UF UNIVERSITY of FLORIDA

(HOS 6932)– Survey of Breeding Tools and Methods

Statistical Learning and Whole-Genome Regression Models

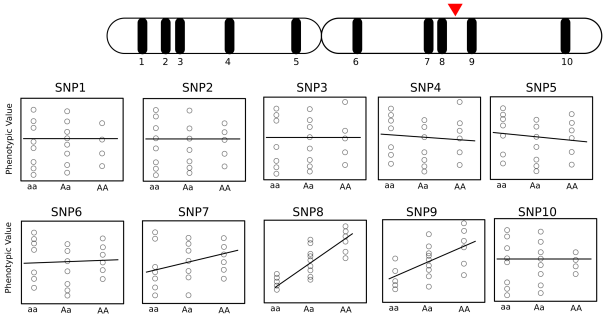
Felipe Ferrão
Research Assistant Scientist
lferrao@ufl.edu
lfelipeferrao

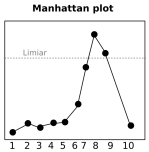
January, 2022

Introduction

In the last class ...

- Single Marker Regression Analyses
- Benefits and Disadvantages

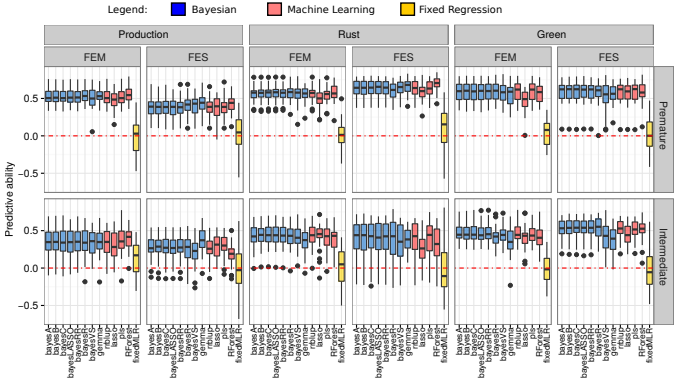




Introduction

Motivation

- Why not use simple linear regression for prediction?
- Example: coffee research, multiple traits and populations



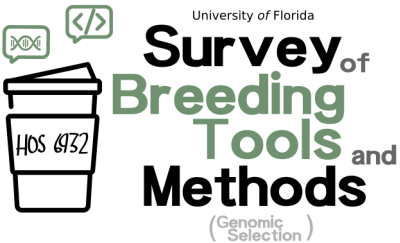
Ferrão, et al 2018. (https://doi.org/10.1038/s41437-018-0105-y)

Introduction

In this class we will discuss

- Introduction to Statistical Learning
- Linear Models: Multiple Regression
- Regularization: the curse of dimensionality
- RR-BLUP and GBLUP

Introduction



SCAN ME

https://lfelipe-ferrao.github.io/teaching/

Statistical Learning

What is Statistical Learning?

- Suppose that we observe a quantitative response Y and p different predictors $(X_1, X2, ..., X_n)$.
- We assume that there is some relationship between Y and X: $Y = f(X) + e$
- In this formulation, f represents the systematic information that X provides about Y.
- Our main goal is estimating f for two purposes:
  - ▷ Prediction
  - ▷ Inference

Statistical Learning

Prediction

- In many situations, a set of inputs X are readily available, but the output Y cannot be easily obtained
- *f()* is often treated as a black box, in the sense that one is not typically concerned with the its exact form, but its predictive performance
- Example: genomic prediction
- Statistical Models: Regularization, Bayesian Models, Deep Learning

Inference

- We are often interested in understanding the way that Y is affecting X
- *f()* cannot be treated as a black box. We need to know its exact form
- Example: QTL mapping, GWAS studies
- Statistical Model: least squares, stepwise regression.

Statistical Learning

Prediction

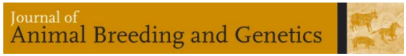
- In many situations, a set of inputs X are readily available, but the output Y cannot be easily obtained
- *f()* is often treated as a black box, in the sense that one is not typically concerned with the its exact form, but its predictive performance
- Example: genomic prediction
- Statistical Models: Regularization, Bayesian Models, Deep Learning

Inference

- We are often interested in understanding the way that Y is affecting X
- *f()* cannot be treated as a black box. We need to know its exact form
- Example: QTL mapping, GWAS studies
- Statistical Model: least squares, stepwise regression.

