

**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**

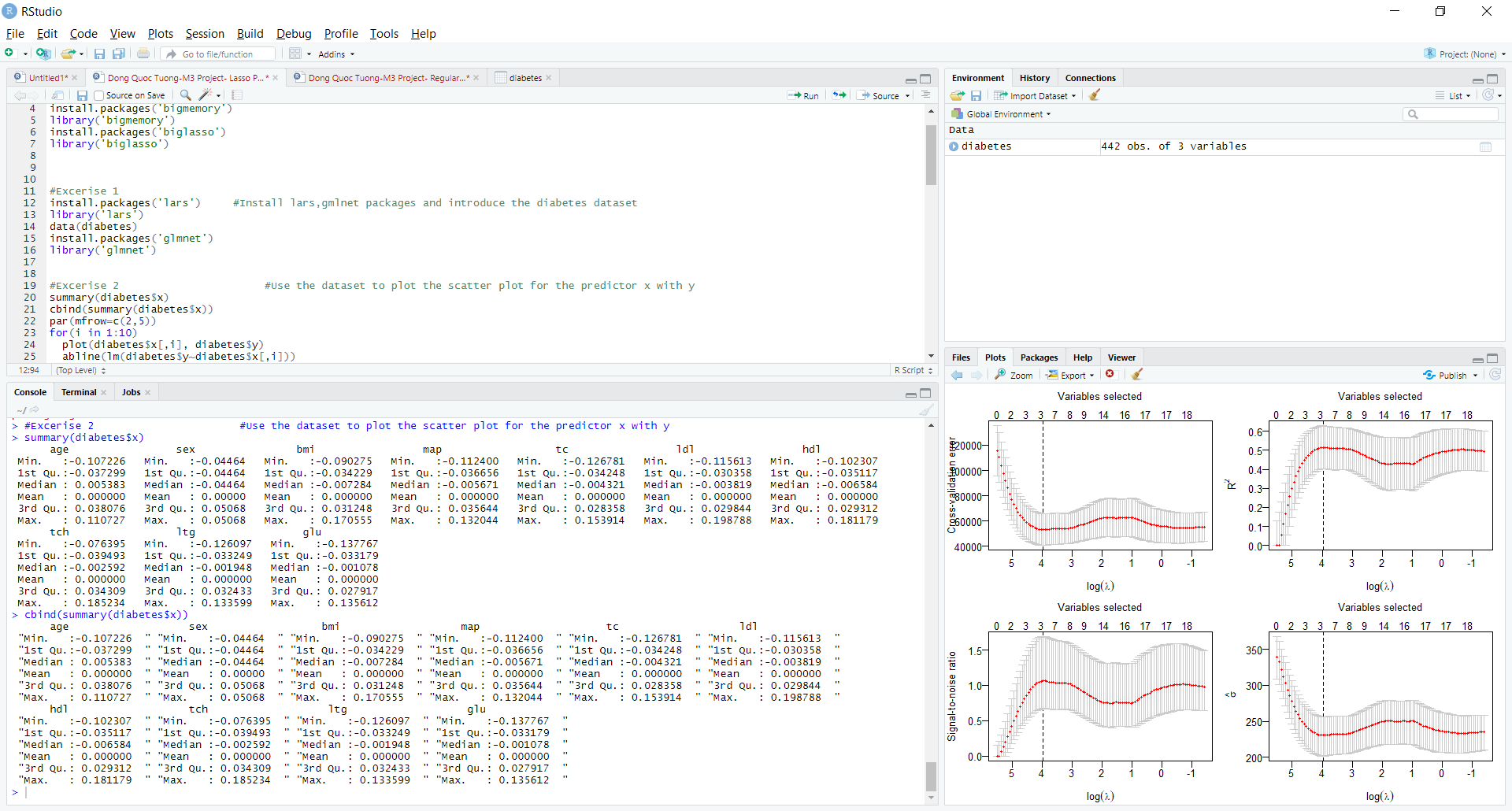
|  |
| --- |
| **Module 3: Lasso Practice**  **Completion Date: January 26th Due Time:12:00am** |

