Yuyi Zhang

Apr 28, 2019

Spring 2019 CPS Quarter Term A

Instructor: Steward Huang

Assignment 3

ALY 6015\_Intermediate Analytics

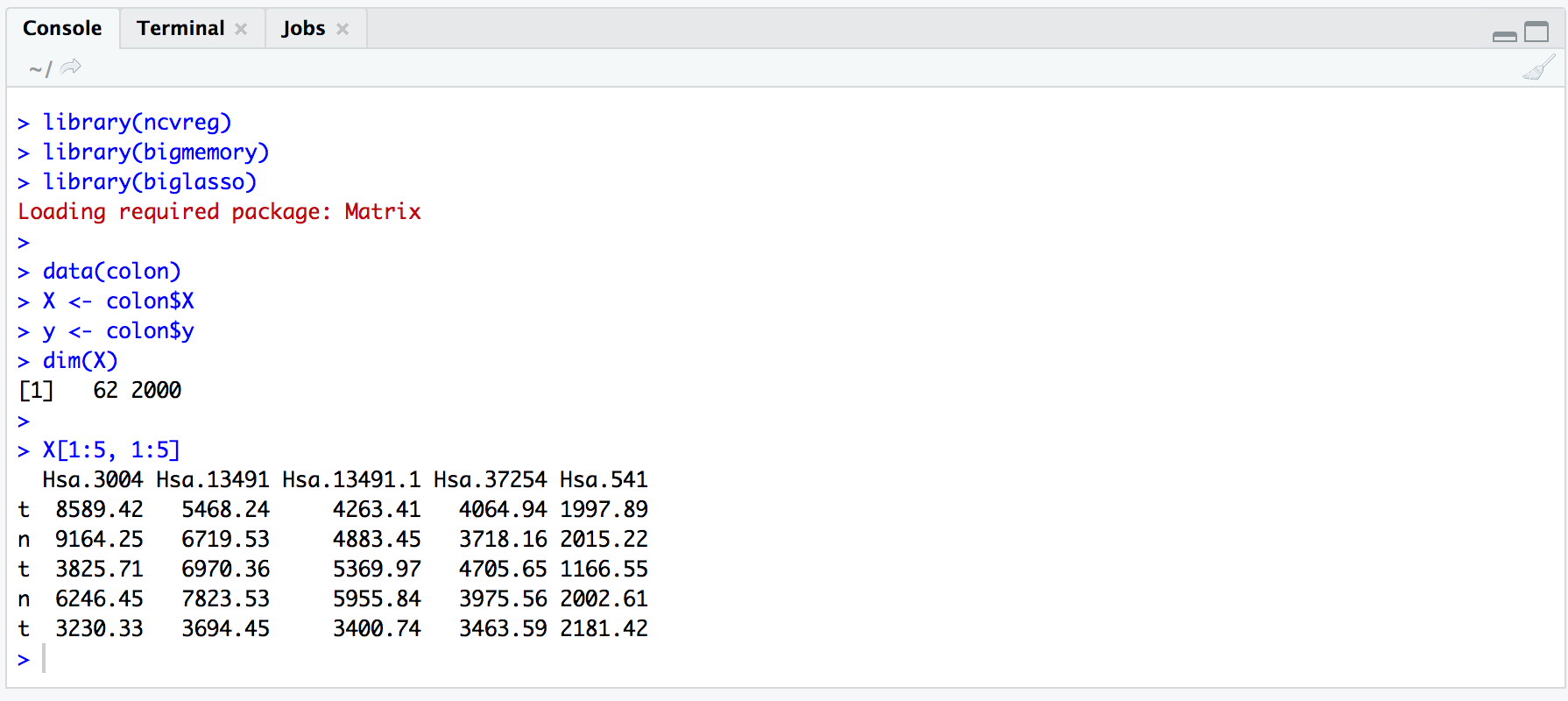
# **Introduction**

In this report, we are going to go through an extending lasso model applying in a small data example. The dataset contains X in matrix of 62\*2000 and y in numbers of 62. Instead of glmnet, this coding sample used biglasso package. The new model is aiming for large size of data. We are going to use it in a real data set colon and see what the conclusion would it be.