Statistical Learning

- Nice 1 page comment !
- Prediction vs. Inference at the genomic level



EDITORIAL | Free Access

Breeding beyond genomics

Miguel Pérez-Enciso

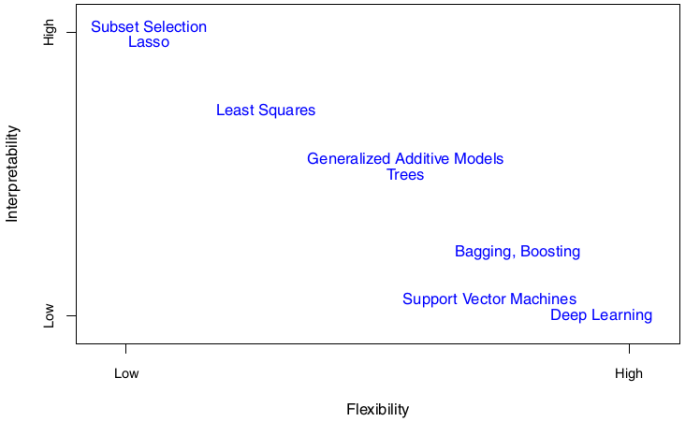
First published: 22 April 2021 | https://doi.org/10.1111/jbg.12547

Check for Full Text

PDF TOOLS SHARE

From the Statistical point of view, I tend to think of Quantitative Genetics and related fields as domains where two main "pillars" cohabitate: Inference and Prediction. Even if the same tool, for example penalized linear models, can be used for both tasks and inference and prediction may reinforce each other, they are distinct concepts. It is interesting to observe how these two pillars have reacted to big data, that is the *large p small n paradigm*. While

Statistical Learning



James, et al 2021. An introduction to Statistical Learning

Statistical Learning

Some Questions:

- Many methods: from simple regression to deep learning
- Why is it necessary to introduce so many different approaches?
- How should I choose which one to use?

My personal opinion !!

- There is not a right answer !!
- No one method dominates all others over all possible data sets
- In the past 20 years, we learned some trends from genomic prediction studies
- Selecting the best approach can be one of the most challenging parts of performing statistical learning in practice.

Statistical Learning

Some Questions:

- Many methods: from simple regression to deep learning
- Why is it necessary to introduce so many different approaches?
- How should I choose which one to use?

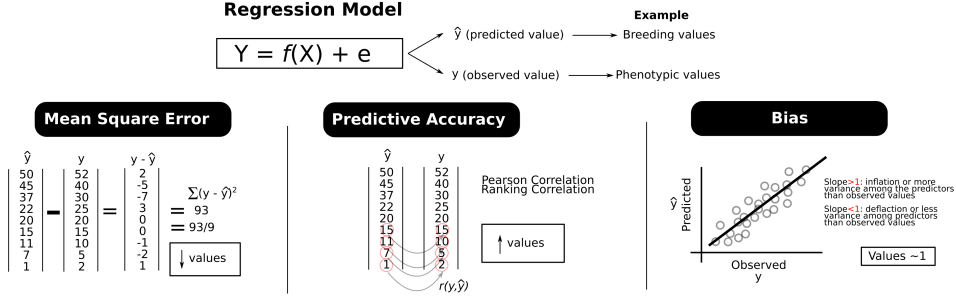
My personal opinion !!

- There is not a right answer !!
- No one method dominates all others over all possible data sets
- In the past 20 years, we learned some trends from genomic prediction studies
- Selecting the best approach can be one of the most challenging parts of performing statistical learning in practice.

Statistical Learning

Measuring the Quality of Fit

- How well its predictions actually match the observed data ?
- What do breeders care most about?

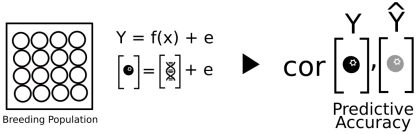


Statistical Learning

Training and Test Population

- For prediction, we are interested in the accuracy of the predictions that we obtain when we apply our method to unseen data.
- Example: plants that I have the genotype, but not the phenotype
- We need a method that improve our prediction in the test data set

Training Data Set

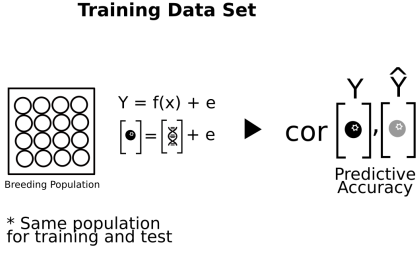


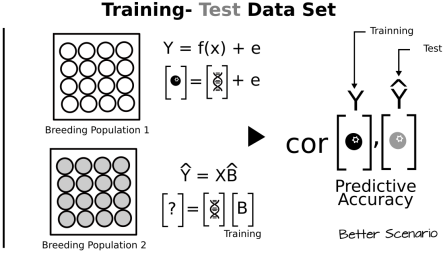
* Same population for training and test

Statistical Learning

Training and Test Population

- For prediction, we are interested in the accuracy of the predictions that we obtain when we apply our method to unseen data.
- Example: plants that I have the genotype, but not the phenotype
- We need a method that improve our prediction in the test data set