**Statement of Authorship**

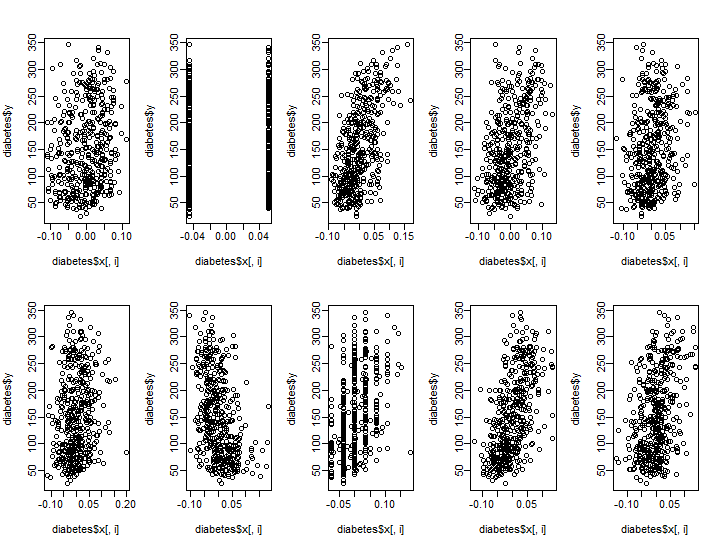
*I confirm that this work is my own. Additionally, I confirm that no part of this coursework, except where clearly quoted and referenced, has been copied from material belonging to any other person e.g. from a book, handout, another student. I am aware that it is a breach of Northeastern University’s 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

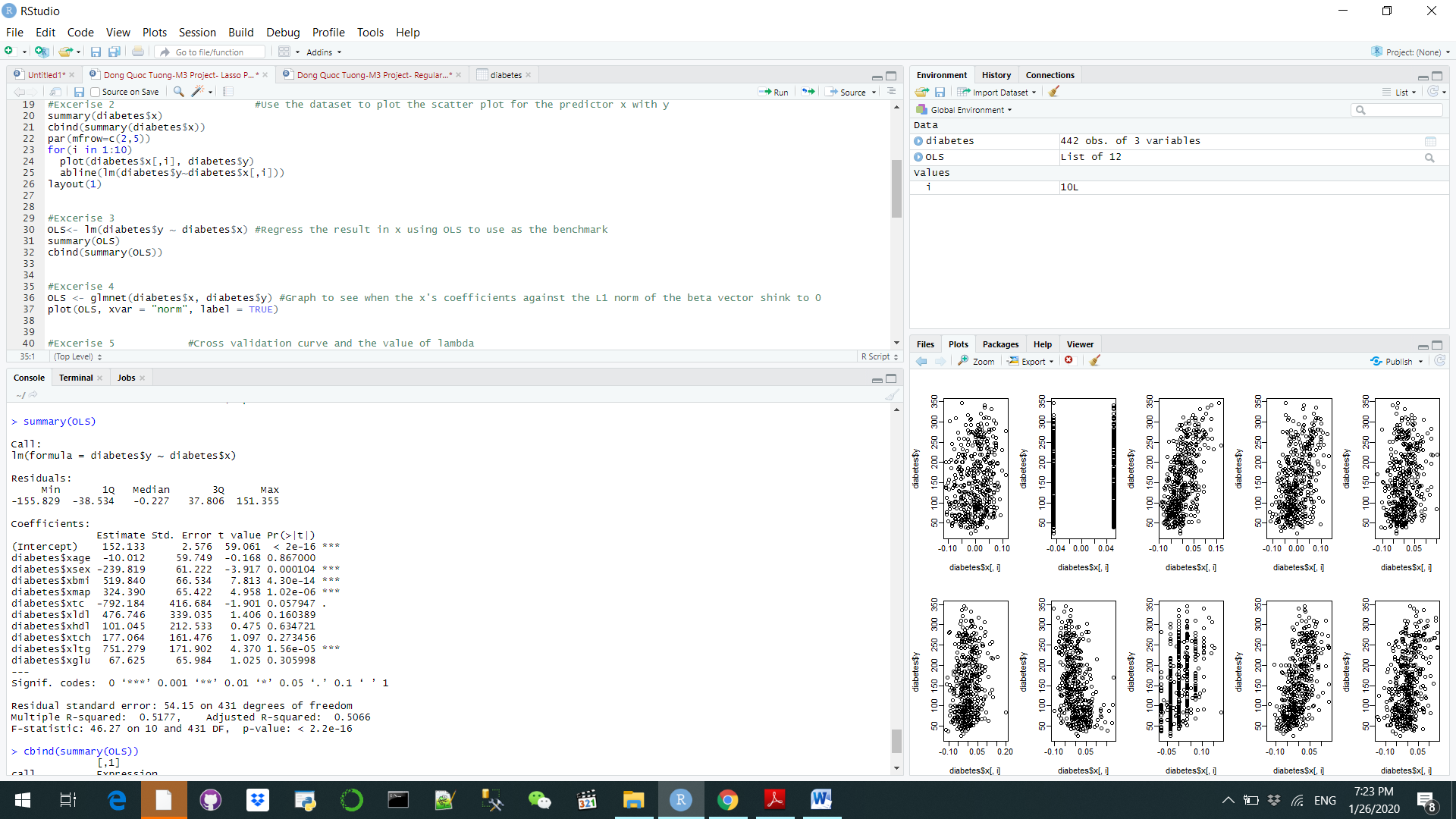
In this excersie, we will be buidling the regularization models with Lasso (least Asoolute skninakeg and selection operator). First of all, we will load the package lars, glmnet and the diabetes dataset from R. The dataset is divided into 3 columns, each columns into 10 different sections, namely: age, sex, bmi, map, tc, ldl, hdl, tch, ltg, and gl.



