

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:

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

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

x_i = 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

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

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 coefficients' significance, 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

A comparison across all statistics presented in Table 1 indicated that race (white) was the strongest direct predictor of gop across multiple 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.

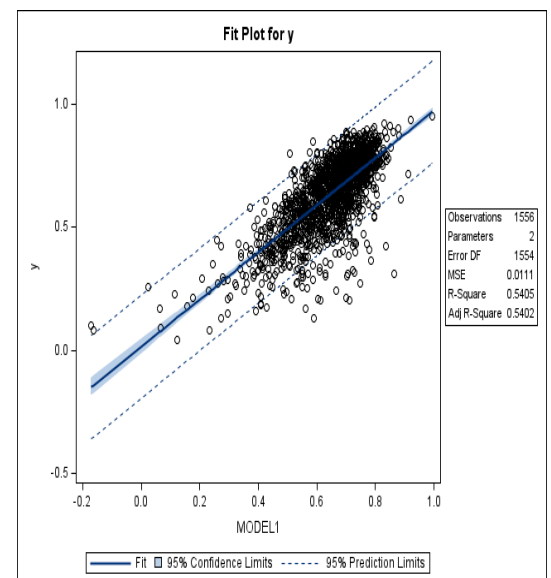
A good and standard statistical package has been used in this model, the R-squared value of the final multiple linear model is .5716 with all $p < 0.01$, which is quite an excellent, highly statistically significant result. About 58% ($R^2\%$) of variations in y can be attributed to variations in x.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

Table 1 Statistical results of final model

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	0.66781	0.00292	228.53	<.0001
(x ₁) Income	1	-0.00000585	5.656001E-7	-10.35	<.0001
(x ₂) White	1	0.00365	0.00019506	18.73	<.0001
(x ₃) Veterans	1	-0.02913	0.00300	-9.71	<.0001
(x ₄) Firms	1	-0.00793	0.00191	-4.15	<.0001
(x ₅) Household*Education	1	0.00798	0.00202	3.94	<.0001
(x ₆) White*Education	1	-0.00026287	0.00002623	-10.02	<.0001
(x ₇) Income*Firms	1	-0.00000121	2.679091E-7	-4.51	<.0001
(x ₈) Income*Household	1	0.00001319	0.00000265	4.98	<.0001
(x ₉) Income*Gender	1	6.895835E-7	2.030597E-7	3.40	0.0007
(x ₁₀) Education*Veterans	1	-0.00152	0.00034197	-4.46	<.0001
(x ₁₁) White*Firms	1	0.00028353	0.00009679	2.93	0.0034
Variance Inflation					
					1.43514
					1.48792
					3.06874
					2.99856
					1.65466
					1.43653
					1.94733
					1.74923
					1.26568
					1.58034
					1.39653



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 research 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 zero¹⁻⁵.

General Linear Model (GLM)

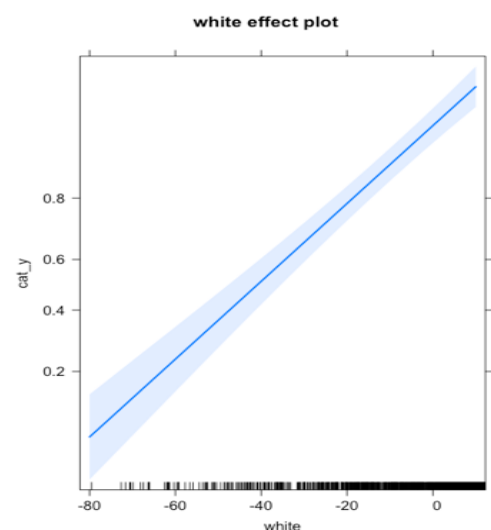
We dichotomized the response variable into two categories: win (vote rate greater than 50%) and lose (vote rate less than 50%), representing by 1 and 0 respectively. The GLM is based on the final model we have found in continuous part, fitting by the dichotomized training dataset. The formula of logistic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

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

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

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

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 coefficients of our fitted model, the interaction of household and education (estimate coefficient 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 $D_{hat4} = 0.87$ 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. Nevertheless, 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 predicted values corresponding to each observation. Afterwards, we categorized 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 candidate 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:

