of linear unbiased estimators

B. Derivation

$$E(\boldsymbol{\lambda}'\mathbf{y}) = \boldsymbol{\lambda}'\mathbf{X}\boldsymbol{\beta} = \mathbf{t}'\boldsymbol{\beta} \quad for\ all\ \boldsymbol{\beta}$$

$$\boldsymbol{\lambda}'\mathbf{X} = \mathbf{t}'\ must\ hold$$

Unbiasedness restriction

$$BLUE(\mathbf{t}'\boldsymbol{\beta}) = \mathbf{t}'\widehat{\boldsymbol{\beta}}$$

$$BLUE(\boldsymbol{\beta}) = \mathbf{t}'\widehat{\boldsymbol{\beta}}$$

$$BLUE(\mathbf{K}'\boldsymbol{\beta}) = \mathbf{K}'\widehat{\boldsymbol{\beta}}$$

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$$

BLUE= Best linear unbiased estimator
BLUE(**β**) = GLS(**β**)= ML(**β**) under normality
VARIANCE AND COVARIANCE COMPONENTS ASSUMED KNOWN!

$$Var\left(\mathbf{t}'\widehat{\boldsymbol{\beta}}\right) = \mathbf{t}'Var\left(\widehat{\boldsymbol{\beta}}\right)\mathbf{t}$$
$$= \mathbf{t}'(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{t}$$

Basis for confidence intervals and hypothesis testing, e.g., F-tests

After some algebra can find BLUE($\boldsymbol{\beta}$) by solving the system:

HENDERSON'S MIXED MODEL EQUATIONS

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix}$$

ML($\boldsymbol{\beta}$) under normality
GLS($\boldsymbol{\beta}$) always

Estimated E(u|y) under normality
BLUP(u) always

Then, the u-solution of the system is BLUP(**u**)
NOTE: IF MULTIVARIATE NORMALITY ASSUMED

$$\hat{\mathbf{u}} = \widehat{E}(\mathbf{u}|\mathbf{y}) = \mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}\left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)$$

Let

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{C}_{\beta\beta} & \mathbf{C}_{\beta u} \\ \mathbf{C}_{u\beta} & \mathbf{C}_{\mathrm{u}u} \end{bmatrix}$$

Can show (after tedious algebra) that:

$$\begin{bmatrix} \mathbf{C}_{\beta\beta} & \mathbf{C}_{\beta u} \\ \mathbf{C}_{u\beta} & \mathbf{C}_{\mathrm{u}u} \end{bmatrix} = \begin{bmatrix} Var\left(\widehat{\boldsymbol{\beta}}\right) & Var\left(\widehat{\boldsymbol{\beta}}, (\mathbf{u} - \hat{\mathbf{u}})'\right) \\ Var\left((\mathbf{u} - \hat{\mathbf{u}}), \widehat{\boldsymbol{\beta}}'\right)\mathbf{C}_{u\beta} & Var(\mathbf{u} - \hat{\mathbf{u}}) \end{bmatrix}$$

Variance-covariance of ML(**β**) or GLS(**β**)

Covariance matrix of prediction errors

EXAMPLE: BLUP OF 599 WHEAT LINES

```
rm(list=ls(all=TRUE))
library(BGLR)

set.seed(1234567)

###LOAD DATA

data(wheat)
Y<-wheat.Y
X<-wheat.X
y<-Y[,1]

####A=PEDIGREE MATRIX
A<-wheat.A

###WORK WITH FIRST 5 MARKERS AS FIXED EFFECTS

X<-X[,1:5]
n<-nrow(X)
p<-ncol(X)


###DEFINITIONS: FOUR LEVELS OF HERITABILITY

h2<-c(0.10,0.25,0.50,0.75)
length<-length(h2)
```