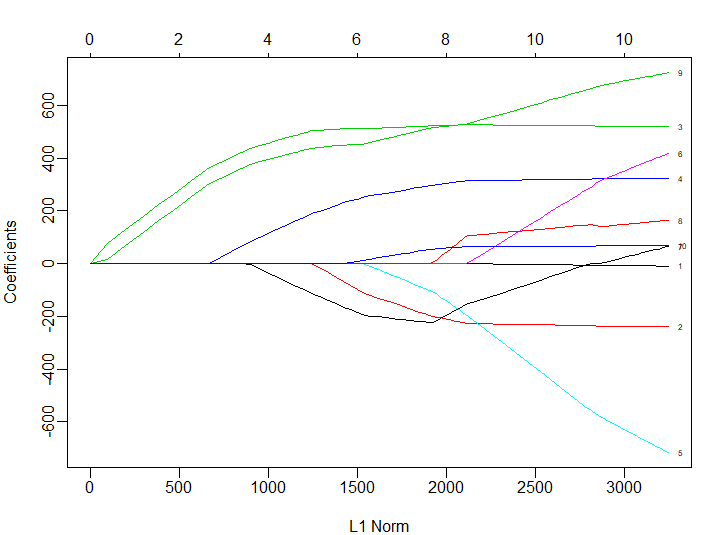
Then we divided the exercise into two parts, the first is we tested dependent variable y with independent factor x then after that with independent factor x2. Because x2 comprises of quadratic and interaction terms, it is easy to over fitting the second one. Thus, we will generate the scatterplots with the best fit line for all the predictors in x and y down below. Most of the scatter plots’ results are similar to each other, but only the Sex one is somewhat different because it is a Clustering model with binary data.