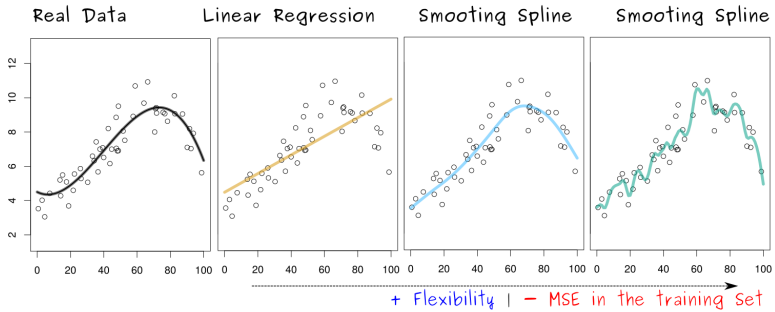




Statistical Learning

More mathematically

- Simulated data and three $f(x)$ or regressions models
- Which model gives me the most flexibility and the lowest MSE?
- What happen with new data is collected?

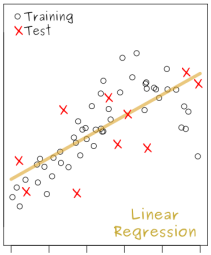


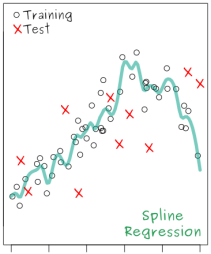
Statistical Learning

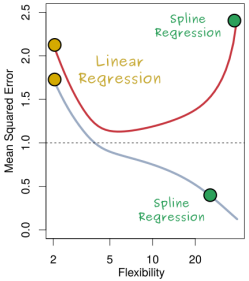
More mathematically

- Simulated data and three $f(x)$ or regressions models
- Which model gives me the most flexibility and the lowest MSE?
- What happen with new data is collected?

Statistical Learning







MSE -- test data
MSE --training data

In the test data set, MSE is lower for less flexible models
linear regression

Some preliminary conclusions

- Model complexity has an influence in the predictive capacity
- More complex models are not necessarily better
- Linear models is a useful approximation (almost always a good starting point)

Connecting the dots ... What do we know so far?

- Statistical Learning involves a $f()$ to connect a response variable and predictors
- Could be used for prediction or inference
- We need training and data sets and metrics to measuring the quality of fit

What are we missing?

- We still don't know the theoretical formulation of $Y = f(x) + e$
- In another words: simple regression model, regression trees, Deep Learning, Bayesian approach, Mixed Model ???

Connecting the dots ... What do we know so far?

- Statistical Learning involves a $f()$ to connect a response variable and predictors
- Could be used for prediction or inference
- We need training and data sets and metrics to measuring the quality of fit

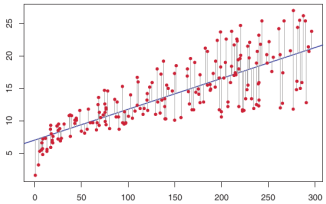
What are we missing?

- We still don't know the theoretical formulation of $Y = f(x) + e$
- In another words: simple regression model, regression trees, Deep Learning, Bayesian approach, Mixed Model ???

Linear Models

Background

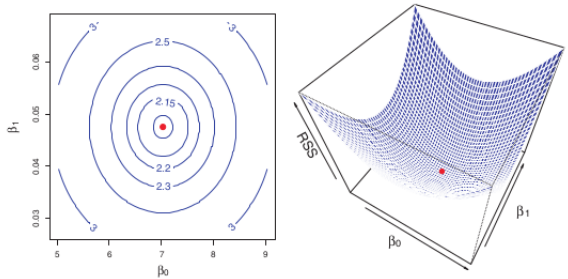
- Last class we presented a simple linear regression model
- It assumes that there is approximately a linear relationship between X and Y
- Intercept and slope could be estimated in a training data by minimizing the least squares criterion.