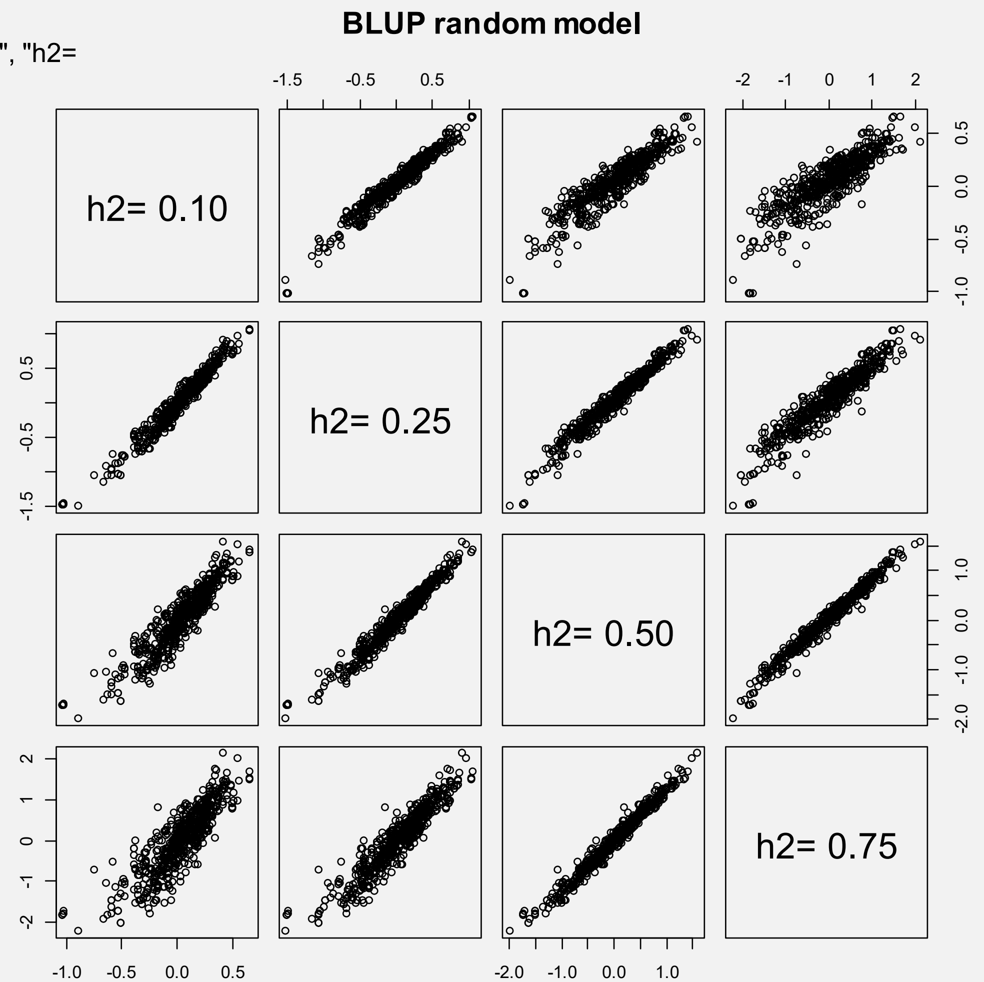
```
####DEFINE FIXED REGRESSION COEFFICIENTS AT EACH h2 VALUE

bgls<-matrix(nrow=length,ncol=p)

####DEFINE PHENOTYPIC VARIANCE-COVARIANCE MATRIX
####(RESIDUAL VARIANCE FACTORED-OUT)

Vstar<-matrix(nrow=n,ncol=n)

####For computing GBLUP assume Var(y)=1, so 1-h2=residual variance

####BLUP calculated with each of the 20 markers
####treated as fixed one at a time

BLUP<-array(dim=c(length,n,p))

#####BLUPrand has length x n; each row contains BLUPS of genetic value
#####for model without fixed effects

BLUPrand<-matrix(nrow=length,ncol=n)

###Incidence matrix for fixed effects. First one will be a column of 1's

Xblup<-matrix(nrow=n,ncol=2)

####VECTOR OF ONES
J<-rep(1,n)
###############
```

```
for (i in 1:length(h2)){

Vstar<-A*h2[i]/(1-h2[i])+diag(n)

C<-chol(Vstar)
CT<-t(C)
Cprimeinv<-chol2inv(chol(CT))
Vinv<-chol2inv(chol(Vstar))


BLUPrand[i,]<-t(A%*%Vinv%*%y)
BLUPrand[i,]<-(h2[i]/(1-h2[i]))*BLUPrand[i,]

###COMPUTE THE GLS OF MARKER EFFECT USING CHOLESKY

z<-Cprimeinv%*%y
K<-Cprimeinv%*%X

for (j in 1:p){

GWASgls<-lm(z~K[,j])

bgls[i,j]<-GWASgls$coefficients[2]

#####GIVES BLUE(x*BETA)=X*bgls

Xblup<-cbind(J,X[,j])%*%GWASgls$coefficients

#### FOR EACH HERITABILITY AND MARKER COMPUTE:

BLUP[i,,j]<-Vinv%*%(y-Xblup)
BLUP[i,,j]<-(h2[i]/(1-h2[i]))*A%*%BLUP[i,,j]
BLUP[i,,j]<-Xblup+BLUP[i,,j]

}
}
```

```
write.table(BLUPrand,file='BLUPrand.txt',row.names=F,col.names=F)
write.table(BLUP[1,,],file='BLUP10.txt', row.names=F, col.names=F)
write.table(BLUP[2,,],file='BLUP25.txt', row.names=F, col.names=F)
write.table(BLUP[3,,],file='BLUP50.txt', row.names=F, col.names=F)
write.table(BLUP[4,,],file='BLUP75.txt', row.names=F, col.names=F)
```