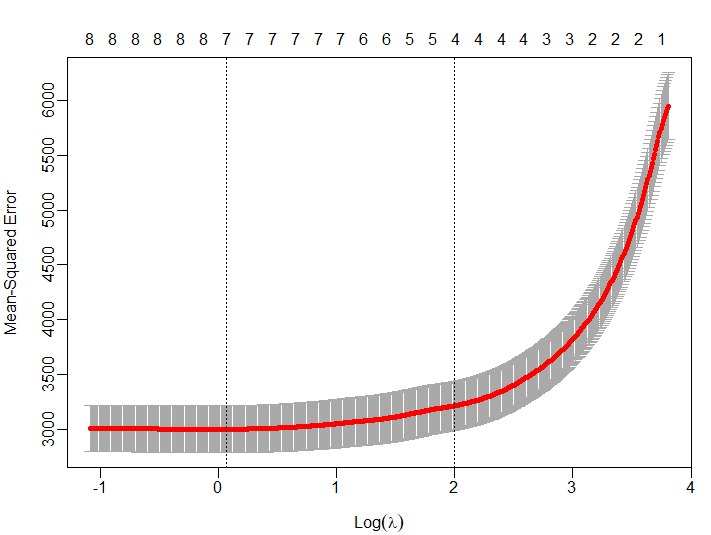


Then we regress y on the predictors in x using OLS and using this as the benchmark for our analysis. The multiple R-Squared and the Adjusted R-squared indicated that at least half of the variability of the responses around its mean can be explained by the model.(“Regression analysis: how do I interpret R-squared and assess the goodness-of-fit?,” 2013) Even the Residual Standard of error is 54.15 with the degrees of freedom of 431. So the model might not be the best fit for the diabetes dataset at this time.