- Model: $y_i = \beta_0 + \beta_1 x_i + e$
- $\hat{\beta}_1 = \dfrac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$
- $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

Linear Models

Simple Linear Regression



Contour and three-dimensional plots of the least square criterion. The red dots correspond to the least squares estimates intercept and slope

Linear Models

Simple Linear Regression

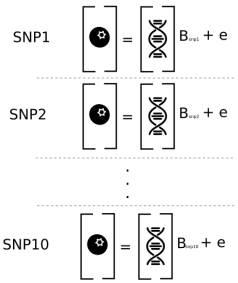
- Simple linear regression is a useful approach if you have a single predictor
- In the first genomic model, we had 10 SNPs (or ten predictors)
- The approach of fitting a separate simple linear regression model for each predictor is not entirely satisfactory.
  - ▷ Poor predictive ability (coffee example)
  - ▷ It is unclear how to make a single SNPs given levels of the nine other markers, since each of the phenotype is associated with a separate regression equation.
  - ▷ Each of the 10 regression equations ignores the other the regression coefficients.

Linear Models

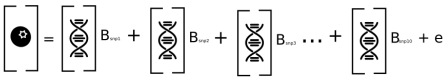
Question

Why not accommodate all predictors (SNPs) in a single regression model?

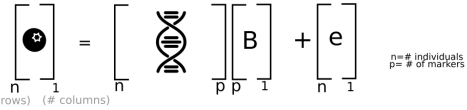
Simple Linear Regression



Multiple Linear Regression



Matrix Format: Y = XB + e

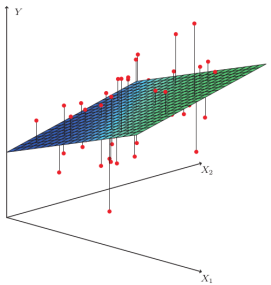


Linear Models

Multiple Regression Model

- It is a good alternative, and used for MAS application
- Interpretation for the coefficients is conditional on the others in the model
- Regression could not be represented for a straight line
- We can also use the least square theory, but in the matrix notation.

$$\arg\min_{\beta} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2$$



- Multidimensional plan
- Matrix: $y = X\beta + e$
- $\hat{\beta} = (X'X)^{-1}X'y$

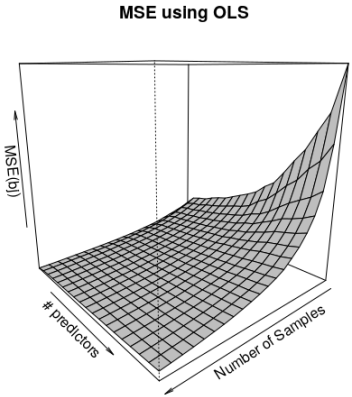
Linear Regression

Question

What happen if we have thousand of individuals and millions of markers?

Multiple Regression Model

- Typical "Big Data" case
- Common scenario in GS studies
- *There is no free lunch in statistics !!*
- Multiple regression works when p is relatively small, and certainly small compared to n.
- If $p > n$ there are more coefficients to estimate than observations



Linear Regression

$$\hat{\beta} = (X'X)^{-1}X'y$$

- In the equation, $(X'X)^{-1}$ is singular and not unique. We cannot inverse the matrix and there is an infinite number of solutions for the β's
- Multicollinearity: situation in which two or more predictor are closely related to one another, it reduces the accuracy of the estimates. For example: SNPs in higher LD
- At the inferential level, we don't have enough degrees of freedom for the statistical test
- Curse of Dimensionality

Ridge Regression

Regularized Regression

- Assuming that multiple regression is not an option
- We can constraining or shrinking the estimated coefficients
- Under two different point of view:
  - ▷ Frequentist: is attained via *ad hoc* penalty functions during the estimation
  - ▷ Bayesian: regularization is part of the prior definition

Naive analogy

- During your high school. Math exams.
- If the teacher gave to you 2 equations and 3 unknowns (a,b and c).
- How to solve this? You need some external information !

Ridge Regression

Regularized Regression

- Assuming that multiple regression is not an option
- We can constraining or shrinking the estimated coefficients
- Under two different point of view:
  - ▷ Frequentist: is attained via *ad hoc* penalty functions during the estimation
  - ▷ Bayesian: regularization is part of the prior definition

Naive analogy

- During your high school. Math exams.
- If the teacher gave to you 2 equations and 3 unknowns (a,b and c).
- How to solve this? You need some external information !