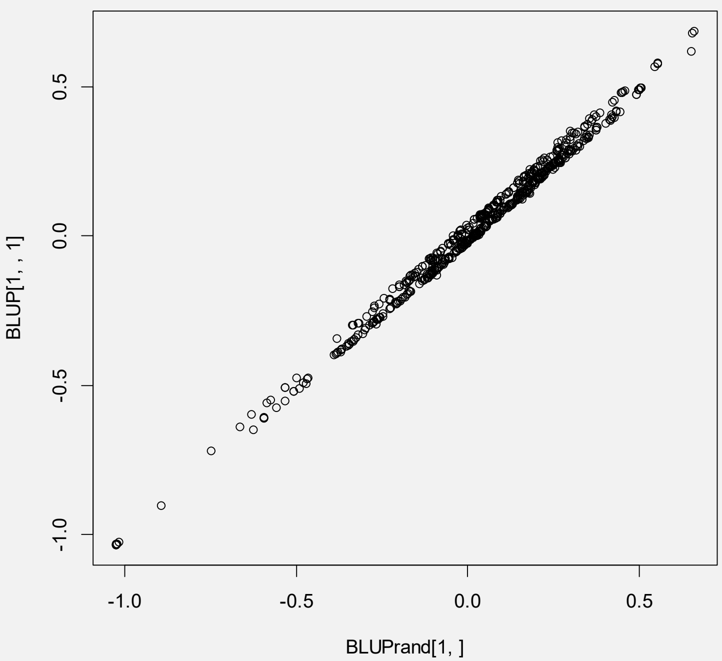
#####PLOT BLUPs UNDER RANDOM MODEL

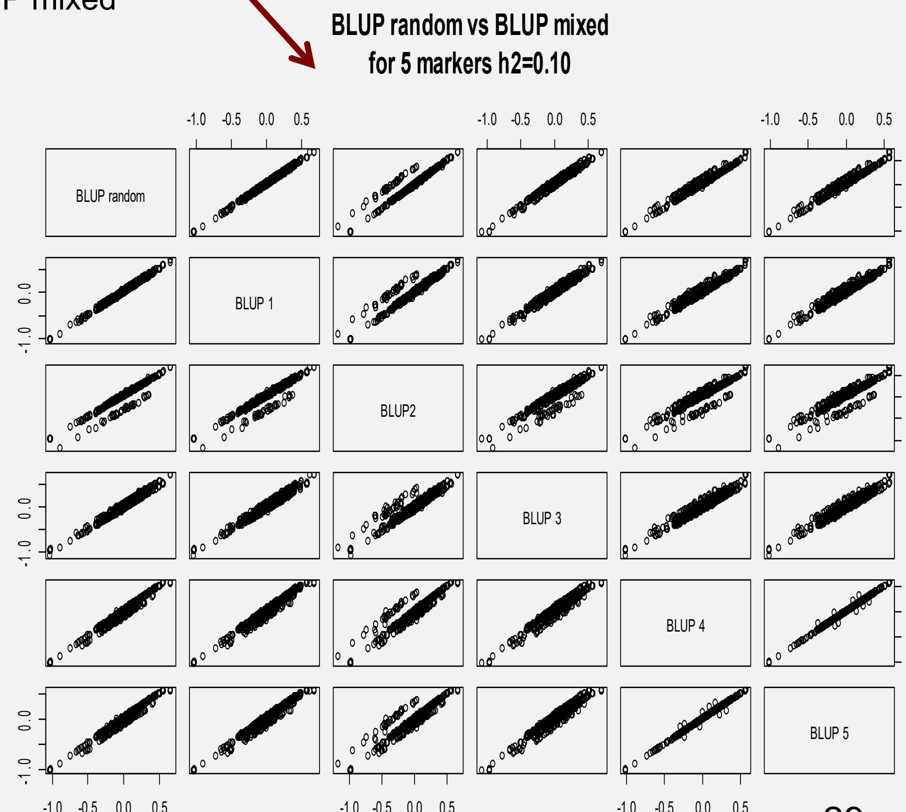
```
MATBLUPrand<-
rbind(BLUPrand[1,],BLUPrand[2,],BLUPrand[3,],BLUPrand[4,])
dim(MATBLUPrand)

pairs(t(MATBLUPrand),labels=c("h2= 0.10","h2= 0.25", "h2= 0.50", "h2=
0.75"),
main="BLUP random model")
```



