**Introduction**

In the final project, we are going to predict the total wealth based on the given information such as one's income and home mortgage. I'm going to apply the feature selection (stepwise AIC, Ridge, and Lasso) to choose the variable that I would use to predict the total wealth. I would consider the outliers and unbalanced data, and how they going to affect the prediction. Furthermore, I'm going to consider transformed the data by adding polynomials and interaction terms in my model. Then, select the model and variables that gives me the least prediction error.

**Body**

**Data**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | IRA | e401 | nifa | income | Home mortgage | Home value | Home equity |
| 25%Quantile | 0 | 0 | 200 | 19413 | 0 | 0 | 0 |
| 50%Quantile | 0 | 0 | 1687 | 31575 | 8000 | 50000 | 10000 |
| 75%Quantile | 0 | 1 | 8875 | 48615 | 52000 | 95000 | 48000 |
| Max | 100000 | 1 | 1425115 | 242124 | 150000 | 300000 | 300000 |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | male | Two earn | No HS | HS | Some college | college | age | Family size | married |
| 25%Quantile | 0 | 0 | 0 | 0 | 0 | 0 | 32 | 2 | 0 |
| 50%Quantile | 0 | 0 | 0 | 0 | 0 | 0 | 40 | 3 | 1 |
| 75%Quantile | 0 | 1 | 0 | 1 | 0 | 0 | 48 | 4 | 1 |
| Max | 1 | 1 | 1 | 1 | 1 | 1 | 64 | 13 | 1 |

Source: the data is from the 1991 Survey of Income and Program Participation (SIPP).

This is the summarized data. The data set contains number of feature variables that could be used to predict one's total wealth. from this summary, by looking at the mean(50% quantile) of variable "male" and "nohs", we can see there is unbalanced data where there is more data on female than on male and more data on data on people went to high school. So, we need to see if the data we tried to predict also share the similar unbalanced data. Besides, "ira" seems to have large outliers since 75% data is 0, however, the max value is way larger than the most of data. I could consider dropping outliers in "ira".

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | IRA | e401 | nifa | income | Home mortgage | Home value | Home equity |
| 25%Quantile | 0 | 0 | 201 | 19420 | 0 | 0 | 0 |
| 50%Quantile | 0 | 0 | 1582.5 | 30987 | 0 | 43000 | 8000 |
| 75%Quantile | 0 | 1 | 8375 | 48020 | 49000 | 90000 | 44000 |
| Max | 100000 | 1 | 1430298 | 199041 | 150000 | 300000 | 300000 |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | male | Two earn | No HS | HS | Some college | college | age | Family size | married |
| 25%Quantile | 0 | 0 | 0 | 0 | 0 | 0 | 32 | 2 | 0 |
| 50%Quantile | 0 | 0 | 0 | 0 | 0 | 0 | 40 | 3 | 1 |
| 75%Quantile | 0 | 1 | 0 | 1 | 1 | 1 | 48 | 4 | 1 |
| Max | 1 | 1 | 1 | 1 | 1 | 1 | 64 | 13 | 1 |

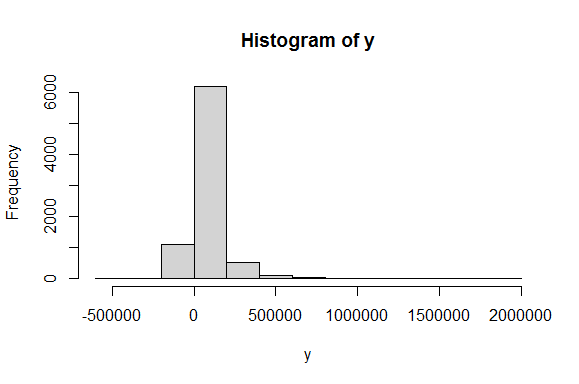
After looking at the test data summary, the test data on the "male" and "nohs" also appears to be unbalanced like the training set. So I will the problem of unbalanced data is not a problem. But I will exclude some extreme values in the "ira".

