Forecasting elections of US2020

Xu Xinyi 1003941577, Minchen Cai 1000523800, Haona Yang 1004949531, Mingkai Zhang 1004903063

2020/10/31

Forecasting elections of US2020

Xu Xinyi 1003941577, Minchen Cai 1000523800, Haona Yang 1004949531, Mingkai Zhang 1004903063

2020/10/31

Model

Here we are interested in predicting the popular vote outcome of the 2020 American federal election (include citation). To do this we are employing a post-stratification technique. In the following sub-sections I will describe the model specifics and the post-stratification calculation.

Model Specifics

I fit a multilevel logistic regression, which is used to model a binary dependent variable, to predict the voting support for Donald Trump and Joe Biden. I use age as a numeric variable, gender and hispanic as dummy variables, to model the probability of voting for both candidates. The model is given by:

$$p(Yi = Donald\ Trump|state_j) = logit^{-1}(\beta_{j0} + \beta_{ji}^{age} + b_{[i]}^{gender} + b_{[i]}^{hispanic})$$

β_{j0} represents the baseline intercept of the variable in group j. Here group j stands for the level 2 variable of state. β_{j1} represents the slope for voters' age fraction among all voters in each state. When the respondent's state changes, the coefficient of both the intercept and the slope will be varying as well. Moreover, each categorical variable is represented as dummy variables and corresponds to the changing coefficients. $b_{[i]}^{gender}$ represents female and male and $b_{[i]}^{hispanic}$) represents dummy variables for Cuban, Mexican, Not hispanic, Other hispanic and Puerto Rican respectively.

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: vote_trump ~ (1 + age | state) + gender + hispanic
##    Data: futher_survey_cleaning
##
##      AIC      BIC   logLik deviance df.resid
##   7325.4   7391.4  -3652.7   7305.4     5440
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.8133 -0.8902 -0.6407  0.9939  2.0686
##
## Random effects:
```

```
##  Groups Name        Variance  Std.Dev. Corr
##  state  (Intercept) 0.1827446 0.42749
##         age         0.0001432 0.01197  -0.90
## Number of obs: 5450, groups:  state, 51
##
## Fixed effects:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -0.09795    0.64679  -0.151    0.880
## genderMale              0.52623    0.05674   9.275   <2e-16 ***
## hispanicCuban           0.33886    0.75675   0.448    0.654
## hispanicMexican        -0.99562    0.65021  -1.531    0.126
## hispanicNot Hispanic   -0.35737    0.64327  -0.556    0.579
## hispanicOther Hispanic -0.55955    0.65625  -0.853    0.394
## hispanicPuerto Rican   -1.10647    1.40415  -0.788    0.431
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) gndrMl hspncC hspncM hspnNH hspnOH
## genderMale  -0.021
## hispanicCbn -0.841  0.011
## hispancMxcn -0.981 -0.015  0.837
## hspncNtHspn -0.992 -0.024  0.846  0.986
## hspncOthrHs -0.971 -0.007  0.830  0.966  0.976
## hspncPrtRcn -0.454 -0.019  0.388  0.451  0.457  0.447
## optimizer (Nelder_Mead) convergence code: 0 (OK)
## Model is nearly unidentifiable: very large eigenvalue
##  - Rescale variables?
## Model is nearly unidentifiable: large eigenvalue ratio
##  - Rescale variables?
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: vote_biden ~ (1 + age | state) + gender + hispanic
##    Data: futher_survey_cleaning
##
##      AIC      BIC   logLik deviance df.resid
##   7397.1   7463.2  -3688.6   7377.1     5440
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.7121 -0.9339 -0.5776  0.9516  1.8132
##
## Random effects:
##  Groups Name        Variance  Std.Dev. Corr
##  state  (Intercept) 1.017e-01 0.318928
##         age         8.495e-05 0.009217 -0.80
## Number of obs: 5450, groups:  state, 51
##
## Fixed effects:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -0.33332    0.65762  -0.507   0.6123
## genderMale             -0.51810    0.05628  -9.206   <2e-16 ***
## hispanicCuban          -0.12874    0.77026  -0.167   0.8673
## hispanicMexican         1.08917    0.66187   1.646   0.0998 .
## hispanicNot Hispanic    0.57067    0.65583   0.870   0.3842
## hispanicOther Hispanic  0.81303    0.66822   1.217   0.2237
## hispanicPuerto Rican    1.45515    1.40743   1.034   0.3012
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) gndrMl hspncC hspncM hspnNH hspnOH
## genderMale  -0.019
## hispanicCbn -0.844  0.010
## hispancMxcn -0.985 -0.014  0.840
## hspncNtHspn -0.995 -0.024  0.848  0.987
## hspncOthrHs -0.975 -0.007  0.833  0.969  0.978
## hspncPrtRcn -0.463 -0.018  0.397  0.460  0.465  0.456
## optimizer (Nelder_Mead) convergence code: 0 (OK)
## Model failed to converge with max|grad| = 0.00372646 (tol = 0.002, component 1)
## Model is nearly unidentifiable: very large eigenvalue
##  - Rescale variables?
## Model is nearly unidentifiable: large eigenvalue ratio
##  - Rescale variables?
```