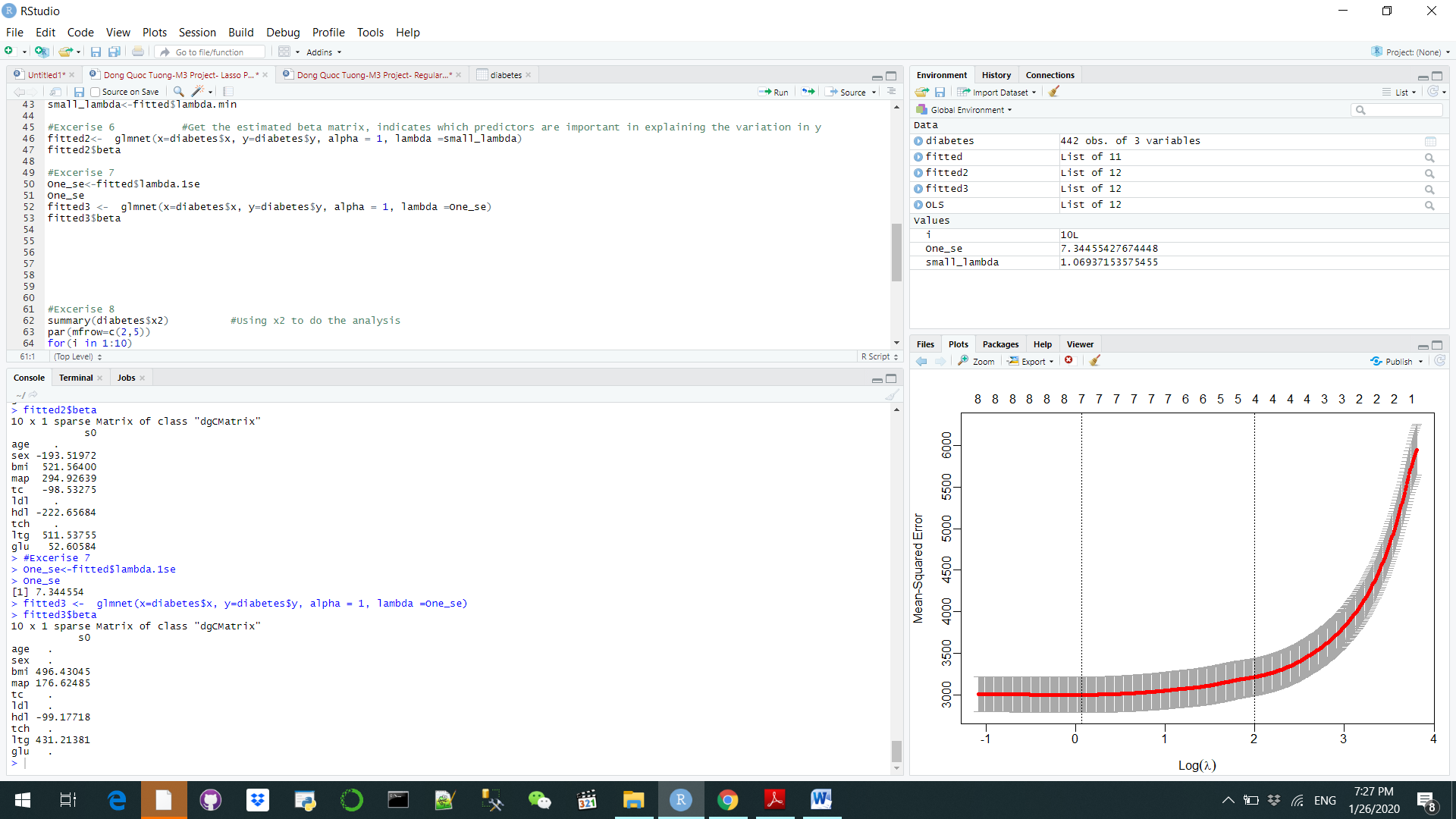
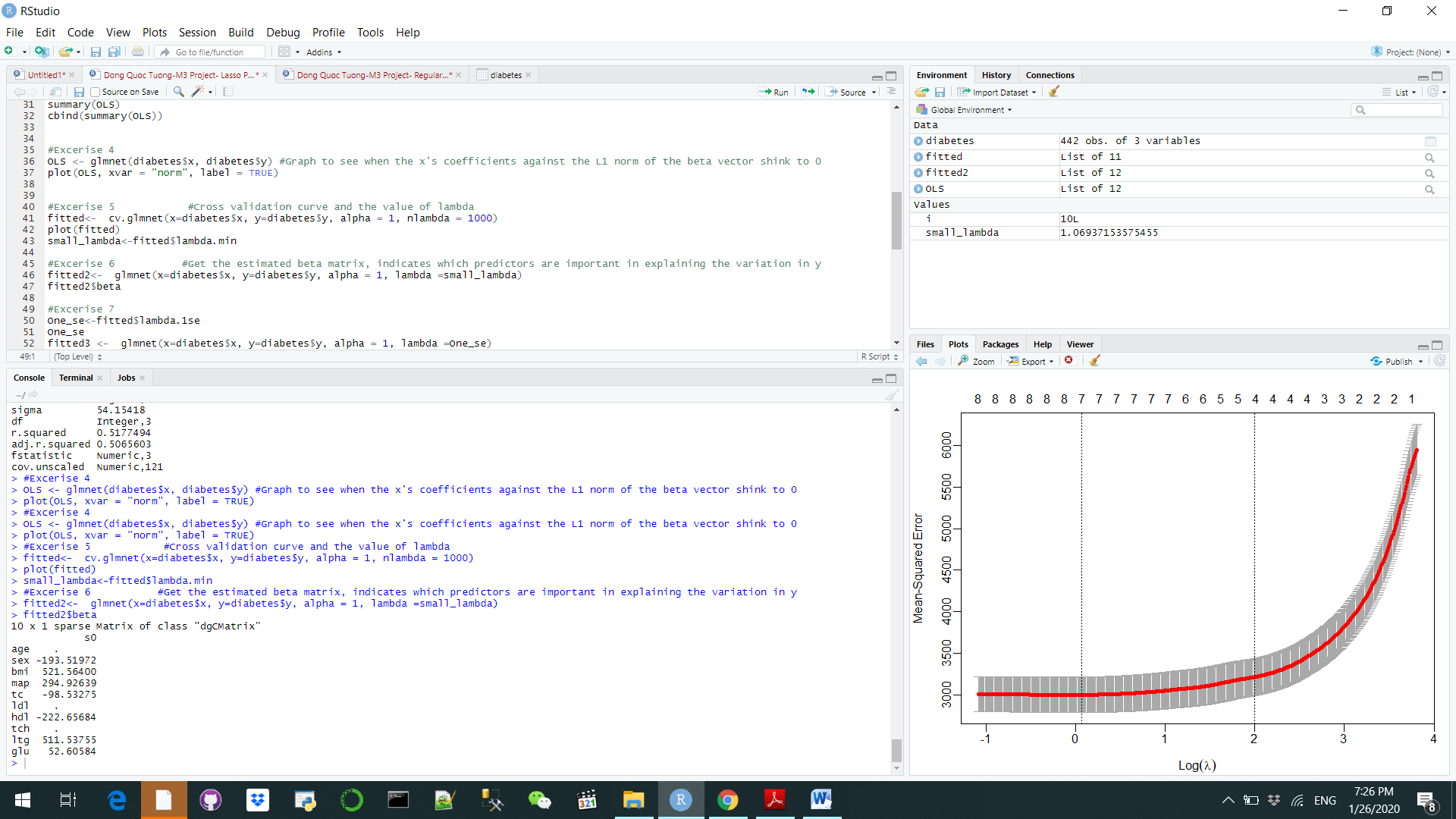
Next, we compared the x’s variable coefficients with the L1 norm of the beta vector. Each color represents the value by the different coefficient of the model running from 1-10 of the dataset we have. Since λ is defined as the weight given to L1 norm, so as λ arrives zero, the loss function of our model also approaches the OLS loss function. Hence, when λ is small, all of the lines are nearly zero (lots of regularization or empty model). But once λ increases, variables will have an effect on the model as their coefficients does not stand at 0 anymore, starting with line 9 (ltg) and ending with line 1 (age)