###DOES BLUP(RANDOM) DIFFER FROM BLUP(mixed)?

```
plot(BLUPrand[1,],BLUP[1,,1])

MATPAIR<-cbind(BLUPrand[1,],BLUP[1,,1],BLUP[1,,2],
BLUP[1,,3],BLUP[1,,4],BLUP[1,,5])

pairs(MATPAIR,labels=c("BLUP random","BLUP 1", "BLUP2",
"BLUP 3",
"BLUP 4", "BLUP 5"),
main="BLUP random vs BLUP mixed
for 5 markers h2=0.10")
```





IS BLUP OVERFITTING?

```
FITCOR<-cbind(y,MATPAIR)

cor(FITCOR)
```

Y	BLUPR	BLUP1	BLUP2	BLUP3	BLUP4	BLUP5
1.0000000	**0.7391247**	**0.7384688**	**0.7159003**	**0.7275751**	**0.7295317**	**0.7287069**
0.7391247	1.0000000	0.9963397	0.9508149	0.9841684	0.9863276	0.9852552
0.7384688	0.9963397	1.0000000	0.9444558	0.9812627	0.9814759	0.9806144
0.7159003	0.9508149	0.9444558	1.0000000	0.9348012	0.9338615	0.9328132
0.7275751	0.9841684	0.9812627	0.9348012	1.0000000	0.9688824	0.9680241
0.7295317	0.9863276	0.9814759	0.9338615	0.9688824	1.0000000	0.9979560
0.7287069	0.9852552	0.9806144	0.9328132	0.9680241	0.9979560	1.0000000

NO EVIDENCE. BLUPS CLOSELY INTER-CORRELATED

####HOW ABOUT TRAINING MEAN-SQUARED ERROR?

```
mse<-numeric(6)
for (i in 1:6){
mse[i]<-sum((y-MATPAIR[,i])**2)/n
}
mse
```

[1] **0.6877218 0.6866927 0.6834385 0.6869435 0.6875240 0.6875065**

```
> MSE<-c(0.6877218, 0.6866927, 0.6834385, 0.6869435, 0.6875240, 0.6875065)
> R2approx<-1-MSE
> coryblup<-c(0.7391247, 0.7384688, 0.7159003, 0.7275751, 0.7295317, 0.7287069)
> R2train<-coryblup^2
> R2approx
[1] 0.3122782 0.3133073 0.3165615 0.3130565 0.3124760 0.3124935
> R2train
[1] 0.5463053 0.5453362 0.5125132 0.5293655 0.5322165 0.5310137
```

- ➔ **DIFFERENCES IN TRAINING MEAN SQUARED ERROR ARE MINIMAL**
- ➔ **DIFFERENCES IN R2 ARE MINIMAL, IRRESPECTIVE OF METRIC**
- ➔ **R2 APPROX POSSIBLY SMALLER BECAUSE MSE=SQUARED BIAS+VARIANCE**

COMPUTATION OF BLUP 'RELIABILITIES' FOR THIS MODEL

$$\begin{bmatrix} \frac{1}{\sigma_e^2}\mathbf{X}'\mathbf{X} & \frac{1}{\sigma_e^2}\mathbf{X}'\mathbf{Z} \\ \frac{1}{\sigma_e^2}\mathbf{Z}'\mathbf{X} & \frac{1}{\sigma_e^2}\mathbf{Z}'\mathbf{Z}+\frac{1}{\sigma_g^2}\mathbf{A}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{g}} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sigma_e^2}\mathbf{X}'\mathbf{y} \\ \frac{1}{\sigma_e^2}\mathbf{Z}'\mathbf{y} \end{bmatrix}$$

$$\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{g}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z}+\frac{\sigma_e^2}{\sigma_g^2}\mathbf{A}^{-1} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{bmatrix}$$

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z}+\frac{\sigma_e^2}{\sigma_g^2}\mathbf{A}^{-1} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{C}_{\beta\beta} & \mathbf{C}_{\beta u} \\ \mathbf{C}_{\beta u} & \mathbf{C}_{\mathrm{u}} \end{bmatrix}$$

$$PEV_{g_i} = c_{g_i g_i}\sigma_e^2$$

$$REL_{g_i} = 1 - \frac{PEV_{g_i}}{\sigma_g^2} = 1 - c_{g_i g_i}\frac{1-h^2}{h^2}$$

Note that "reliability"or "prediction error variance" depend on data structure and partition on variance but not on performance