**Analysis of Social Capital of US**

**1.Goal and Objective**

The objective of this project is to dig into the Social Capital index of Counties in United States and see how it change across time. We are also interested in what is the driving factor for a high social capital index. In order to do so, we are going to build models for

1.Lasso for variable selection

2.Tree-based Classification for quantify feature importance (Based on the variable selected by step 1)

**2.Data preprocessing**

**2.1 How I select first batch of X**

Most of the variables in Cencus\_Data.csv are irrelevant for our target. However, they are encoded in a specific pattern: HC00\_VC00

VC00 represent the variable, the prefix

HC01 is the Estimate which is a real positive number

HC02 is the Margin of the error

**HC03 is HC01/Population (percent)**

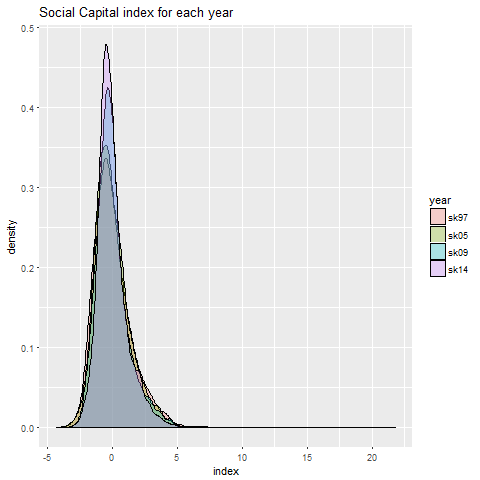
HC04 is HC02/Population

Then we extract all the variables with prefix HC03 and there are about a hundred potential variables that might be contributing to our response variable.

**2.2 How I treat those data**

All of these variables are percent number. Particularly, for all the insurance coverage, I divided them by population: **HC01\_VC03 Population 16 years and over**. After doing so we have to scale and center those variables which means subtracting the means and divided by its standard deviation. This process is necessary, although we will lose the ability to interpret about these coefficient quantity, the Interpretation about the sign of coefficient still holds.

**3.Exploratory and Concerns**

****

On the left hand side, Social Capital Index seems to be more concentrated each year. And the mean social capital index for each year is 0. Here is my explanation:

This dataset for each year **might not give us what we desire.** The actual change of absolute stock value might not be obtained from this dataset. Although it could happens that the county with higher stock value shift towards left, and the county with lower stock value shift towards right. But the chances are very very small.

After dig into the data source, I found:

1. Since such index was obtained **by principle component analysis**, they said in the report **they first centered the data**, that is to subtract the mean of each column, thus we are unable to see the real difference of absolute stock value, but we could see **relative comparison**.

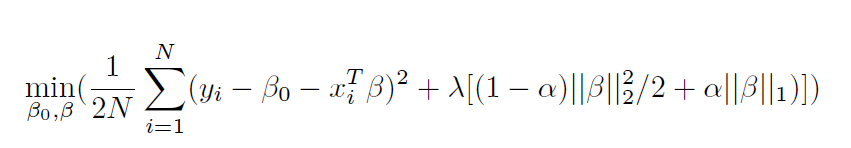
2. I also have concerns regarding the way they do PCA, if the value of the later year is not obtained by projection onto the previous basis, then those values are not comparable. To be more specific, if they do PCA separately each year, then the number of index for these years are not under a consistent basis.

**4. Model Building**

**4.1 Elastic Net for variable selection**

**4.1.1 Method**

Elastic net is a combination of lasso regression and ridge regression. It is a powerful tool for variable selection. The loss function is listed below. We can see on the left part is ordinary least square, on the right-hand side is the regularization with L1 and L2 penalty. Those are controlled by the parameter alpha. This alpha controls weight of for each penalty.



