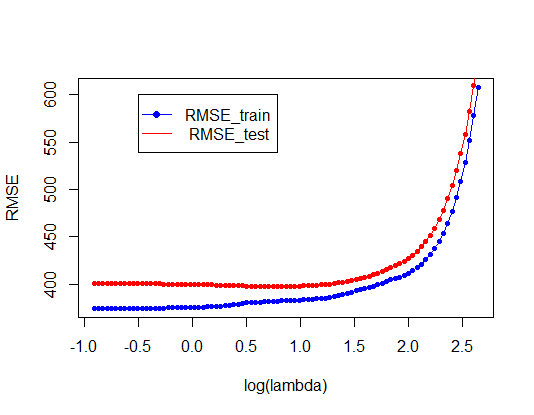
The goal of this project is to determine the 3 main factors that determine crime rate in communities in the US. The dataset description is available here:

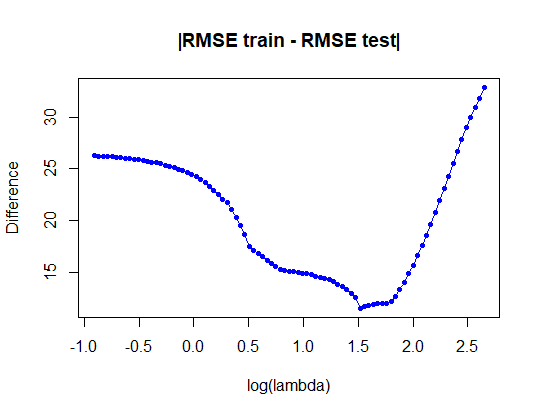
<https://archive.ics.uci.edu/ml/datasets/Communities+and+Crime+Unnormalized>

1. **Plotting RMSE for Lasso Regression on training and testing data sets with suitable regularization parameters λ**



The **RMSE decreases** for both train and testing data as **log(lambda)** goes to 0

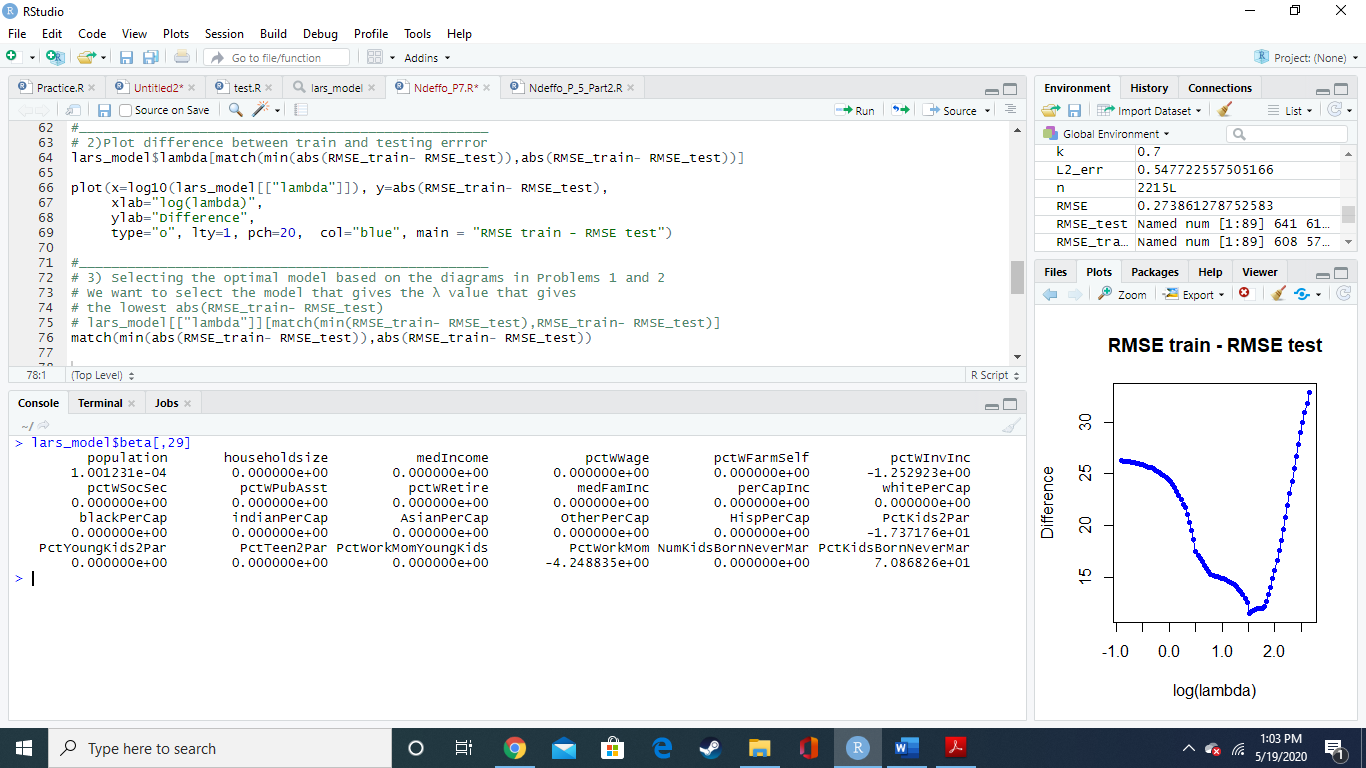
1. **Plotting the curve for the difference of the training error and testing error**



Using the picture above, we can notice as the **difference gets smaller lambda gets closer to 0 especially when λ = 32.84557**

1. **Selecting the optimal model based on the diagrams in Problems 1 and 2**

The optimal model is one for which the absolute value of training error – testing error will be minimal. In that case λ = 32.84557 and thus βλ = 32.84557 will have the following coefficients ranging from βi=1…. βm=24 as show in the image below



Therefore

This can be translated to Y = [1.00123\*10^-4…. 70.8626]

Also something important to see to get the above result, we were looking to extract the 24 coefficients matching λ = 32.84557 which is at the 29 position in the list that contains all the lambda values. Thus, we use the command lars\_model$beta[,29] to return corresponding beta values.

1. **Determining the 3 main factors that influence crime rate**

Those factors will be ones that will have the highest coefficients or beta values positive or negative.

Looking at the model above those factors are:

* **PctKidsBornNeverMar** (Percentage of Kids born never married) with **β24=70.8626**
* **PctKids2Par** (percentage of kids in family housing with two parents) with **β18= -17.37176**
* **PctWorkMom** (percentage of moms of kids under 18 in labor force ) with **β22= -4.248835**