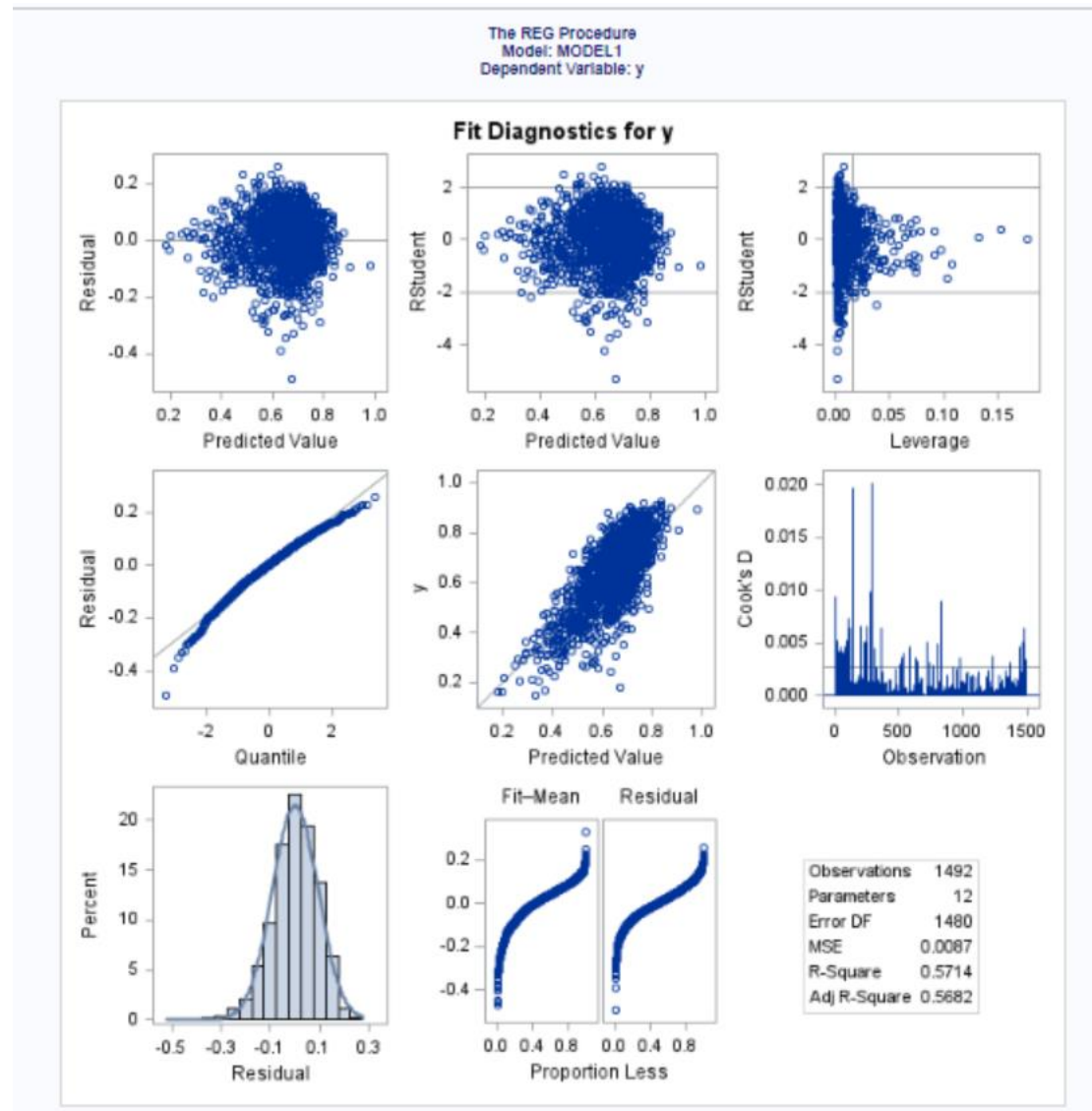
- 2020.11.28 Get an overall understanding of the research topic and work for the protocol together.
- 2020.11.28 - 2020.12.6 Feihong Du, Lin Tang and Shengmin Zhang work for the code part.
- 2020.12.7 Discuss and solve problems for the code part and work together to form the final version.
- 2020.12.7 – 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	

```

1  /*Statistic Molding Project 2020*/
2
3
4
5  %let dir='/folders/myfolders/StatmodProject2020/';
6  libname moldproj &dir;
7
8  proc import out=moldproj.county
9      datafile="/folders/myfolders/StatmodProject2020/county_facts_dictionary.csv" dbms=csv replace;
10     getnames=yes;
11     datarow=2;
12 run;
13
14 proc import out=moldproj.election
15     datafile="/folders/myfolders/StatmodProject2020/US_County_Level_Presidential_Results_12-16_.csv"
16     dbms=csv replace;
17     getnames=yes;
18     datarow=2;
19 run;
20
21 proc print data=moldproj.election (obs=5);
22 run;
23
24 /*Choose relevant variables and create a new dataset*/
25
26 %let x = white income household gender education firms veterans;
27
28 data election (keep=county y &x);
29     set moldproj.election;
30     rename county_name=county per_gop_2016=y RHI125214=white INC910213=income
31           HSD310213=household SEX255214=gender EDU635213=education SB0001207=firms
32           VET605213=veterans;
33 run;
34
35 proc print data=election (obs=10);
36 run;
37
38 #####;
39 ### STEP 1: Descriptives Analysis ###;
40 #####;
41
42 /*1. Univariate exploration of the data*/
43 proc univariate plot data=election;
44     var y &x;
45 run;
46
47 /* Log-transform the right skewed data: */
48 /* firms (skewness: 16.1494055) */
49 /* veterans (skewness: 8.26233385) */
50 data election;
51     set election;
52
53     if firms ne 0 then
54         firms=log(firms);
55     veterans=log(veterans);
56 run;
57
58 proc means data=election nmiss;
59     var y &x;
60 run;
61
62 proc univariate plot data=election;
63     var &x;
64 run;
65
66 /*2. Bivariate relationships*/
67 title "Correlation matrix and scatter-plots";
68
69 proc sgscatter data=election;
70     matrix &x / diagonal=(histogram normal);
71 run;
72
73 title;
74
75 data election;
76     set election;
77     deel=firms/veterans;
78 run;

```

```

73 proc sgscatter data=election;
74     matrix &x deel/ diagonal=(histogram normal);
75 run;
76
77 title;
78
79 proc corr data=election nosimple;
80     var &x;
81 run;
82
83 #####;
84 *** STEP 2: Data splitting 50%:50% ***;
85 #####;
86
87 proc surveyselect data=election out=train_test_split method=srs samprate=0.5
88     outall seed=123 noprint;
89     samplingunit y;
90 run;
91
92 data train_set;
93     set train_test_split;
94     where Selected=1;
95 run;
96
97 data test_set;
98     set train_test_split;
99     where Selected=0;
100 run;
101
102 /* Standardize data! */
103 proc stdize data=train_set method=mean out=cent_train;
104     var &x;
105 run;
106
107 proc stdize data=test_set method=mean out=cent_test;
108     var &x;
109 run;
110
111 proc print data=cent_test;
112 run;
113
114 data cent_train;
115     set cent_train;
116     intwhiteincome=white*income;
117     intwhitehousehold=white*household;
118     intwhitegender=white*gender;
119     intwhiteeducation=white*education;
120     intwhitefirms=white*firms;
121     intwhiteveterans=white*veterans;
122     intincomehousehold=income*household;
123     intincomegender=income*gender;
124     intincomeeducation=income*education;
125     intincomefirms=income*firms;
126     intincomeveterans=income*veterans;
127     inhouseholdgender=household*gender;
128     inhouseholdeducation=household*education;
129     inhouseholdfirms=household*firms;
130     inhouseholdveterans=household*veterans;
131     intgendereducation=gender*education;
132     intgenderfirms=gender*firms;
133     intgenderveterans=gender*veterans;
134     inteducationfirms=education*firms;
135     inteducationveterans=education*veterans;
136     intfirmsveterans=firms*veterans;
137 run;
138
139 data cent_test;
140     set cent_test;
141     intwhiteincome=white*income;
142     intwhitehousehold=white*household;
143     intwhitegender=white*gender;
144     intwhiteeducation=white*education;
145     intwhitefirms=white*firms;
146     intwhiteveterans=white*veterans;
147     intincomehousehold=income*household;
148     intincomegender=income*gender;
149     intincomeeducation=income*education;
150     intincomefirms=income*firms;
151     intincomeveterans=income*veterans;
152     inhouseholdgender=household*gender;
153     inhouseholdeducation=household*education;
154     inhouseholdfirms=household*firms;