```
##
## Attaching package: 'sjlabelled'
```

```
## The following object is masked from 'package:forcats':
##
##     as_factor
```

```
## The following object is masked from 'package:dplyr':
##
##     as_label
```

```
## Learn more about sjPlot with 'browseVignettes("sjPlot")'.
```

```
##
## Attaching package: 'sjmisc'
```

```
## The following objects are masked from 'package:sjlabelled':
##
##     to_character, to_factor, to_label, to_numeric
```

```
## The following object is masked from 'package:purrr':
##
##     is_empty
```

```
## The following object is masked from 'package:tidyr':
##
##     replace_na
```

```
## The following object is masked from 'package:tibble':
##
##     add_case
```

vote trump

Predictors

Odds Ratios

CI

p

(Intercept)

0.91

0.26 – 3.22

0.880

gender [Male]

1.69

1.51 – 1.89

<0.001

hispanic [Cuban]

1.40

0.32 – 6.18

0.654

hispanic [Mexican]

0.37

0.10 – 1.32

0.126

hispanic [Not Hispanic]

0.70

0.20 – 2.47

0.579

hispanic [Other Hispanic]

0.57

0.16 – 2.07

0.394

hispanic [Puerto Rican]

0.33

0.02 – 5.18

0.431

Random Effects

2

3.29

00 state

0.18

11 state.age

0.00

01 state

-0.90

ICC

0.05

N state

51

Observations

5450

Marginal R2 / Conditional R2

0.032 / 0.083

vote biden

Predictors

Odds Ratios

CI

p

(Intercept)

0.72

0.20 – 2.60

0.612

gender [Male]

0.60

0.53 – 0.67

<0.001

hispanic [Cuban]

0.88

0.19 – 3.98

0.867

hispanic [Mexican]

2.97

0.81 – 10.87

0.100

hispanic [Not Hispanic]

1.77

0.49 – 6.40

0.384

hispanic [Other Hispanic]

2.25

0.61 – 8.35

0.224

hispanic [Puerto Rican]

4.29

0.27 – 67.60

0.301

Random Effects

2

3.29

00 state

0.10

11 state.age

0.00

01 state

-0.80

ICC

0.03

N state

51

Observations

5450

Marginal R2 / Conditional R2

0.029 / 0.058

Post-Stratification

We use our multilevel regression model to predict the response voting rate for Trump or Biden in each state. Post stratification is a popular statistical technique to correct estimates when there are known differences between target population and study population. By applying post stratification to our census data, first all of is to partition the population into groups based on different attributes. Then, we may use samples to estimate the response variable in each group. Finally, it's time to aggregate group level values by weighting each group by its corresponding proportion in population. In our model we grouped the population by their states, because the geographic location can influence an individual's voting preferences. Then, we are interested in individuals' voting intent and within each state we have 3 different variables: age, gender, and hispanic. We are able to predict a certain age group of female or male, that is considered to be a certain type of hispanic their preference in voting. Furthermore, by summing up all the outcomes over the entire population we are able to forecast the voting preference by this specific group of individuals.

