Analysis on the popular vote of the 2020 American federal election

Yena Joo, Woolim Kim, Guemin Kim

Nov 2, 2020

Predictions on the proportions of voters for Donald Trump and Joe Biden in the 2020 US Presidential Election based on the voter survey responses.

Code and data supporting this analysis is available at: https://github.com/Guemin/Problem_Set_3

Model

As the 2020 presidential election of the United States approaches, people across the world are interested in to which candidate the vote of the US citizens will be concentrated, either to Donald Trump or to Joe Biden. Since the election outcome will also affect our community in Canada, we are going to analyze and predict the winner of the popular vote in the 2020 American federal election.

Using the survey and census data obtained from Democracy Fund + UCLA Nationscape and IPUMS USA, we are going to predict the popular vote outcome of the election. To be more specific, we are going to use two logistic regression models, one for each candidate, and employ a post-stratification technique[1] with the models. Then, we will predict the winner of the election in each state, using a post-stratification outcome and compare it with the popular vote prediction.

In the following sub-sections, we will describe the model specifics, the post-stratification calculation, and the result of the analysis.

Model specifics

As already mentioned, we will be using the logistic regression models[2] and post-stratification technique with R software to predict the proportions of voters who will vote for either Donald Trump or Joe Biden. Specifically, we will create two models, each for proportions of voters for Trump or Biden, using 6 different variables(age_group, gender, race, education, household_income , and state)[3].

To briefly explain, we will include demographic variables such as age group, gender, race, and education attainment in the models; here, we will organize ages by categorizing them into different age groups. Also,

[1] Post-stratification is a technique used in sample survey design to improve the quality of population estimates. In the post-stratification analysis, the population is partitioned into subgroups, and estimates are predicted within the subgroups. Then, the sum of the estimate times the respective population size in each group is calculated, and finally, it is divided by the sum of the total population size. Detailed procedures on post-stratification for our analysis will be shown in the following sub-sections.

[2] glm() function in the "lme4" package is used to make the logistic regression model.

[3] * age_group is divided into 4 different groups: "18-29 years old", "30-44 years old", "45-64 years old", "65 years and older".
* gender indicates either "Male" or "Female".
* race is divided into 5 different categories: "White", "Black", "Native", "Asian", "Other".
* education is divided into 4 different categories: "Didn't graduate from high school", "High school graduate", "Some college or associate degree", "Bachelor's degree or higher".
* household income consists of 9 categories range from "Less than $14,999" to "$150,000 and over".
* state indicates abbreviated names of 52 states in the US.

we will include household income variable to see how the campaign promises of each candidate affect the voters with different income, and state variable to compare the winner in each state.

Since our response variables, vote_Trump and vote_Biden, are binary(either 'vote for' or 'not vote/not sure'), the logistic regression model[4] is a suitable model to be used.
The equation for logistic regression models we are using is:

$$log(\frac{p_i}{1-p_i}) = \beta_0 + \beta_1 x_{age\ group} + \beta_2 x_{gender} + \beta_3 x_{race} + \beta_4 x_{education} + \beta_5 x_{household\ income} + \beta_6 x_{state}$$

where $log(\frac{p_i}{1-p_i})$ represents log odds in each model, and p_i is the proportion of voters who will vote for Donald trump or Joe Biden. Similarly, β_0 represents the intercept, and β_1,..., β_6 indicate the slope parameters of the model. (Detailed descriptions on the x variables can be found in the footnote[5]).

Data cleaning process

Prior to the modelings, we mutated variables in the survey data to create new variables that could be used in the analysis. Our response variables, vote_Trump and vote_Biden are also mutated from a variable named "vote_2020", which provides a name of a candidate that the respondent supports[6]. Also, the predictor variables, age_group, gender, race, education, household_income, and state are mutated in the data cleaning process so that the categories in each variable in the survey data match with those in the census data.
Since only those who are 18 years old or older are eligible to vote, we removed the observations obtained from the respondents who are younger than 18 years old in the data cleaning process. Similarly, we removed the observations of respondents who answered "No, I am not eligible to vote" as vote_intention, since their responses to vote_Trump and vote_Biden will not count in the actual election. Also, we removed people who are "less than 1 year old" or "90 (90+ in 1980 and 1990)" since their responses are unrealistic or not necessary in our analysis.

Model Diagnostics

With the logistic regression models we created above, we are going to study diagnostics of the models. First, we need to keep in mind that logistic regressions are well performed under the following assumptions:

1. Linearity between the log odds and the predictor variables
   (independent variables should be linearly related to the log odds)

2. Binary response variable
   (Binary logistic regression requires the response variable to be binary)

3. Large sample size

4. Multicollinearity among predictors is not too high
   (predictor variables should be independent to each other)

In our models, we do not need to worry about the violation of the first assumption since all of our predictor variables are categorical; hence, no further categorization of the independent variables is necessary.

[4]Logistic regression is a mathematical model used to estimate the probability of an event occurring using binary data.

[5]* $x_{age\ group}$ represents one of the four age groups that the respondent is in.
* x_{gender} indicates the gender of the respondent(either "Male" or "Female").
* x_{race} indicates the race ethnicity of the respondent.
* $x_{education}$ indicates the education attainment of the respondent.
* $x_{household\ income}$ indicates the total pre-tax income of the respondent's household.
* x_{state} indicates the state in which the respondent is located.

[6]vote_Trump is 1 when vote_2020 is "Donald Trump", and 0 otherwise; vote_Biden is 1 when vote_2020 is "Joe Biden", and 0 otherwise.

Similarly, since our response variables, vote_Trump and vote_Biden are binary, and the size of the survey data is large enough, we can confirm that the second and the third assumptions are also satisfied.

Now, we want to check if the multicollinearity among predictor variables is not too high. This can be done by calculating the variance inflation factor(VIF) for each predictor variable, which measures the amount of multicollinearity in a set of multiple regression variables; the bigger the VIF, the bigger the multicollinearity is. When the variance inflation factor is greater than 5, the corresponding predictor is said to be highly correlated with other predictors. Here are the values of variance inflation factors for predictors in each model:

Table 1: VIF models

model_trump_predictor	VIF	model_biden_predictor	VIF
age_group	1.210003	age_group	1.246368
gender	1.068826	gender	1.072452
race	1.240050	race	1.353103
education	1.477839	education	1.452881
household_income	1.555977	household_income	1.564889
state	1.468859	state	1.461147

As shown above, VIF values do not exceed 2 in both models for Trump and Biden, which suggest that there is no sign of multicollinearity among predictors. Therefore, it is safe to say that the last assumption is also satisfied.

Post-Stratification

Using the log odds estimates, we are going to find vote_Trump and vote_Biden (the proportions of voters each for Donald Trump and Joe Biden) in every possible combination of categories in our predictor variables, age_group, gender, race, education, household_income, and state.

In order to estimate the proportions of voters for both Donald Trump and Joe Biden, we are going to perform a post-stratification analysis. First, we need to subdivide the population having similar characteristics into cells. Hence, we are going to create a total of 55,325 cells based on different age groups, gender, race-ethnicity, education attainment, household income, and state.
Using the logistic regression models presented in the previous sub-section, we will estimate the proportions of voters in each cell for each candidate. Then, we will weight each estimate within each cell by the respective population size of the cell, and sum those values, and divide that by the entire population size. This process can also be described by the expression:

$$\hat{y}^{ps} = \frac{\sum N_j * \hat{y}_j}{\sum N_j}$$

where $\hat{y}_j$ is the estimate of the proportion of voters voting for either Trump or Biden in each cell, and N_j is the population size of the j^{th} cell based off demographics.

Results

In the previous sub-sections, we have created the logistic regression models on proportions of voters voting for Donald Trump and Joe Biden using 6 different variables such as age_group, gender, race, education, household_income, and state, and employed the post-stratification technique using the models.

Model Summary
Figure N is the summarized reult of the logistic regression, and it is going to be used to find the significance

of the independent variables using the p-value[7]. "(age_group)30-44 year olds" shows a p-value of 1.43e-09 which is the biggest p-value among the age variable, and the smallest p-value among them is 2.59e-15 in "(age_group)45-64 year olds". The age variable is highly significant to the model. Similarly, the gender variable shows a p-value of 4.10e-13 which is significant at a 0% significance level, as well as race, income, and education variables. The state variable has at least one state that is statistically significant, so we can conclude every variable is statistically significant and is going to be used in further analysis.

Results of Post-Stratification

Based on the result from the post-stratification analysis in Table 2, we can estimate that the proportion of voters voting for Donald Trump is 0.433 (43.3%) and Joe Biden to be 0.394(39.4%).

Table 2: Comparison of predicted estimate between Trump and Biden

total_predict_trump	total_predict_biden
0.4334444	0.3944298

Table 3: Comparison of predicted estimate grouped by household income level

household_income	predict_trump	household_income	predict_biden
$100,000 to $149,999	0.4788905	$100,000 to $149,999	0.3644576
$15,000 to $24,999	0.3819243	$15,000 to $24,999	0.4089684
$150,000 and over	0.5013108	$150,000 and over	0.3746106
$25,000 to $34,999	0.3917869	$25,000 to $34,999	0.3918452
$35,000 to $44,999	0.4036812	$35,000 to $44,999	0.4117412
$45,000 to $54,999	0.4363749	$45,000 to $54,999	0.3898337
$55,000 to $74,999	0.4219549	$55,000 to $74,999	0.4162607
$75,000 to $99,999	0.4191757	$75,000 to $99,999	0.4305907
Less than $14,999	0.3219549	Less than $14,999	0.3930797

Income

In Table 3, it is noticeable that the estimated proportion of voters for Trump increases as the household income range increases. The lowest predicted value is 0.3219 (32.2%) in the income category "Less than $14,999" and the highest predicted estimate is in "$150,000 and over" which is 0.5013 (50.13%).

For Biden, there is no strong deviation shown between the estimates in different household income categories. The proportions of voters voting for Biden in each income level sit in the range between 0.3644 (Income level "$100,000 to $149,999") and 0.4305907 ("$75,000 to $99,999").

The result shows individuals with income "Less than $14,999" are more likely to vote for Biden (39.31%) over Trump(32.19%), and the individuals with income "$150,000 and over" are more likely to vote for Trump (50.13%) over Biden (37.46%).

[7]The p-value for each independent variable tests the null hypothesis that the variable has no correlation with the dependent variable. If the p-value is smaller than the significance level, you can reject the null hypothesis measures the significane of the independent variable on the dependent variable

Figure 1: Predicted Win Counts Per State

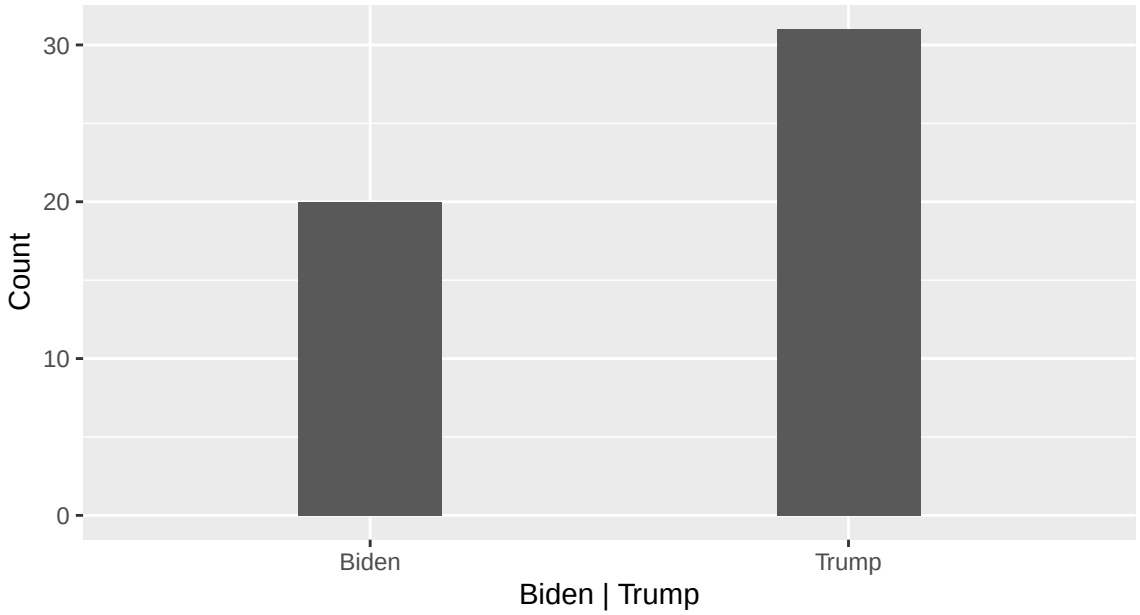

Table 4: Comparison of predicted estimates between Trump and Biden in swing states

state	predict_trump2	state	predict_biden2
AZ	0.4970620	AZ	0.3532261
FL	0.4677497	FL	0.3841190
MI	0.4074363	MI	0.4562030
NC	0.4647314	NC	0.4126773
OH	0.4457807	OH	0.3743578
PA	0.4706141	PA	0.3096477
WI	0.3970139	WI	0.4124047

State

Since the post-stratification result includes proportions of voters for each candidate in all of the 51 states in America[8], we are only going to talk about some noteworthy outcomes instead of going over estimates in every single state. More specifically, for each state, we will compare Trump and Biden's $\hat{y}^{ps}$ values, and consider whoever has the greater proportion of voters wins in the state.

The histogram(Figure 1) above shows the predicted win counts per state. Here, we can see that Trump is expected to win in 31 states, whereas Biden has a higher proportion of voters in the other 20 states.

Swing State

Swing States such as North Carolina, Florida, Pennsylvania, Michigan, Arizona, Wisconsin, and Ohio are the key battleground states where it is unclear who the winner will be. Table 4 shows a brief result of the proportions of voters for each candidate in each swing state; here, we can observe that Trump is expected to get the inside track in most of the states, except for Michigan(0.4562030) and Wisconsin(0.4124047), where

[8]Check Appendix for the full result

Biden is expected to have a higher proportion to get voted with 0.4562(45.62%) and 0.4124(41.24%). Table 4 also expects Trump to win by a landslide, since he wins 4 states out of 6 major swing states.

Discussion

Using the 2020 survey data and 2018 census data obtained from Democracy Fund + UCLA Nationscape and IPUMS USA, we have predicted the popular vote outcome of the 2020 presidential election in the USA. In the "Model Specifics" section, logistic regression models are used to predict the candidate who is more likely to win the popular vote in the election, with explanatory variables age_group, gender, race, education, household_income, and state. However, one thing to note regarding the models is that there is a possibility of having omitted variable bias[9] and measurement error bias since some people could try to hide their political orientation and give false information.

After fitting the models, we employed the post-stratification analysis in which the census data is partitioned into 55,325 cells - based on the same 6 variables used in the logistic regression models - and the proportions of voters for each candidate are estimated within each cell. Using the estimates in each cell, the total proportions of voters for both Donald Trump and Joe Biden, $\hat{y}^{ps}$, are measured to predict the winner of the popular vote.
The result shows that the estimated value $\hat{y}^{ps}$ for Joe Biden is 39.4% and for Donald Trump is 43.34%, which suggests that Trump is more likely to win the popular vote.

Moving on, the result in Table 3 (where we grouped the estimates of proportion by household income level) shows that individuals with income "Less than $14,999" are more likely to vote for Biden (39.31%) over Trump(32.19%). On the other hand, Trump is expected to have a higher proportion of getting voted from individuals with higher household income level; individuals with the income "$150,000 and over" tend to vote for Trump by 50.13%. These outcomes may possibly be affected by the campaign promises of each candidate; for example, Biden's promise of raising taxes for those with income greater than $400,000 could affect people with higher income to not vote for him.

Furthermore, we grouped the cell estimates by states and predicted the candidate who is expected to win in each state. The result shows that Trump has greater probabilities to win in 31 states, whereas Biden has higher probabilities to win in 20 states; Trump is ahead of Biden by 11 states. Also, as already explained in the Result section, the outcome in the Swing States shows that Trump is expected to win in 4 out of 6 swing states. !!!!The "grouping by states" method provides us with better intuition of what the expected result is in the actual election since the method is more accurate to the electoral voting method than the popular vote previously done in the analysis.!!!! Likewise, both popular vote and "grouping by states" vote predict Trump to win in the 2020 presidential election.

To sum up, the overall results from the post-stratification analysis suggest that Donald Trump is more likely to win the popular vote in the 2020 US presidential election. However, this is just an estimation based on the given data sets which do not provide enough information required for predicting the winner of the electoral vote[10]. Historically, in 2016, Hillary Clinton won in the popular vote but lost the election, because Trump won the electoral College. Therefore, winning the popular vote does not determine the next president of USA.

Weaknesses

write about improvement

One of the weaknesses in our analysis is regarding the omitted variables. In the data cleaning process, some of the variables were removed from the data sets prior to the modeling, because either the census data or

[9]* explanation of omitted variable bias is described in "Weakness section"

[10]The US presidential election actually uses the electoral college vote. There are 538 electors in the electoral college, divided among each state.

the survey data did not include the particular variables. If there were any important variables among the omitted ones, that could affect the vote outcome and there might also exist an omitted variable bias in our models. (The omitted variables could be correlated with the dependent variable in the model). This can be improved by choosing a new census data that contains predictor variables that could potentially affect the election outcome.

Moving on, another weakness in our analysis is regarding the census data. Since the census data used in the analysis is the 2018 data, it might not reflect the population in 2020 or predict the election outcome most accurately. For example, in our analysis, those who are not eligible to vote(back in 2018) were omitted from the data set, however, they could be eligible to vote in the 2020 election. Hence, if the 2020 census data was available, it should be more suitable for our analysis.

Lastly, we should note that our prediction on the winner of the popular vote could not match with the winner of the electoral college. Even if a candidate wins in the public vote, it is the result of the electoral college that determines the next president of USA. Therefore, analyzing the winner of the popular vote is not the most accurate way to predict the winner of the presidential election. One way to mitigate this issue is to additionally include a new data set that contains survey responses of electors in the electoral college, and predict the winner of the electoral college.

Next Steps

The analysis does not take into account the possible effect of other factors - such as an individual's Health insurance state - in the vote result. Analyzing the vote outcome by focusing more on the election promise would give a more realistic and reasonable prediction of the election. Also, since the 2018 census data used for the analysis does not reflect the most accurate population, we can try using the 2020 census data in the future, so that we could estimate the proportion of voting for each candidate by the factors that are closely related to the election promises such as health care, market industry, and etc.

Also, for future analysis, we can get survey and census data about the electoral college and do a similar analysis using them. Then, we can compare it with the original analysis that was done in this report.

Lastly, we can compare our predictions with the 2020 presidential election outcomes and see if our predictions were accurate enough compared to the actual results.

References (MLA8)

1. Survey data: "Insights into the Beliefs and Behaviors of American Voters." Democracy Fund Voter Study Group, www.voterstudygroup.org/downloads?key=9337162e-e5ef-49d7-96fd-48a5c5dba31c.

2. Census data: "Census Data." IPUMS USA: Extract Summary, usa.ipums.org/usa-action/extract_requests/summary.

3. Post-Stratification technique: Wang, Wei, et al. "Forecasting Elections with Non-Representative Polls." International Journal of Forecasting, vol. 31, no. 3, 2015, pp. 980–991., doi:10.1016/j.ijforecast.2014.06.001.

4. Logit Regression Assumptions source 1: https://rpubs.com/guptadeepak/logit-assumptions Gupta, Deepak. "Logit Reression Assumptions." RPubs, 18 May 2018, rpubs.com/guptadeepak/logit-assumptions.

5. Logit Regression Assumptions source 2: "Assumptions of Logistic Regression." pp.1-2., https://www.statisticssolutions.com/wp-content/uploads/wp-post-to-pdf-enhanced-cache/1/assumptions-of-logistic-regression.pdf

6. Variance Inflation Factor(VIF): Stephanie. "Variance Inflation Factor." Statistics How To, 9 July 2020, www.statisticshowto.com/variance-inflation-factor/.

7. Tables side by side: Yihui Xie, Christophe Dervieux. "R Markdown Cookbook." 10.1 The Function Knitr::Kable(), 21 Sept. 2020, bookdown.org/yihui/rmarkdown-cookbook/kable.html.

8. Comparing the campaign promises of Trump and Biden: D'Souza, Deborah. "Comparing the Economic Plans of Trump and Biden." Investopedia, Investopedia, 23 Oct. 2020, www.investopedia.com/comparing-the-economic-plans-of-trump-and-biden-4843240.

9. Making plots side by side: "R Multiple Plot Using Par() Function." DataMentor, 8 Oct. 2018, www.datamentor.io/r-programming/subplot/.

10. Hold kable position: Justas MundeikisJustas Mundeikis, et al. "Rmarkdown Setting the Position of Kable." 1 Feb. 1968, stackoverflow.com/questions/53153537/rmarkdown-setting-the-position-of-kable.

Appendix

Table 5: Comparison of predicted estimate grouped by states

state	predict_trump2	state	predict_biden2
AK	0.6178175	AK	0.2121776
AL	0.5294578	AL	0.3429311
AR	0.5690489	AR	0.2163157
AZ	0.4970620	AZ	0.3532261
CA	0.3500102	CA	0.4605909
CO	0.4748248	CO	0.3723088
CT	0.2840064	CT	0.5343268
DC	0.2715509	DC	0.7314580
DE	0.3901171	DE	0.5308874
FL	0.4677497	FL	0.3841190
GA	0.4716816	GA	0.3827232
HI	0.3371692	HI	0.5260513
IA	0.4501696	IA	0.3894466
ID	0.6617140	ID	0.2276048
IL	0.4152228	IL	0.3984838
IN	0.4497127	IN	0.3483643
KS	0.5724607	KS	0.2903427
KY	0.4997506	KY	0.4122667
LA	0.4574786	LA	0.4197140
MA	0.2894075	MA	0.5138926
MD	0.3519287	MD	0.4938823
ME	0.4062306	ME	0.4861133
MI	0.4074363	MI	0.4562030
MN	0.4807777	MN	0.4625594
MO	0.4489350	MO	0.3787824
MS	0.4849136	MS	0.3758363
MT	0.5407824	MT	0.3513251
NC	0.4647314	NC	0.4126773
ND	0.5234047	ND	0.1747900
NE	0.4228730	NE	0.3367801
NH	0.4164238	NH	0.4753414
NJ	0.4045109	NJ	0.4240766
NM	0.2288712	NM	0.5074545
NV	0.5159548	NV	0.3375036
NY	0.3888962	NY	0.4355101
OH	0.4457807	OH	0.3743578
OK	0.4921180	OK	0.2217403
OR	0.4084833	OR	0.4259771
PA	0.4706141	PA	0.3096477
RI	0.3574909	RI	0.4515775
SC	0.5061444	SC	0.2786692
SD	0.5185028	SD	0.3387865
TN	0.5126872	TN	0.2784533
TX	0.5087421	TX	0.3048833
UT	0.4055765	UT	0.2632457
VA	0.3846518	VA	0.4476426
VT	0.1026603	VT	0.7455384
WA	0.3766525	WA	0.4615358
WI	0.3970139	WI	0.4124047
WV	0.5386586	WV	0.3194931
WY	0.1878584	WY	0.2659662

```
## # A tibble: 70 x 5
##    term                                  estimate std.error statistic  p.value
##    <chr>                                    <dbl>     <dbl>     <dbl>    <dbl>
##  1 (Intercept)                             -0.707    0.741     -0.954 3.40e- 1
##  2 as.factor(age_group)30-44 year olds      0.575    0.0950     6.05  1.43e- 9
##  3 as.factor(age_group)45-64 year olds      0.743    0.0940     7.91  2.59e-15
##  4 as.factor(age_group)65 years and older   0.782    0.108      7.25  4.10e-13
##  5 as.factor(gender)Male                    0.422    0.0612     6.90  5.25e-12
##  6 as.factor(race)Black                    -1.42     0.209     -6.79  1.12e-11
##  7 as.factor(race)Native                    0.483    0.285      1.70  8.99e- 2
##  8 as.factor(race)Other                    -0.132    0.200     -0.661 5.08e- 1
##  9 as.factor(race)White                     0.589    0.160      3.68  2.37e- 4
## 10 as.factor(education)Didn't graduate fr~  0.357    0.119      3.01  2.61e- 3
## # ... with 60 more rows
```

```
## # A tibble: 70 x 5
##    term                                estimate std.error statistic    p.value
##    <chr>                                  <dbl>     <dbl>     <dbl>      <dbl>
##  1 (Intercept)                          -0.418      0.839    -0.498     6.18e-1
##  2 as.factor(age_group)30-44 year olds  -0.200      0.0856   -2.34      1.93e-2
##  3 as.factor(age_group)45-64 year olds  -0.287      0.0855   -3.35      8.01e-4
##  4 as.factor(age_group)65 years and old~ -0.125     0.101    -1.24      2.14e-1
##  5 as.factor(gender)Male                -0.302      0.0592   -5.11      3.27e-7
##  6 as.factor(race)Black                  0.999      0.166     6.03      1.67e-9
##  7 as.factor(race)Native                -0.442      0.278    -1.59      1.12e-1
##  8 as.factor(race)Other                  0.00339    0.173     0.0196    9.84e-1
##  9 as.factor(race)White                 -0.449      0.143    -3.14      1.66e-3
## 10 as.factor(education)Didn't graduate ~ -0.667     0.114    -5.86      4.61e-9
## # ... with 60 more rows
```