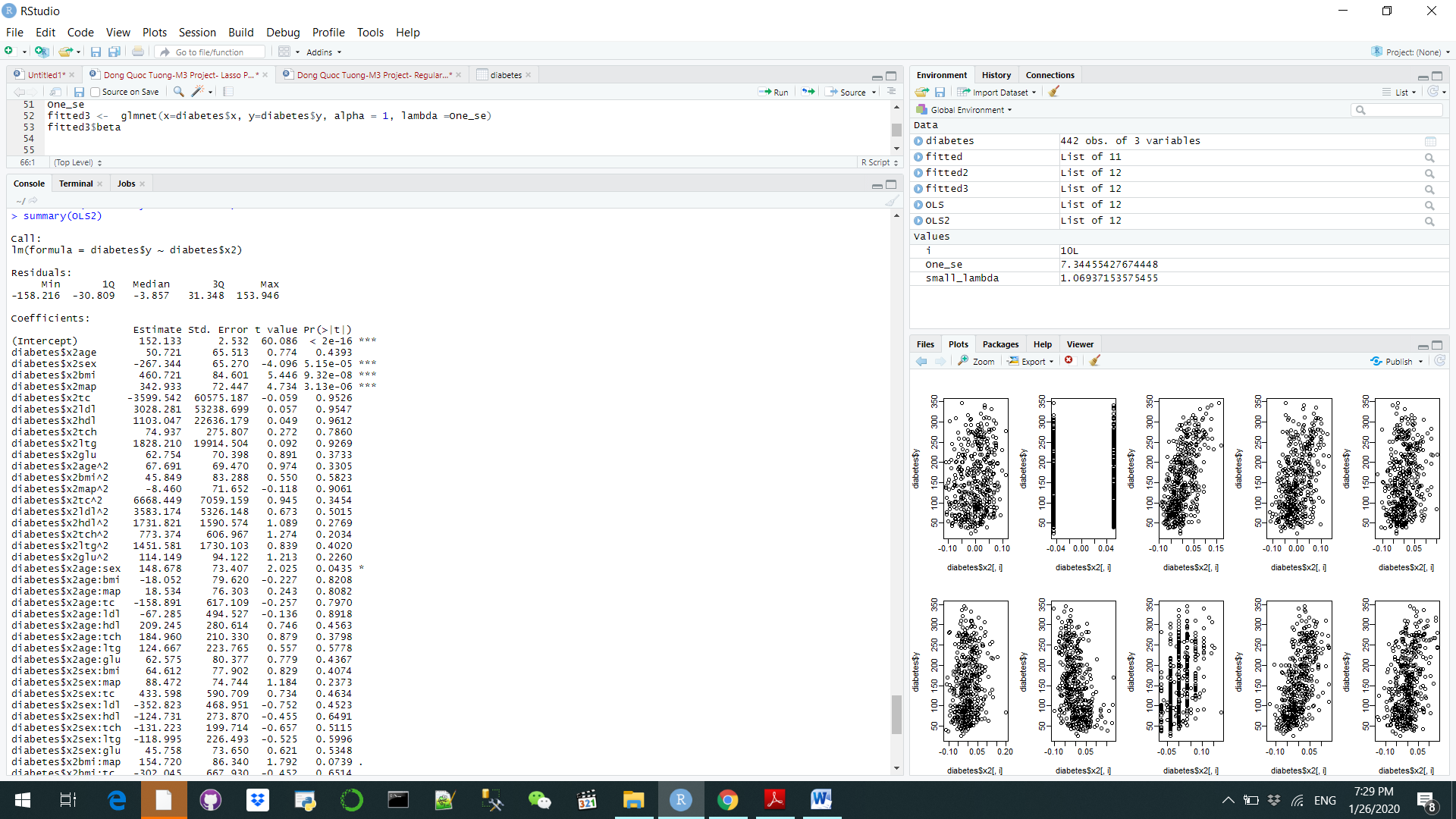




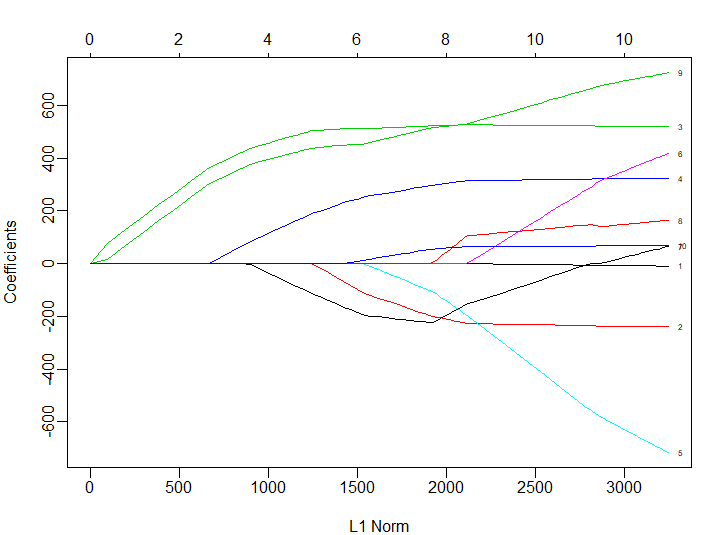
After that, we will illustrate the cross-validation curve as well as minimize the cross validation error. Applying the minimum value of lamba that we have found previously to this, we can get the estimated matrix. Usually the value of vary between 1/-1 as perfect correlation/ perfect negative correlation. But when the r = 0 in this case for ldl and tch (left picture) in this case, that means there is zero correlation between those predictors with the dependent variable y and we can take them out. The lower the number, the less relevant the factor is to the dependent variable. For example, in the left picture, bmi is much more important in predicting y than glu.

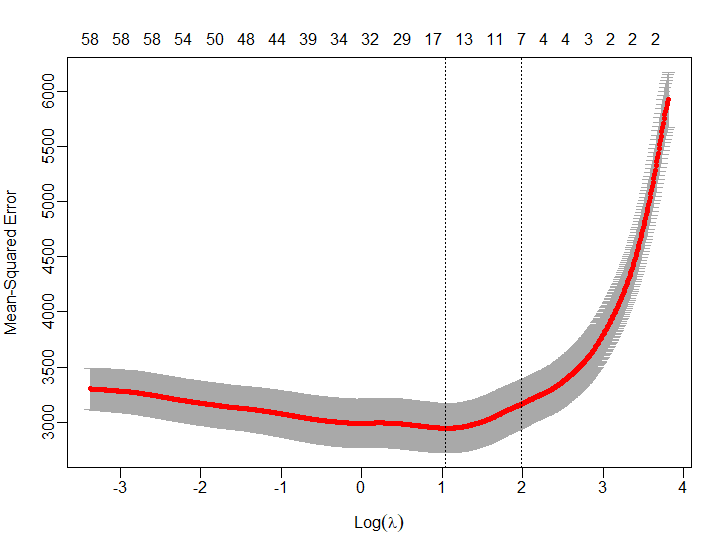
Since we want to optimize our model to the fullest, we will use the higher value of λ (7.344 ) with different beta coefficients. We will notice that compared to the left picture below, there are only 4 predictors left that can drastically change the value of depdenent variable y:n bmi, map, hdl and ltg