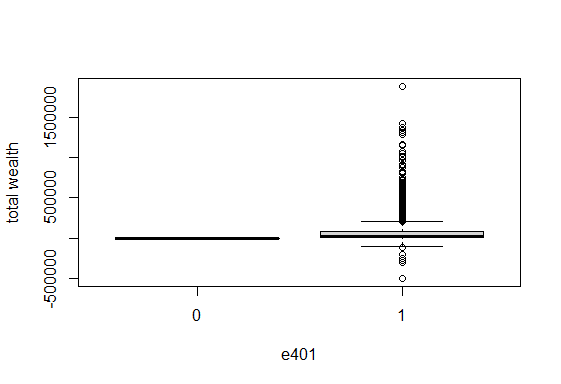
**Method**

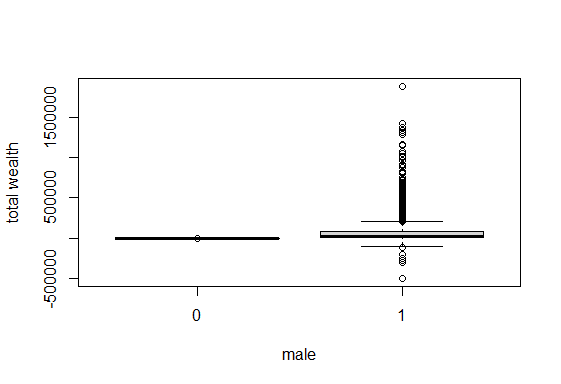
I will draw boxplots to show the categorical variable's sample mean and sample variance (I left out the boxplot for "nohs","hs","smcol", because all those variables are basically falls into the category of level of education, and in the category of level of education, I want to focus on looking at the mean difference between people went to college or not). Then, I will draw level plot to show the correlation between each continuous variables with dependent variable. From the plots, I could look at which variables are most associated for predicting the total wealth.

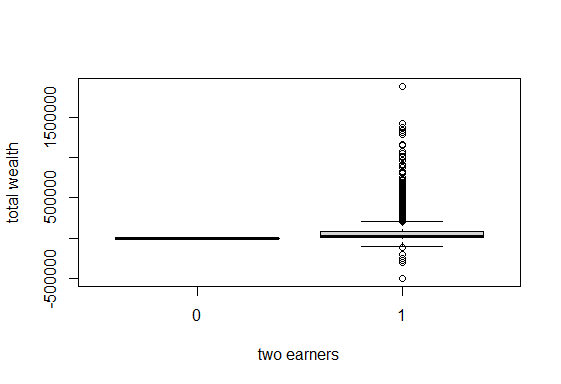
I will then exclude any outliers that significantly negatively affects the prediction accuracy. Then apply stepwiseAIC, Lasso, and Ridge model to predict. After doing cross validations for both models, the model that yield the smallest prediction error across all fold would be a better model to predict. Then I will transform the original data by adding polynomial degree and interaction terms to improve the prediction result, and redo the cross validation based on those models.

