Study of *Oryza sativa* genotypes involved in plant height using genome-wide association

Adib HABBOU, Alae KHIDOUR – Group 1 – Bio Project

Variable Selection:

After dealing with all the missing values in our datasets we still have too many variables with only few observations, so we need to do some variable selection to have regression models with significant results.

Simple Regression:

Applying a simple linear regression to our data shows us that some estimated coefficients have missing values because of an exact linear relationship between the variables known as perfect multicollinearity. It's mainly because of the pseudo-inverse matrix computation. We decided to drop the variables with missing values for coefficients to avoid having highly correlated variables. The computation of the coefficient follows: $\beta_{OLS} = argmin_{\beta} \sum_{i=0}^{n} (y_i - \beta . x_i)^2$

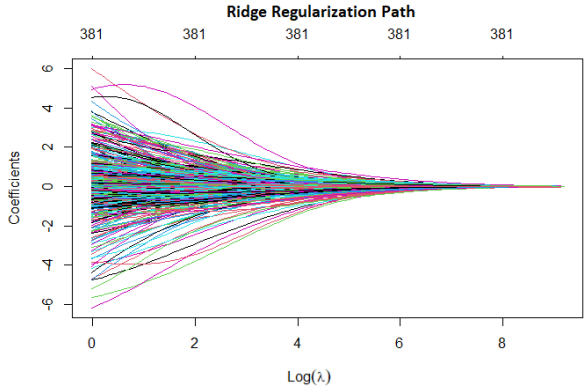
Forward and Stepwise Regression:

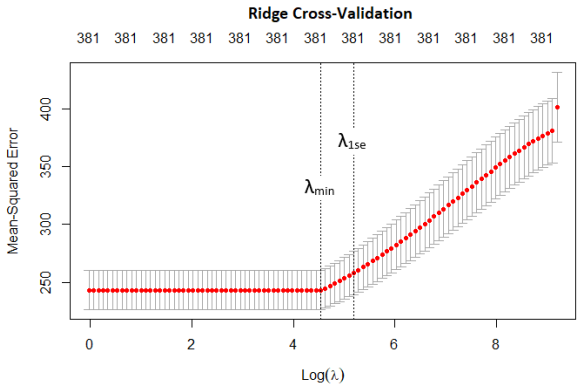
Both methods give us an R^2 equal to 1 which means that our model is overfitting since he perfectly fit the target data he trained on. So, we need to go further and use penalized regression methods to avoid overfitting and obtain a better model. Knowing that both models give us the same output, we are going to keep only the variables that the forward and stepwise regression has selected. On the other hand, the backward was not possible due to the number of variables.

Penalized Methods:

Ridge Regression:

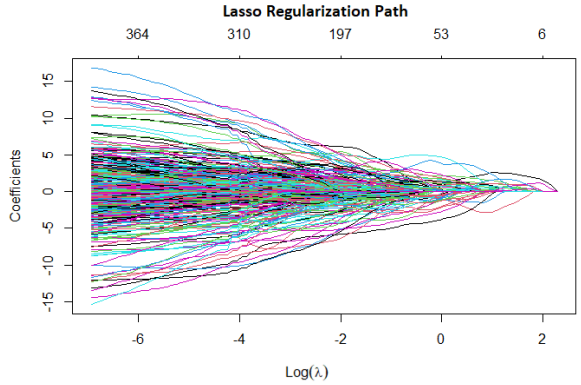
All coefficients decrease until being all equal to zero for the same value λ. Which means the resulting model always includes all the variables and that can be a problem for large datasets specifically in our case where the goal is to identify the genotypes involved in a selected phenotype. The coefficients are computed using the following formula for various values of λ with an L2-norm penalization: $\beta_{RIDGE} = argmin_{\beta} \sum_{i=0}^{n} (y_i - \beta . x_i)^2 + \lambda . ||\beta||_2^2$

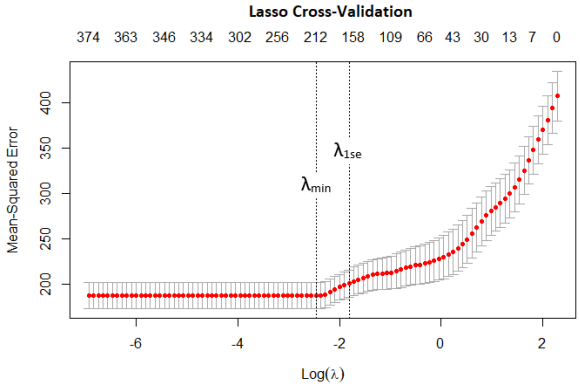




Lasso Regression:

Lasso is mainly useful in high dimensional datasets, where there are more variables than observations, but we only expect a small part of the variables to be truly meaningful. The coefficients become equal to zero one by one. The advantage of Lasso is that it does variable selection, for a given value λ some variable coefficients will be equal to zero. We penalize the function to minimize with an L1-norm such as: $\beta_{LASSO} = argmin_{\beta} \sum_{i=0}^{n} (y_i - \beta . x_i)^2 + \lambda . ||\beta||_1$

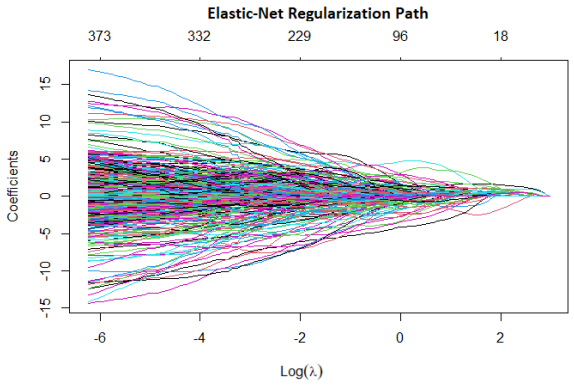


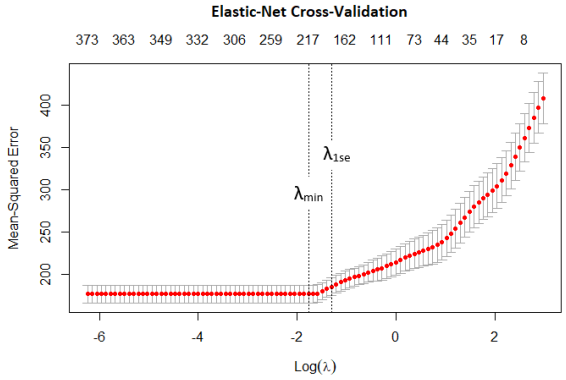


Sophisticated Methods:

Elastic-Net:

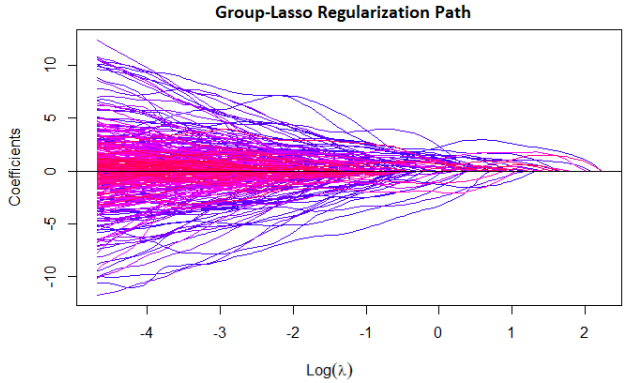
After going through various values for α and λ we found that the couple that minimizes the most the R^2 value is α = 1 (which corresponds to Lasso) with λ = λ_{min}. However, we have chosen α = 0.5 penalize the likelihood with an L1-term and an L2-term at the same time following the formula: $\beta_{ELASTIC-NET} = argmin_{\beta} \sum_{i=0}^{n} (y_i - \beta . x_i)^2 + \lambda \left(\alpha . ||\beta||_1 + \frac{1-\alpha}{2} . ||\beta||_2^2 \right)$

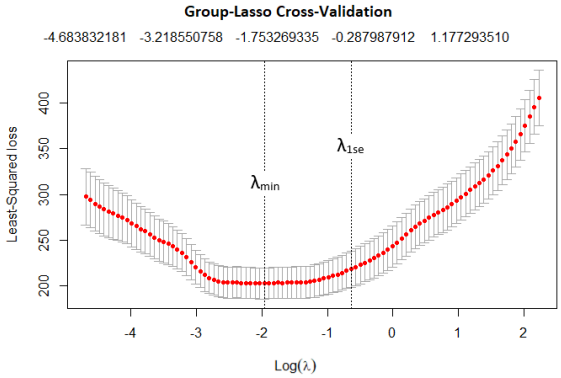




Group-Lasso:

Lasso gives us quite good results. However, there are situations in which the variables have a natural grouped structure. Group-Lasso is built as the sum of squares of coefficients belonging to the same group. This way it considers the possible grouped structure of variables, and it sends to zero whole groups of variables. Where G_k is the k-st group found within our variables and p_k is the number of elements in this group: $\beta_{GROUP-LASSO} = \text{argmin}_\beta \sum_{i=0}^{n}(y_i - \beta.x_i)^2 + \lambda \sum_{k=1}^{K} \sqrt{p_k} . \left|\left|\beta_{G_k}\right|\right|_1$





Support Vector Regression:

SVR finds an appropriate hyperplane to fit the data. We look for the hyperplane that fits the best to our observations by looking for the one with the highest margin. The margin is the distance that separates a hyperplane from the closest observation. In contrast to Ordinary Least Square, the objective function of SVR is to minimize the coefficients, more specifically, the L2-norm of the coefficient vector with respect to a constraint: $min_\beta \frac{1}{2}\left|\left|\beta\right|\right|_2^2 \; with\; constraint \;\; |y_i - \beta_i x_i| \leq \varepsilon$

Summary:

To evaluate the performance of our several models, we use the coefficient of determination: $R^2 = 1 - \frac{\left|\left|y - X\hat{\beta}\right|\right|_2^2}{\left|\left|y - \bar{y}\right|\right|_2^2}$

	Ridge	Lasso	Elastic-Net	Group-Lasso	Support Vector Regression
R^2	0.51	0.92	0.85	0.91	0.81

We clearly see that Ridge Regression is a bad choice for our specific case and that the highest R^2 is obtained using Lasso or Group-Lasso which is implied by the fact that only few genotypes are really involved in plant height. Therefore, the R^2 value tells us how good our model is at making predictions of plant height given the information about genotypes.

List of genotypes mainly involved in plant height obtained using Lasso Regression with λ_{min}

"id1000852" "ud7001348" "id7004052" "ud1000154" "id1004393" "id1000051" "id1000994" "id3015622" "wd7002954" "wd3001877" "id7004166" "id3012431"
"id3008445" "id7002916" "id7003169" "id3008476" "ud7001215" "id3011930" "id3014217" "id3011218" "id7003112" "ud7001352" "id3013258" "id1000026"
"id1003248" "id7003361" "id7003332" "id7003377" "ud7001201" "id3014412" "id7002708" "id1000399" "id7002717" "id1000660" "id1002920" "id7002946"
"id1000685" "id1000858" "id1001692" "id3011075" "id4000010" "id3015885" "id7003201" "id1002509" "id7003704" "ud7001582" "id3009868" "id7004086"
"id1000528" "id1000529" "id3008355" "ud7001515" "id1003839" "id3012912" "id3008721" "id3009920" "id7002795" "id7003442" "id1003840" "id7002799"
"id3009602" "id1001049" "ud7001237" "id7002775" "id7003947" "id3011059" "id4000756" "wd1000189" "id1001664" "id1001588" "id7004032" "id1002087"
"id7003555" "wd7003065" "id3016604" "id3009692" "id7003576" "id1002919" "id7002918" "id7002896" "id3017376" "id7003475" "id3010236" "id7003195"
"id3011813" "id3009243" "wd3000757" "id3009276" "id3009331" "id1000007" "id3008454" "id3008833" "id3008397" "id3011640" "id3013989" "id7002923"
"id3012384" "id1001981" "id1001915" "id3008500" "id1001509" "id3011423" "id1001366" "id1000857" "id3011035" "ud7001419" "id7003190" "id4000164"
"ud7001180" "id7003709" "id7003983" "id3010511" "ud3001031" "id1000980" "id3008682" "ud3000950" "id3013397" "id3011115" "id1001516" "id3012248"
"ud7001238" "id7002838" "id3017406" "wd7002936" "dd3000953" "id3017627" "id1000661" "id3010370" "id3017179" "id3011960" "id1000027" "id1001599"
"id1001430" "id1001224" "id7003931" "id1003919" "id3009037" "id3011044" "id1005297" "id1000030" "id3008411" "id3015746" "ud1000187" "id1001266"
"id1001638" "id1004041" "id1002227" "wd7002914" "id3017002" "id7003064" "id7002749" "ud1000074" "id7003060" "id3014618" "id1000731" "id7003641"
"id7004040" "id3009233" "ud7001500" "id7003174" "id7003039" "id7003877" "id1004156" "id3008674" "id1003877" "id3017136" "id3016668" "id7004170"
"id7003467" "id3014850" "id1001128" "id7002788" "id3009780" "wd7002589" "fd10" "id1003368" "id3016043" "id1004633" "id1003465" "id1003115"
"id1001003" "id1001015" "id3008432" "id7003272" "id7003078" "id3009379" "id3009824" "id7002999" "id3016447" "id1001332" "id3017403" "id1004872"
"id3008453" "id1002730" "id3012473" "id3018096" "id1000929" "id1000423" "id3009934" "id1000841" "id1002622" "ud7001431" "id1000011" "id3016565"
"id7003040" "id3009130" "id3008844"

Our final goal was to create a model of the problem. Using the results obtained with Lasso Regression we can select the most important genotypes that are mainly involved in the *Oryza sativa* plant height. We finally obtained a list of 207 genotypes selected among our initial 36 901 genotypes using λ_{min} which is the value that gives us the best R^2. Still, if we want to have fewer genotypes, we can use the model with λ_{1se} which gives a simpler model with only 164 genotypes.