# **Analysis**

## **Data Preparation**

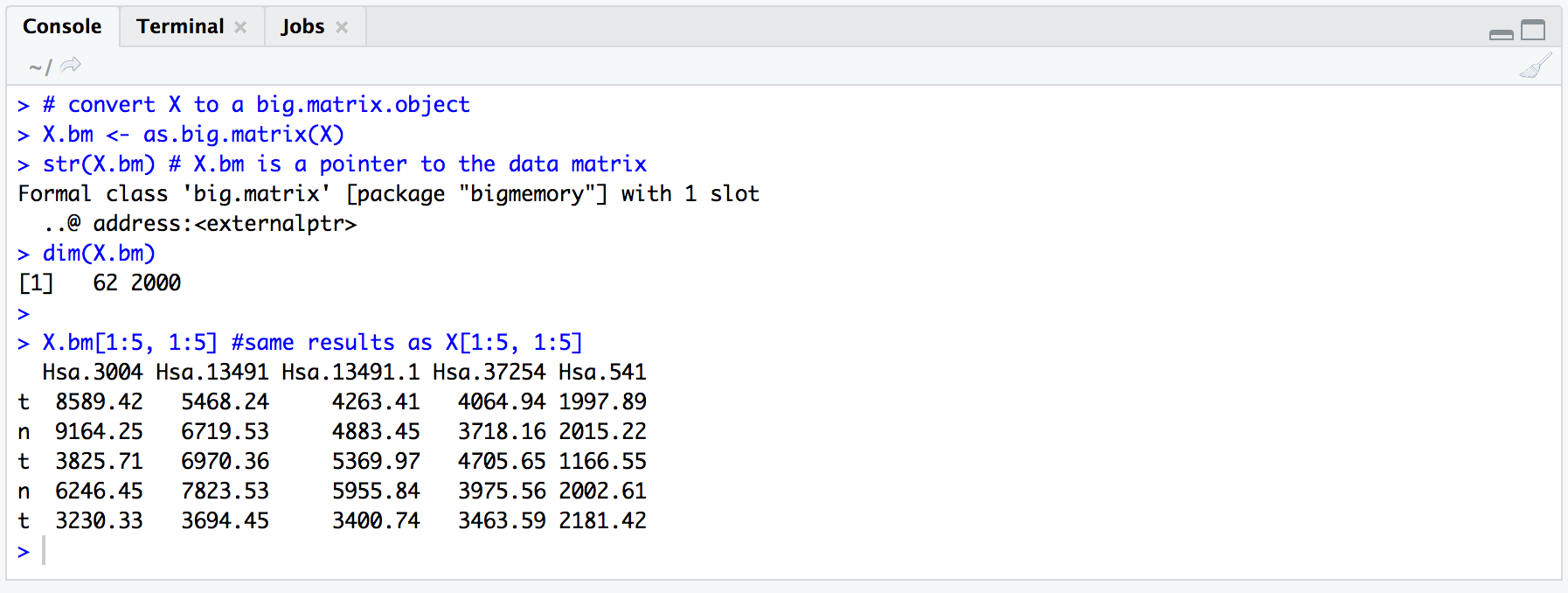
In this phase, we are going to load data set colon and get to see the output of the snippet.



*Figure 1*. Prepare for data

## **Convert X to a big.matrix.object**

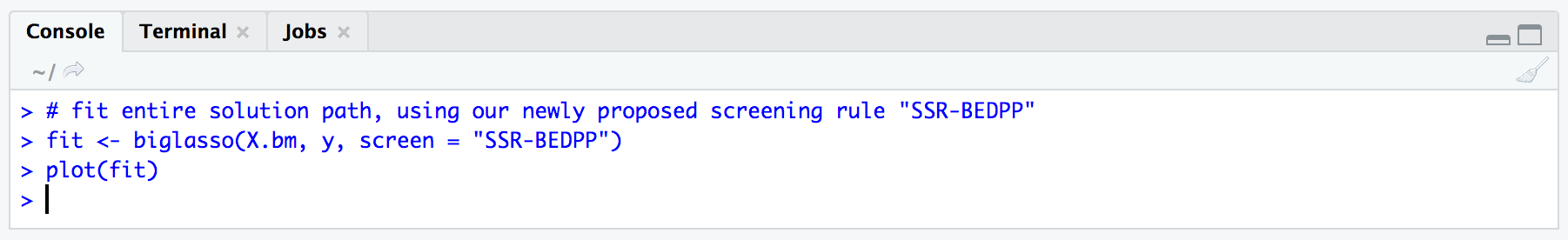
Since we are using the biglasso package, we have to use a big.matrix.object instead of a matrix. As we can see from the figure 2, it didn’t change the size of the original matrix.