Ridge Regression

Ridge Regression Estimator

$$\arg\min_{\beta} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{i=1}^{n} \beta_j^2$$

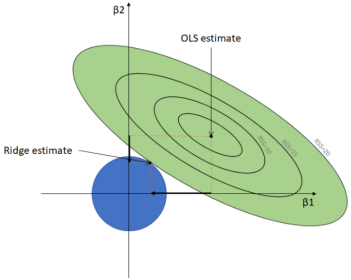
$$\hat{\beta} = (X'X + \lambda I)^{-1} X'y$$

- It is very similar to the original OLS equation
- $\lambda \sum_{i=1}^{n} \beta_j^2$ is a shrinkage penalty
- λserves to control the penalty
- If $\lambda = 0$, penalty term has no effect, and RR will produce OLS estimates
- If $\lambda \longrightarrow \infty$ coefficient estimates will approach zero.
- Penalty is included in the diagonal of the matrix

Ridge Regression

Geometrical Representation

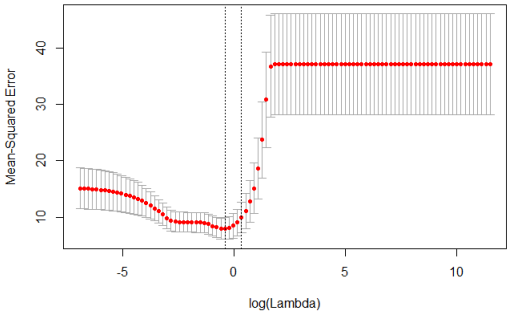
- RR coefficients are shrunken towards zero relative to OLS estimates.
- Good statistic properties: shrinkage has the effect of reducing variance
- Ridge relatively drops more quickly to zero than OLS.



Ridge Regression

Penalty

- One of the problems is that we need to select a value for λ
- Cross-validation scheme:
  - ▷ Generating a grid of potential values for λ
  - ▷ Split the data into training and test data sets
  - ▷ Fit the model in the train and predict the test under different lambda values
  - ▷ Use a metric to check quality of fit (ex: MSE)



Connecting the dots

- Ridge Regression is a machine learning approach !
- Lambda can be selected via cross-validation
- What is the connection between Ridge Regression, Mixed Model and all the other literature in the animal and plant breeding

Connection

- Mixed model is also a shrinkage procedure !!
- When lambda has a "specific form" we can connect both worlds: quantitative genetics and statistical learning

Connecting the dots

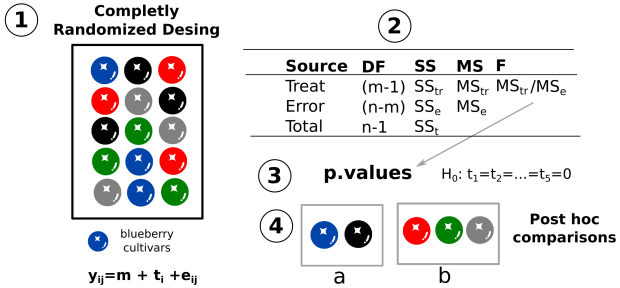
- Ridge Regression is a machine learning approach !
- Lambda can be selected via cross-validation
- What is the connection between Ridge Regression, Mixed Model and all the other literature in the animal and plant breeding

Connection

- Mixed model is also a shrinkage procedure !!
- When lambda has a "specific form" we can connect both worlds: quantitative genetics and statistical learning

Ridge Regression

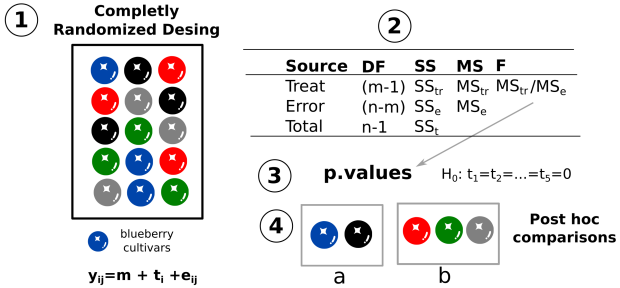
Traditional ANOVA



- Experiments are concerning with making comparisons among specific factor levels
- Involves comparing treatment means in an attempt to detect differences.
- Experimenter's attention is fixed upon to certain levels of interest and no thought for any other level in the analysis

Ridge Regression

Traditional ANOVA



- Experiments are concerning with making comparisons among specific factor levels
- Involves comparing treatment means in an attempt to detect differences.
- Experimenter's attention is fixed upon to certain levels of interest and no thought for any other level in the analysis

Ridge Regression

When certain effects are fixed or random?