```

```

146     inhouseholdveterans=household*veterans;
147     intgendereducation=gender*education;
148     intgenderfirms=gender*firms;
149     intgenderveterans=gender*veterans;
150     inteducationfirms=education*firms;
151     inteducationveterans=education*veterans;
152     intfirmsveterans=firms*veterans;
153 run;
154 %let int_x = intwhiteincome intwhitehousehold intwhitegender intwhiteeducation intwhitefirms intwhiteveterans
155
156     intincomehousehold intincomegender intincomeeducation intincomefirms intincomeveterans
157
158     inhouseholdgender inhouseholdeducation inhouseholdfirms inhouseholdveterans
159
160     intgendereducation intgenderfirms intgenderveterans
161
162     inteducationfirms inteducationveterans
163
164     intfirmsveterans;
165 #####;
166 *** STEP 3: Forward selection ***;
167 #####;
168 /* 3.1 Simple Linear regression */
169 /* t:0.0376 F:61.83 */
170 proc reg data=train_set;
171     model y=income / vif;
172     title "Simple linear regression";
173 run;
174 quit;
175
176 title;
177
178 /* First extra predictor*/
179 /* income need to stay in the model */
180 /* 3.2 determine the second fixed variable via t-value and p-value */
181 /* t:28.40 F:449.92 */
182 proc reg data=cent_train;
183     model y=income white;
184 run;
185 quit;
186
187 /* t:-9.36 F:76.41 */
188 proc reg data=cent_train;
189     model y=income household;
190 run;
191 quit;
192
193 /* t:-5.10 F:44.42 */
194 proc reg data=cent_train;
195     model y=income gender;
196 run;
197 quit;
198
199 /* t:3.22 F:36.28 */
200 proc reg data=cent_train;
201     model y=income education;
202 run;
203 quit;
204
205 /* t:-15.92 F:162.64 */
206 proc reg data=cent_train;
207     model y=income firms;
208 run;
209 quit;
210
211 /* t:-18.27 F:204.34 */
212 proc reg data=cent_train;
213     model y=income veterans;
214 run;
215 quit;
216
217 /* we choose white as a candidate variable, now check for it */
218 /* r_square:0.3626 vif < 10*/
219 proc reg data=cent_train;
220     model y=income white / vif;
221 run;
222 quit;
223
224 /* 3.3 Add white to model, determine the third fixed variable via t-value and p-value */
225 /*t:-2.28 F:302.47 */

```

```

219 proc reg data=cent_train;
220     model y=income white household;
221     run;
222 quit;
223
224 /*t:-5.05 F:315.56 */
225 proc reg data=cent_train;
226     model y=income white gender;
227     run;
228 quit;
229
230 /*t:-1.25 F:300.57 */
231 proc reg data=cent_train;
232     model y=income white education;
233     run;
234 quit;
235
236 /*t:-14.67 F:412.37 */
237 proc reg data=cent_train;
238     model y=income white firms;
239     run;
240 quit;
241
242 /*t:-17.04 F:451.66 */
243 proc reg data=cent_train;
244     model y=income white veterans;
245     run;
246 quit;
247
248 /* we choose veterans as a candidate variable, now check for it */
249 /* r_square:0.4605 vif<10*/
250 proc reg data=cent_train;
251     model y=income white veterans / vif;
252     run;
253 quit;
254
255 /* 3.3 Add veterans to model, determine the fourth fixed variable via t-value and p-value */
256 /*t:0.09 F:338.54 p:0.9249*/
257 proc reg data=cent_train;
258     model y=income white veterans household;
259     run;
260 quit;
261
262 /*t:-1.15 F:339.15 p:0.2485*/
263 proc reg data=cent_train;
264     model y=income white veterans gender;
265     run;
266 quit;
267
268 /*t:0.52 F:338.66 p:0.6029*/
269 proc reg data=cent_train;
270     model y=income white veterans education;
271     run;
272 quit;
273
274 /*t:-3.33 F:343.69 */
275 proc reg data=cent_train;
276     model y=income white veterans firms;
277     run;
278 quit;
279
280 /* we choose firms as a candidate variable, now check for it */
281 /* r_square:0.4639 vif<10*/
282 proc reg data=cent_train;
283     model y=income white veterans firms/vif;
284     run;
285 quit;
286
287 /* 3.5 Add firms, determine the fifth fixed variable via t-value f-value and p-value (reject */
288 /*r:0.22 F:274.79 p:0.8236*/
289 proc reg data=cent_train;
290     model y=income white veterans firms household;
291     run;
292 quit;
293
294 /*r:-0.94 F:275.11 p:0.3458*/
295 proc reg data=cent_train;
296     model y=income white veterans firms gender;
297     run;
298 quit;
299
300 /*r:-0.78 F:275.01 p:0.4330*/

```

```
292 proc reg data=cent_train;
293     model y=income white veterans firms education;
294     run;
295 quit;
296
297 /* 3.6 fix model with income white veterans and firms, r_square:0.4639 vif < 10 */
298 /* Add one interaction variable */
299 proc reg data=cent_train;
300     model y=income white veterans firms intwhiteincome;
301     run;
302 quit;
303
304 proc reg data=cent_train;
305     model y=income white veterans firms intwhitehousehold;
306     run;
307 quit;
308
309 proc reg data=cent_train;
310     model y=income white veterans firms intwhitegender;
311     run;
312 quit;
313
314 proc reg data=cent_train;
315     model y=income white veterans firms intwhiteeducation;
316     run;
317 quit;
318
319 proc reg data=cent_train;
320     model y=income white veterans firms intwhitefirms;
321     run;
322 quit;
323
324 proc reg data=cent_train;
325     model y=income white veterans firms intwhiteveterans;
326     run;
327 quit;
328
329 proc reg data=cent_train;
330     model y=income white veterans firms intincomehousehold;
331     run;
332 quit;
333
334 proc reg data=cent_train;
335     model y=income white veterans firms intincomegender;
336     run;
337 quit;
338
339 proc reg data=cent_train;
340     model y=income white veterans firms intincomeeducation;
341     run;
342 quit;
343
344 proc reg data=cent_train;
345     model y=income white veterans firms intincomefirms;
346     run;
347 quit;
348
349 proc reg data=cent_train;
350     model y=income white veterans firms intincomeveterans;
351     run;
352 quit;
353
354 proc reg data=cent_train;
355     model y=income white veterans firms inthouseholdgender;
356     run;
357 quit;
358
359 proc reg data=cent_train;
360     model y=white income veterans firms inthouseholdeducation;
361     run;
362 quit;
363
364 proc reg data=cent_train;
365     model y=income white veterans firms inthouseholdfirms;
366     run;
367 quit;
368
369 proc reg data=cent_train;
370     model y=income white veterans firms inthouseholdveterans;
371     run;
372 quit;
```

```

365 proc reg data=cent_train;
366     model y=income white veterans firms intgendereducation;
367     run;
368 quit;
369
370 proc reg data=cent_train;
371     model y=income white veterans firms intgenderfirms;
372     run;
373 quit;
374
375 proc reg data=cent_train;
376     model y=income white veterans firms intgenderveterans;
377     run;
378 quit;
379
380 proc reg data=cent_train;
381     model y=income white veterans firms inteducationfirms;
382     run;
383 quit;
384
385 proc reg data=cent_train;
386     model y=income white veterans firms inteducationveterans;
387     run;
388 quit;
389
390 proc reg data=cent_train;
391     model y=income white veterans firms intfirmsveterans;
392     run;
393 quit;
394
395 /* we choose inhouseholdeducation as a candidate interaction variable, now check for it */
396 /* r_square:0.5173 vif<10 */
397
398 proc reg data=cent_train;
399     model y=income white veterans firms inhouseholdeducation / vif;
400     run;
401 quit;
402
403 /* 3.7 add inhouseholdeducation, determine the second interaction variable*/
404
405 proc reg data=cent_train;
406     model y=income white veterans firms inhouseholdeducation intwhiteincome;
407     run;
408 quit;
409
410 proc reg data=cent_train;
411     model y=income white veterans firms inhouseholdeducation intwhitehousehold;
412     run;
413 quit;
414
415 proc reg data=cent_train;
416     model y=income white veterans firms inhouseholdeducation intwhitegender;
417     run;
418 quit;
419
420 proc reg data=cent_train;
421     model y=income white veterans firms inhouseholdeducation intwhiteeducation;
422     run;
423 quit;
424
425 proc reg data=cent_train;
426     model y=income white veterans firms inhouseholdeducation intwhitefirms;
427     run;
428 quit;
429
430 proc reg data=cent_train;
431     model y=income white veterans firms inhouseholdeducation intwhiteveterans;
432     run;
433 quit;
434
435 proc reg data=cent_train;
436     model y=income white veterans firms inhouseholdeducation intincomehousehold;
437     run;
438 quit;
439
440 proc reg data=cent_train;
441     model y=income white veterans firms inhouseholdeducation intincomegender;
442     run;
443 quit;
444
445 proc reg data=cent_train;
446     model y=income white veterans firms inhouseholdeducation intincomeeducation;
447     run;
448 quit;