*Figure 2*. Converting matrix

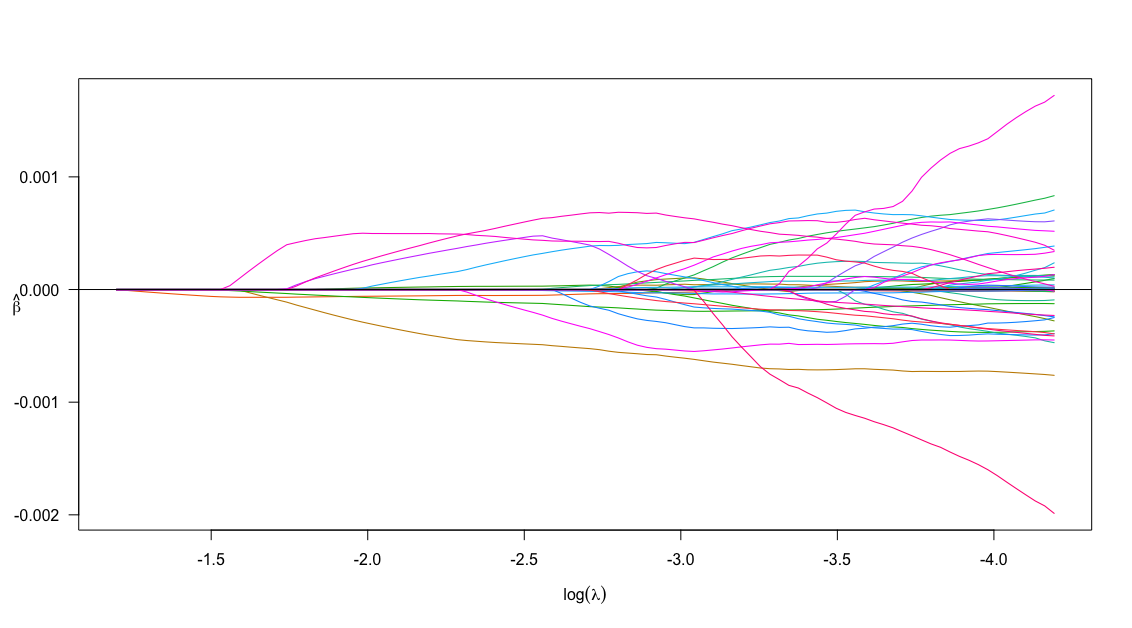
## **Biglasso**

After all data has been properly setup, now we are going to use biglasso model to apply for a single model fitting.



*Figure 3*. Biglasso model

Then we get a fit plot shown as below.

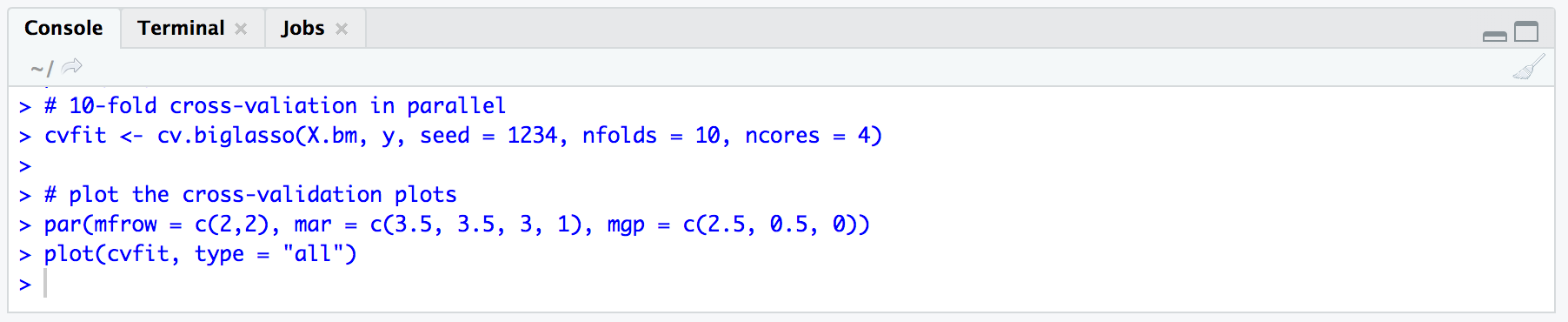


*Figure 4*. plot of fit

In this plot we could see that there are many beta-hats stay in 0, which means they are the ones to be eliminated from the model as unrelative parameters.