- The objectives are different
  - ▷ Fixed effects: compare specific levels of a certain factor
  - ▷ Random effects: conclusions draw for a much larger universe of interest
- Sampling are different
  - ▷ Fixed effects: treatment levels are selected by the investigator.
  - ▷ Random effects: treatment levels are randomly sampled.
- Statistically
  - ▷ Fixed effects: hypothesis tests on the mean $H_0 : t_1 = t_2 = ... = t_n = 0$
  - ▷ Random effects: hypothesis tests on the variance component $H_0 : \sigma_t^2 = 0$

Mixed Models

- Contain fixed and random terms in the same framework
- When we need mixed models ?

Ridge Regression

When certain effects are fixed or random?

- The objectives are different
  - ▷ Fixed effects: compare specific levels of a certain factor
  - ▷ Random effects: conclusions draw for a much larger universe of interest
- Sampling are different
  - ▷ Fixed effects: treatment levels are selected by the investigator.
  - ▷ Random effects: treatment levels are randomly sampled.
- Statistically
  - ▷ Fixed effects: hypothesis tests on the mean $H_0 : t_1 = t_2 = ... = t_n = 0$
  - ▷ Random effects: hypothesis tests on the variance component $H_0 : \sigma_t^2 = 0$

Mixed Models

- Contain fixed and random terms in the same framework
- When we need mixed models ?

Ridge Regression

ANOVA assumptions

- Experimental errors are normally distributed (normality)
- Equal variances (homogeneity)
- Independence

Other points

- Unbalanced Data
- More than one random effect or residual term
- Observational and hierarchical structure
- Correlated measures (spacial analyzes, pedigree information and etc)

Ridge Regression

ANOVA assumptions

- Experimental errors are normally distributed (normality)
- Equal variances (homogeneity)
- Independence

Other points

- Unbalanced Data
- More than one random effect or residual term
- Observational and hierarchical structure
- Correlated measures (spacial analyzes, pedigree information and etc)

Ridge Regression

Mixed Model

$$y = Xb + Zu + e$$

Henderson's equation:

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + G^{-1}\lambda \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix}$$

Solution for fixed and random terms:

$$\hat{b} = (X'V^{-1}X)^{-1}XV^{-1}y$$

$$\hat{u} = GZ'V^{-1}(y - X\hat{b})$$

Ridge Regression

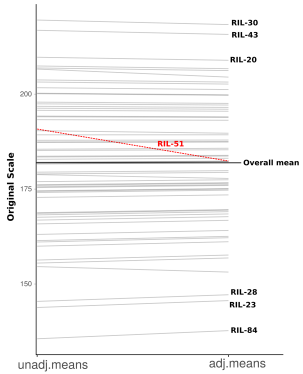
Why predicted random values are shrunk (regularized) estimates?

- $\hat{\beta}$ is the ordinary difference between a treatment mean and the overall mean. It is called the best linear unbiased estimate (BLUE).
- Prediction of random effects ($\hat{u}$) is called the best linear unbiased predictor (BLUPs)
- BLUPs are penalized for the h^2 and number of observations

$$BLUP_k = BLUE_k \times shrinkagefactor_k$$

$$BLUP_k = (\hat{\mu}_k - \hat{\mu}) \times \left(\frac{\hat{\sigma}_g^2}{\hat{\sigma}_g^2 + \frac{\hat{\sigma}_e^2}{r_k}} \right)$$

Ridge Regression



What is the importance of such penalties?

- Mixed models provides a way of building the pessimism of the plant breeder into a formal analysis by using the shrinkage factor
- Amount of shrinkage is large when:
  - ▷ Genetic variance is small
  - ▷ Residual variance is large
  - ▷ Number of reps is small
- Example: experiment in block design, where I excluded some reps for the RIL-51

RR-BLUP

Introduction

- Several methods that fit SNP effects as random effects have been presented in the literature
- So-called RR-BLUP or SNP-BLUP
- Same idea of Ridge Regression but the penalty has a specific form: $\lambda = \frac{\sigma_e^2}{\sigma_a^2}$

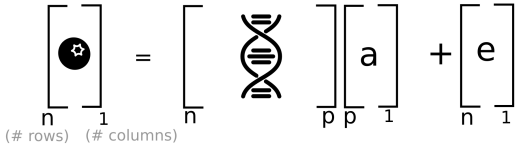
$$y = Xb + Za + e$$

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + I^{-1}\lambda \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{a} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix}$$

RR-BLUP

RR-BLUP

Matrix Format: Y = XB + e

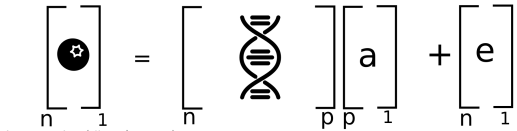


n=# individuals
p= # of markers

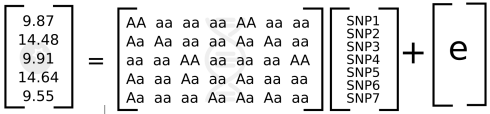
RR-BLUP

RR-BLUP

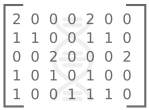
Matrix Format: Y = XB + e



n=# individuals
p= # of markers



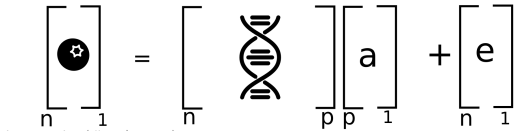
Numeric format, "number of reference alleles"



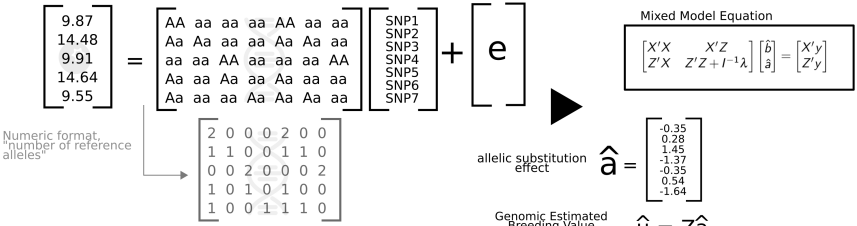
RR-BLUP

RR-BLUP

Matrix Format: Y = XB + e



n=# individuals
p= # of markers



RR-BLUP

RR-BLUP

Matrix Format: Y = XB + e

$$\begin{bmatrix} \\ \\ \\ \end{bmatrix}_{n\ 1} = \begin{bmatrix} \\ \\ \\ \end{bmatrix}_{n} \begin{bmatrix} \\ \\ \\ \end{bmatrix}_{p\ p} \begin{bmatrix} a \\ \\ \\ \end{bmatrix}_{p\ 1} + \begin{bmatrix} e \\ \\ \\ \end{bmatrix}_{n\ 1}$$

(# rows) (# columns)

n=# individuals
p= # of markers

$$\begin{bmatrix} 9.87 \\ 14.48 \\ 9.91 \\ 14.64 \\ 9.55 \end{bmatrix} = \begin{bmatrix} AA & aa & aa & aa & AA & aa & aa \\ Aa & Aa & aa & aa & Aa & Aa & aa \\ aa & aa & AA & aa & aa & aa & AA \\ Aa & aa & Aa & aa & Aa & aa & aa \\ Aa & aa & aa & Aa & Aa & Aa & aa \end{bmatrix} \begin{bmatrix} SNP1 \\ SNP2 \\ SNP3 \\ SNP4 \\ SNP5 \\ SNP6 \\ SNP7 \end{bmatrix} + \begin{bmatrix} e \\ \\ \\ \\ \end{bmatrix}$$

Numeric format, "number of reference alleles"

$$\begin{bmatrix} 2 & 0 & 0 & 0 & 2 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 & 2 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 1 & 0 \end{bmatrix}$$

Mixed Model Equation

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + I^{-1}\lambda \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{a} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix}$$

allelic substitution effect

$$\hat{a} = \begin{bmatrix} -0.35 \\ 0.28 \\ 1.45 \\ -1.37 \\ -0.35 \\ 0.54 \\ -1.64 \end{bmatrix}$$

Genomic Estimated Breeding Value (GEBV)

$$\hat{u} = Z\hat{a}$$

Example:

$$Ind_x = \begin{bmatrix} 2 & 1 & 0 & 0 & 2 & 1 & 1 \end{bmatrix}$$

$\hat{u}$ = [2 x (-0.35)] + [1 x 0.28] + [0 x 1.45] + [0 x (-1.37)] + [2 x (-0.35)] + [1 x 0.54] + [1 x (-1.64)]

RR-BLUP

Notes about RR-BLUP

- GEBV are predicted in two steps:
  - ▷ Estimate the marker effects ($\hat{a}$)
  - ▷ We compute the GEBV by multiplying the vector of marker effect and the genomic matrix ($\hat{u} = Z\hat{a}$)
- Assumes normal distribution for SNP effects and constant variance, which means all SNP explain the same proportion of variance on the trait.

Mixed Model for computing SNP Effects

There are two equivalent models in the genomic literature, when markers are assumed normally distributes

- SNP-BLUP: marker and GEBV are estimated separately, in two steps.
- GBLUP: breeding values are estimated directly

RR-BLUP

Notes about RR-BLUP

- GEBV are predicted in two steps:
  - ▷ Estimate the marker effects ($\hat{a}$)
  - ▷ We compute the GEBV by multiplying the vector of marker effect and the genomic matrix ($\hat{u} = Z\hat{a}$)
- Assumes normal distribution for SNP effects and constant variance, which means all SNP explain the same proportion of variance on the trait.

Mixed Model for computing SNP Effects

There are two equivalent models in the genomic literature, when markers are assumed normally distributes

- SNP-BLUP: marker and GEBV are estimated separately, in two steps.
- GBLUP: breeding values are estimated directly

RR-BLUP

GBLUP

$$y = \mu + Zu + e$$

$$u \sim N(0, \sigma_u^2 G)$$

$$\begin{bmatrix} 1'1 & 1'Z \\ 1'X & Z'Z + \sigma_e^2 \sigma_u^2 G^{-1} \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{a} \end{bmatrix} = \begin{bmatrix} X'y \\ 1'y \end{bmatrix}$$

$$\hat{u} = (Z'Z + G^{-1} \frac{\sigma_e^2}{\sigma_u^2})^{-1} Z'(y - 1\hat{\mu})$$

RR-BLUP

$$y = \mu + \sum M_i a_i + e$$

$$a_i \sim N(0, \sigma^2)$$

$$\begin{bmatrix} 1'1 & 1'Z \\ 1'X & Z'Z + \sigma_e^2 (var(\sum M_i a_i))^{-1} \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{a} \end{bmatrix} = \begin{bmatrix} X'y \\ 1'y \end{bmatrix}$$

$$var(\sum M_i a_i) = \sum var(M_i a_i) = \sum M_i' M_i \sigma_a^2 = \sigma_u^2 G$$

Important

- GBLUP and RR-BLUP are equivalent models
- GBLUP has the advantage that existing software can be used
- First step, in this case is to compute the **G** matrix
- When **G** matrix is plugged in the mixed model equation, the GEBV are predicted directly.

RR-BLUP

GBLUP

$$y = \mu + Zu + e$$

$$u \sim N(0, \sigma_u^2 G)$$

$$\begin{bmatrix} 1'1 & 1'Z \\ 1'X & Z'Z + \sigma_e^2 \sigma_u^2 G^{-1} \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{a} \end{bmatrix} = \begin{bmatrix} X'y \\ 1'y \end{bmatrix}$$

$$\hat{u} = (Z'Z + G^{-1} \frac{\sigma_e^2}{\sigma_u^2})^{-1} Z'(y - 1\hat{\mu})$$

RR-BLUP

$$y = \mu + \sum M_i a_i + e$$

$$a_i \sim N(0, \sigma^2)$$

$$\begin{bmatrix} 1'1 & 1'Z \\ 1'X & Z'Z + \sigma_e^2 (var(\sum M_i a_i))^{-1} \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{a} \end{bmatrix} = \begin{bmatrix} X'y \\ 1'y \end{bmatrix}$$

$$var(\sum M_i a_i) = \sum var(M_i a_i) = \sum M_i' M_i \sigma_a^2 = \sigma_u^2 G$$

Important

- GBLUP and RR-BLUP are equivalent models
- GBLUP has the advantage that existing software can be used
- First step, in this case is to compute the **G** matrix
- When **G** matrix is plugged in the mixed model equation, the GEBV are predicted directly.

Conclusions

What we learned

- Discussed the importance to divide our data set in training and test when studying $Y = f(x) + e$
- To define $f(x)$, multiple regression does not work when $p \gg n$
- We need to use some kind of regularization
- Ridge regression is the most common
- RR is equivalent to GBLUP

Next class

- Discuss theoretical and practical aspects related to GS implementation
- Hands-on: implementing RRBLUP

Conclusions

What we learned

- Discussed the importance to divide our data set in training and test when studying $Y = f(x) + e$
- To define $f(x)$, multiple regression does not work when $p \gg n$
- We need to use some kind of regularization
- Ridge regression is the most common
- RR is equivalent to GBLUP

Next class

- Discuss theoretical and practical aspects related to GS implementation
- Hands-on: implementing RRBLUP

References

References

- Lynch and Walsh, 1998. Genetics and Analysis of Quantitative Traits
  Book
- Bernardo, 2010. Breeding for Quantitative Traits in Plants
  Book – Second edition
- Falconer and Mackay, 1996. Introduction to Quantitative Genetics
  Book – Fourth edition
- Cruz, 2005. Princípios de Genética Quantitativa.
  Book – Portuguese version
- Hamilton, 2009. Population Genetics.
  Book – First edition