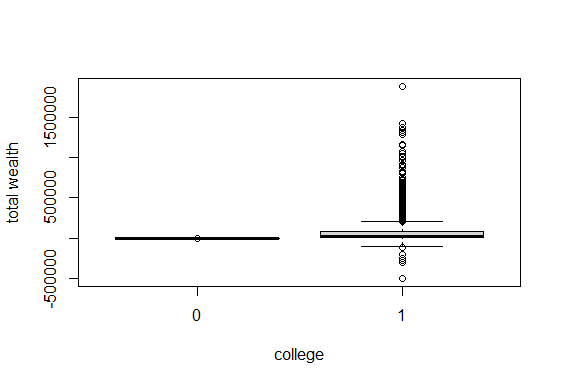
**Before excluding outliers:**

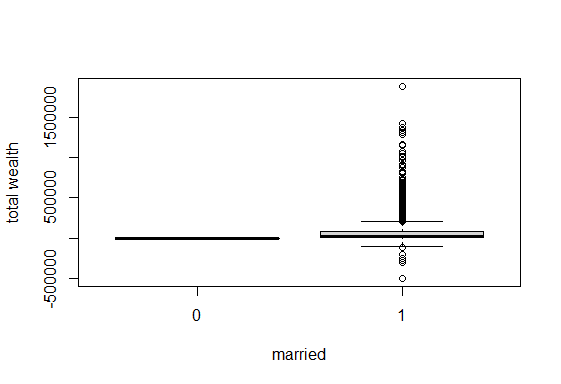
 we can see the distribution of wealth is heavily skewed to the right with outliers who have extremely high wealth, which I would consider excluding them. I now further explore how those outliers affects my residuals of predictions by drawing a simple regression model:



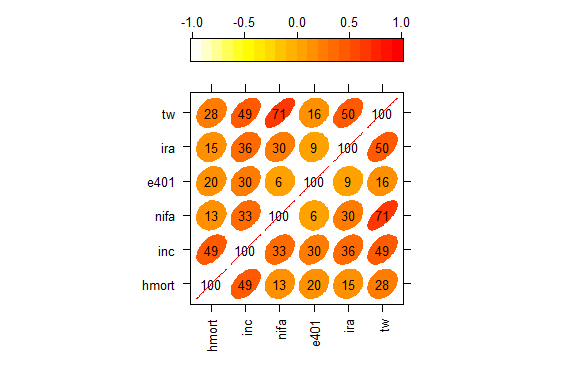






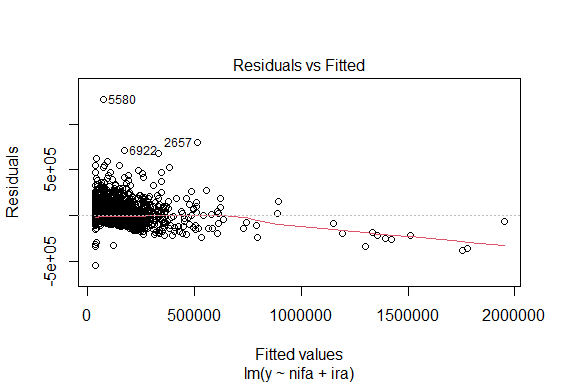


Those are boxplots for the dummy variables against the dependent variable "total wealth". We could see the outlier significantly impacted the value of variables. So, I'm going to exclude the outliers in total wealth for prediction in the later part.



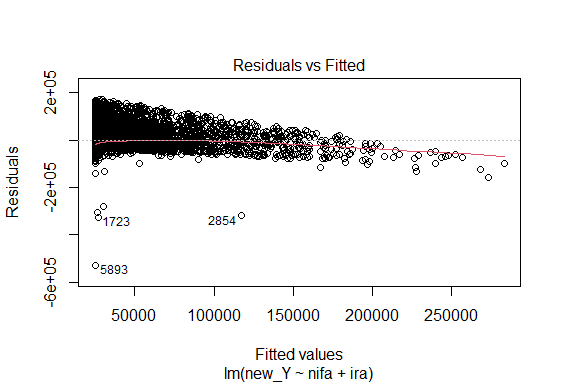
From the correlation plot, the more eclipse the shape is, the more correlation they are (red means positive association). So, we can see there is a strong positive association between the variable "nifa" and "total wealth", and a relatively strong positive association between "ira" and "total wealth", while other variables are less associated with each other.

