Comparação entre os modelos de seleção genômica RRblup e GBlup

Renato Gonçalves de Oliveira

23 de junho de 2021

Dentre os modelos disponíveis para a construção de modelos para a seleção genômica, os modelos RRblup e GBlup forma escolhidos para que suas saídas fossem comparadas neste trabalho. A equação de relação entre os dois modelos é descrita abaixo:

Onde:

* : efeito estimado dos marcadores;
* Z: matriz dos estados dos marcadores para cada genótipo;
* : genomic estimated breeding values (GEBV´s).

O modelo RRblup fornece com saída os valores de . Esses valores correspondem aos efeitos estimados para os marcadores utilizados no modelo. O modelo GBlup apresenta como saída os valores de , que são os valores dos Blups estimados pelo modelo, ou GBlups. Esses valores correspondem aos valores genômicos preditos, ou GEBV´s. Para que se obtenham os valores os valores de GEBV´s à partir do modelo RRblup é preciso multiplicar os valores estimados dos efeitos dos marcadores, , pelo estado desses marcadores nos genótipos. O somatório desses efeitos deve ser acrescido do valor do instercepto estimado, . Dessa forma, a relação entre os valores da saída do modelo RRblup e GBlup pode ser descrita pela equação abaixo:

# Pacotes utilizados na análise

library(rrBLUP)   
library(bWGR)  
library(lattice)

source('gbuild.R')

# Arquivos de dados

Dois conjuntos de dados serão utilizados para a comparação dos modelos de seleção genômica. O arquivo de dados gentípicos é composto pelo resultado da genotipagem de 2025 indivíduos utilizando 453 marcadores. O arquivos de dados fenotípicos é composto por 5 caráteres avaliados em 1000 indivíduos.

gen <- read.table("genotype.txt",header=TRUE)  
phe <- read.table("phenotype.txt",header=TRUE)

## Arquivo de dados genotípicos

Corresponde aos dados dos estados dos marcadores dos indvíduos genotipados.

gen[1:10, 1:10]

## m1 m2 m3 m4 m5 m6 m7 m8 m9 m10  
## 301 1 0 0 0 0 0 0 0 2 2  
## 302 0 0 0 0 0 0 0 0 1 2  
## 303 0 0 0 0 0 0 0 0 1 1  
## 304 0 0 0 0 1 0 0 0 1 2  
## 305 0 0 0 0 1 0 0 0 0 2  
## 306 0 0 0 0 1 0 0 0 1 1  
## 307 0 0 0 0 1 0 0 0 1 2  
## 308 0 0 0 0 2 0 0 0 0 1  
## 309 0 0 0 0 0 0 0 0 1 1  
## 310 0 0 0 0 0 0 1 1 2 1

## Arquivo de dados fenotípicos

Dos 5 caráteres avaliados do arquivos de dados genotípicos o que será utilizado será o p530.

phe[1:10, ]

## ind p0 p132 p265 p397 p530  
## 1 346 0.2999 1.3938 4.0470 8.9365 14.4663  
## 2 347 0.4265 1.9578 6.6809 15.9458 27.3269  
## 3 348 0.4991 2.0284 6.0664 13.7166 22.7103  
## 4 349 0.1739 1.2515 4.4695 11.0793 18.7735  
## 5 350 0.3712 1.8365 5.9575 14.4277 23.8408  
## 6 351 0.2727 1.3336 3.9884 8.7238 14.1380  
## 7 352 1.1542 3.7294 9.8721 20.2459 32.2920  
## 8 353 0.3175 1.7614 5.6780 13.8240 22.7556  
## 9 354 0.1726 1.2156 4.4640 11.2814 19.6790  
## 10 355 0.6935 2.8703 8.4873 19.1791 30.8544

# Informações do arquivo de dados

O conjunto de dados utilizados para comparação dos dois modelos de seleção genômica possui as seguintes informações:

* n.° de machos: 5
* n.° de fêmeas: 20
* n.° de indivíduos fenotipados: 1000
* n.° de indivíduos genotipados: 2025
* tamanho efetivo da população (Ne): 16
* herdabilidade para o fenótipo : 0.47
* acurácia da seleção fenotípica: 0.73
* comprimento do genoma: 500 cM
* acurácia esperada para a seleção genômica: 0.67

# Modelo RRblup

## Matriz M

A matriz M, também conhecida como matriz Z, é a matriz que corresponde aos estados dos marcadores para cada um dos indivíduos genotipados. Para criar a matriz M foram colocados o nome dos genòtipos na linhas, usando os nomes do arquivos de dados fenotípicos, e o nome dos marcadores nas colunas.

