Predicting 2020 Federal Presidential Election using Logistic Model

Jia Yuan Liu, Gen Cao, Yuanjie Ji

2 November 2020

Model

The purpose of this study is to predict the popular vote outcome of the 2020 American federal election. We used R (R Core Team, 2019) with lme4 (Bates et al.,2015) to construct a logistic regression with random intercepts model using Democracy Fund and UCLA Nationscape survey data (Tausanovitch et al., 2020). In our model, the outcome of vote for Trump or not is constructed as a binary outcome which is modeled by both individual and group leveled predictors "agegroup", "sex", "race", "region" and "education". We select a logistic regression with random intercepts model to effectively analyze the log odds of the binary outcome variable as a linear combination of the predictor variables to account for both fixed and random effects from individual and group-leveled data. A post-stratification technique is employed to adjust the weights of the survey sample based on 2018 IPSUMS USA Census data, so that nonresponse and underrepresented groups in the population are accounted (Ruggles et al., 2020). In the following sub-sections, we will describe the model specifics and the post-stratification calculation.

Model Specifics

Before constructing the model, we hypothesize that the probability of voting Trump in the election for people with certain characteristics is lower than the average. For instance, female, Asian, African-American race, high educational degree earner and young people are less likely to vote for Trump in the upcoming federal presidential election. Therefore, we include sex, educational level, race , region and age in our model in order to perform a precise prediction.

Additional data manipulations include removing observations of those who are not "registered" to decrease prediction bias. We considered only keeping the observations who have vote intentions, but we decide to include the probability that people change their minds given that there is still a five month period between the survey and election. We choose age groups over age because we think the political preference for people who are in the same age group is very similar. Voters of the same age have experienced a period of time under the leadership of different political parties. Thus, they would have roughly the same expectation of the new president and the political party.

According to regression assumptions, observations should be independent from each other in order to perform an appropriate linear relationship. However, it might not be the case because voters who are from the same region are most likely to have the same political preferences. Using the random intercept model and accounting the variable "region" into the group level could potentially help us to better understand the true relationship between the political preferences and other characteristics of the voter from different regions. Since sex, educational level, race and age are all different for each voter, we classify these variables as individual level or level 1. Equation 1 is at the individual-respondent level:

$$Level\ 1 : log\ (\frac{p_{ij}}{1-p_{ij}}) = \beta_{0j} + \beta_1 X_{sex\ i} + \beta_2 X_{agegroup\ i} + \beta_3 X_{race\ i} + \beta_4 X_{educd\ i} + r_{ij}$$

$$Such\ that\ i = \{1, \ldots, N\}\ and\ j = \{1, \ldots, 4\}$$

It says that the probability of some individual i in region j will vote for Trump depends on their gender, age group, race, education. β_{0j} is the intercept term, it is different for the voters based on the different region that they come from.
β_1 is the coefficient for the slope of $X_{sex\ j}$ and it is only dependent on each voters, not the region that the voter comes from.
$X_{sex\ j}$ is associated with the sex and it is pooled over everyone.
β_2 is the coefficient for the slope of $X_{agegroup\ j}$, it is only dependent on each voter.
$X_{agegroup\ j}$ is associated with the age of each individual. Based on their age, each individual is separated into different age groups. The age groups are: "18-24","25-29","30-34","35-39","40-44","45-49","50-54","55-59","60-64","65-69","70-74","75-79","80-84","85+".
β_3 is the coefficient for the slope of $X_{race\ j}$.
$X_{race\ j}$ is associated with the race of each individual, voters are separated into 7 different groups based on their race.
β_4 is the coefficient for the slope of $X_{educd\ j}$.
$X_{educd\ j}$ is associated with the educational level of each voter, we separate the voters into 8 different groups based on the educational degree they have completed.
r_{ij} is the error term.