So, I'm going to apply a simple linear regression on the most associate variables ("nifa" and "ira")to look at the prediction:

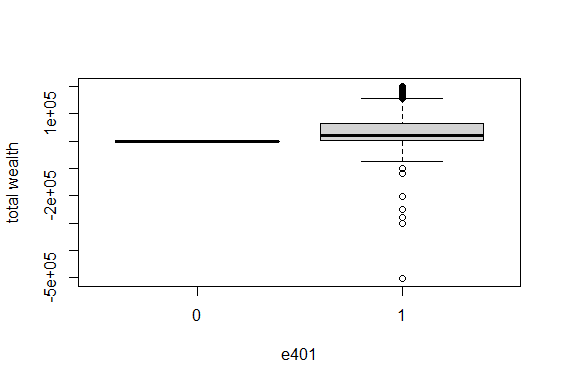


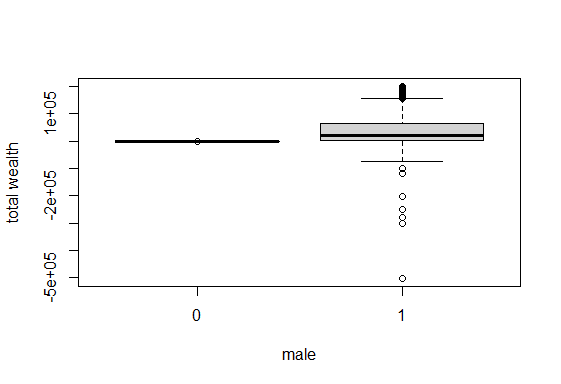
We could from the plot that the residual is significantly affected by those outliers; therefore, I 'm going to exclude outliers in the predicator (total wealth), which increase my accuracy of my model and in order to fit most of data. By calculating the interquartile range(IQR), I created a new table that excluded outliers that is out of 1.5\*IQR bound. This would help me better predict the total wealth because majority of the data fall within the IQR boundary and it will make my model to be a more accurate fit on the data. I now draw the simple regression model based on the same variables as the previous one:

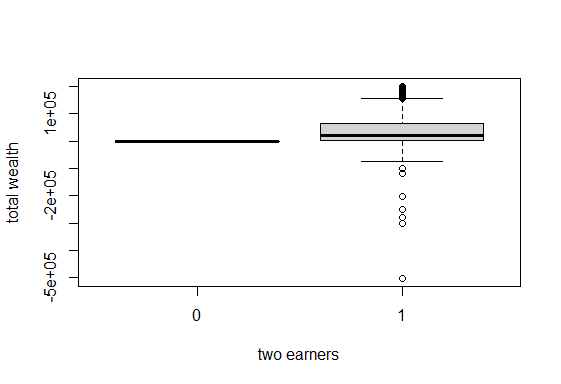
**After excluding outliers:**

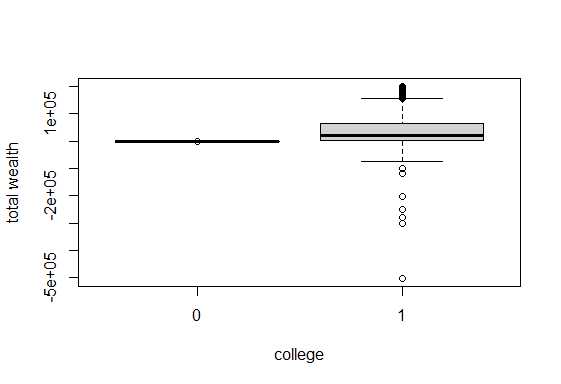
****

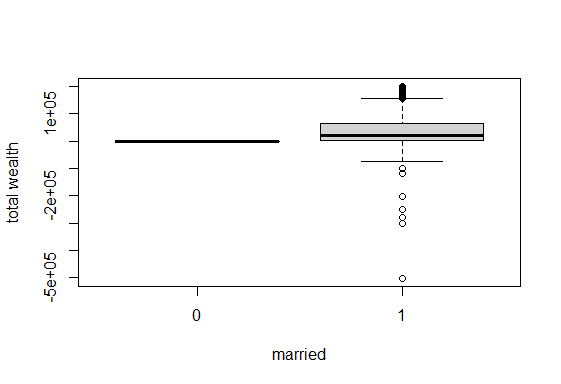
From the plot, we can see the points are more clustered together with just a few outliers in the residual plot, which is good.