M <- gen[as.character(phe$ind),]  
M[1:15,1:15]

## m1 m2 m3 m4 m5 m6 m7 m8 m9 m10 m11 m12 m13 m14 m15  
## 346 1 0 0 0 1 0 0 0 1 1 0 1 1 2 0  
## 347 1 0 0 0 1 0 0 0 1 1 0 1 1 2 0  
## 348 1 0 0 0 0 0 0 0 1 2 0 2 2 2 0  
## 349 0 0 0 0 1 0 0 0 1 1 0 1 1 2 0  
## 350 1 0 0 0 1 0 0 0 1 1 0 1 1 2 0  
## 351 0 0 0 0 1 0 0 0 1 1 0 1 1 2 0  
## 352 0 0 0 0 0 0 0 0 1 2 0 2 2 2 0  
## 353 1 0 0 0 0 0 0 0 1 2 0 2 2 2 0  
## 354 0 0 0 0 1 0 0 0 1 2 0 2 2 2 0  
## 355 0 0 0 0 1 0 0 0 1 1 0 1 1 2 0  
## 356 0 0 0 0 1 0 0 0 1 1 0 1 1 2 0  
## 357 0 0 0 0 0 0 0 0 1 2 0 2 2 2 0  
## 358 1 0 0 0 0 0 0 0 1 2 0 2 2 2 0  
## 359 0 0 0 0 0 0 0 0 1 2 0 2 2 2 0  
## 360 1 0 0 0 0 0 0 0 1 2 0 2 2 2 0

## População de treinamento

Para a formação da população de treinamento foram selecionados 800 indivíduos, do indivíduo 201 até o indivíduo 1000 da matriz M.

poptr <- rownames(M)[201:1000]

## População de validação

Para a população de validação foram selecionados os indivíduos de 1 a 200 da matriz M.

popvl <- rownames(M)[1:200]

## Valores fenotípicos da população de treinamento

Valores fenotípicos para o caráter p530 para a população de treinamento.

g.tr <- phe[phe$ind%in%poptr,]$p530

## Valores fenotípicos da população de validação

Valores fenotípicos para o caráter p530 para a população de validação.

g.vl <- phe[phe$ind%in%popvl,]$p530

## Treinamento do modelo

Treinamento do modelo de selção genômica utilizando a população de treinamento

train <- mixed.solve(g.tr,M[poptr,])  
str(train)

## List of 5  
## $ Vu : num 0.186  
## $ Ve : num 21.7  
## $ beta: num [1(1d)] 23  
## $ u : num [1:453(1d)] -0.16 -0.1344 0.0419 0.3001 0.1029 ...  
## ..- attr(\*, "dimnames")=List of 1  
## .. ..$ : chr [1:453] "m1" "m2" "m3" "m4" ...  
## $ LL : num -2457

## Catálogo de marcas

Vetor contendo os efeitos estimados para cada um dos marcadores utilizados.

m.cat <- train$u

## GEBV´s

Através da multiplicação da matriz dos estados pela matriz dos efeitos dos marcadores tem-se a matriz com o somatório dos efeitos dos marcadores em função do estado desses marcadores em cada indivíduo. O inertcepto é dado pela média dos GEBV´s estimados na população de treinamento. A soma do valor do efeito dos marcadores corresponde ao GEBV dos indivíduos da popualção de treinamento, ou .

g.hat <- as.matrix(M[popvl,])%\*%m.cat + mean(g.tr)

## Acurácia realizada

A acurácia realizada corresponde a correlção entre os valores dos GEBV´s, ou , ,a população de teste e os valores fenotípicos da população de validação.

cor(g.hat,g.vl)

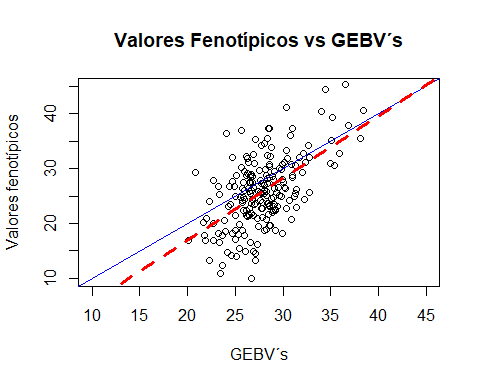
## [,1]  
## [1,] 0.56465

Para esse conjunto de dados esse valor foi 0.56. Esse valor ficou próximo do valor da acurácia esperada para esse conjunto de genótipos que foi de 0.67.

