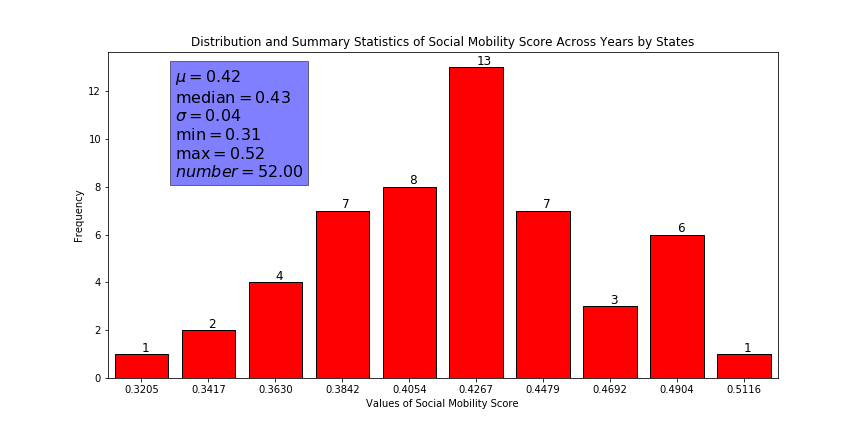
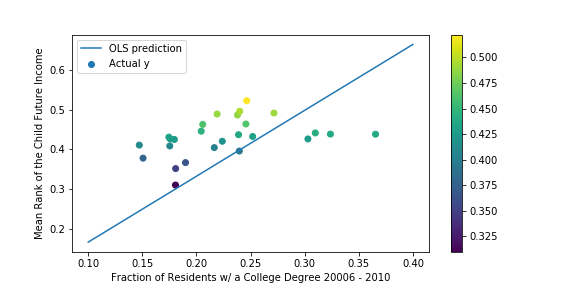
**Research Question:**

In this project, I would like to predict the social mobility in US. To do this, I used OLS prediction with cross-validation. I also choose the variable with the highest weight (absolute value) to do the prediction again.

**Data Set:**

I mainly used three data sets. Two of them are named as “adltas\_training” and “atlas\_test” and they are Stata dta file. The main dependent variable is included in these two datasets and the name is “kfr\_pooled\_p25”. It is the mean percentile rank of people in the national distribution of household income in 2014 – 2015 for those whose parents are at the 25th percentile of the national income distribution in each US county. The above two datasets have already included so many independent variables. But, to make the result more robust, I also used another data set, “edu.xls”. This dataset contains several variables of people’s education level at each US county.

**Estimation Steps:**

1. Because the adltas dataset can only be downloaded as zip file, I used ZipFile package with request to download and extract all the documents from zip file. The file’s name is “file.zip”. It will be downloaded directly to your working directory. These two datasets are in the “file.zip”. If you unzip this file, you will get “project4” folder. I cannot change the name of this folder because it is fixed. Because there are other files in the “project4” folder, I took the two main datasets out of the “project4” folder. I also used request method to download the “edu.xls” from the website.
2. For “adltas\_training” dataset, I only used several variables and this is shown in the python file. I also save the label of these variables in Stata so that I could check the meaning of them later. The “atlas\_test” only has two columns, county fips id and “kfr\_pooled\_p25” value. I also only kept several important columns in “edu.xls”.
3. After all datasets are imported, I merged them into one dataset. I first merged pd\_train(“adltas\_training”) with pd\_edu(“edu.xls”). Then, I merged it with pd\_test(“atlas\_test”). Finally, I get my mainly used dataset, “all\_data”.
4. After I get the main dataset, I clean it and move the columns to facilitate further analysis. I also dropped all observations with NA values.
5. I grouped the all\_data by “state” to draw a histogram with summary statistics of the dependent variable, “social\_mobi”. “social\_mobi” is the “kfr\_pooled\_p25” I mentioned above. Each observation is a US state in the below graph. 
6. I then define several functions to do the cross-validation test. I do all these parts manually with numpy package. I used OLS regression. The group size is 100 observations. That is to say, each time there will be 26 groups (the number of observations is 2517). Each time, 1 group will be chosen as a test group and the rest of the groups (25 groups) will be the training group. I will use the training group to get the weight of all independent variables and apply that weight to the test group to get the predicted y value (“social\_mobi”). Because the true y value is less than 1, I calculated the changing rate of the prediction value, i.e. abs((y\_test – y\_hat) / y\_test). I then take the mean of all the changing rate. I looped the cross-validation 26 times because there are 26 groups in total. Each group will be a test group. The final changing rate is 59.61%, which is very high. The prediction is not very accurate.
7. I also use all data and calculate the weight. Then, I choose the variable with the highest weight and estimate y\_hat only with that variable. So, I get w\_hat from this regression and apply w\_hat to that single variable to get y\_hat. Then, I plot them together. 
8. The line is the prediction line and the scatter points are the real value. The x-axis is the variable with the highest weight and the y axis is the value of “social\_mobi”.
9. I also outputted the main dataset, “all\_data” as “Output\_Data.csv” in the working directory.

**Limitation:**

As I said above, the changing rate is very high, so the prediction error is very high. I guess this is mainly because the variables I choose are probably not very related to the dependent variable. So, they cannot explain very much the variation of the dependent variable. However, in the second graph, when I only used the variable with the highest weight, I got a better prediction. The scatter points are all around the OLS prediction line.