```
## # A tibble: 1 x 1
##   trump_na_pre
##          <dbl>
## 1        0.467
```

```
## # A tibble: 1 x 1
##   biden_na_pre
##          <dbl>
## 1        0.492
```

```
## # A tibble: 1 x 1
##   trump_pre
##       <dbl>
## 1     0.442
```

```
## # A tibble: 1 x 1
##   biden_pre
##       <dbl>
## 1     0.521
```

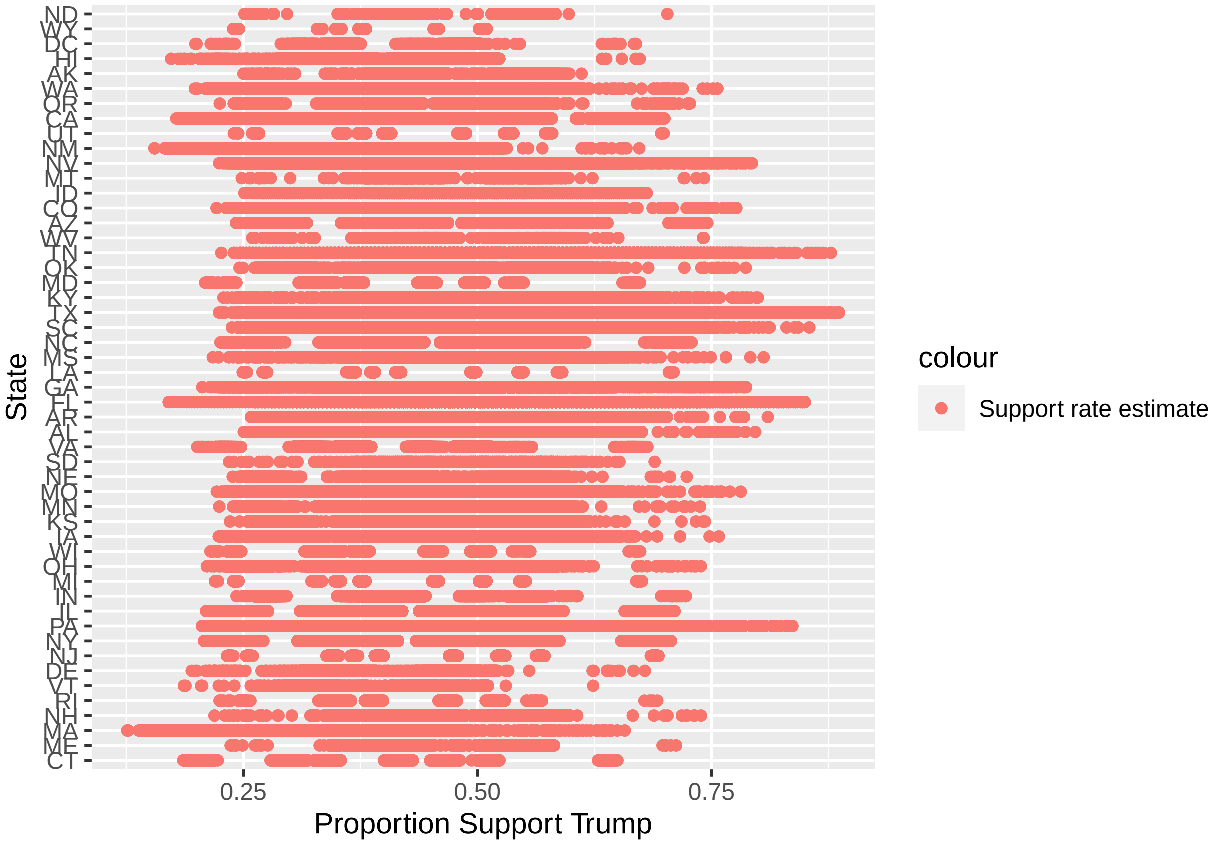
```
## # A tibble: 2 x 2
##   winner       total_votes
##   <chr>              <dbl>
## 1 Donald Trump          87
## 2 Joe Biden            451
```

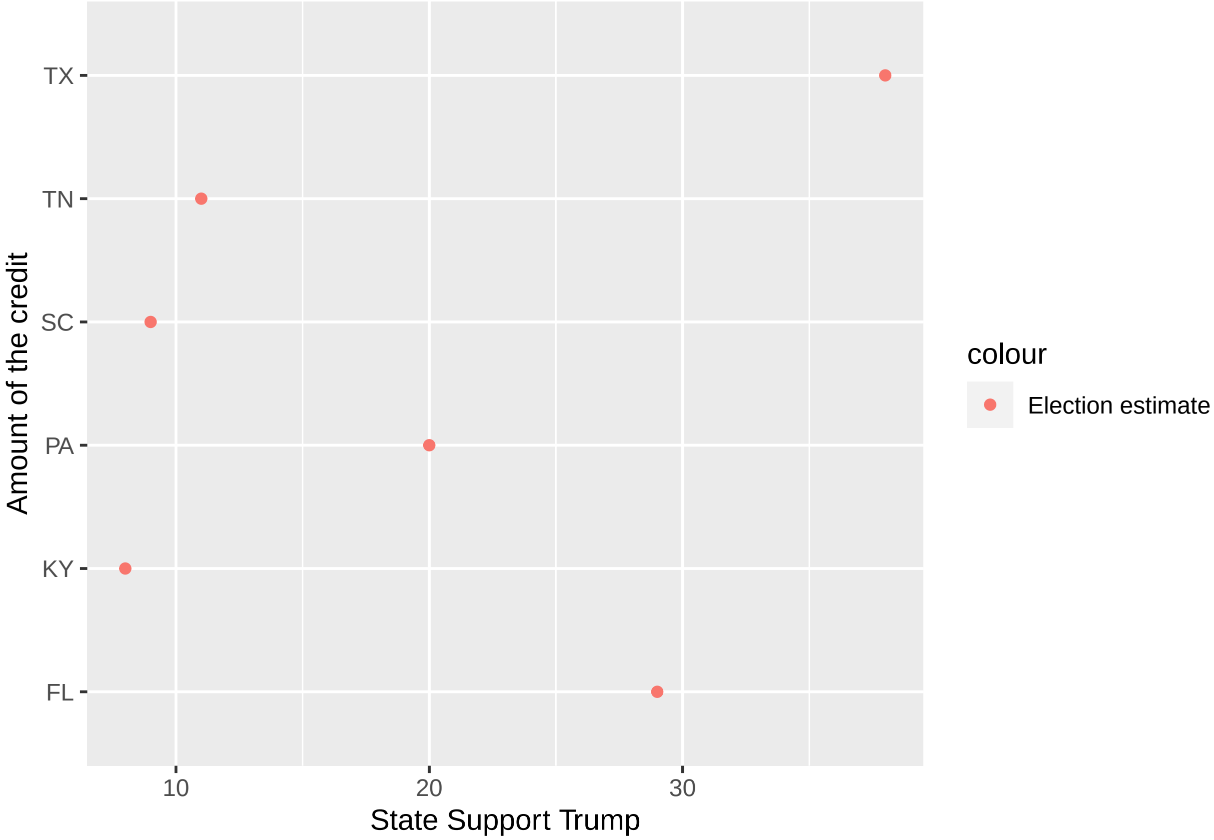
```
## # A tibble: 2 x 2
##   winner       total_votes
##   <chr>              <dbl>
## 1 Donald Trump         115
## 2 Joe Biden            423
```

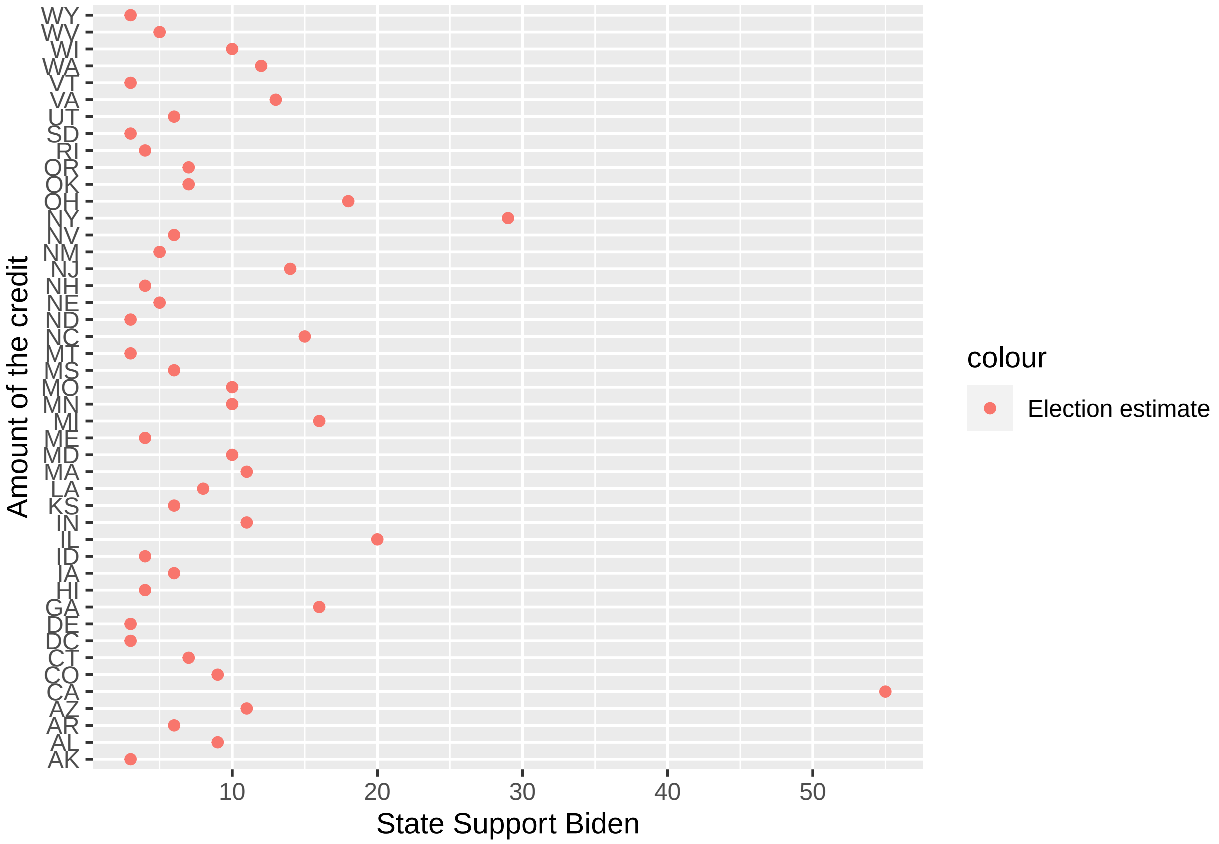
```
## [1] 0.2137546
```