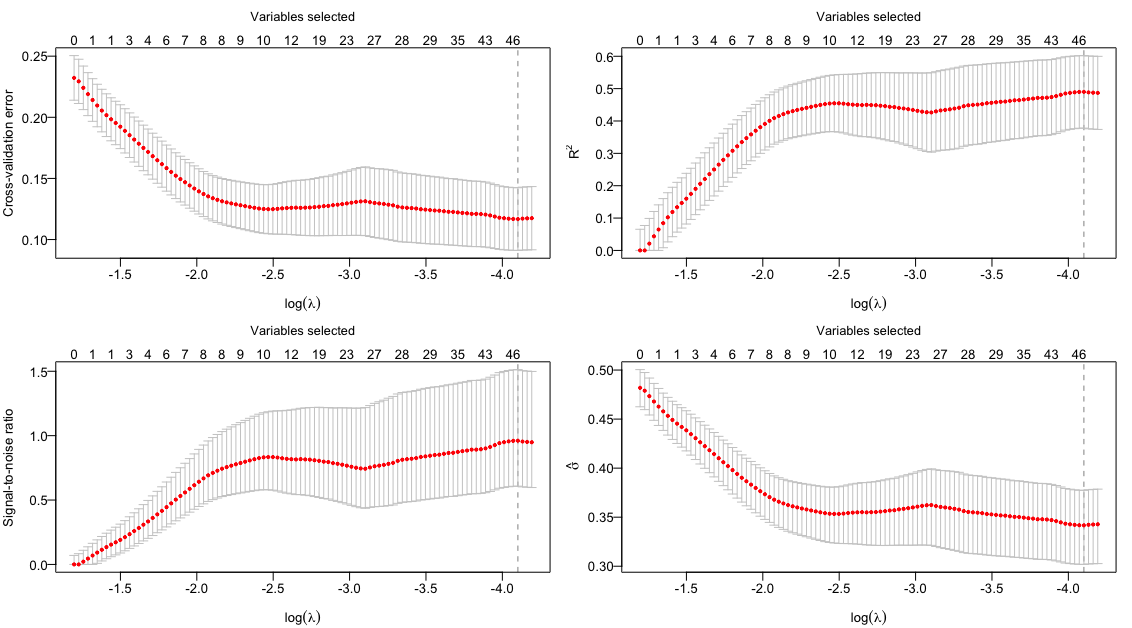
## **10-fold Cross-validation**

After we get the lasso model, now are going to apply a cross validation test for it, and select the parameter λ.



*Figure 5*. 10-fold cross-validation

Then we got 4 cross-validation curves with vertical dashed red line indicates the λ value.