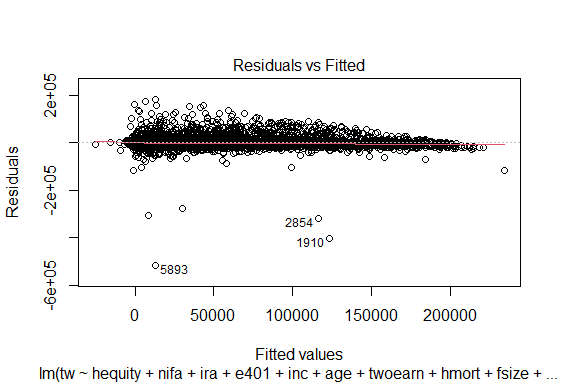




Now the boxplots also looks much better as I excluded outliers. The mean for different categories is much comparable and clear. We can see that people who are eligible for 401k, male, with two earner in the household, went to college, and married have slightly higher sample median in the total wealth.

**Analysis**

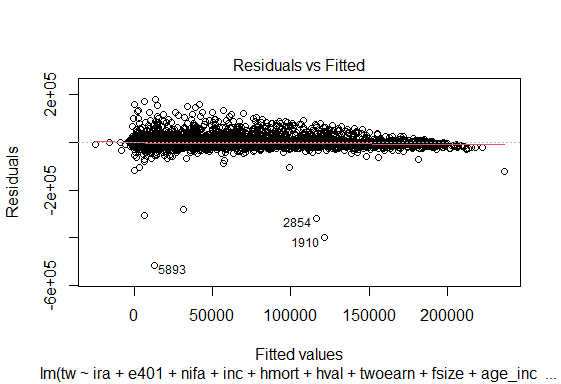
Now, I am going to further investigate the model that is going to predict the best for one's total wealth based on both those variables. In the new dataset excluding the outliers, I can do feature selections to see which variables I should include for my model. So, I did stepwise, backward stepwise AIC, Ridge, and Lasso model, and compute their predicted MSPE. Among the four model, the Lasso model gave the lowest MSPE. (Mspe\_step\_backward, mspe\_step\_forward, mspe.Lasso, mspe.ridge are following: 414495915 414188746 467650517 474164932). Then, by looking at the backward stepwise in detail, I found each variables' coefficients and plot the residuals as show below:



From the residuals plot, we could see the predictions fit the actual points pretty well with a few outliers.

Data Transformation:

Now I'm going to add some polynomial degree on some variables and add some interaction terms to improve the prediction result. I added the interaction between "inc" and "educ", "inc" and "age", "inc" and "male", "male" and "marr", and ira squared in the both training dataset and test dataset. Then redo the cross validation analysis on stepwise, backward stepwise AIC, Ridge, and Lasso model, and compute their predicted MSPE. This time, all four model generated a smaller prediction error then before data transformation. Backward and forward stepAIC generated a similar result. I'm going to predict the final dataset with the backward stepAIC because it relatively generates the smallest MSPE. (mspe\_step\_backward, mspe\_step\_forward, mspe.Lasso, mspe.ridge are the following: 413828018 413845114 456971150 465416541). And I plotted the residual of fitted line:



From the plot, it has pretty good predictions with a few outliers.

**Result**

Finally, I predicted my final predictions on test dataset using the backward stepAIC model after the data transformation. Because backward stepAIC yield the smallest MSPE among all other models.

**Conclusion**

In this project, I did exploratory analysis by drawing boxplots and level plots, and excluded the outliers that is going to negatively affect the prediction result. Then I applied different models(backward stepAIC, forward stepAIC, Lasso, and Ridge) for predicting the dependent variable "total wealth" and cross validated the mean squared predicted errors. Among the four models, the forward stepAIC gave a best prediction result before the data transformation. However, after I added some polynomial degrees and interaction terms, all four model's prediction error decreased, and backward stepAIC yield a slightly better result. Thus, I choose the backward stepAIC to predict the total wealth in the test dataset.