We consider the variable "region" in the group level / level 2 because it is playing a role which contributes to a group of individuals, and the region-effect is modeled as:

$$Level\ 2 : \beta_{0j} = r_{00} + r_{01}W_j + u_{0j}$$

r_{00} is the intercept term for the intercept model.
W_j is the group level variable in the intercept model.
r_{01} is the slope coefficient of the W_j for the intercept model.
u_{0j} is the error term for the intercept model.

```r
# random intercept model - logistic regression
model1_trump <- glmer(vote_trump ~ sex  + agegroup + race + educd+ (1|region),
            data=survey_data, family = binomial)

model2_trump <- glmer(vote_trump ~ sex + race + (1|region),
            data=survey_data, family = "binomial")

model1_bidon <- glmer(vote_bidon ~ sex  + agegroup + race + educd+ (1|region),
            data=survey_data, family = "binomial")

model2_bidon <- glmer(vote_bidon ~ sex  + race + (1|region),
            data=survey_data, family = "binomial")
```

```r
# ANOVA
anova_compare <- anova(model1_trump, model2_trump)
kable(anova_compare, caption = 'ANOVA Comparing Full and Reduced Model')
```

Table 1: ANOVA Comparing Full and Reduced Model

	npar	AIC	BIC	logLik	deviance	Chisq	Df	Pr(>Chisq)
model2_trump	9	6375.563	6434.320	-3178.781	6357.563	NA	NA	NA
model1_trump	29	6328.106	6517.434	-3135.053	6270.106	87.45652	20	0

```r
#multicollinearity
model_vif <- vif(model1_trump)
kable(model_vif, caption = 'VIF of Full Model')
```

Table 2: VIF of Full Model

	GVIF	Df	GVIF^(1/(2*Df))
sex	1.049928	1	1.024660
agegroup	1.218715	13	1.007637
race	1.088494	6	1.007091
educd	1.204465	7	1.013377

Model diagnostics are performed to ensure the analysis is informative. With regard to this, we have completed the ANOVA chi-square test for both the reduced model which only contains the variables "sex", "race" and "region" and the full model which contains all the variables we are interested in. According to Table 1, we see that the full model has a lower AIC (full: 6238.1 < reduced: 6375.6), lower deviance (full: 6270.1 < reduced: 6357.6) and a p-value close to 0. This indicates that the full model is statistically significant compared to the reduced model, thus it supports our model selection decision. Variance Inflation Factor test is also performed to detect the problem of to multicollinearity. According to Table 2, predictor variables "sex", "agegroup", "race", and "educd" all have VIF close to 1, which suggests that the regressors are not correlated with each other.

Post-Stratification

```r
# Here I will perform the post-stratification calculation - Trump
census_data$logodds_estimate_trump <-
  model1_trump %>%
  predict(newdata = census_data)

census_data$estimate_trump <-
  exp(census_data$logodds_estimate_trump)/(1+exp(census_data$logodds_estimate_trump))

census_data %>%
  mutate(alp_predict_prop_trump = estimate_trump*n) %>%
  summarise(alp_predict_trump  = sum(alp_predict_prop_trump)/sum(n))
```

```
## # A tibble: 1 x 1
##   alp_predict_trump
##               <dbl>
## 1             0.424
```

```r
# Here I will perform the post-stratification calculation - Biden
census_data$logodds_estimate_biden <-
  model1_bidon %>%
  predict(newdata = census_data)

census_data$estimate_biden <-
  exp(census_data$logodds_estimate_biden)/(1+exp(census_data$logodds_estimate_biden))

census_data %>%
  mutate(alp_predict_prop_biden = estimate_biden*n) %>%
  summarise(alp_predict_biden = sum(alp_predict_prop_biden)/sum(n))
```

```
## # A tibble: 1 x 1
##   alp_predict_biden
##               <dbl>
## 1             0.440
```

$$\hat{y}^{PS} = \frac{\sum N_j \ \hat{y_j}}{\sum N_j}$$

$$Such\ that\ j = \{1, \ldots, n\}$$

$\hat{y}^{PS}$ is the estimated probability of voting certain candidate on the total population size perspective N_j is the population size of each cell $\hat{y_j}$ is the estimated probability of voting certain candidate for each cell

In order to predict the probability of voting for the two candidates that we are interested in for each cell, we apply the random intercept logistic model to each cell. We put the variables that we are interested in from the census data set into the random intercept logistic model that we have created in the previous subsection for Trump and Biden. Then we create two new variables called "logodds_estimate_Trump" and "logodds_estimate_Biden" for Trump and Biden, "logodds_estimate_Trump" and "logodds_estimate_Biden" contain the estimated log odds of voting for Trump and Biden for each cell division. Since we are using logistic regression, we convert from odds to probability in a new variable called "estimate". The variable "estimate_trump" and "estimate_biden" contains the proportion of people voting for Trump and Biden for each cell division. From this point, we multiply each estimated proportion of people by the population size of the corresponding cell, and then add those proportions together. Lastly, we divide the added proportion by the total population size of all the cells. Hence, we have the estimated proportion of people voting for Trump and Biden on the total population size perspective. We exclude sex during the cell division process because sex has a weak correlation with non-response bias

Results

Exploratory Analysis

```r
library(ggplot2)
#Region bar plot
survey_data %>%
  group_by(region) %>%
  ggplot(aes(vote_trump)) +geom_bar() +facet_wrap(~region)+
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))+
  labs(title = "Figure 1: Votes for Trump based on Region", subtitle = "Data based on June 2020 Survey
```

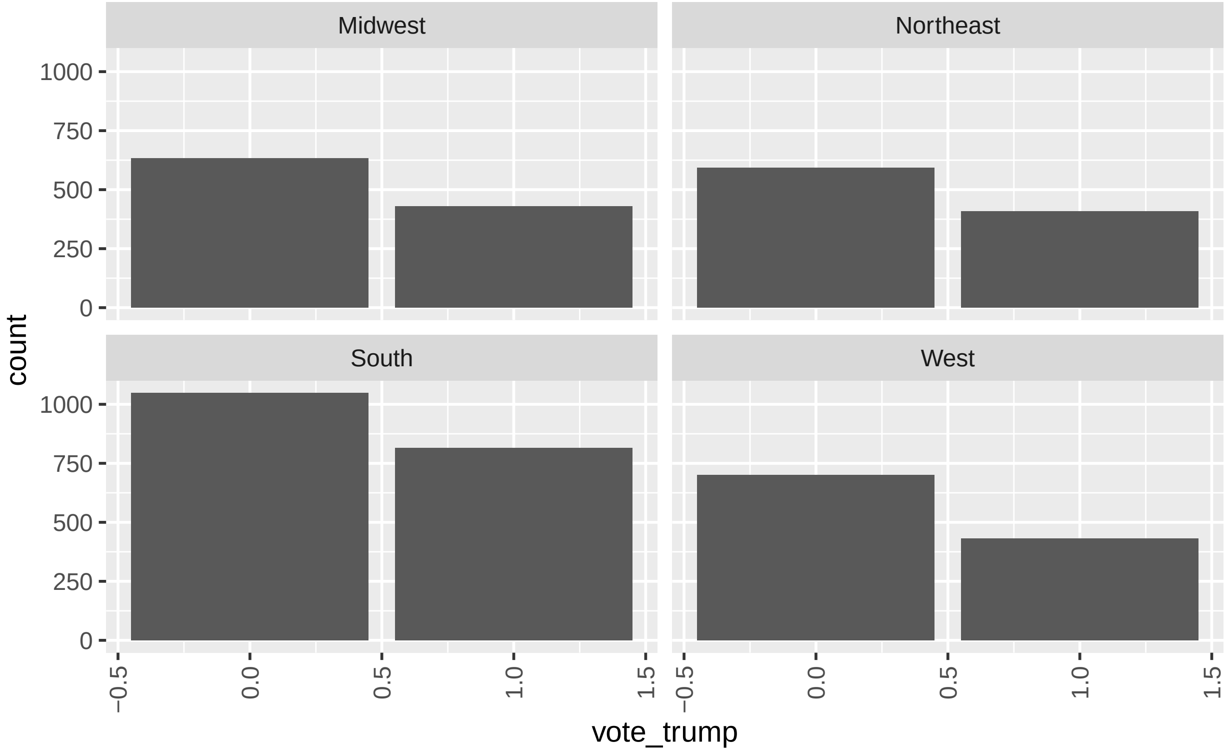
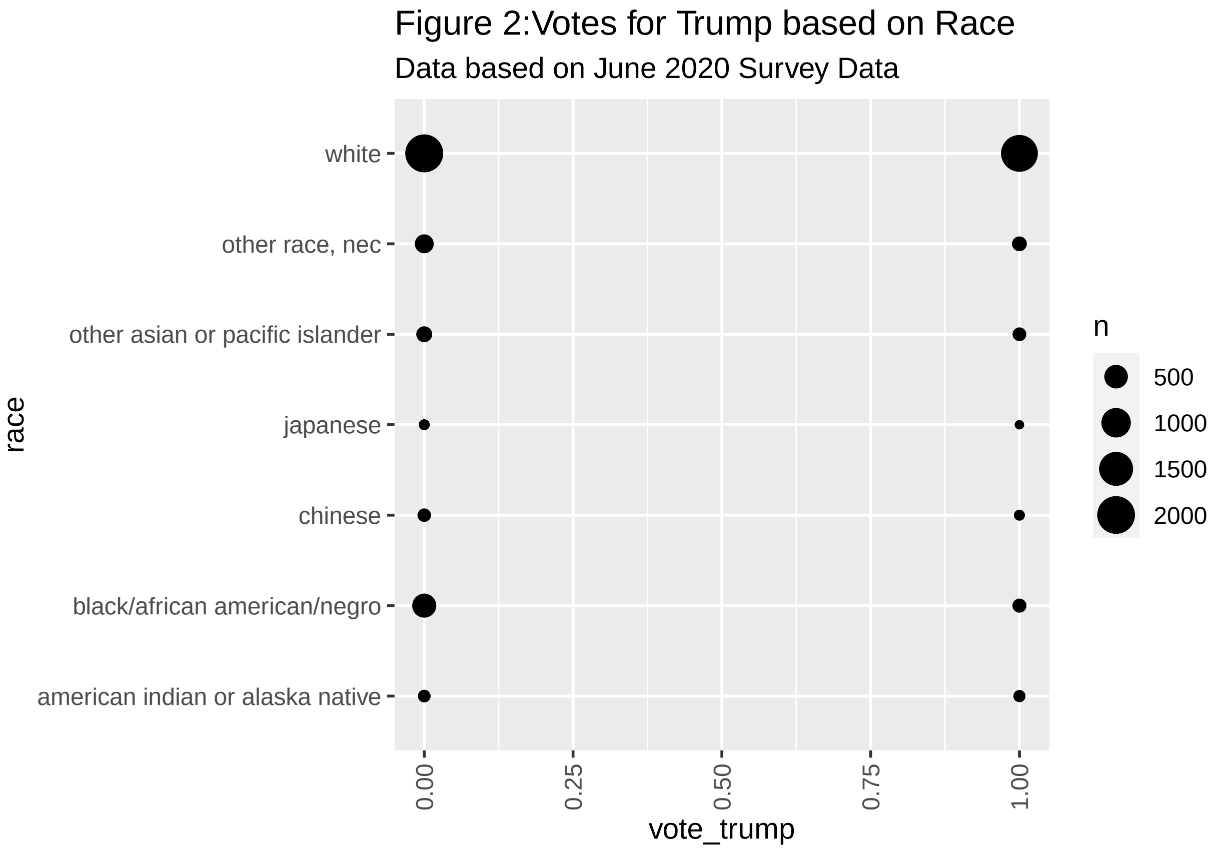


Figure 1: Votes for Trump based on Region

Data based on June 2020 Survey Data

```
#Race Count plot
survey_data %>%
  ggplot(aes(vote_trump, race)) +geom_count() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))+
  labs(title = "Figure 2:Votes for Trump based on Race", subtitle = "Data based on June 2020 Survey Dat
```



Exploratory plots show that individuals from the South and White ethinicity are make up the majority of the survey population.From Figure 1, we can see that there are more people who do not support Trump in all four regions, and West region have the highest difference in count between those who do no support Trump and those who do support Trump. From Figure 2, it can be seen that a lot more Black/African American/Negro do not support Trump.

Table 3: Intercepts Summary for Region

Region	Intercept
Midwest	-1.408
Northeast	-1.430
South	-1.134
West	-1.430

Table 4: Statistical Summary of Logistic Model for Trump

Estimated_Parameters	Estimated_coefficient	pvalue
sexmale	0.4378	0.0000
agegroup25-29	0.2421	0.3278
agegroup30-34	0.5305	0.0005
agegroup35-39	0.7871	0.0000
agegroup40-44	0.8153	0.0000
agegroup45-49	0.8499	0.0000
agegroup50-54	0.8345	0.0000
agegroup55-59	0.8211	0.0000

Estimated_Parameters	Estimated_coefficient	pvalue
agegroup60-64	0.7890	0.0000
agegroup65-69	0.9400	0.0000
agegroup70-74	0.7308	0.0001
agegroup75-79	1.0502	0.0000
agegroup80-84	1.1880	0.0032
agegroup85+	1.3658	0.0964
raceblack/african american/negro	-1.7731	0.0000
racechinese	-1.1482	0.0119
racejapanese	-0.5438	0.4555
raceother asian or pacific islander	-0.2755	0.4163
raceother race, nec	-0.4974	0.0388
racewhite	0.2442	0.5926
Eductional level Below Highschool	0.2364	0.0559
College Degree	0.0968	0.4776
Incompleted college	0.0869	0.5319
educdDoctorate degree	0.6241	0.0065
educdHigh school graduate	0.1843	0.0217
educdMasters degree	0.2297	0.1401
Other post high school vocational training	0.3374	0.0252

With higher values represent increased log odds of voting for Trump, the random intercepts column of table 3 shows that South have the highest intercept (-1.134), while Northeast and West have low intercpets (-1.43). Logistic model summary from table 4 results that region, sex, age group 30-79, and Black/African American/Negro race are the most significant factors in vote for Trump.The coefficient for male is 0.41638, age groups 40-54,65-69, 75-79 all have coefficients higher than 0.8. Black/African American/Negro race has a coefficient of -1.92.

We calculated that the proportion of people voting for Trump is 42.38%, and the proportion of people voting for Biden is 43.97%. The calculation is based off our post-stratification analysis of the proportion of voters in favor of Trump modeled by a logistic regression model with random intercepts, which accounted for "sex", "region", "age group", "race", and "education" variables.

Discussion

Summary

In this study, we demonstrate the application of the random intercept logistic regression model with post-stratification techniques in predicting the overall popular vote of the 2020 American federal election. First, we load the Democracy Fund and UCLA Nationscape survey data (Tausanovitch et al., 2020) to get a sample of people's voting preferences. Then, several predictor variables are identified from the sample data, including gender, race, education level, age group, and region. To better establish the logistic regression model, we classify region as a level 2 (group) variable and all others as level 1 (individual) variables. Subsequently, a logistic regression model is established using R Studio (R Core Team, 2019b). The interpretation of the coefficients of the model is presented in the Conclusion part. Then, another dataset, 2018 IPSUMS USA Census data, is loaded (Ruggles et al., 2020). The post-stratification technique is performed on this dataset to adjust the weights of the survey sample, and a prediction of the 2020 USA presidential election is calculated, which is shown in the Result part. Finally, the weaknesses and future steps of our study are discussed.

Interpreting the Model

By looking at the coefficients of the predictor variables of voting for Trump, we find that the estimations are very close to what we mentioned early in our hypothesis. The coefficients of the random intercepts reflect the log odds of an individual in a specific region voting for Trump. For example, the log odds of someone from the West will vote for Trump is -1.43. For the fitted logistic model, it can be inferred that holding all else constant, the log odds of a male who will vote for Trump is 0.416 high. Additionally, we find that the coefficient rises steadily with the age group, the log odds of someone in the age group 40-54,65-69, 75-79 who will vote for Trump is 0.8+. The proportion of older aged people voting for Trump is surprisingly high, which is not what we expected to see. However, if holding all else constant, the log odds of Black/African American/Negro race individual will vote for Trump is as low as -1.92, which relates to allegations of racism during his presidency (Miller, 2020). These statistics show that midaged to old aged male are more likely to be supporters for Trump, and Black/African American/Negro race people are less likely to be supporters for Trump. Region also plays a significant group factor in vote for Trump.

Conclusion

Based on our model's estimation, the proportion of people voting for Donald Trump is 42.38%, and the proportion of people voting for Joe Biden is 43.97%. We fitted a model of votes in favor of Trump and a model of votes in favor of Biden to account for voters who do not have a certain decision yet. The calculation is based off our post-stratification analysis of the proportion of voters in favor of Trump modeled by a logistic regression model with random intercepts, which accounted for "sex", "region", "age group", "race", and "education" variables. As a result, we predict that Joe Biden, representing the Democrats, will win the election by a nose.

Weaknesses

One weakness of our analysis is that our data come from surveys and census. The population of the surveys and census may be different from the population of voters. According to the Nationscape-User-Guide (Tausanovitch et al., 2020), the surveys were conducted online, and the only requirement is a networked device. Therefore, People who completed the survey or census might not be qualified voters or be qualified voters but choose not to vote. Therefore, the prediction of our model might not perfectly represent the true preferences of the voters.

Another weakness can be the fact that the political preferences are not identical to final vote decisions. The Nationscape-User-Guide (Tausanovitch et al., 2020) says that the latest data were from July 2020, thus there is still a period of time between the time people finished surveys and the final vote date. Consequently, if some breaking news or scandals leaked during this time period, people still got enough time to change their time. Since almost every voter has access to the news, the actual result then could be hugely different from the prediction of our model.

Next Steps

Although we have included a number of predictor variables in our analysis, there are still a few to further consider. For instance, Jérôme Viala-Gaudefroy suggests that religion plays a strong role in the 2020 election (Viala-Gaudefroy, 2020). He points out that Joe Biden is considered a "rather religious" candidate. He has won most religious voters by building an empathetic bond with them. Therefore, we can predict that people believing in a certain religion (especially Christian) are likely to vote for Joe Biden, and one future step of our study is to add religion as a new predictor variable into our model.

Furthermore, a follow-up survey can be conducted after the election. In this survey, people will be asked about their final voting decision. Then we can verify the significance of our predictor variables by comparing people's preferences before the election and their final voting decisions. If people with the same characteristic do not change their ideas much, then the variable related to this characteristic is significant. Conversely, if the majority of a certain group change their minds, then this variable is not significant. For example, if many aged people said that they were going to vote for Donald Trump, but voted for Joe Biden finally in the election, the model's prediction that aged people favored Trump was not true. The reason might be that Trump was giving speeches favored by the aged people at the time surveys were conducted.

References

Steven Ruggles, Sarah Flood, Ronald Goeken, Josiah Grover, Erin Meyer, Jose Pacas and Matthew Sobek. IPUMS USA: Version 10.0 [usa_00001.dta]. Minneapolis, MN: IPUMS, 2020. https://doi.org/10.18128/D010.V10.0

Tausanovitch, Chris and Lynn Vavreck . 2020 . Democracy Fund + UCLA Nationscape, October 10- 17, 2019 (version 20200814) . Retrieved from https://www.voterstudygroup.org/publication/nationscape-data-set

Wu, Changbao, and Mary E. Thompson. "Basic Concepts in Survey Sampling." Sampling Theory and Practice. Springer, Cham, 2020. 3-15.

R Core Team (2019b). R: A Language and Environment for Statistical Computing.R Foundation for Statistical Computing, Vienna, Austria.

Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H (2019). "Welcome to the tidyverse." Journal of Open Source Software, 4(43), 1686. doi: 10.21105/joss.01686.

Wickham H (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org.

Bates D, Mächler M, Bolker B, Walker S (2015). "Fitting Linear Mixed-Effects Models Using lme4." Journal of Statistical Software, 67(1), 1–48. doi: 10.18637/jss.v067.i01.

Fox J, Weisberg S (2019). An R Companion to Applied Regression, Third edition. Sage, Thousand Oaks CA. https://socialsciences.mcmaster.ca/jfox/Books/Companion/.

Xie Y (2020). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.30, https://yihui.org/knitr/.

Demnati, A., and J. N. K. Rao. 2004. Linearization variance estimators for survey data. Survey Methodology 30: 17–26.

Jérôme Viala-Gaudefroy (2020). "How strong a role does religion play in US elections?". https://theconversation.com/how-strong-a-role-does-religion-play-in-us-elections-133224.