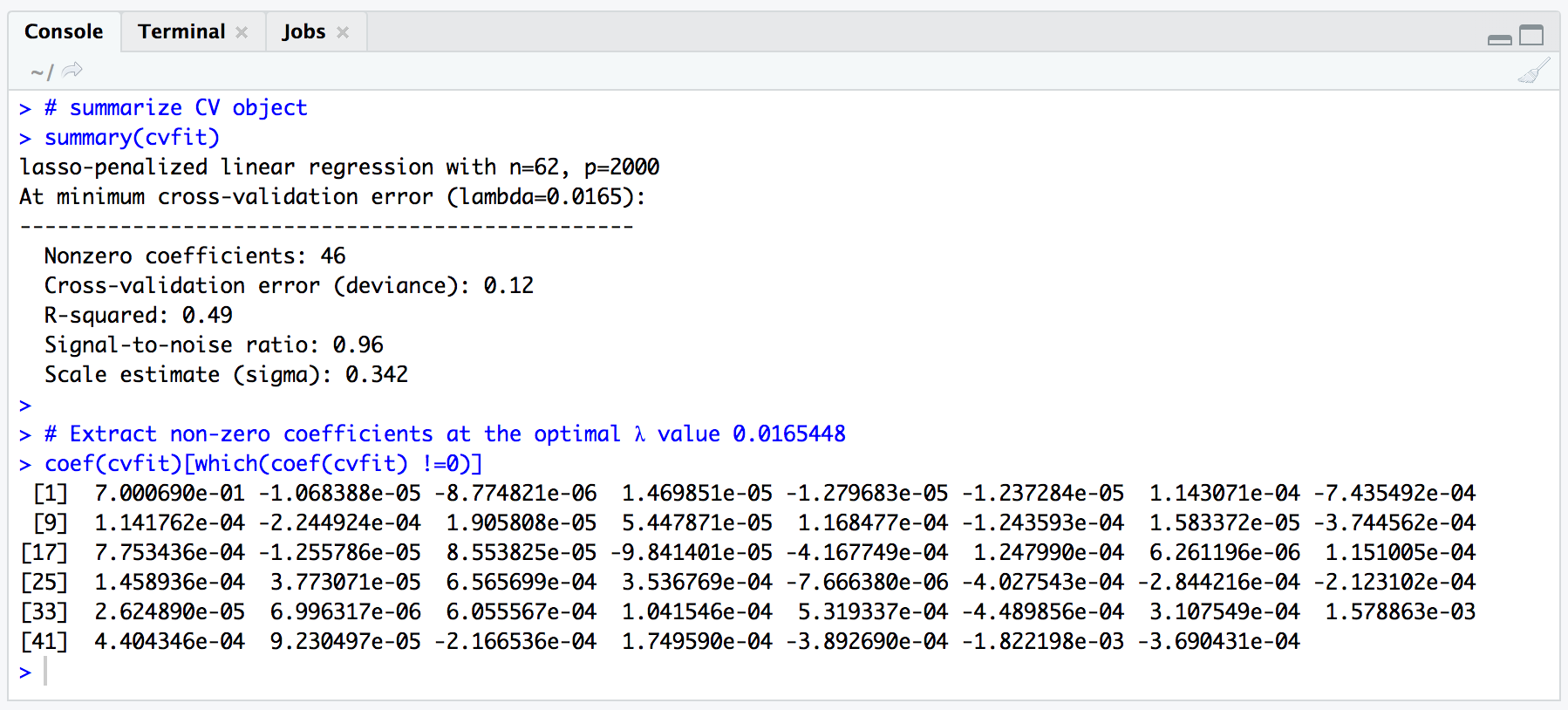


*Figure 6*. plots of cv

In these plots, we can see that the parameters were down size from 2,000 to 46. And all the red line indicators point out the same log(λ) value. We chose the λ value corresponding to the minimum cross-validation error, maximum R2, maximum SNR, and minimum scale estimate sigma.

## **Cv object and λ**

In the last phase, we are going to take a look at the summary of cv indicators and the output of all the nonzero coefficients.



*Figure 7*. CV summary and non-zero coefficients

In the summary we could know that there were 2000 parameters in the initial linear regression. After lasso-penalized, there are only 46 with nonzero coefficients stay. The R-squared is 0.49, which means this model is going to explain almost half of the y according to X in forecast. All the coefficient values are very small (e-04), so the parameter cannot be eliminated anymore.

# **Conclusion**

After tuning, the optimal λ value that we get is 0.0165. Low coefficient value indicates that we’ve already get minimum parameters as 46.

Reference

1. Maindonald, J. H. (2008). *Using R for Data Analysis and Graphics.* Retrieved from <https://cran.r-project.org/doc/contrib/usingR.pdf>
2. Zeng, Y. Breheny, P. (2016). *Biglasso: extending lasso model to Big Data in R*. Retrieved from <https://github.com/YaohuiZeng/biglasso/blob/master/vignettes/biglasso.pdf>
3. Zeng, Y. Breheny, P. (2017). *The biglasso Package: A Memory- and Computation-Efficient Solver for Lasso Model Fitting with Big Data in R*. Retrieved from <https://arxiv.org/pdf/1701.05936.pdf>