```



```

438 proc reg data=cent_train;
439     model y=income white veterans firms inhouseholdeducation intincomefirms;
440     run;
441 quit;
442
443 proc reg data=cent_train;
444     model y=income white veterans firms inhouseholdeducation intincomeveterans;
445     run;
446 quit;
447
448 proc reg data=cent_train;
449     model y=income white veterans firms inhouseholdeducation inhouseholdgender;
450     run;
451 quit;
452
453 proc reg data=cent_train;
454     model y=income white veterans firms inhouseholdeducation inhouseholdfirms;
455     run;
456 quit;
457
458 proc reg data=cent_train;
459     model y=income white veterans firms inhouseholdeducation inhouseholdveterans;
460     run;
461 quit;
462
463 proc reg data=cent_train;
464     model y=income white veterans firms inhouseholdeducation intgendereducation;
465     run;
466 quit;
467
468 proc reg data=cent_train;
469     model y=income white veterans firms inhouseholdeducation intgenderfirms;
470     run;
471 quit;
472
473 proc reg data=cent_train;
474     model y=income white veterans firms inhouseholdeducation intgenderveterans;
475     run;
476 quit;
477
478 proc reg data=cent_train;
479     model y=income white veterans firms inhouseholdeducation inteducationfirms;
480     run;
481 quit;
482
483 proc reg data=cent_train;
484     model y=income white veterans firms inhouseholdeducation inteducationveterans;
485     run;
486 quit;
487
488 proc reg data=cent_train;
489     model y=income white veterans firms inhouseholdeducation intfirmsveterans;
490     run;
491 quit;
492
493 proc reg data=cent_train;
494     model y=income white veterans Firms inhouseholdeducation intincomeeducation;
495     run;
496
497 proc reg data=cent_train;
498     model y=income white veterans firms inhouseholdeducation intincomeeducation;
499     run;
500
501 /* we choose intwhiteeducation as a candidate interaction variable, now check for it */
502 /*r:0.5300 vif<10*/
503
504 proc reg data=cent_train;
505     model y=income white veterans firms inhouseholdeducation intwhiteeducation /
506         vif;
507     run;
508
509 /* 3.8 add intwhiteeducation, determine the third interaction variable*/
510
511 proc reg data=cent_train;
512     model y=income white veterans firms inhouseholdeducation intwhiteeducation
513         intwhiteincome;
514     run;
515 quit;
516
517 proc reg data=cent_train;
518     model y=income white veterans firms inhouseholdeducation intwhiteeducation
519         intwhitehousehold;
520     run;
521 quit;
522

```

```
511
512 proc reg data=cent_train;
513     model y=income white veterans firms inhouseholdeducation intwhiteeducation
514           intwhitegender;
515     run;
516 quit;
517
518 proc reg data=cent_train;
519     model y=income white veterans firms inhouseholdeducation intwhiteeducation
520           intwhitefirms;
521     run;
522 quit;
523
524 proc reg data=cent_train;
525     model y=income white veterans firms inhouseholdeducation intwhiteeducation
526           intwhiteveterans;
527     run;
528 quit;
529
530 proc reg data=cent_train;
531     model y=income white veterans firms inhouseholdeducation intwhiteeducation
532           intincomehousehold;
533     run;
534 quit;
535
536 proc reg data=cent_train;
537     model y=income white veterans firms inhouseholdeducation intwhiteeducation
538           intincomegender;
539     run;
540 quit;
541
542 proc reg data=cent_train;
543     model y=income white veterans firms inhouseholdeducation intwhiteeducation
544           intincomeeducation;
545     run;
546 quit;
547
548 proc reg data=cent_train;
549     model y=income white veterans firms inhouseholdeducation intwhiteeducation
550           intincomefirms;
551     run;
552 quit;
553
554 proc reg data=cent_train;
555     model y=income white veterans firms inhouseholdeducation intwhiteeducation
556           inhouseholdgender;
557     run;
558 quit;
559
560 proc reg data=cent_train;
561     model y=income white veterans firms inhouseholdeducation intwhiteeducation
562           inhouseholdfirms;
563     run;
564 quit;
565
566 proc reg data=cent_train;
567     model y=income white veterans firms inhouseholdeducation intwhiteeducation
568           inhouseholdveterans;
569     run;
570 quit;
571
572 proc reg data=cent_train;
573     model y=income white veterans firms inhouseholdeducation intwhiteeducation
574           intgendereducation;
575     run;
576 quit;
577
578 proc reg data=cent_train;
579     model y=income white veterans firms inhouseholdeducation intwhiteeducation
580           intgenderfirms;
581     run;
582 quit;
583
584 proc reg data=cent_train;
585     model y=income white veterans firms inhouseholdeducation intwhiteeducation
586           intgenderveterans;
```