## Valores fenotípicos vs preditos

O gráfico abaixo mostra os valores fenotípicos versus os valores preditos, GEBV´s, para a população de validação. a linha azul marca a exatidão dos valores preditos e a linha vermelha a precisão desses valores. Como os valores foram estimados utilizando modelos mistos há um encurtamento esperado para os vlores preditos em relação aos valores fentípicos.

plot(g.hat,g.vl,xlim=c(10,45),ylim=c(10,45), xlab = 'GEBV´s', ylab = 'Valores fenotípicos', main = 'Valores Fenotípicos vs GEBV´s')  
abline(0,1,col="blue")  
abline(lm(g.vl~g.hat), col="red", lwd=3, lty=2)



## Eficiência do modelo RRblup

A eficiência do modelo de seleção genômica é dada pela relação entre a acurácia otida pelo modelo e a acurácia esperada para o conjunto de dados. Se um modelo consegue capturar toda a acurácia esperada sua eficiênica é de 100%. No caso do modelo RRblup para esse conujnto de dados a eficiência foi de 0.84, ou seja, o modelo conseguiu capturar 84.27% da acurácia esperada para esse conjunto de dados.

cor(g.hat,g.vl)/0.67

## [,1]  
## [1,] 0.8427612

# Modelo GBlup

Para o cálculo dos valores preditos para o modelo GBlup foi utilizada a função **emML** do pacote **bWGR**. Essa função estima o valor dos dos efeitos dos marcadores e estima os valores preditos utilizando a matriz M.

## Ajuste do modelo

train1 <- emML(g.tr,as.matrix(M[poptr,]))  
str(train1)

## List of 7  
## $ mu : num 25.6  
## $ b : num [1:453] -0.157 -0.134 0.041 0.301 0.101 ...  
## $ hat: num [1:800] 26.6 22.9 22 24.7 20.6 ...  
## $ h2 : num 0.445  
## $ Vb : num 0.188  
## $ Va : num 17.3  
## $ Ve : num 21.6

## Catálogo de marcas

Vetor contendo os efeitos estimados para cada um dos marcadores utilizados.

m.cat1 <- train1$b

## GEBV´s

Através da multiplicação da matriz dos estados pela matriz dos efeitos dos marcadores tem-se a matriz com o somatório dos efeitos dos marcadores em função do estado desses marcadores em cada indivíduo. O inertcepto é dado pela média dos GEBV´s estimados na população de treinamento. A soma do valor do efeito dos marcadores corresponde ao GEBV dos indivíduos da popualção de treinamento, ou .

g.hat1 <- as.matrix(M[popvl,])%\*%m.cat1 + mean(g.tr)

## Acurácia realizada

A acurácia realizada corresponde a correlção entre os valores dos GEBV´s, ou , ,a população de teste e os valores fenotípicos da população de validação.

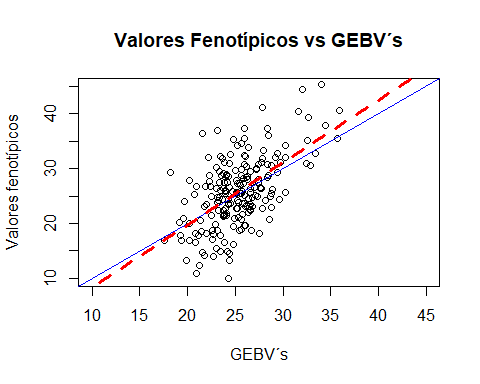
cor(g.hat1, g.vl)

## [,1]  
## [1,] 0.5668707

## Valores fenotípicos vs preditos

O gráfico abaixo mostra os valores fenotípicos versus os valores preditos, GEBV´s, para a população de validação. a linha azul marca a exatidão dos valores preditos e a linha vermelha a precisão desses valores. Assim como no modelo anterior os valores foram estimados utilizando modelos mistos há um encurtamento esperado para os valores preditos em relação aos valores fentípicos.

plot(g.hat1, g.vl, xlim=c(10,45), ylim=c(10,45), xlab='GEBV´s', ylab='Valores fenotípicos', main='Valores Fenotípicos vs GEBV´s')  
abline(0,1,col="blue")  
abline(lm(g.vl~g.hat1), col="red", lwd=3, lty=2)



## Eficiência do modelo GBlup

No caso do modelo GBlup para esse conujnto de dados a eficiência foi de 0.86, ou seja, o modelo conseguiu capturar 86.55% da acurácia esperada para esse conjunto de dados.

cor(g.hat1, g.vl)/0.67

## [,1]  
## [1,] 0.8460756

# Correalção entre os valores preditos

A correlação entre os valores preditos pelo modelo RRblup e GBlup foi de 0.99.

cor(g.hat, g.hat1)

## [,1]  
## [1,] 0.9995376

# Correlação entre os efeitos dos marcadores

A correlação entre os valores estimados para os efeitos dos marcadores pelo modelo RRblup e GBlup foi de 0.99.

cor(m.cat, m.cat1)

## [1] 0.9972198

# Conclusões

Os valores de acurácia alcançada e de eficiência do modelo obtidos pelo modelo RRblup foram de 0.56465 e 0.8427612, respectivamente. Enquanto para o modelo GBlup esses valores foram de 0.5668707 e 0.8460756. Os valores foram muito próximos e só divergiram após a 3° casa decimal.

A correlção entre os valores estimados para os efeitos dos marcadores para os dois modelos foi de 0.99. E a correlação entre os valores preditos para os dois modelos foi de 0.99.

Diante disso, pode-se dizer que os dois modelos chegaram a resultados bastante próximos.