```
## [1] 0.7862454
```

```
## Rows: 5
## Columns: 4
## $ hispanic <chr> "Cuban", "Mexican", "Not Hispanic", "Other Hispanic", "Pue...
## $ Trump    <dbl> 360526, 87209, 2241954, 253491, 1471
## $ Biden    <dbl> 9561, 1563711, 3335231, 1482586, 82800
## $ winner   <chr> "Donald Trump", "Joe Biden", "Joe Biden", "Joe Biden", "Jo...
```


ND
WY
DC
HI
AK
WA
OR
CA
UT
NM
NV
MT
ID
CO
AZ
WV
TN
OK
MD
KY
TX
SC
NC
MS
LA
GA
FL
AR
AL
VA
SD
NE
MO
MN
KS
IA
WI
OH
MI
IN
IL
PA
NY
NJ
DE
VT
RI
NH
MA
ME
CT
State
0.25
0.50
0.75
Proportion Support Trump
colour
● Support rate estimate


TX
TN
SC
PA
KY
FL
Amount of the credit
10
20
30
State Support Trump
colour
● Election estimate



Results

$\hat{y}^{PS}$ is the weighted average of the estimate in each cell over the entire population size. By using the post-stratification process we are able to calculate $\hat{y}^{PS}$ for both Trump and Biden. Our estimate of the proportion of voters who supports Trump is 0.4423359, based on our post stratification model of proportion of voters who support Trump by a logistic regression model, which contains the variables age, gender and hispanic. Likewise for Biden the $\hat{y}^{PS}$ is 0.5214624 using the same logistic regression model and variables. The model of Trump has a lower value of AIC which evaluates how well a model fits the data, so we prefer to choose Trump's model. In addition, we estimate, out of 538 votes cast in all 50 states, Trump will get 423 to Biden's 115.

Discussion

Summary: Using the census data and the survey data that was collected based on the US citizen's preference on voting, we are able to conduct some analysis and predict who might have a higher chance in winning the election. Our interests of variables are age, gender, hispanic, geographic location, and preference in voting for Trump or Biden. First we fitted a multilevel regression model, to investigate the relationship between voting for Trump and an individual's age, gender, hispanic type and state of living, similar process is applied for voting Biden as well. Then we grouped the population by states, and studied the voting preference in each group to forecast the voting result of the entire population.

Conclusion: Based on our post stratification model, by partition the population based on states and calculated the total outcome by the weighted outcome in each state over the entire sample size. We are able to conclude our estimated proportion of voters in favor of Biden is around 52%, while the voters who favor

Trump is around 44%, with the other 4% favoring someone else. Therefore we are able to predict that Biden will probably win the election. However, American election count votes based on state instead of individual, thus we performed an calculation based on state, and through calculation, we found that Trump is estimated to get 21.4% of the total votes, while Biden may get 78.6% of the votes. According to the analysis above, we would conclude that Biden is more likely to win the election.

Weaknesses

The survey data was collected on June,25,2020, which fails to inform anything new at the first time. We delete several variables like people who are not voting or not sure to vote from the raw dataset. However, those voters might change their mind and vote later. The removal of variables causes a weakness to sample observations and it may affect the accuracy of our modeling while doing predictions. Besides, the survey data records individuals' voting counts, though the President is not elected directly by the voters, but by the states. So, the candidate with a high proportion of individual voters may not absolutely win the election at the end.

Next Steps

We will compare our estimation with actual election results. For the future election estimation, we could combine historical election poll results with present survey data, and investigate if each state prefers Republican party or Democratic party presidential candidates.

References

1. Tausanovitch, Chris and Lynn Vavreck. 2020. Democracy Fund + UCLA Nationscape, October 10-17, 2019 (version 20200814). https://www.voterstudygroup.org/downloads?key=bdb6ff88-27ff-444c-bac1-f8f7c7e87c41
2. Steven Ruggles, Sarah Flood, Ronald Goeken, Josiah Grover, Erin Meyer, Jose Pacas and Matthew Sobek. IPUMS USA: Version 10.0 [dataset]. Minneapolis, MN: IPUMS, 2020. https://doi.org/10.18128/D010.V10.0
3. Wang W., et al., Forecasting elections with non-representative polls. International Journal of Forecasting (2014). http://doi:10.1016/j.ijforecast.2014.06.001