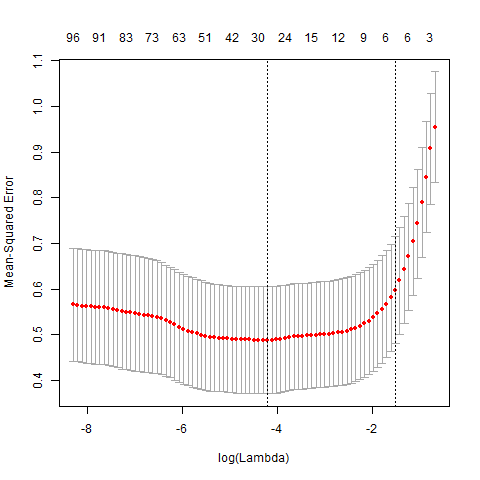
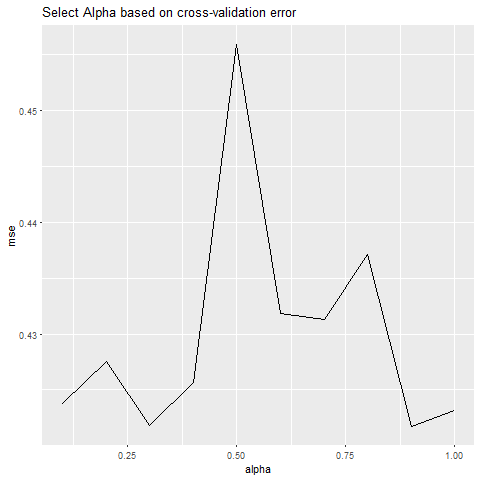
The advantages of such a model:

1. This method can perform **variable selection**, because like lasso it has L1 penalty which is the norm of the beta. By saying variable selection, I means we could see all the variables importance in a decreasing order, thus gives us freedom to **choose whatever number of variables we want**. Such as top 5 or top 10.
2. Such shrinkage method could partially **solve the collinearity problem**. Collinearity of X will result in high variance of our estimate, this will make the sign opposite to theory because high variance makes it unstable.
3. It will not make same mistakes **like excluding the variables with only interaction effect** as stepwise algorithm do.
4. Computation convenience. Greedy algorithms like stepwise selection are computationally intense if dimension increases and will probably not yield the optimal solution.

**4.1.2 Test\_vs\_split and cross-validation**

A rule of thumb is to split the dataset 20% for test and 80% for train. And then fine-tune the parameter on the train data with cross-validation to avoid overfitting. based on the 5-fold cross-validation error we select alpha=0.1

The two parameters control the severity we penalize the length of the parameter. The intuition behind the penalization is we would like to decrease the variance significantly. The larger lambda, the less variables are selected, and the higher train error is.



By controlling the lambda we can select what ever number of parameters we want. For example if we would like to select top 9 variables which are highly correlated with the y, then the corresponding lambda is 0.1932824, this specific number will shrink all the variable towards zero because they are relatively not “important”.

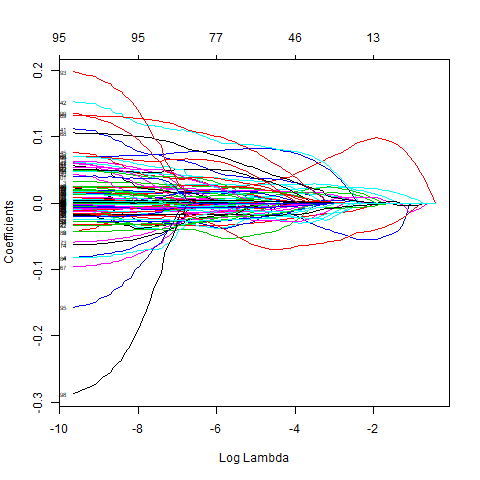
**4.1.3 Result of Variable Selection**

This model is my previous one.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Code | name | Variable value | Coefficient Sign | Interpretation |
| HC03\_VC07 | Unemployed rate | 3.4 | Negative  -5.2e-02 | More people unemployed, less social captital |
| HC03\_VC23 | Percent of all parents in the labor | 66.7 | Positive  1.93e-02 | This is a composite of employment rate |
| HC03\_VC50 | Percent of people Agriculture Forest and fishing Hunting and mining Industry | 1.2 | Positive  1.12e-02 | This is interesting.  This might result from geological specific variables.  Maybe people in this areas have more labor unions. |
| HC03\_VC69 | Percentage of Self-employed | 5.6 | Positive  9.18e-02 | A composite of employment rate |
| HC03\_VC101 | Food stamp benefits | 12.8 | Negative  -4.35e-05 | This might indicate the poverty |
| HC03\_VC131 | Health insurance  Public | 91.1 | Positive  2.42e-03 |  |
| HC03\_VC132 | Health insurance  private | 72.7 | Positive  1.60e-03 |  |
| HC03\_VC134 | No Health insurance no coverage | 8.9 | Negative  -3.50e-03 |  |
| HC03\_VC164 | Percentage of all people is below the poverty level | 3.7 | Negative  -3.52e-02 | More people are below the poverty less social capital. |

