May 2023

Linear Regression with Regularization on the Genetic Architecture of Maize Flowering Time

Roland Fiagbe
University of Central Florida, fiagberoland@Knights.ucf.edu

Linear Regression with Regularization on the Genetic Architecture of Maize Flowering Time*

Roland Fiagbe
Department of Statistics and Data Science
University of Central Florida
Orlando, United States
fiagberoland@knights.ucf.edu

***Abstract*—Over a century, the maize crop has been one of the most important crop species that is targeted for genetic investigations and experiments. One of the major experiments that have been a topic of interest is crossing inbred lines to produce better offspring through a process called heterosis. Crossing the inbred lines create numerous SNP markers that determine the time to male flowering. This project seeks to explore the SNP markers to select the most relevant ones for predicting time to male flowering using linear regression with regularization methods due to the fact that $p > n$ in our dataset. Various regularization methods were employed and compared. The l_1-norm regularization method (LASSO) was chosen as the best regularization method for our data.**

***Index Terms*—SNPs, maize, regularization, crossing**

I. Introduction

For over a century, maize has been one of the most important crop species that is considered to be the target of genetics experiments and investigations [2]. The term maize is frequently used synonymously with corn in some parts of the world. However, the term maize is often referred to as its plant. Maize is one of the largest grain plants that emerged from its wild-grass ancestors due to human agriculture intervention [4]. There are many varieties that can be distinguished by their physical characteristics but generally, they grow as a single-stalk plant to approximately 8 feet tall with about 20 long narrow single leaves.

However, maize is one of the species that are naturally outcrossing and its genetic architecture is not different from other outcrossing organisms like human beings [5]. Through a process called *heterosis*, two inbred strains can be crossed to produce better offspring. This process has been adapted in an attempt to map the genetic loci but few among those attempts came out successful [3]. Buckler et al. [2] in their paper created a genomic map of maize that shows the combined genome structures. Their research discovered that the difference in maize flowering time between the inbred strains was caused by the additive effects of many quantitative trait loci and each of them has a minor effect on the trait [1].

In this project, due to the vast amount of SNP markers resulting from crossing inbred lines, we want to perform variable selection to select the more relevant variables that contribute to the time to male flowering (measured in days) and perform linear regression. We will perform the linear regression with regularization due to the fact that $p > n$ for our dataset.

II. Data

The data used in this project is a popular dataset from [2] at https://www4.stat.ncsu.edu/~boos/var.select/maize.html. The dataset was originally provided by the Funda Ogut of the NC State Department of Crop Science. From 25 crosses, each with about 200 recombinant inbred lines, the time to male flowering (dtoa) was measured along with marker data. The dataset contains 4981 observations and 7393 variables with some missing values. However, in the dataset, there are 7389 independent variables representing the SNP markers and a response variable DtoA (time to male flowering) recorded in a number of days. For the purpose of analysis, we will drop the following variables; gene_code, pop, and Entry.

III. Exploratory Data Analysis

In this section, we will conduct some exploratory data analysis to have a visual understanding of the response variable (DtoA) and how it can be dealt with to create the best predictive model. We begin by inspecting the distribution of the response variable (DtoA). Table I gives some basic descriptions of the response variable.

Min	Q1	Median	Mean	Q3	Max
66.02	74.85	77.33	77.16	79.43	91.23

TABLE I
Summary Statistics of Days to Male Flowering (DtoA)

From the summary statistics (table I), the mean and median are approximately equal which suggests the normality of the response variable. However, to further visually investigate the distribution of the target variable $(DtoA)$, we want to look at its histogram and Q-Q plot. Figure 1 and 2 show the histogram and density curve and the Q-Q plot of the response. These confirm that the variable is normally distributed and this will work best for the linear regression model.

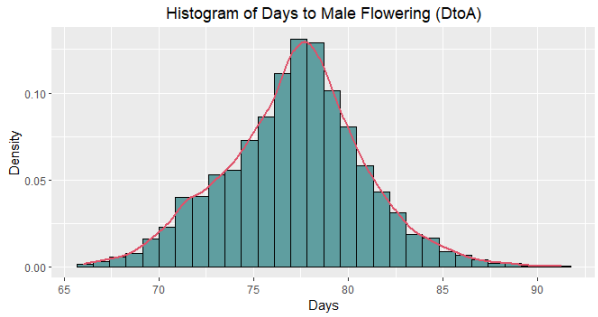

Fig. 1. Histogram of the Response Variable (DtoA)

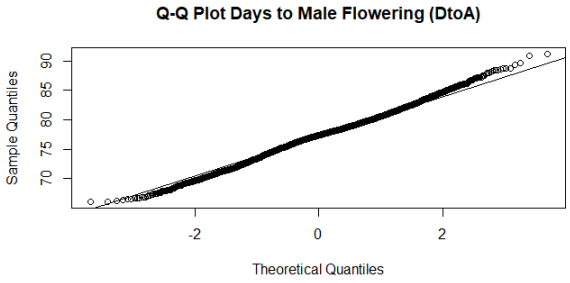

Fig. 2. Q-Q Plot of the Response Variable (DtoA)

IV. METHODOLOGY

In this project, we will employ the Regression model using regularization (LASSO, Ridge and Elastic Net) to perform variable selection and predict the number of days to male flowering based on the SNP markers.

A. Linear Regression Model

The regression model assumes

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon \tag{1}$$

where
Y is the response variable (DtoA)
β_0 is the intercept
β_i is the regression coefficient associated with X_i
ϵ is the error term

The regression model is optimized by using the Least Square Estimation to find the parameter estimates $\hat{\beta}_i$ associated with X_i.
The estimates of coefficients are estimated by minimizing the residual sum of squares (RSS). By using matrix representation,

$$Y = X\beta + \epsilon \tag{2}$$

The least square method minimizes

$$\begin{aligned} RSS(\beta) &= \sum_{i=1}^{n} (y_i - \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p)^2 \\ &= ||y - X\beta||^2 \end{aligned} \tag{3}$$

$$\begin{aligned} \frac{\partial RSS(\beta)}{\partial \beta} &= X^T(y - X\beta) = 0 \\ \implies \hat{\beta} &= (X^T X)^{-1} X^T y \end{aligned} \tag{4}$$

Now the estimated regression model is

$$\hat{y} = X\hat{\beta} \tag{5}$$

B. LASSO

Considering the standard linear regression model (explained in IV-A)

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

and considering the fact that $p > n$ in our dataset, the parameters β can be estimated by the linear regression with regularization method. The penalized regression function is given as

$$\hat{\beta}_\lambda = \underset{\beta}{\operatorname{argmin}} ||y - X\beta||^2 + P(\lambda, \beta)$$

where $P(\lambda, \beta)$ is a general penalty function with regularization parameter λ. Lasso applies l_1-norm penalty ($P(\lambda, \beta) = \lambda||\beta||$) to regularize the regression coefficients. The lasso penalized least squares is

$$\hat{\beta}_\lambda = \underset{\beta}{\operatorname{argmin}} ||y - X\beta||^2 + \lambda||\beta||_1$$

The function is a convex optimization function and it results in a non-linear regression problem in y. λ is the regularized parameter (tuning parameter) that controls the amount of shrinkage of the coefficients. In lasso, the l_1-norm shrinks redundant or less useful coefficients to zero [6].

C. Ridge Regression

On the other hand, ridge regression applies l_2-norm penalty ($P(\lambda, \beta) = \lambda||\beta||^2$) to regularize the regression coefficients. The ridge penalized least squares is

$$\hat{\beta}_\lambda = \underset{\beta}{\operatorname{argmin}} ||y - X\beta||^2 + \lambda||\beta||_2^2$$

Here, λ is the regularized parameter (tuning parameter) that controls the amount of shrinkage of the coefficients. This method shrinks coefficients of redundant predictors towards zero but no variable is set to zero. This shrinkage results in biased estimates with low variance [6].

D. Elastic Net

The elastic net method is made up of a combination of lasso and ridge regression penalties. Its penalized least squares is given by

$$\underset{\beta}{\operatorname{argmin}} ||y - X\beta||^2 + \lambda[(1-\alpha)/2||\beta||_2^2 + \alpha||\beta||_1]$$

where $||\beta||_1 = \sum_{j=1}^{p} |\beta_j|$ and $||\beta||_2^2 = \sqrt{\sum_{j=1}^{p} \beta_j^2}$
These two penalties control the amount of shrinkage of the coefficients. The new penalty applies $\frac{\lambda(1-\alpha)}{2}$ to the ridge penalty and $\lambda\alpha$ to the lasso penalty. These two penalties λ and α are tuned to select the best choice. α takes value between 0 and 1. Setting $\alpha = 1$ results in the lasso and $\alpha = 0$ results in the ridge.

V. Variable Selection

In this section, we applied lasso, ridge, and elastic net methods to select relevant features for the regression model. These regularization methods were applied individually to the data in aid of selecting the best possible variables. We applied cross-validation in all the methods to select the best optimal parameters (λ and α) for the models. Unlike the direct variable selection methods (i.e lasso and elastic net), we performed variable screening for ridge regression by setting a cutoff point to the coefficients to eliminate less useful features. The dataset contains 487 missing values for all the features hence was removed and further partitioned into $80, 20$ for the training set (3595 obs) and testing set (899 obs) respectively.

A. LASSO

In this algorithm, we set up the model using the glmnet function by setting $\alpha = 1$. We first performed 10-fold cross-validation with 500 possible values of λ to select the best penalty for the data. The two lasso penalty values, lambda.1se and lambda.min were both considered where $lambda.1se = 0.2441375$ and $lambda.min = 0.05374275$. Lambda.1se selected 19 features and lambda.min selected 132 features including the intercept. The variables selected were used to train the regression model.

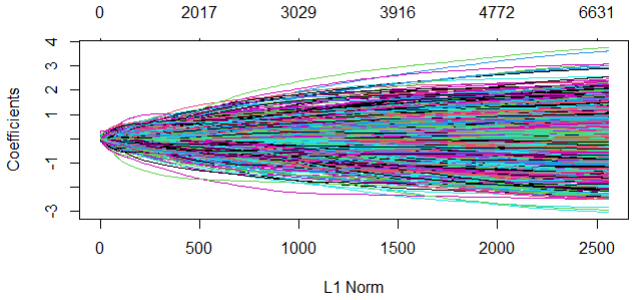

Fig. 3. LASSO plot of variable shrinkage

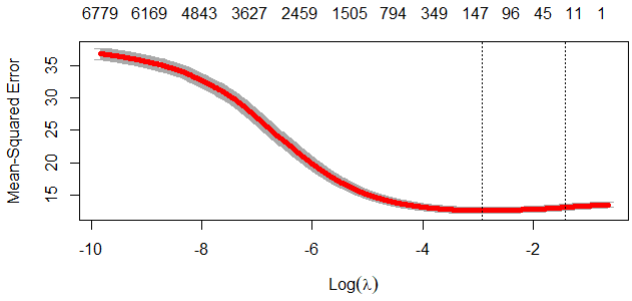

Fig. 4. Plot of the Mean-Squared Error (MSE) and the number of features (SNP) in the model for 10-fold cross-validation for LASSO

B. Ridge Regression

The ridge algorithm was also set up using the glmnet function by setting $\alpha = 0$. 10-fold cross-validation was performed to select the best λ value. Again, both lambda.1se and lambda.min were considered. In this case, we performed variable screening by setting a cutoff point of 0.002 to lambda.1se and 0.005 to lambda.min. 17 and 33 features (SNPs) were selected respectively to train the regression model.

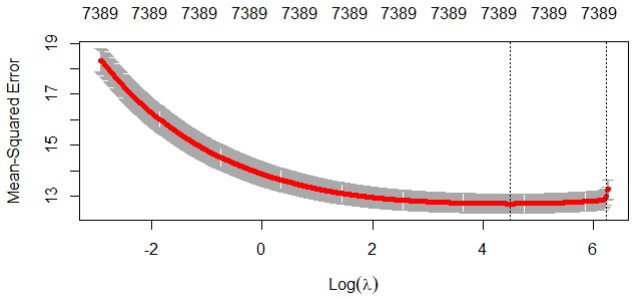

Fig. 5. Plot of the Mean-Squared Error (MSE) and the number of features (SNP) in the model for 10-fold cross-validation

C. Elastic Net

This algorithm was also set up using the glmnet function. Instead of specifying the hyperparameter α, we performed a 10-fold cross-validation on a wide range of α values to select the optimal α with the lowest MSE. The best α with the lowest MSE is $\alpha = 0.7$. Now, we proceed to perform 10-fold cross-validation to select the best λ value. Again, both lambda.1se and lambda.min were considered for feature selection where $lambda.1se = 0.2899854$ and $lambda.min = 0.07537127$. Lambda.1se selected 38 features and lambda.min selected 154 features including the intercept.

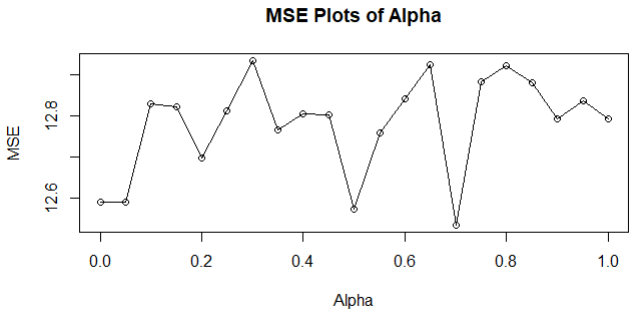

Fig. 6. Plot of the Mean-Squared Error (MSE) against α values using 10-fold cross-validation

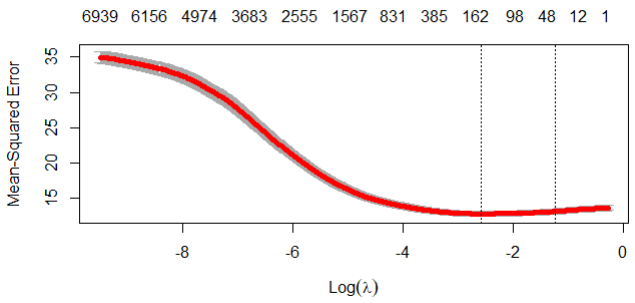

Fig. 7. Plot of the Mean-Squared Error (MSE) and the number of features (SNP) in the model for 10-fold cross-validation for Elastic Net

VI. RESULTS

Table II gives the results for comparing the performance of the three regularization methods. For the models' setup, the best tuning parameters λ for lasso and ridge regression are 0.2441375 and 510.8293 respectively for 1 standard error (lambda.1se) and 0.05374275 and 90.10913 respectively for minimum MSE (lambda.min). For elastic net, $\alpha = 0.7$ and the best λ is 0.2899854 and 0.07537127 for 1 standard error (lambda.1se) and minimum MSE (lambda.min) respectively.

Table II shows the number of features (SNPs) selected by each model with their Mean Square Errors (MSE) and Root Mean Square Errors (RMSE). These two measures were employed to explore the prediction accuracy (performance) of our models. From the results, we could see that LASSO with 1 standard error (lambda.1se) method produced the smallest MSE and RMSE, hence will be considered the best performing model, although there is no significant difference between the MSE and RMSE of all the models.

Models	Methods	SNPs Selected	Model MSE	Model RMSE
LASSO	Lmabda.1se	19	12.44893	3.528304
	Lambda.min	132	12.90047	3.591722
Ridge	Lmabda.1se	17	13.09461	3.618647
	Lambda.min	33	13.20899	3.634418
Elastic Net	Lmabda.1se	38	12.61774	3.552146
	Lambda.min	154	12.48514	3.533432

TABLE II
RESULTS FROM THE ANALYSIS USING THE THREE MODELS

VII. CONCLUSION

In this project, we have developed a suitable regression model with regularization that predicts the phenotype (time to male flowering (DtoA)) of maize plants based on its useful SNP markers from different inbred lines. Various models applied in this study performed approximately the same but the lasso regularization method was selected as the best-performing model with fewer features (SNPs). The SNPs selected by this model have a higher contribution in predicting the phenotype (time to male flowering (DtoA)).

REFERENCES

[1] Amir Alipour Yengejeh. "Genome-Wide Association Study of The Maize Crop by The Lasso Regression Analysis". In: (2023).

[2] Edward S Buckler et al. "The genetic architecture of maize flowering time". In: *Science* 325.5941 (2009), pp. 714–718.

[3] Trudy FC Mackay. "A-maize-ing diversity". In: *Science* 325.5941 (2009), pp. 688–689.

[4] Natalie J Nannas and R Kelly Dawe. "Genetic and genomic toolbox of Zea mays". In: *Genetics* 199.3 (2015), pp. 655–669.

[5] Antoni Rafalski and Michele Morgante. "Corn and humans: recombination and linkage disequilibrium in two genomes of similar size". In: *TRENDS in Genetics* 20.2 (2004), pp. 103–111.

[6] Patrik Waldmann et al. "Evaluation of the lasso and the elastic net in genome-wide association studies". In: *Frontiers in genetics* 4 (2013), p. 270.