2020 – 2021 Project Continuous Data Analysis/Statistical Modelling

Group 3: Dereje Mengist Belete, Feihong Du, Echo Capwell Forbang, Serge Martin Nkoumnga, Lin Tang, Shengmin Zhang

Introduction

United States Presidential election is widely concerned throughout the world. The republican candidate Trump Donald defeated the democratic candidate Hillary Clinton and won the election in 2016, which is far away from the result of the public opinion poll. In this study, we would like to investigate the factors, like socio-economic and demographic characteristics, which might impact the vote rate to Trump in 2016 election.

The dataset we used in this study is from the Kaggle website. In this dataset, we have 3141 observation indicating the election result and status of different areas of America. We implemented the regression approach which constituted a linear regression and a categorical regression of selected variables on the percentage of votes for the Republican party (GOP), to gain some insight in the association of certain variables with the vote rate of Trump.

Protocol

Research Question

For this project we will analyse part of the “2016 US Presidential Election Dataset”to examine the effect of the per capital money income obtained in the past 12 months on the amount of the percentage of GOP votes during the 2016 presidential elections, while accounting for other potential covariables, using a multiple linear regression model.

Did the economic characteristic (per capital money income obtained in the past 12 months) influence the percentage of GOP votes during the 2016 presidential elections?

What is the impact of socio-economic and demographic characteristics including race, gender, education, household, veterans and number of firms on the percentage of GOP votes during the 2016 presidential elections?

Moreover, we would like to find a way to predict the undecided counties (vote rate: 48%-52%) where our candidate may put more effort to on next election.

Study Method

A new dataset will be created based on the variables we choose to study. And then the distribution of each individual variable will be examined by univariate procedure, if it is necessary the missing value will be removed and the skewed data will be transformed. Bivariate relationships between the continuous variables will be examined with correlation and scatterplot matrix. Then we will split the dataset to training dataset and testing dataset (50:50) and normalize them respectively. The interactions between two variables will be created as new columns.

After the data preprocessing, the forward stepwise selection will be applied on the training dataset to find our "final model". After all the available variables access to the model, we will check whether the "final model" can fulfill all the assumptions.

In order to build logistic regression model, we will categorize the vote rate into 2 level: win (>50%) and lose(<50%). Then Compare your results/conclusions with those of the linear model to adapt the final model.

Fit the final linear model on the "testing" dataset to evaluate the model performance.

Data Exploration

We choose the income (per capital money income obtained in the past 12 months) as our key variable and other variables are white race, household, gender, education, firms and veterans. The distribution of all individual variables is examined by the univariate procedure. Since the firms and veterans are heavily right skewed (skewness: 16.15, 8.27 respectively), the log-transform is applied to those two variables. Missing value is checked by the mean procedure and there is no missing value in our dataset. Bivariate relationships between the continuous variables will be examined with correlation and scatterplot matrix.

We split the whole dataset data into training (1571 observations) and test (1570 observations) datasets with a proportion of 50:50. Both datasets were then standardized independently to have uniformity in interpreting our coefficients.

Simple linear regression

Simple linear regression (SLR) was used to assess the performance of key predictor (income) affecting the percentage of votes obtained by GOP candidate, the formula is as follows:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad \varepsilon_i \sim N(0, \delta^2),\ i = 1, \dots, n$$

$y_i = \ vote\ rate$ (Percentage of votes for Republican Party)

$x_i = \mathrm{income}$ (Per capita money income in past 12 months (dollars), 2009-2013)

The results showed that income affect the votes obtained by GOP candidate marginally significant (p-value <0.0001), our model then can be written as

$$y_i = 0.767 - 0.0000055x_i + \varepsilon_i \qquad \varepsilon_i \sim N(0, \delta^2)\ ,\ i = 1, \ldots, n$$

which means income increase one unit will result in the decrease of GOP support of 0.0000055%.

Multiple linear regression

We used forward stepwise regression method to build our model, absolute t-value, F-value and p-value were considered as the access criterions when we judge whether a new variable could be included or removed. Accordingly, newly added variable is of lowest residual sum of squares (SSR) and of highest R square, absolute t-values, F-values and p-value at 0.05 significance level in each step. Multicollinearity and confounding effect were checked by VIFs and p-value after each selection. With the addition of a new relevant variable included into our model, the already fitted variables did not change much in the significance and magnitude of the coefficients, hence no confounding effects. Variables which did not fulfill those criteria were omitted from the model.

According to the criterion we set, our model was built with main effects of income, white race, veterans, and firms. A similar approach was implemented to include influential two-way interactions terms into the model. The following interactions were included: household*education, white*education, income*firms, white*household, income*household, income*gender, education*veterans, white*firms and firms*veterans.

Checking assumptions

Multicollinearity: There is no multicollinearity amongst the predictors according to the VIF method, where we set VIF < 10. Normality: We used the QQ-plot of (studentized) residuals to check for the normality. There were slight deviations at the tails observed. However, normality can be assumed. Linearity: It was verified through the plot of (studentized) residuals versus predicted values. Points look randomly scattered around 0. No evidence of nonlinear pattern, hence linearity could be assumed. Furthermore, equality of variance (Homoscedasticity) was checked by examining the plot of (squared) residuals versus predicted values. Plot indicates that points equally randomly scattered around 0 point and therefore no evidence of heteroscedasticity.

Checking for outliers

Using the Cook's Distance threshold (4/n = 0.0025, n denotes the number of observations) to remove the outliers till convergence. The t-value, F-value and VIF value are the basis of our judgement. There were 93 outliers detected and removed in first time. Then the model was refitted by the new dataset without outliers. When checking the estimate coeffecients' siginificance, the p-value of white*household and firms*veterans are 0.4253 and 0.5833. These two variables were then removed from our model. Checking the outlier again and we found only 2 outliers which are slight beyond the threshold. The assumptions for linear regression were assessed once more and arrived at the conclusion that, all assumptions seem to be fulfilled (Appendix II).

Interpretation

Table 1 Statistical results of final model

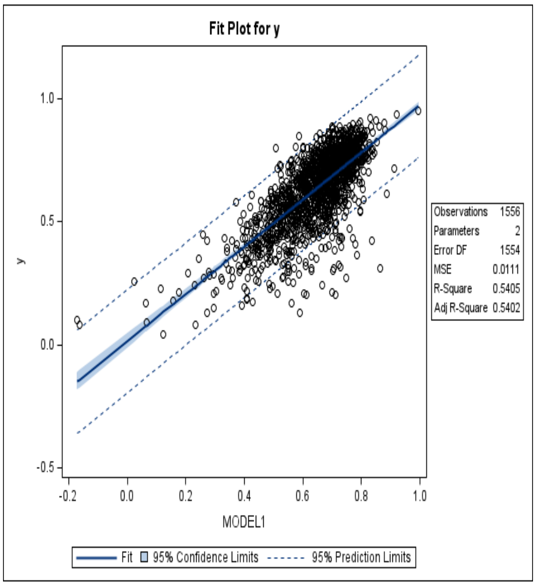
Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	0.66781	0.00292	228.53	<.0001	0
(x_1) Income	1	-0.00000585	5.656001E-7	-10.35	<.0001	1.43514
(x_2) White	1	0.00365	0.00019506	18.73	<.0001	1.48792
(x_3) Veterans	1	-0.02913	0.00300	-9.71	<.0001	3.06874
(x_4) Firms	1	-0.00793	0.00191	-4.15	<.0001	2.99856
(x_5) Household*Education	1	0.00798	0.00202	3.94	<.0001	1.65466
(x_6) White*Education	1	-0.00026287	0.00002623	-10.02	<.0001	1.43653
(x_7) Income*Firms	1	-0.00000121	2.679091E-7	-4.51	<.0001	1.94733
(x_8) Income*Household	1	0.00001319	0.00000265	4.98	<.0001	1.74923
(x_9) Income*Gender	1	6.895835E-7	2.030597E-7	3.40	0.0007	1.26568
(x_{10}) Education*Veterans	1	-0.00152	0.00034197	-4.46	<.0001	1.58034
(x_{11}) White*Firms	1	0.00028353	0.00009679	2.93	0.0034	1.39653

A comparation across all statistics presented in Table 1 indicated that race (white) was the strongest direct predictor of gop across mutiple indices, followed by income, Intwhiteeducation, and Veterans, Intincomehousehold, Inteducationveterans, Intincomefirms, firms, Inthouseholdeducation, Intincomegender, Intwhitefirms. Race obtained the largest T-value (t-value = 18.73, p < .0001), demonstrating that it made the largest contribution to the regression equation, while holding all other predictor variables constant. The squared structure coefficient (R^2 = .2755) demonstrated that race (white) explained the largest amount (27.55%) of variance in y, the predicted values of gop.

A good and standard statistical package has been used in this model, the R-squared value of the final multiple linear model is .5716 with all p < 0.01, which is quite an excellent, highly statistically significant result. About 58% (R^2%) of variations in y can be attributed to variations in x. Therefore y is reliably predictably with the multiple linear regression model:

$$Y_i = 0.67 - 5.85 * 10^{-6}x_1 + 3.65 * 10^{-3}x_2 - 0.03x_3 - 0.01x_4 + 0.01x_5 - 2.63 * 10^{-4}x_6 - 1.21 * 10^{-6}\,x_7 + 1.32 * 10^{-5}x_8 + 6.9 * 10^{-7}x_9 - 1.52 * 10^{-3}x_{10} + 2.84 * 10^{-4}x_{11}$$

By applying the final model we built on the testing dataset, the results (R^2=.5405, MSE = .0111) still shows its robust and reliance. The corresponding plot is shown as below.



The overall findings supported how both race (white) was the most significant direct contributor and income was the second most important direct contributor to predicting variance in gop, as reflected across different T-values, F-values and VIF. This might be because the white people are currently the major communities of voters, and income has a negative contribution on y, the predicted values of gop, which is corresponding to the reasearch that indicates, in poor counties, income is associated with Republican voting, while in many rich states, the relation between income and vote choice is nearly zero[1-5].

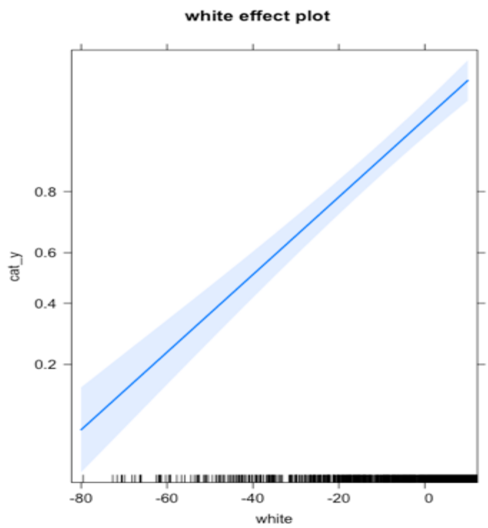
General Linear Model (GLM)

We dichotomized the response variable into two categories: win (vote rate greater than 50%) and lose (vote rate less than 50%), representing by 1 and 0 respectively. The GLM is based on the final model we have found in continuous part, fitting by the dichotomized training dataset. The formula of logistic regression can be written as following:

$$P(Y=1) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_p x_p)}}$$

For easier understanding, it also can be transformed as following equation:

$$log(\frac{P(Y=1)}{1-P(Y=1)}) = log(\frac{P(Y=1)}{P(y=0)}) = \beta_0 + beta_1 x_1 + \ldots + \beta_p x_p$$



Obviously, the left part of the equation is the log odds. We choose white race as an example to interpret our GLM. The effect of white race has linear relationship with the cat_y, displayed as right figure, in which the slope is the estimate coefficient of white race (0.04949103). The conditional odds of the white race would be exp(0.04949103) = 1.050736, indicating that when other variables keep constants the odds would increase 5.0736% if the white race increase one unit. What should be noticed here is that the firms and veterans had been log transformed in previous data process, so when discussing about those two variables, the conditional odds ratios would be the estimate coefficients directly. Using the testing dataset for prediction, the accuracy of our model is 87.66% and the precision of win and lose are 88.8% and 79% respectively.

Based on the estimate coeffecients of our fitted model, the interaction of househould and education (estimate coeffecient 0.219) shows highest positive effect on the vote rate to our candidate in 2016, while the log-transformed firms (estimate coefficient -0.8772) shows the highest negative effect on it. Moreover, the income and log-transformed veterans do not show significant effects to the vote rate in our model.

Discussion

Our study performed on county level data on the US elections 2016. In our study, we fitted the final model with training dataset and then applied it to test dataset for prediction. Overlap, t-test and variance test were used to test the similarity and difference. Overlap ranges from 0 (no overlap) to 1 (complete overlap). Since our sample size is greater than 75, we use the Dhat4 = 0.87 of overlap coefficient which implies, the boundaries of both response variables from actual and predicted dataset greatly coincide with one another. Significant difference was found when conducting t- and var- test, which means the prediction is marginally different from observations. Nevertheness, we may improve our model by including more relevant variables in the future.

In order to get the "undecided" counties which we need to focus for the next election, the linear regression model was applied to train the whole dataset with the variables already introduced. We then performed function of 'predict' to get perdicted values corresponding to each observation. Afterwards, we categoried the predicted values and defined interval located between 48%-52% as "undecided" situations. According to those criterion, we got 145 counties among 1738 counties, where the candiate may put more effort in the next election (see Appendix III).

Reference

1. Why do we need to re-use training parameters to transform test data?
https://sebastianraschka.com/faq/docs/scale-training-test.html

2. Jim Frost (2019), Using Confidence Intervals to Compare Means.
https://statisticsbyjim.com/hypothesis-testing/confidence-intervals-compare-means/

3. Dr. Frank Wood (2010), Inference in Regression Analysis.
http://www.stat.columbia.edu/~fwood/Teaching/w4315/Spring2010/lecture_4.pdf

4. Nathans, Laura L., Oswald, Frederick L. and Nimon, Kim. "Interpreting Multiple Linear Regression: A Guidebook of Variable Importance." *Practical Assessment, Research & Evaluation,* 17, no. 9 (2012) Practical Assessment, Research & Evaluation: https://hdl.handle.net/1911/71096

5. Andrew Gelman, Lane Kenworthy, Yu-Sung Su(2010). Income Inequality and Partisan Voting in the United States. 91(5):1204-1219.

Schedule

The distribution of tasks of the whole program is as follow:

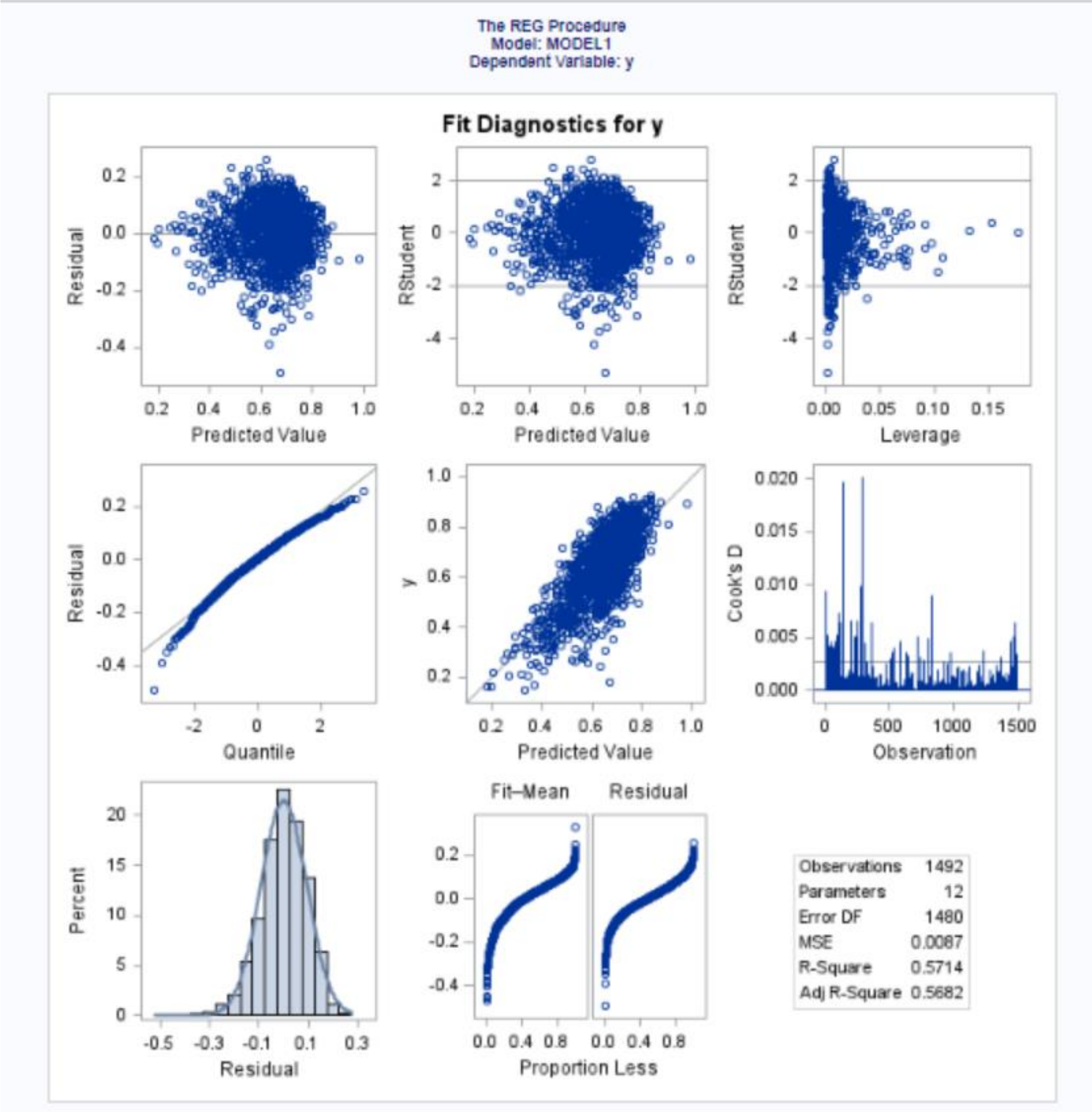
- 2020.11.28 Get an overall understanding of the research topic and work for the protocol together.
- 2020.11.28 - 2020.12.6 Feihong Du, Lin Tang and Shengmin Zhang work for the code part.
- 2020.12.7 Discuss and solve problems for the code part and work together to form the final version.
- 2020.12.7 – 2020.12.10 Dereje Mengist Belete, Echo Capwell Forbang and Serge Martin Nkoumnga work for the manuscript.
- 2020.12.10 Feihong Du, Lin Tang and Shengmin Zhang work together to revise the manuscript.
- 2020.12.11 All members work together to form the final report.

APPENDIX

Appendix I: Variables selected and their descriptions

Variable		Codes	Description
Outcome	GOP	Per vote gop	"Percentage of votes for Republican party"
Covariates (X)	**Income**	INC910213	"Per capita money income in past 12 months (2013 dollars), 2009-2013"
	Household	HSD310213	"Household, 2009-2013"
	Gender	SEX255214	"Female persons, percent, 2014"
	Education	EDU635213	"High school graduate or higher, percent of persons age 25+, 2009-2013"
	Firms	SBO001207	"Total number of firms, 2007"
	Veterans	VET605213	"Veterans, 2009-2013"
	White	RHI125214	"White alone, percent, 2014"

Appendix II: Plots of linear model assumptions



Appendix III: Counties need to be treated seriously for the next election

Adams	Brazor	Clevel	Greenv	Indian	Linn C	Nantuc	Richmo	St. He	Warren
Aiken	Calcas	Cumber	Grenad	Island	Lownde	Nash C	Roosev	St. Ja	Washin
Alachu	Calhou	Danvil	Hamilt	Jasper	Lucas	Nevada	Russel	St. Ma	Washoe
Alaska	Cape M	Darlin	Hampsh	Jeffer	Macomb	Ocean	San Ju	St. Ta	Wayne
Albema	Cass C	Delawa	Hardee	Kern C	Madera	Okaloo	Sangam	Talbot	Westmo
Anoka	Champa	Early	Harris	Kings	Madiso	Olmste	Santa	Talbot	Will C
Anson	Chicka	El Dor	Haywoo	Knox C	Manate	Orange	Scotla	Taliaf	Willia
Atkins	Chicot	Escamb	Hendry	La Sal	Mareng	Ouachi	Sedgwi	Terreb	Wilson
Atlant	Chitte	Fairfa	Hidalg	Lake C	Martin	Passai	Semino	Thurst	York C
Attala	Cibola	Fayett	Hoke C	Lancas	Meriwe	Peoria	Shelby	Tollan	
Bell C	Clacka	Frankl	Holmes	Larime	Minera	Pima C	Sonoma	Troup	
Berkel	Claibo	Galves	Hopewe	Lauder	Monroe	Pitt C	Spaldi	Tulare	
Bernal	Claren	Genese	Housto	Lee Co	Monter	Platte	Spokan	Tuscal	
Bexar	Clarke	Glouce	Iberia	Lenoir	Montgo	Provid	St. Ch	Ventur	
Bossie	Clear	Greene	Imperi	Lexing	Moreho	Rapide	St. Fr	Waltha	

```
/*Statistic Molding Project 2020*/


%let dir='/folders/myfolders/StatmodProject2020/';
libname moldproj &dir;

proc import out=moldproj.county
        datafile="/folders/myfolders/StatmodProject2020/county_facts_dictionary.csv" dbms=csv replace;
    getnames=yes;
    datarow=2;
run;

proc import out=moldproj.election
        datafile="/folders/myfolders/StatmodProject2020/US_County_Level_Presidential_Results_12-16_.csv"
        dbms=csv replace;
    getnames=yes;
    datarow=2;
run;

proc print data=moldproj.election (obs=5);
run;

/*Choose relevant variables and create a new dataset*/


%let x = white income household gender education firms veterans;

data election (keep=county y &x);
    set moldproj.election;
    rename county_name=county per_gop_2016=y RHI125214=white INC910213=income
        HSD310213=household SEX255214=gender EDU635213=education SBO001207=firms
        VET605213=veterans;
run;

proc print data=election (obs=10);
run;

*#####################################;
*### STEP 1: Descriptives Analysis ###;
*#####################################;

/*1. Univariate exploration of the data*/
proc univariate plot data=election;
    var y &x;
run;

/* Log-transform the right skewed data:  */
/* firms (skewness: 16.1494055) */
/* veterans (skewness: 8.26233385) */
data election;
    set election;

    if firms ne 0 then
        firms=log(firms);
    veterans=log(veterans);
run;

proc means data=election nmiss;
    var y &x;
run;

proc univariate plot data=election;
    var &x;
run;

/*2. Bivariate relationships*/
title "Correlation matrix and scatter-plots";

proc sgscatter data=election;
    matrix &x / diagonal=(histogram normal);
run;

title;

data election;
    set election;
    deel=firms/veterans;
run;

```

```
proc sgscatter data=election;
    matrix &x deel/ diagonal=(histogram normal);
run;

title;

proc corr data=election nosimple;
    var &x;
run;

*####################################;
*### STEP 2: Data spliting 50%:50% ###;
*####################################;

proc surveyselect data=election out=train_test_split method=srs samprate=0.5
        outall seed=123 noprint;
    samplingunit y;
run;

data train_set;
    set train_test_split;
    where Selected=1;
run;

data test_set;
    set train_test_split;
    where Selected=0;
run;

/* Standardize data! */
proc stdize data=train_set method=mean out=cent_train;
    var &x;
run;

proc stdize data=test_set method=mean out=cent_test;
    var &x;
run;

proc print data=cent_test;
run;

data cent_train;
    set cent_train;
    intwhiteincome=white*income;
    intwhitehousehold=white*household;
    intwhitegender=white*gender;
    intwhiteeducation=white*education;
    intwhitefirms=white*firms;
    intwhiteveterans=white*veterans;
    intincomehousehold=income*household;
    intincomegender=income*gender;
    intincomeeducation=income*education;
    intincomefirms=income*firms;
    intincomeveterans=income*veterans;
    inthouseholdgender=household*gender;
    inthouseholdeducation=household*education;
    inthouseholdfirms=household*firms;
    inthouseholdveterans=household*veterans;
    intgendereducation=gender*education;
    intgenderfirms=gender*firms;
    intgenderveterans=gender*veterans;
    inteducationfirms=education*firms;
    inteducationveterans=education*veterans;
    intfirmsveterans=firms*veterans;
run;

data cent_test;
    set cent_test;
    intwhiteincome=white*income;
    intwhitehousehold=white*household;
    intwhitegender=white*gender;
    intwhiteeducation=white*education;
    intwhitefirms=white*firms;
    intwhiteveterans=white*veterans;
    intincomehousehold=income*household;
    intincomegender=income*gender;
    intincomeeducation=income*education;
    intincomefirms=income*firms;
    intincomeveterans=income*veterans;
    inthouseholdgender=household*gender;
    inthouseholdeducation=household*education;
    inthouseholdfirms=household*firms;
```

```
    inthouseholdveterans=household*veterans;
    intgendereducation=gender*education;
    intgenderfirms=gender*firms;
    intgenderveterans=gender*veterans;
    inteducationfirms=education*firms;
    inteducationveterans=education*veterans;
    intfirmsveterans=firms*veterans;
run;

%let int_x = intwhiteincome intwhitehousehold intwhitegender intwhiteeducation intwhitefirms intwhiteveterans

       intincomehousehold intincomegender intincomeeducation intincomefirms intincomeveterans

       inthouseholdgender inthouseholdeducation inthouseholdfirms inthouseholdveterans

       intgendereducation intgenderfirms intgenderveterans

       inteducationfirms inteducationveterans

       intfirmsveterans;
*######################################;
*### STEP 3: Forward selection     ###;
*######################################;

/* 3.1 Simple Linear regression */
/* t:0.0376 F:61.83 */
proc reg data=train_set;
    model y=income / vif;
    title "Simple linear regression";
    run;
quit;

title;

/* First extra predictor*/
/* income need to stay in the model */
/* 3.2 determine the second fixed variable via t-value and p-value */
/* t:28.40 F:449.92 */
proc reg data=cent_train;
    model y=income white;
    run;
quit;

/* t:-9.36 F:76.41 */
proc reg data=cent_train;
    model y=income household;
    run;
quit;

/* t:-5.10 F:44.42  */
proc reg data=cent_train;
    model y=income gender;
    run;
quit;

/* t:3.22 F:36.28 */
proc reg data=cent_train;
    model y=income education;
    run;
quit;

/* t:-15.92 F:162.64 */
proc reg data=cent_train;
    model y=income firms;
    run;
quit;

/* t:-18.27 F:204.34 */
proc reg data=cent_train;
    model y=income veterans;
    run;
quit;

/* we choose white as a candidate varible, now check for it */
/* r_square:0.3626 vif < 10*/
proc reg data=cent_train;
    model y=income white / vif;
    run;
quit;

/* 3.3 Add white to model, determine the third fixed variable via t-value and p-value */
/*t:-2.28 F:302.47 */
```

```
proc reg data=cent_train;
    model y=income white household;
    run;
quit;

/*t:-5.05 F:315.56 */
proc reg data=cent_train;
    model y=income white gender;
    run;
quit;

/*t:-1.25 F:300.57 */
proc reg data=cent_train;
    model y=income white education;
    run;
quit;

/*t:-14.67 F:412.37 */
proc reg data=cent_train;
    model y=income white firms;
    run;
quit;

/*t:-17.04 F:451.66 */
proc reg data=cent_train;
    model y=income white veterans;
    run;
quit;

/* we choose veterans as a candidate varible, now check for it */
/* r_square:0.4605 vif<10*/
proc reg data=cent_train;
    model y=income white veterans / vif;
    run;
quit;

/* 3.3 Add veterans to model, determine the fourth fixed variable via t-value and p-value */
/*t:0.09 F:338.54 p:0.9249*/
proc reg data=cent_train;
    model y=income white veterans household;
    run;
quit;

/*t:-1.15 F:339.15 p:0.2485*/
proc reg data=cent_train;
    model y=income white veterans gender;
    run;
quit;

/*t:0.52 F:338.66 p:0.6029*/
proc reg data=cent_train;
    model y=income white veterans education;
    run;
quit;

/*t:-3.33 F:343.69 */
proc reg data=cent_train;
    model y=income white veterans firms;
    run;
quit;

/* we choose firms as a candidate varible, now check for it */
/* r_square:0.4639 vif<10*/
proc reg data=cent_train;
    model y=income white veterans firms/vif;
    run;
quit;

/* 3.5 Add firms, determine the fifth fixed variable via t-value f-value and p-value (reject */
/*r:0.22 F:274.79 p:0.8236*/
proc reg data=cent_train;
    model y=income white veterans firms household;
    run;
quit;

/*r:-0.94 F:275.11 p:0.3458*/
proc reg data=cent_train;
    model y=income white veterans firms gender;
    run;
quit;

/*r:-0.78 F:275.01 p:0.4330*/
```

```
proc reg data=cent_train;
    model y=income white veterans firms education;
    run;
quit;

/* 3.6 fix model with income white veterans and firms, r_square:0.4639 vif < 10 */
/* Add one interaction variable */
proc reg data=cent_train;
    model y=income white veterans firms intwhiteincome;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms intwhitehousehold;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms intwhitegender;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms intwhiteeducation;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms intwhitefirms;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms intwhiteveterans;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms intincomehousehold;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms intincomegender;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms intincomeeducation;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms intincomefirms;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms intincomeveterans;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdgender;
    run;
quit;

proc reg data=cent_train;
    model y=white income veterans firms inthouseholdeducation;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdfirms;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdveterans;
    run;
quit;
```

```
proc reg data=cent_train;
    model y=income white veterans firms intgendereducation;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms intgenderfirms;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms intgenderveterans;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inteducationfirms;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inteducationveterans;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms intfirmsveterans;
    run;
quit;

/* we choose inthouseholdeducation as a candidate interaction varible, now check for it */
/* r_square:0.5173 vif<10 */
proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation/ vif;
    run;

    /* 3.7 add inthouseholdeducation, determine the second interaction variable*/
proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteincome;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhitehousehold;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhitegender;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhitefirms;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteveterans;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intincomehousehold;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intincomegender;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intincomeeducation;
    run;
quit;
```

```
proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intincomefirms;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intincomeveterans;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation inthouseholdgender;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation inthouseholdfirms;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation inthouseholdveterans;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intgendereducation;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intgenderfirms;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intgenderveterans;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation inteducationfirms;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation inteducationveterans;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intfirmsveterans;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans Firms inthouseholdeducation intincomeeducation;
    run;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intincomeeducation;
    run;

    /* we choose intwhiteeducation as a candidate interaction varible, now check for it */
    /*r:0.5300 vif<10*/
proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation /
        vif;
    run;

    /* 3.8 add intwhiteeducation, determine the third interaction variable*/
proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intwhiteincome;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intwhitehousehold;
    run;
quit;
```

```
proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intwhitegender;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intwhitefirms;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intwhiteveterans;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomehousehold;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomegender;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomeeducation;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomeveterans;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        inthouseholdgender;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        inthouseholdfirms;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        inthouseholdveterans;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intgendereducation;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intgenderfirms;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intgenderveterans;
```

```
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        inteducationfirms;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        inteducationveterans;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intfirmsveterans;
    run;
quit;

/* we choose intincomefirms as a candidate interaction varible, now check for it */
/*r_square:0.5345 vif<10*/
proc reg data=cent_train;
    model y=white income veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms / vif;
    run;

    /* 3.9 add intincomefirms, determine the fourth interaction variable*/
proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhiteincome;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitegender;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitefirms;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhiteveterans;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intincomehousehold;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intincomegender;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intincomeeducation;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intincomeveterans;
    run;
quit;
```

```
proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms inthouseholdgender;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms inthouseholdfirms;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms inthouseholdveterans;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intgendereducation;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intgenderfirms;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intgenderveterans;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms inteducationfirms;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms inteducationveterans;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intfirmsveterans;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans Firms inthouseholdeducation intwhiteeducation
        intincomefirms intincomeeducation;
    run;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intincomeeducation;
    run;

    /* we choose intwhitehousehold as a candidate interaction varible, now check for it */
    /*r:0.5412 vif<10*/
proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold/ vif;
    run;

    /* 3.10 add intwhitehousehold, determine the fifth interaction variable*/
proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intwhiteincome;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intwhitegender;
```

```
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intwhitefirms;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intwhiteveterans;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intincomehousehold;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intincomegender;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intincomeeducation;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intincomeveterans;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold inthouseholdgender;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold inthouseholdfirms;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold inthouseholdveterans;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intgendereducation;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intgenderfirms;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intgenderveterans;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold inteducationfirms;
    run;
quit;

proc reg data=cent_train;
```

```
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold inteducationveterans;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intfirmsveterans;
    run;
quit;

proc reg data=cent_train;
    model y=white Income veterans Firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intincomeeducation;
    run;

    /* we choose intincomehousehold as a candidate interaction varible, now check for it */
proc reg data=cent_train;
    model y=white income veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intincomehousehold/ vif;
    run;

    /* 3.11 add intincomehousehold, determine the sisth interaction variable*/
proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intincomehousehold intwhiteincome;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intincomehousehold intwhitegender;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intincomehousehold intwhitefirms;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intincomehousehold intwhiteveterans;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intincomehousehold intincomegender;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intincomehousehold intincomeeducation;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intincomehousehold intincomeveterans;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intincomehousehold inthouseholdgender;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intincomehousehold inthouseholdfirms;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intincomehousehold inthouseholdveterans;
    run;
quit;
```

```
proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intincomehousehold intgendereducation;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intincomehousehold intgenderfirms;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intincomehousehold intgenderveterans;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intincomehousehold inteducationfirms;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intincomehousehold inteducationveterans;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intincomehousehold intfirmsveterans;
    run;
quit;

/* we choose  intincomegender as a candidate interaction varible, now check for it */
/*r:0.5531 vif<10*/
proc reg data=cent_train;
    model y=white income veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intincomehousehold intincomegender / vif;
    run;

    /* 3.12 add intincomegender, determine the seventh interaction variable*/
proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intincomehousehold intincomegender
        intwhiteincome;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intincomehousehold intincomegender
        intwhitegender;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intincomehousehold intincomegender
        intwhitefirms;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intincomehousehold intincomegender
        intwhiteveterans;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intincomehousehold intincomegender
        intincomeeducation;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
```

```
        intincomefirms intwhitehousehold intincomehousehold intincomegender
        intincomeveterans;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intincomehousehold intincomegender
        inthouseholdgender;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intincomehousehold intincomegender
        inthouseholdfirms;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intincomehousehold intincomegender
        inthouseholdveterans;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intincomehousehold intincomegender
        intgendereducation;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intincomehousehold intincomegender
        intgenderfirms;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intincomehousehold intincomegender
        intgenderveterans;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intincomehousehold intincomegender
        inteducationfirms;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intincomehousehold intincomegender
        inteducationveterans;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intincomehousehold intincomegender
        intfirmsveterans;
    run;
quit;

/* we choose inteducationveterans as a candidate interaction varible, now check for it */
/*r_square:0.5555 vif<10 */
proc reg data=cent_train;
    model y=white income veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intincomehousehold intincomegender
        inteducationveterans/ vif;
    run;

    /* 3.13 add inteducationveterans, determine the eighth interaction variable*/
proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intincomehousehold intincomegender
        inteducationveterans intwhiteincome;
    run;
```

```
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intincomehousehold intincomegender
        inteducationveterans intwhitegender;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intincomehousehold intincomegender
        inteducationveterans intwhitefirms;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intincomehousehold intincomegender
        inteducationveterans intwhiteveterans;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intincomehousehold intincomegender
        inteducationveterans intincomeeducation;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intincomehousehold intincomegender
        inteducationveterans intincomeveterans;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intincomehousehold intincomegender
        inteducationveterans inthouseholdgender;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intincomehousehold intincomegender
        inteducationveterans inthouseholdfirms;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intincomehousehold intincomegender
        inteducationveterans inthouseholdveterans;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intincomehousehold intincomegender
        inteducationveterans intgendereducation;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intincomehousehold intincomegender
        inteducationveterans intgenderfirms;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intincomehousehold intincomegender
        inteducationveterans intgenderveterans;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intincomehousehold intincomegender
```

```
        inteducationveterans inteducationfirms;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intincomehousehold intincomegender
        inteducationveterans intfirmsveterans;
    run;
quit;

/* we choose intwhitefirms as a candidate interaction varible, now check for it */
/*r_square:0.5577 vif<10*/
proc reg data=cent_train;
    model y=white income veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intincomehousehold intincomegender
        inteducationveterans intwhitefirms/ vif;
    run;

    /* 3.14 add intwhitefirms, determine the ninth interaction variable*/
proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intincomehousehold intincomegender
        inteducationveterans intwhitefirms intwhiteincome;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intincomehousehold intincomegender
        inteducationveterans intwhitefirms intwhitegender;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intincomehousehold intincomegender
        inteducationveterans intwhitefirms intwhiteveterans;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intincomehousehold intincomegender
        inteducationveterans intwhitefirms intincomeeducation;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intincomehousehold intincomegender
        inteducationveterans intwhitefirms intincomeveterans;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intincomehousehold intincomegender
        inteducationveterans intwhitefirms inthouseholdgender;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intincomehousehold intincomegender
        inteducationveterans intwhitefirms inthouseholdfirms;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intincomehousehold intincomegender
        inteducationveterans intwhitefirms inthouseholdveterans;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intincomehousehold intincomegender
        inteducationveterans intwhitefirms intgendereducation;
    run;
quit;
```

```
proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intincomehousehold intincomegender
        inteducationveterans intwhitefirms intgenderfirms;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intincomehousehold intincomegender
        inteducationveterans intwhitefirms intgenderveterans;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intincomehousehold intincomegender
        inteducationveterans intwhitefirms inteducationfirms;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intincomehousehold intincomegender
        inteducationveterans intwhitefirms intfirmsveterans;
    run;
quit;

/* we choose intfirmsveterans as a candidate interaction varible, now check for it */
/*r_square:0.5589 vif<10 */
proc reg data=cent_train;
    model y=white income veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intincomehousehold intincomegender
        inteducationveterans intwhitefirms intfirmsveterans/ vif;
    run;

    /* 3.15 add intfirmsveterans, determine the tenth interaction variable*/
proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intincomehousehold intincomegender
        inteducationveterans intwhitefirms intfirmsveterans intwhiteincome;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intincomehousehold intincomegender
        inteducationveterans intwhitefirms intfirmsveterans intwhitegender;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intincomehousehold intincomegender
        inteducationveterans intwhitefirms intfirmsveterans intwhiteveterans;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intincomehousehold intincomegender
        inteducationveterans intwhitefirms intfirmsveterans intincomeeducation;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intincomehousehold intincomegender
        inteducationveterans intwhitefirms intfirmsveterans intincomeveterans;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intincomehousehold intincomegender
        inteducationveterans intwhitefirms intfirmsveterans inthouseholdgender;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
```

```
        intincomefirms intwhitehousehold intincomehousehold intincomegender
        inteducationveterans intwhitefirms intfirmsveterans inthouseholdfirms;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intincomehousehold intincomegender
        inteducationveterans intwhitefirms intfirmsveterans inthouseholdveterans;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intincomehousehold intincomegender
        inteducationveterans intwhitefirms intfirmsveterans intgendereducation;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intincomehousehold intincomegender
        inteducationveterans intwhitefirms intfirmsveterans intgenderfirms;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intincomehousehold intincomegender
        inteducationveterans intwhitefirms intfirmsveterans intgenderveterans;
    run;
quit;

proc reg data=cent_train;
    model y=income white veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intincomehousehold intincomegender
        inteducationveterans intwhitefirms intfirmsveterans inteducationfirms;
    run;
quit;

/* all the p value > 0.05 and r_square is 0.5589, so the final model is  */
proc reg data=cent_train;
    model y=white income veterans firms inthouseholdeducation intwhiteeducation
        intincomefirms intwhitehousehold intincomehousehold intincomegender
        inteducationveterans intwhitefirms intfirmsveterans/ vif;
    run;
    *#####################################;
    *### STEP 4: Final Model verify    ###;
    *#####################################;
    %let reg_x= white income veterans firms inthouseholdeducation intwhiteeducation intincomefirms intwhitehouseho

    /*1. Check multicollinearity*/
proc reg data=cent_train;
    model y=&reg_x / vif;
    run;
quit;

/*2. Check the assumptions */
/* Normality: verify qqplot of (studentised) residuals */
proc reg data=cent_train;
    model y=&reg_x;
    output out=resid r=rman p=pman student=student;
    run;
quit;

/* Linearity: verify plot of (studentised) residuals vs predicted values*/
/* Homoscedasticity: verify (squared) residuals vs predicted values */
data resid2;
    set resid;
    rman2=rman**2;
run;

proc sgplot data=resid2;
    scatter x=pman y=rman2;
    refline 0 / axis=y lineattrs=(color=red);
run;

/*3. Check for outliers */
proc reg data=cent_train noprint;
    model y=&reg_x / r;
    output out=cookdis cookd=cdist;
    run;
```

```
quit;

data cookdis2;
    set cookdis;
    n=_n_;
run;

title "Cook's distance threshold 4/n=0.00255";

proc sgplot data=cookdis2;
    scatter x=n y=cdist;
    refline 0.0025 / axis=y lineattr=(color=red);
run;

title;

/*Remove outlier according to Cook's distance threshold 4/n=0.00255;*/
data cent_training_outlier_removed;
    set cookdis2;
    where cdist<0.00255;

    /* 93 outlier removed*/
run;

/*Build model after removing outlier*/
/*r_square:0.5714,intwhitehousehold(p=0.4253) and intfirmsveterans(p=0.5833)*/
proc reg data=cent_training_outlier_removed;
    model y=&reg_x / vif;
    run;
quit;

/*Remove intwhitehousehold and intfirmsveterans and check outlier again*/
/*Only 2 outliers and they are very close to the transhold, r_square:0.5714*/

%let reg_x= white income veterans firms inthouseholdeducation intwhiteeducation intincomefirms intincomehousehold

proc reg data=cent_training_outlier_removed noprint;
    model y=&reg_x / r;
    output out=cookdis cookd=cdist;
    run;
quit;

data cookdis2;
    set cookdis;
    n=_n_;
run;

title "Cook's distance threshold 4/n=0.00255";

proc sgplot data=cookdis2;
    scatter x=n y=cdist;
    refline 0.0025 / axis=y lineattr=(color=red);
run;

title;

/*Export the dataset for Logistic Regression*/
proc export data=cent_training_outlier_removed (keep=y county &reg_x)
        outfile="/folders/myfolders/StatmodProject2020/cent_train.csv" dbms=csv replace;
run;

proc export data=cent_test (keep=y county &reg_x)
        outfile="/folders/myfolders/StatmodProject2020/cent_test.csv" dbms=csv replace;
run;

*######################################;
*### STEP 5: Testing Final Model   ###;
*######################################;

proc reg data=cent_training_outlier_removed outest=train_estimate noprint;
    model y=&reg_x;
    run;

proc score data=cent_test score=train_estimate out=test_result type=parms
        predict;
    var y &reg_x;
run;

ods listing gpath=&dir;
ods graphics on;
ods select FitPlot;
```

```
proc reg data=test_result;
    model y=model1;
    plot model1*y / pred conf;
    run;
    ods graphics off;
    ods listing;
```

US_election_final

DU Feihong, Tang Lin, Zhang Shengmin

2020/12/11

```r
#import data#
mypath = '.'
setwd(mypath)
cent_train<-read.csv('cent_train.csv')
cent_test<-read.csv('cent_test.csv')
whole_dataset<- rbind(cent_train,cent_test)
str(cent_train)
str(cent_test)
plot(cent_train$y,xlab = 'series',ylab = 'vote_percent')

#---------------------------------------------------------------
#6.categorize y#

cent_train$cat_y <-ifelse(cent_train$y>0.5,1,0)
cent_test$cat_y <-ifelse(cent_test$y>0.5,1,0)



#fit with categorized y# state what we have found when compare two models # #
fit.raw.train <-lm(y ~ white+income+veterans+firms+inthouseholdeducation+
                     intwhiteeducation+intincomefirms+intincomehousehold+
                     intincomegender+inteducationveterans+intwhitefirms,
                   data=cent_train)



fit.cat.train <- glm(cat_y ~ white+income+veterans+firms+inthouseholdeducation+
                       intwhiteeducation+intincomefirms+intincomehousehold+
                       intincomegender+inteducationveterans+intwhitefirms,
                     family = binomial,data=cent_train)
fit.raw.train.summary<-summary(fit.raw.train)
fit.cat.train.summary<-summary(fit.cat.train)

beta1 <- fit.cat.train$coefficients[2]
con.odds.white <- exp(beta1)
plot(effects::effect('white', fit.cat.train))
```

```r
#7.---------------------------------------------------------------------------
glm.pred.raw <- predict(fit.raw.train, newdata = cent_test, type = "response")
glm.pred.cat <- ifelse(glm.pred.raw >0.5,1,0)
glm.pred.cat<-as.numeric(glm.pred.cat)
```

```r
#test result#
p_vs_a<-as.matrix(table(actual=cent_test$cat_y,predict=glm.pred.cat))
n = sum(p_vs_a) # number of instances
nc = nrow(p_vs_a) # number of classes
diag = diag(p_vs_a) # number of correctly classified instances per class
rowsums = apply(p_vs_a, 1, sum) # number of instances per class
colsums = apply(p_vs_a, 2, sum) # number of predictions per class
p = rowsums / n # distribution of instances over the actual classes
q = colsums / n # distribution of instances over the predicted classes
accuracy = sum(diag) / n
precision = diag / colsums
recall = diag / rowsums

#overlap---------------------------------------------------------
overlapEst(cent_test$cat_y, glm.pred.cat)
#difference--------------------------------------
t.test(glm.pred.cat,cent_test$cat_y,paired = T)
#variance------------------------------------
var.test(glm.pred.cat, cent_test$cat_y, alternative = "two.sided")
##########################################################################
fit.raw.whole <-lm(y ~ white+income+veterans+firms+inthouseholdeducation+
                     intwhiteeducation+intincomefirms+intincomehousehold+
                     intincomegender+inteducationveterans+intwhitefirms,
                   data=whole_dataset)
glm.pred.whole <- predict(fit.raw.whole, newdata = whole_dataset, type = "response")
glm.pred.whole<-as.numeric(glm.pred.whole)
glm.pred.whole.cat <- ifelse(glm.pred.whole >0.52,'win',
                             ifelse(glm.pred.whole<0.48,'lose','undecided'))
whole_dataset$result<-glm.pred.whole.cat



undecided <- whole_dataset%>%
  subset(result == 'undecided')

factor(whole_dataset$county)
factor(undecided$county)
table(undecided$county,undecided$result)
```