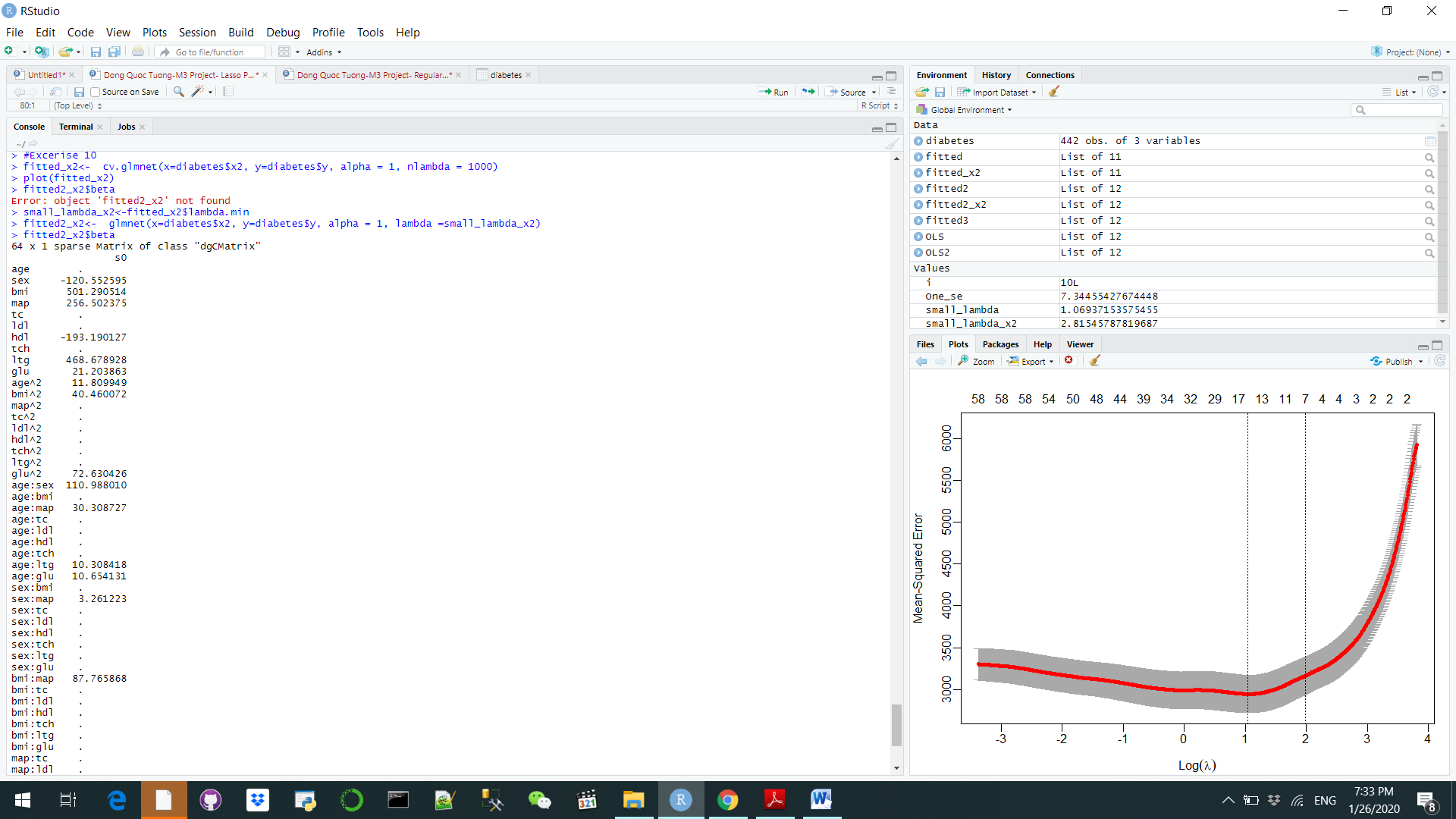




Now, we do the same with the set of predictors in x2, noted that in the snapshot on the right, the number of predictors has increased dramatically to 54. This is an example in real business that anything can be seen as relevant to each other. All the analysis will be the same like above. Then we do the plotting and analysis like usual







But at the end of the day, despite having multiple predictors, there are only 15 predictors that actually make a huge impact on the dependent variable y such as sex, bmi, hdl, etc.

In conclusion, when we do the Lasso analysis, we were able to reduce the amount of irrelevant predictors that we need to take into consideration when making a business decision. This is the epitome of what analytics can help you to achieve

**References**

Regression analysis: how do I interpret R-squared and assess the goodness-of-fit? (2013). Retrieved from https://blog.minitab.com/blog/adventures-in-statistics-2/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit