

**ASSIGNMENT FRONT SHEET**

**Course Name: ALY6015 20904 Intermediate Analytics**

**Professor Name: ChuanLi Jiang,**

**Student Name: Dong Quoc Tuong (Lukas)**

**Student Class: Fall 2019 CPS Term: A. 2020**

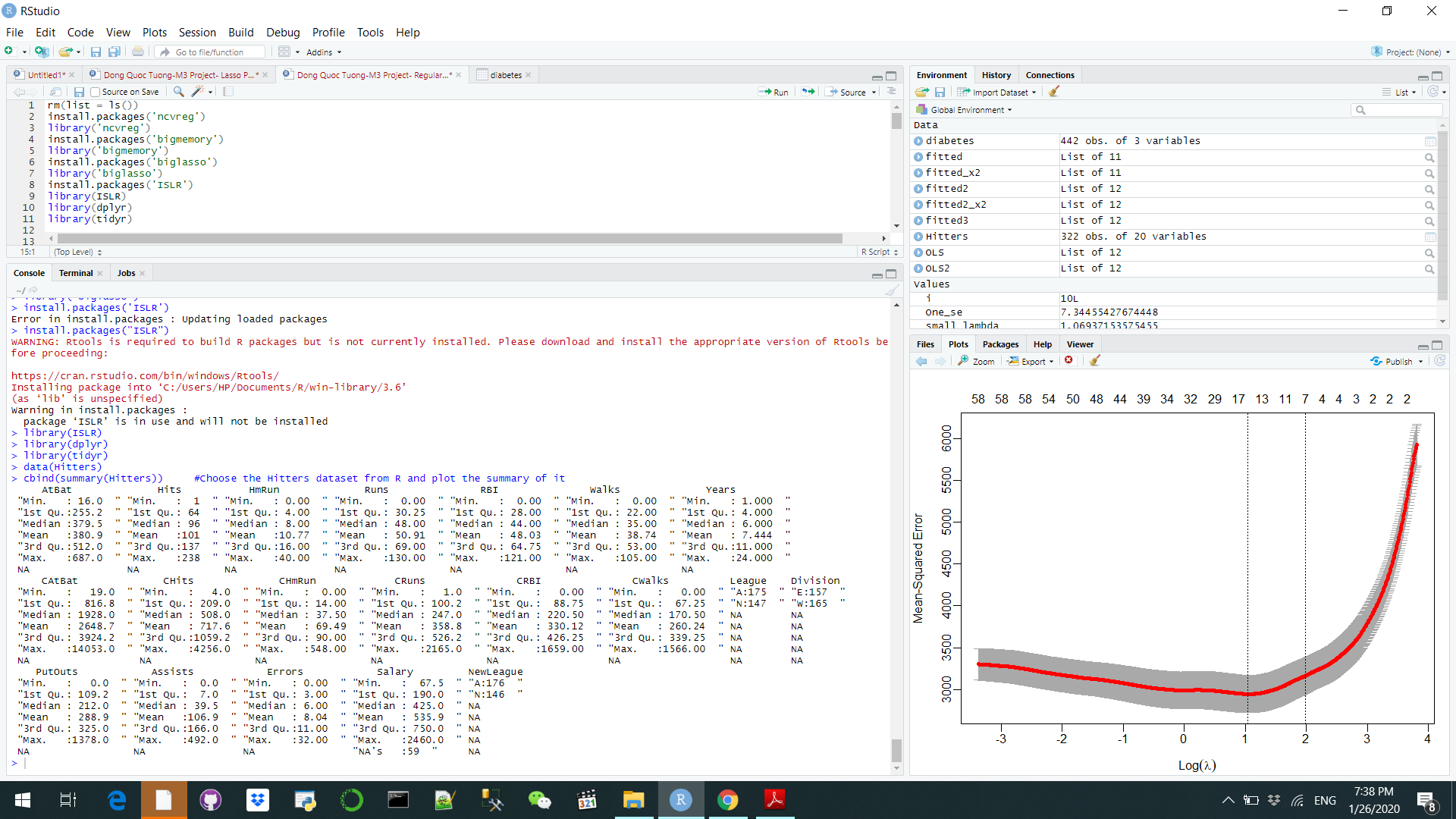
|  |
| --- |
| **Module 3: Regularization Assignment**  **Completion Date: January 26th Due Time:12:00am** |

**Statement of Authorship**

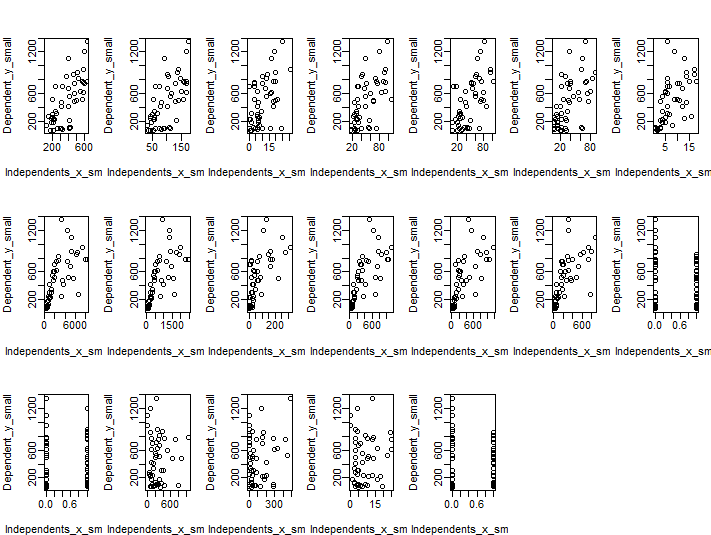
*I confirm that this work is my own. Additionally, I confirm that no part of this coursework, except where clearly quoted and referenced, has been copied from material belonging to any other person e.g. from a book, handout, another student. I am aware that it is a breach of Northeastern University’s regulations to copy the work of another without clear acknowledgement and that attempting to do so renders me liable to disciplinary procedures. To this effect, I have uploaded my work onto Turnitin and have ensured that I have made any relevant corrections to my work prior to submission.*

**Tick here** to confirm that your paper version is identical to the version submitted through Turnitin

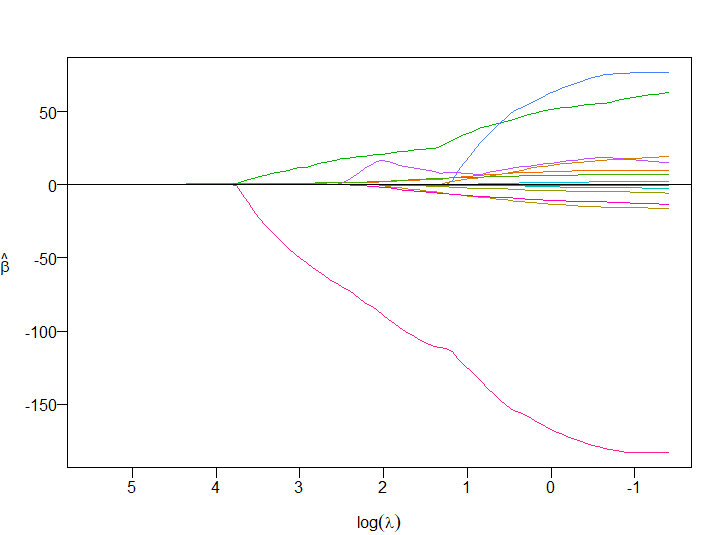
In this paper, we will imploy the Regulization models with Lasso instead of Ridge analysis to analyze the prediction model for the dataset that we have. First and foremost, we will load all the necessary packages in R and choose the dataset Hitters, which has 322 observations of a major leagues players on the following 20 factors: AtBat, Hits,HmRun,etc.



We then use na.omit() function to eliminate all the null values for our analysis. Then we assigned the first column as dependent value in numeric format and then the rest as independent values. Keeping in mind that since we will extend the model to fit the big data on the second examination, we will only take 50 rows for the first test and use everything for the second one. Now we will generate the scatter plot with the best fit line for all the predictors in Dependent\_X\_small and independent\_y\_small down below. There are 3 groups of scatterplots patterns we can see: The binary group, the uncontrolled group and the patterned ones. It is interesting to see what the difference between the uncontrolled group and the patterned ones is

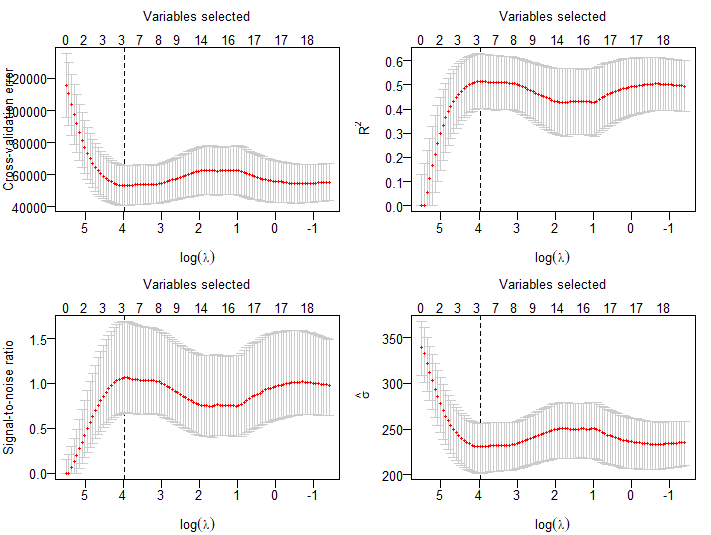


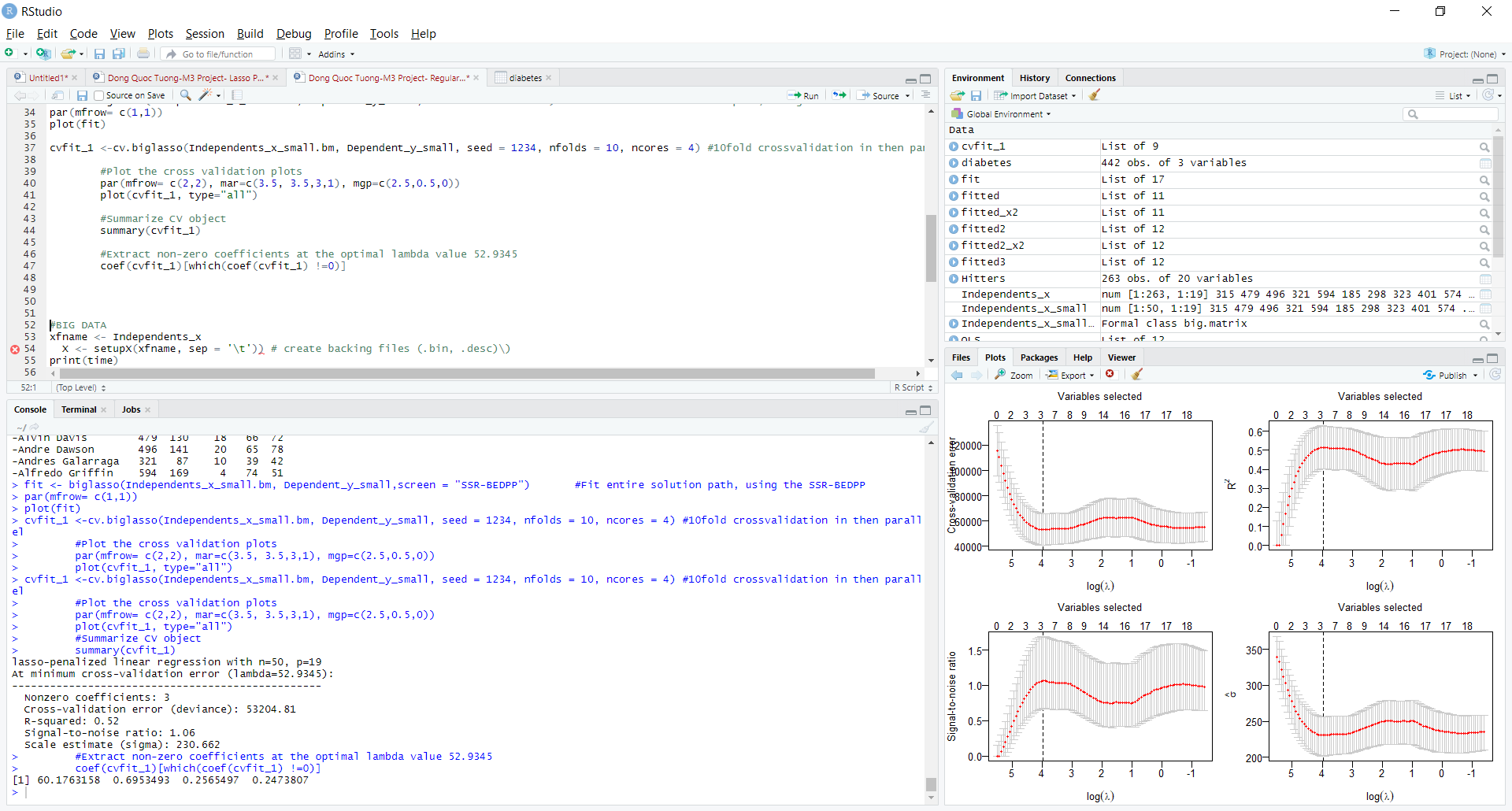
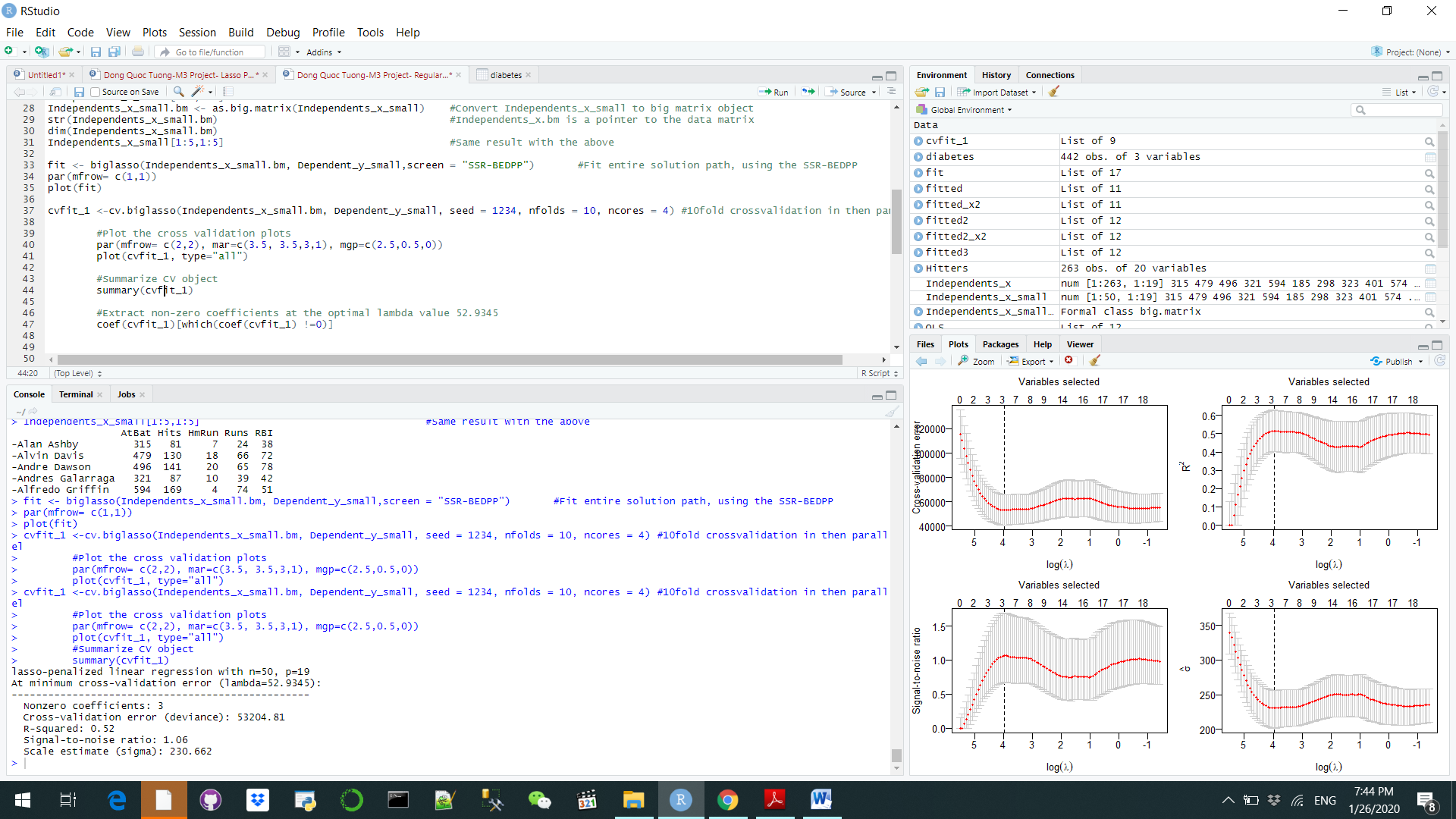
We then convert the Independent\_x\_small to big matrix object and the Independents\_x.bm functions as the pointer to the matrix. Then we see that the dimming result is the same before and after appointing Indepenedents\_x.bm as the pointer. After that, we will regress the whole dataset small using the biglasso instead of the glmnet, using the SSR-BEDPP and plot the x’s variable coefficients with the L1 norm of the beta vector. Each predictor is illustrated by a different color. If lambda approaches 0 it means that the loss functions approaches the OLS loss function. Hence when it starts out everything is 0, but then when lambda reaches 3.5, things starts to change and the variables starts to have an effect on the model as their coefficients derive away from 0.



Now, there are three things that we can do with the dataset from now on.

* Number1: we will illustrate the validation curve as well as minimize the cross validation error. Usually the value fluctuates between 1 as perfect correlation and -1 as perfect negative correlation. All four models have the same shapes no matter what.
* Number 2: Summarize CV object. This summary allows us to evaluate if the model is fitted with the data set or not. Since R-Squared is 0.49, it means that only around half of the dataset can be explained by the model while the rest is not. We need to optimize this model by using the lambda
* Number 3: extract non-zero coefficients at the optimal value of 52.934. After extracting, we are left with only 4 predictors as seen below. Thus, we can conclude that we only need 4 predictors to make our summary count





Now, we will use the all the big raw data to do or analysis, with 342 variables. We put the raw data into the file for design matric and create the backing files both in .bin and .desc. Usually the reading of the data takes a few second but it takes a while for this dataset because the veloume is bigger. Since the operation is just a one-time execution, the data can always be resumed simultaneously in any new R services. At the end, we can do the analysis with the data. The only difference between this analysis and the one above is the use of attach.big.matric() function as normal.