**4.4 AdaBoosting**

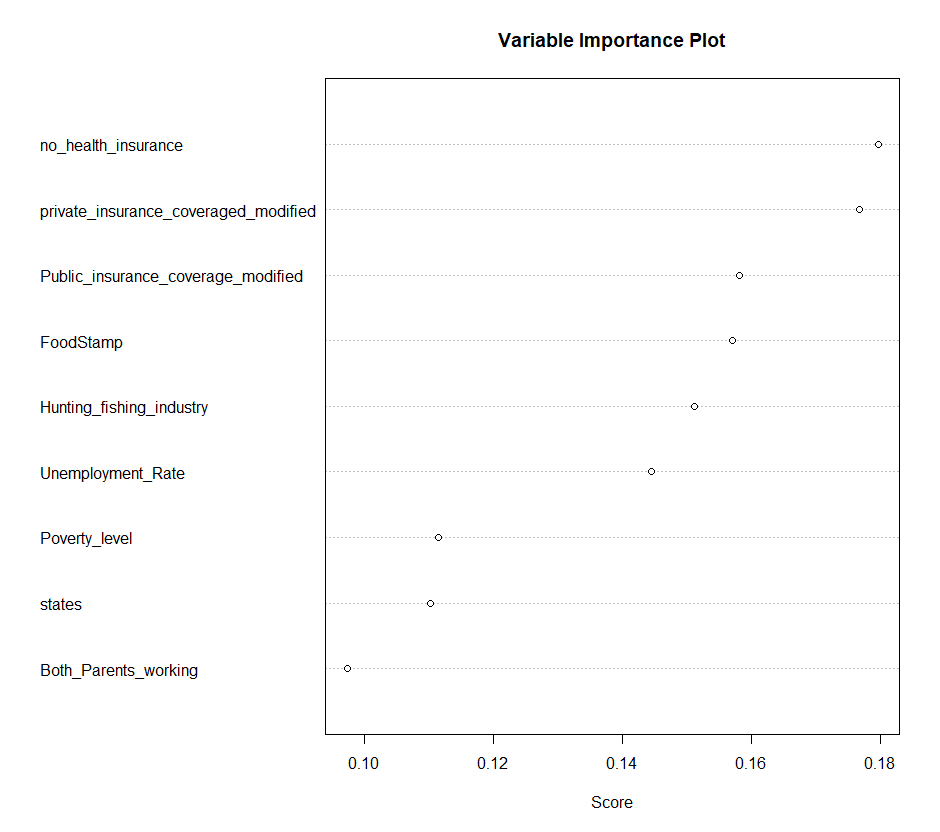
Since the variables listed above is selected by our model. We could see the relative importance by looking at the coefficient convergence graph. We can see how coefficients of each estimator change when we loose the penalization to add new variables.



However, evaluating importance by these absolute value coefficients is somewhat not reliable. Because of the unit are not consistent. The order for these variables show in the model is a good indicator. However, I am interested in how important these variables are going to determine whether this county has high social capital. I would like to quantify the importance.

**4.4.1 Adaboosting Method**

We are first divide those counties **into two classes**, those counties with index above 0 and those counties with index below 0. And we would like to include 9 variables selected by lasso. There are 3 variables are related to health **insurance coverage**. We are going to **divide them by county’s population.**



The scores represent how important each variable contribute to this classifier. There are measured by total sum of error rate will be decreased by introducing the simple split on that variable dimension. However, such greedy tree-based algorithms are prone to correlation. The correlated variables might swap position if we repeat again.