9/22

```

657
658 proc reg data=cent_train;
659     model y=income white veterans firms inhouseholdeducation intwhiteeducation
660           intincomefirms inhouseholdgender;
661     run;
662     quit;
663
664 proc reg data=cent_train;
665     model y=income white veterans firms inhouseholdeducation intwhiteeducation
666           intincomefirms inhouseholdfirms;
667     run;
668     quit;
669
670 proc reg data=cent_train;
671     model y=income white veterans firms inhouseholdeducation intwhiteeducation
672           intincomefirms inhouseholdveterans;
673     run;
674     quit;
675
676 proc reg data=cent_train;
677     model y=income white veterans firms inhouseholdeducation intwhiteeducation
678           intincomefirms intgendereducation;
679     run;
680     quit;
681
682 proc reg data=cent_train;
683     model y=income white veterans firms inhouseholdeducation intwhiteeducation
684           intincomefirms intgenderfirms;
685     run;
686     quit;
687
688 proc reg data=cent_train;
689     model y=income white veterans firms inhouseholdeducation intwhiteeducation
690           intincomefirms intgenderveterans;
691     run;
692     quit;
693
694 proc reg data=cent_train;
695     model y=income white veterans firms inhouseholdeducation intwhiteeducation
696           intincomefirms inteducationfirms;
697     run;
698     quit;
699
700 proc reg data=cent_train;
701     model y=income white veterans firms inhouseholdeducation intwhiteeducation
702           intincomefirms inteducationveterans;
703     run;
704     quit;
705
706 proc reg data=cent_train;
707     model y=income white veterans Firms inhouseholdeducation intwhiteeducation
708           intincomefirms intincomeeducation;
709     run;
710
711 proc reg data=cent_train;
712     model y=income white veterans firms inhouseholdeducation intwhiteeducation
713           intincomefirms intincomeeducation;
714     run;
715
716 /* we choose intwhitehousehold as a candidate interaction variable, now check for it */
717 /*r:0.5412 vif<10*/
718 proc reg data=cent_train;
719     model y=income white veterans firms inhouseholdeducation intwhiteeducation
720           intincomefirms intwhitehousehold/ vif;
721     run;
722
723 /* 3.10 add intwhitehousehold, determine the fifth interaction variable*/
724 proc reg data=cent_train;
725     model y=income white veterans firms inhouseholdeducation intwhiteeducation
726           intincomefirms intwhitehousehold intwhiteincome;
727     run;
728     quit;
729
730 proc reg data=cent_train;
731     model y=income white veterans firms inhouseholdeducation intwhiteeducation
732           intincomefirms intwhitehousehold intwhitegender;

```

```
730     run;
731 quit;
732
733 proc reg data=cent_train;
734     model y=income white veterans firms inhouseholdeducation intwhiteeducation
735           intincomefirms intwhitehousehold intwhitefirms;
736     run;
737 quit;
738
739 proc reg data=cent_train;
740     model y=income white veterans firms inhouseholdeducation intwhiteeducation
741           intincomefirms intwhitehousehold intwhiteveterans;
742     run;
743 quit;
744
745 proc reg data=cent_train;
746     model y=income white veterans firms inhouseholdeducation intwhiteeducation
747           intincomefirms intwhitehousehold intincomehousehold;
748     run;
749 quit;
750
751 proc reg data=cent_train;
752     model y=income white veterans firms inhouseholdeducation intwhiteeducation
753           intincomefirms intwhitehousehold intincomegender;
754     run;
755 quit;
756
757 proc reg data=cent_train;
758     model y=income white veterans firms inhouseholdeducation intwhiteeducation
759           intincomefirms intwhitehousehold intincomeeducation;
760     run;
761 quit;
762
763 proc reg data=cent_train;
764     model y=income white veterans firms inhouseholdeducation intwhiteeducation
765           intincomefirms intwhitehousehold intincomeveterans;
766     run;
767 quit;
768
769 proc reg data=cent_train;
770     model y=income white veterans firms inhouseholdeducation intwhiteeducation
771           intincomefirms intwhitehousehold inhouseholdgender;
772     run;
773 quit;
774
775 proc reg data=cent_train;
776     model y=income white veterans firms inhouseholdeducation intwhiteeducation
777           intincomefirms intwhitehousehold inhouseholdveterans;
778     run;
779 quit;
780
781 proc reg data=cent_train;
782     model y=income white veterans firms inhouseholdeducation intwhiteeducation
783           intincomefirms intwhitehousehold intgendereducation;
784     run;
785 quit;
786
787 proc reg data=cent_train;
788     model y=income white veterans firms inhouseholdeducation intwhiteeducation
789           intincomefirms intwhitehousehold intgenderfirms;
790     run;
791 quit;
792
793 proc reg data=cent_train;
794     model y=income white veterans firms inhouseholdeducation intwhiteeducation
795           intincomefirms intwhitehousehold intgenderveterans;
796     run;
797 quit;
798
799 proc reg data=cent_train;
800     model y=income white veterans firms inhouseholdeducation intwhiteeducation
801           intincomefirms intwhitehousehold inteducationfirms;
802     run;
803 quit;
804
805 proc reg data=cent_train;
```

875

```

876
877 proc reg data=cent_train;
878     model y=income white veterans firms inhouseholdeducation intwhiteeducation
879           intincomefirms intwhitehousehold intincomehousehold intgendereducation;
880     run;
881 quit;
882
883 proc reg data=cent_train;
884     model y=income white veterans firms inhouseholdeducation intwhiteeducation
885           intincomefirms intwhitehousehold intincomehousehold intgenderfirms;
886     run;
887 quit;
888
889 proc reg data=cent_train;
890     model y=income white veterans firms inhouseholdeducation intwhiteeducation
891           intincomefirms intwhitehousehold intincomehousehold intgenderveterans;
892     run;
893 quit;
894
895 proc reg data=cent_train;
896     model y=income white veterans firms inhouseholdeducation intwhiteeducation
897           intincomefirms intwhitehousehold intincomehousehold inteducationfirms;
898     run;
899 quit;
900
901 proc reg data=cent_train;
902     model y=income white veterans firms inhouseholdeducation intwhiteeducation
903           intincomefirms intwhitehousehold intincomehousehold inteducationveterans;
904     run;
905 quit;
906
907 proc reg data=cent_train;
908     model y=income white veterans firms inhouseholdeducation intwhiteeducation
909           intincomefirms intwhitehousehold intincomehousehold intfirmsveterans;
910     run;
911 quit;
912
913 /* we choose intincomegender as a candidate interaction variable, now check for it */
914 /*r:0.5531 vif<10*/
915
916 proc reg data=cent_train;
917     model y=white income veterans firms inhouseholdeducation intwhiteeducation
918           intincomefirms intwhitehousehold intincomehousehold intincomegender / vif;
919     run;
920
921 /* 3.12 add intincomegender, determine the seventh interaction variable*/
922
923 proc reg data=cent_train;
924     model y=income white veterans firms inhouseholdeducation intwhiteeducation
925           intincomefirms intwhitehousehold intincomehousehold intincomegender
926           intwhiteincome;
927     run;
928 quit;
929
930 proc reg data=cent_train;
931     model y=income white veterans firms inhouseholdeducation intwhiteeducation
932           intincomefirms intwhitehousehold intincomehousehold intincomegender
933           intwhitegender;
934     run;
935 quit;
936
937 proc reg data=cent_train;
938     model y=income white veterans firms inhouseholdeducation intwhiteeducation
939           intincomefirms intwhitehousehold intincomehousehold intincomegender
940           intwhitefirms;
941     run;
942 quit;
943
944 proc reg data=cent_train;
945     model y=income white veterans firms inhouseholdeducation intwhiteeducation
946           intincomefirms intwhitehousehold intincomehousehold intincomegender
947           intwhiteveterans;
948     run;
949 quit;
950
951 proc reg data=cent_train;
952     model y=income white veterans firms inhouseholdeducation intwhiteeducation
953           intincomefirms intwhitehousehold intincomehousehold intincomegender
954           intincomeeducation;
955     run;
956 quit;
957
958 proc reg data=cent_train;
959     model y=income white veterans firms inhouseholdeducation intwhiteeducation

```

```

949     intincomefirms intwhitehousehold intincomehousehold intincomegender
950     intincomeveterans;
951     run;
952 quit;

953 proc reg data=cent_train;
954     model y=income white veterans firms inhouseholdeducation intwhiteeducation
955           intincomefirms intwhitehousehold intincomehousehold intincomegender
956           inhouseholdgender;
957     run;
958 quit;

959 proc reg data=cent_train;
960     model y=income white veterans firms inhouseholdeducation intwhiteeducation
961           intincomefirms intwhitehousehold intincomehousehold intincomegender
962           inhouseholdfirms;
963     run;
964 quit;

965 proc reg data=cent_train;
966     model y=income white veterans firms inhouseholdeducation intwhiteeducation
967           intincomefirms intwhitehousehold intincomehousehold intincomegender
968           inhouseholdveterans;
969     run;
970 quit;

971 proc reg data=cent_train;
972     model y=income white veterans firms inhouseholdeducation intwhiteeducation
973           intincomefirms intwhitehousehold intincomehousehold intincomegender
974           intgendereducation;
975     run;
976 quit;

977 proc reg data=cent_train;
978     model y=income white veterans firms inhouseholdeducation intwhiteeducation
979           intincomefirms intwhitehousehold intincomehousehold intincomegender
980           intgenderfirms;
981     run;
982 quit;

983 proc reg data=cent_train;
984     model y=income white veterans firms inhouseholdeducation intwhiteeducation
985           intincomefirms intwhitehousehold intincomehousehold intincomegender
986           intgenderveterans;
987     run;
988 quit;

989 proc reg data=cent_train;
990     model y=income white veterans firms inhouseholdeducation intwhiteeducation
991           intincomefirms intwhitehousehold intincomehousehold intincomegender
992           inteducationfirms;
993     run;
994 quit;

995 proc reg data=cent_train;
996     model y=income white veterans firms inhouseholdeducation intwhiteeducation
997           intincomefirms intwhitehousehold intincomehousehold intincomegender
998           inteducationveterans;
999     run;
1000 quit;

1001 proc reg data=cent_train;
1002     model y=income white veterans firms inhouseholdeducation intwhiteeducation
1003           intincomefirms intwhitehousehold intincomehousehold intincomegender
1004           intfirmsveterans;
1005     run;
1006 quit;

1007 /* we choose inteducationveterans as a candidate interaction variable, now check for it */
1008 /*r_square:0.5555 vif<10 */
1009 proc reg data=cent_train;
1010     model y=white income veterans firms inhouseholdeducation intwhiteeducation
1011           intincomefirms intwhitehousehold intincomehousehold intincomegender
1012           inteducationveterans/ vif;
1013     run;

1014 /* 3.13 add inteducationveterans, determine the eighth interaction variable*/
1015 proc reg data=cent_train;
1016     model y=income white veterans firms inhouseholdeducation intwhiteeducation
1017           intincomefirms intwhitehousehold intincomehousehold intincomegender
1018           inteducationveterans intwhiteincome;
1019     run;
1020
1021

```



```

1022 quit;
1023
1024 proc reg data=cent_train;
1025     model y=income white veterans firms inhouseholdeducation intwhiteeducation
1026           intincomefirms intwhitehousehold intincomehousehold intincomegender
1027           inteducationveterans intwhitegender;
1028 run;
1029 quit;
1030
1031 proc reg data=cent_train;
1032     model y=income white veterans firms inhouseholdeducation intwhiteeducation
1033           intincomefirms intwhitehousehold intincomehousehold intincomegender
1034           inteducationveterans intwhitefirms;
1035 run;
1036 quit;
1037
1038 proc reg data=cent_train;
1039     model y=income white veterans firms inhouseholdeducation intwhiteeducation
1040           intincomefirms intwhitehousehold intincomehousehold intincomegender
1041           inteducationveterans intwhiteveterans;
1042 run;
1043 quit;
1044
1045 proc reg data=cent_train;
1046     model y=income white veterans firms inhouseholdeducation intwhiteeducation
1047           intincomefirms intwhitehousehold intincomehousehold intincomegender
1048           inteducationveterans intincomeeducation;
1049 run;
1050 quit;
1051
1052 proc reg data=cent_train;
1053     model y=income white veterans firms inhouseholdeducation intwhiteeducation
1054           intincomefirms intwhitehousehold intincomehousehold intincomegender
1055           inteducationveterans intincomeveterans;
1056 run;
1057 quit;
1058
1059 proc reg data=cent_train;
1060     model y=income white veterans firms inhouseholdeducation intwhiteeducation
1061           intincomefirms intwhitehousehold intincomehousehold intincomegender
1062           inteducationveterans inhouseholdgender;
1063 run;
1064 quit;
1065
1066 proc reg data=cent_train;
1067     model y=income white veterans firms inhouseholdeducation intwhiteeducation
1068           intincomefirms intwhitehousehold intincomehousehold intincomegender
1069           inteducationveterans inhouseholdfirms;
1070 run;
1071 quit;
1072
1073 proc reg data=cent_train;
1074     model y=income white veterans firms inhouseholdeducation intwhiteeducation
1075           intincomefirms intwhitehousehold intincomehousehold intincomegender
1076           inteducationveterans inhouseholdveterans;
1077 run;
1078 quit;
1079
1080 proc reg data=cent_train;
1081     model y=income white veterans firms inhouseholdeducation intwhiteeducation
1082           intincomefirms intwhitehousehold intincomehousehold intincomegender
1083           inteducationveterans intgendereducation;
1084 run;
1085 quit;
1086
1087 proc reg data=cent_train;
1088     model y=income white veterans firms inhouseholdeducation intwhiteeducation
1089           intincomefirms intwhitehousehold intincomehousehold intincomegender
1090           inteducationveterans intgenderveterans;
1091 run;
1092 quit;
1093
1094 proc reg data=cent_train;
1095     model y=income white veterans firms inhouseholdeducation intwhiteeducation
1096           intincomefirms intwhitehousehold intincomehousehold intincomegender

```

```

1095     inteducationveterans inteducationfirms;
1096 run;
1097 quit;
1098
1099 proc reg data=cent_train;
1100     model y=income white veterans firms inhouseholdeducation intwhiteeducation
1101           intincomefirms intwhitehousehold intincomehousehold intincomegender
1102           inteducationveterans intfirmsveterans;
1103 run;
1104 quit;
1105
1106 /* we choose intwhitefirms as a candidate interaction variable, now check for it */
1107 /*r_square:0.5577 vif<10*/
1108 proc reg data=cent_train;
1109     model y=income white veterans firms inhouseholdeducation intwhiteeducation
1110           intincomefirms intwhitehousehold intincomehousehold intincomegender
1111           inteducationveterans intwhitefirms/ vif;
1112 run;
1113
1114 /* 3.14 add intwhitefirms, determine the ninth interaction variable*/
1115 proc reg data=cent_train;
1116     model y=income white veterans firms inhouseholdeducation intwhiteeducation
1117           intincomefirms intwhitehousehold intincomehousehold intincomegender
1118           inteducationveterans intwhitefirms intwhiteincome;
1119 run;
1120 quit;
1121
1122 proc reg data=cent_train;
1123     model y=income white veterans firms inhouseholdeducation intwhiteeducation
1124           intincomefirms intwhitehousehold intincomehousehold intincomegender
1125           inteducationveterans intwhitefirms intwhitegender;
1126 run;
1127 quit;
1128
1129 proc reg data=cent_train;
1130     model y=income white veterans firms inhouseholdeducation intwhiteeducation
1131           intincomefirms intwhitehousehold intincomehousehold intincomegender
1132           inteducationveterans intwhitefirms intwhiteveterans;
1133 run;
1134 quit;
1135
1136 proc reg data=cent_train;
1137     model y=income white veterans firms inhouseholdeducation intwhiteeducation
1138           intincomefirms intwhitehousehold intincomehousehold intincomegender
1139           inteducationveterans intwhitefirms intincomeeducation;
1140 run;
1141 quit;
1142
1143 proc reg data=cent_train;
1144     model y=income white veterans firms inhouseholdeducation intwhiteeducation
1145           intincomefirms intwhitehousehold intincomehousehold intincomegender
1146           inteducationveterans intwhitefirms inhouseholdgender;
1147 run;
1148 quit;
1149
1150 proc reg data=cent_train;
1151     model y=income white veterans firms inhouseholdeducation intwhiteeducation
1152           intincomefirms intwhitehousehold intincomehousehold intincomegender
1153           inteducationveterans intwhitefirms inhouseholdfirms;
1154 run;
1155 quit;
1156
1157 proc reg data=cent_train;
1158     model y=income white veterans firms inhouseholdeducation intwhiteeducation
1159           intincomefirms intwhitehousehold intincomehousehold intincomegender
1160           inteducationveterans intwhitefirms inhouseholdveterans;
1161 run;
1162 quit;
1163
1164 proc reg data=cent_train;
1165     model y=income white veterans firms inhouseholdeducation intwhiteeducation
1166           intincomefirms intwhitehousehold intincomehousehold intincomegender
1167           inteducationveterans intwhitefirms intgendereducation;
1168 run;
1169 quit;

```

```

1168
1169 proc reg data=cent_train;
1170     model y=income white veterans firms inhouseholdeducation intwhiteeducation
1171           intincomefirms intwhitehousehold intincomehousehold intincomegender
1172           inteducationveterans intwhitefirms intgenderfirms;
1173     run;
1174 quit;
1175
1176 proc reg data=cent_train;
1177     model y=income white veterans firms inhouseholdeducation intwhiteeducation
1178           intincomefirms intwhitehousehold intincomehousehold intincomegender
1179           inteducationveterans intwhitefirms intgenderveterans;
1180     run;
1181 quit;
1182
1183 proc reg data=cent_train;
1184     model y=income white veterans firms inhouseholdeducation intwhiteeducation
1185           intincomefirms intwhitehousehold intincomehousehold intincomegender
1186           inteducationveterans intwhitefirms inteducationfirms;
1187     run;
1188 quit;
1189
1190 proc reg data=cent_train;
1191     model y=income white veterans firms inhouseholdeducation intwhiteeducation
1192           intincomefirms intwhitehousehold intincomehousehold intincomegender
1193           inteducationveterans intwhitefirms intfirmsveterans;
1194     run;
1195 quit;
1196
1197 /* we choose intfirmsveterans as a candidate interaction variable, now check for it */
1198 /* r_square:0.5589 vif<10 */
1199 proc reg data=cent_train;
1200     model y=white income veterans firms inhouseholdeducation intwhiteeducation
1201           intincomefirms intwhitehousehold intincomehousehold intincomegender
1202           inteducationveterans intwhitefirms intfirmsveterans / vif;
1203     run;
1204
1205 /* 3.15 add intfirmsveterans, determine the tenth interaction variable*/
1206 proc reg data=cent_train;
1207     model y=income white veterans firms inhouseholdeducation intwhiteeducation
1208           intincomefirms intwhitehousehold intincomehousehold intincomegender
1209           inteducationveterans intwhitefirms intfirmsveterans intwhiteincome;
1210     run;
1211 quit;
1212
1213 proc reg data=cent_train;
1214     model y=income white veterans firms inhouseholdeducation intwhiteeducation
1215           intincomefirms intwhitehousehold intincomehousehold intincomegender
1216           inteducationveterans intwhitefirms intfirmsveterans intwhitegender;
1217     run;
1218 quit;
1219
1220 proc reg data=cent_train;
1221     model y=income white veterans firms inhouseholdeducation intwhiteeducation
1222           intincomefirms intwhitehousehold intincomehousehold intincomegender
1223           inteducationveterans intwhitefirms intfirmsveterans intincomeeducation;
1224     run;
1225 quit;
1226
1227 proc reg data=cent_train;
1228     model y=income white veterans firms inhouseholdeducation intwhiteeducation
1229           intincomefirms intwhitehousehold intincomehousehold intincomegender
1230           inteducationveterans intwhitefirms intfirmsveterans intincomeveterans;
1231     run;
1232 quit;
1233
1234 proc reg data=cent_train;
1235     model y=income white veterans firms inhouseholdeducation intwhiteeducation
1236           intincomefirms intwhitehousehold intincomehousehold intincomegender
1237           inteducationveterans intwhitefirms intfirmsveterans inhouseholdgender;
1238     run;
1239 quit;
1240
1241 proc reg data=cent_train;
1242     model y=income white veterans firms inhouseholdeducation intwhiteeducation

```

```

1241     intincomefirms intwhitehousehold intincomehousehold intincomegender
1242     inteducationveterans intwhitefirms intfirmsveterans inthouseholdfirms;
1243 run;
1244 quit;
1245
1246 proc reg data=cent_train;
1247     model y=income white veterans firms inthouseholdeducation intwhiteeducation
1248           intincomefirms intwhitehousehold intincomehousehold intincomegender
1249           inteducationveterans intwhitefirms intfirmsveterans inthouseholdveterans;
1250 run;
1251 quit;
1252
1253 proc reg data=cent_train;
1254     model y=income white veterans firms inthouseholdeducation intwhiteeducation
1255           intincomefirms intwhitehousehold intincomehousehold intincomegender
1256           inteducationveterans intwhitefirms intfirmsveterans intgendereducation;
1257 run;
1258 quit;
1259
1260 proc reg data=cent_train;
1261     model y=income white veterans firms inthouseholdeducation intwhiteeducation
1262           intincomefirms intwhitehousehold intincomehousehold intincomegender
1263           inteducationveterans intwhitefirms intfirmsveterans intgenderfirms;
1264 run;
1265 quit;
1266
1267 proc reg data=cent_train;
1268     model y=income white veterans firms inthouseholdeducation intwhiteeducation
1269           intincomefirms intwhitehousehold intincomehousehold intincomegender
1270           inteducationveterans intwhitefirms intfirmsveterans intgenderveterans;
1271 run;
1272 quit;
1273
1274 proc reg data=cent_train;
1275     model y=income white veterans firms inthouseholdeducation intwhiteeducation
1276           intincomefirms intwhitehousehold intincomehousehold intincomegender
1277           inteducationveterans intwhitefirms intfirmsveterans inteducationfirms;
1278 run;
1279 quit;
1280
1281 /* all the p value > 0.05 and r_square is 0.5589, so the final model is */
1282 proc reg data=cent_train;
1283     model y=white income veterans firms inthouseholdeducation intwhiteeducation
1284           intincomefirms intwhitehousehold intincomehousehold intincomegender
1285           inteducationveterans intwhitefirms intfirmsveterans / vif;
1286 run;
1287 #####;
1288 ### STEP 4: Final Model verify ###;
1289 #####;
1290 %let reg_x= white income veterans firms inthouseholdeducation intwhiteeducation intincomefirms intwhitehouseho
1291
1292 /*1. Check multicollinearity*/
1293 proc reg data=cent_train;
1294     model y=&reg_x / vif;
1295 run;
1296 quit;
1297
1298 /*2. Check the assumptions */
1299 /* Normality: verify qqplot of (studentised) residuals */
1300 proc reg data=cent_train;
1301     model y=&reg_x;
1302     output out=resid r=rman p=pman student=student;
1303 run;
1304 quit;
1305
1306 /* Linearity: verify plot of (studentised) residuals vs predicted values*/
1307 /* Homoscedasticity: verify (squared) residuals vs predicted values */
1308 data resid2;
1309     set resid;
1310     rman2=rman**2;
1311 run;
1312
1313 proc sgplot data=resid2;
1314     scatter x=pman y=rman2;
1315     refline 0 / axis=y lineattrs=(color=red);
1316 run;
1317
1318 /*3. Check for outliers */
1319 proc reg data=cent_train noprint;
1320     model y=&reg_x / r;
1321     output out=cookdis cookd=cdist;
1322 run;
1323

```

```

1314 quit;
1315
1316 data cookdis2;
1317     set cookdis;
1318     n=_n_;
1319 run;
1320 title "Cook's distance threshold 4/n=0.00255";
1321
1322 proc sgplot data=cookdis2;
1323     scatter x=n y=cdist;
1324     refline 0.0025 / axis=y lineattr=(color=red);
1325 run;
1326 title;
1327
1328 /*Remove outlier according to Cook's distance threshold 4/n=0.00255;*/
1329 data cent_training_outlier_removed;
1330     set cookdis2;
1331     where cdist<0.00255;
1332
1333     /* 93 outlier removed*/
1334 run;
1335
1336 /*Build model after removing outlier*/
1337 /*r_square:0.5714,intwhitehousehold(p=0.4253) and intfirmsveterans(p=0.5833)*/
1338 proc reg data=cent_training_outlier_removed;
1339     model y=&reg_x / vif;
1340 run;
1341 quit;
1342
1343 /*Remove intwhitehousehold and intfirmsveterans and check outlier again*/
1344 /*Only 2 outliers and they are very close to the transhold, r_square:0.5714*/
1345
1346 %let reg_x= white income veterans firms inhouseholdeducation intwhiteeducation intincomefirms intincomehousehold
1347
1348 proc reg data=cent_training_outlier_removed noprint;
1349     model y=&reg_x / r;
1350     output out=cookdis cookd=cdist;
1351 run;
1352 quit;
1353
1354 data cookdis2;
1355     set cookdis;
1356     n=_n_;
1357 run;
1358 title "Cook's distance threshold 4/n=0.00255";
1359
1360 proc sgplot data=cookdis2;
1361     scatter x=n y=cdist;
1362     refline 0.0025 / axis=y lineattr=(color=red);
1363 run;
1364 title;
1365
1366 /*Export the dataset for Logistic Regression*/
1367 proc export data=cent_training_outlier_removed (keep=y county &reg_x)
1368     outfile="/folders/myfolders/StatmodProject2020/cent_train.csv" dbms=csv replace;
1369 run;
1370
1371 proc export data=cent_test (keep=y county &reg_x)
1372     outfile="/folders/myfolders/StatmodProject2020/cent_test.csv" dbms=csv replace;
1373 run;
1374
1375 #####;
1376 *### STEP 5: Testing Final Model ###;
1377 #####;
1378
1379 proc reg data=cent_training_outlier_removed outest=train_estimate noprint;
1380     model y=&reg_x;
1381 run;
1382
1383 proc score data=cent_test score=train_estimate out=test_result type=parms
1384     predict;
1385     var y &reg_x;
1386 run;
1387
1388 ods listing gpath=&dir;
1389 ods graphics on;
1390 ods select FitPlot;

```

```
1387 proc reg data=test_result;  
1388     model y=model1;  
1389     plot model1*y / pred conf;  
1390     run;  
1391     ods graphics off;  
1392     ods listing;  
1393  
1394  
1395  
1396  
1397  
1398  
1399  
1400  
1401  
1402  
1403  
1404  
1405  
1406  
1407  
1408  
1409  
1410  
1411  
1412  
1413  
1414  
1415  
1416  
1417  
1418  
1419  
1420  
1421  
1422  
1423  
1424  
1425  
1426  
1427  
1428  
1429  
1430  
1431  
1432  
1433  
1434  
1435  
1436  
1437  
1438  
1439  
1440  
1441  
1442  
1443  
1444  
1445  
1446  
1447  
1448  
1449  
1450  
1451  
1452  
1453  
1454  
1455  
1456  
1457  
1458  
1459
```

US_election_final

DU Feihong, Tang Lin, Zhang Shengmin

2020/12/11

```
#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+inhouseholdeducation+
  intwhiteeducation+intincomefirms+intincomehousehold+
  intincomegender+inteducationveterans+intwhitefirms,
  data=cent_train)

fit.cat.train <- glm(cat_y ~ white+income+veterans+firms+inhouseholdeducation+
  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))

#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)
```

```

#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$cat_y, alternative = "two.sided")
#####
fit.raw.whole <-lm(y ~ white+income+veterans+firms+inhouseholdeducation+
                  intwhiteeducation+intincomefirms+intincomehousehold+
                  intincomegender+inteducationveterans+intwhitefirms,
                  data=whole_dataset)
glm.pred.whole <- predict(fit.raw.whole, newdata = whole_dataset, type = "response")
glm.pred.whole<-as.numeric(glm.pred.whole)
glm.pred.whole.cat <- ifelse(glm.pred.whole >0.52,'win',
                             ifelse(glm.pred.whole<0.48,'lose','undecided'))
whole_dataset$result<-glm.pred.whole.cat

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

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

```