GR5018 Assignment #1

Yun Choi

2/19/2022

Model Selection

I choose a multiple multinomial logistic regression model to examine how different party affiliations have differential impacts on one's selection of voting method. The specification and variables for my model are as below:

$$VotingMethod = \beta 1 Party + \beta 2 Age + \beta 3 White + \beta 4 Gender + \beta 5 TravelDistance + e$$

- Variables:
- Dependent: VotingMethod (categorical) - 'IN-PERSON' if voted in person on Election Day; 'EARLY IN-PERSON' if voted early in person; 'MAIL' if casted a mail ballot
- Main independent: Party (categorical) - 'REP' if Republican; 'DEM' if Democrat; 'UNA' if unaffiliated
- Control 1: Age (continuous) - age of voter as of 2020
- Control 2: White (binary) - '1' if non-Hispanic white; 0 if not non-Hispanic white
- Control 3: Gender (categorical) - 'M' if male; 'F' if female
- Control 4: TravelDistance (continuous) - Distance between residence and a designated polling location in mile

A multiple multinomial logistric regression model is the right choice for this regression task for two reasons. First, it allows the use of an unordered categorical dependent variable and independent variables of different types. The dependent variable VotingMethod in my model consists of three categories: 'IN-PERSON', 'EARLY IN-PERSON', and 'MAIL'. The model also has two types of independent variables. 'Age' and 'TravelDistance' are continuous, and 'Party', 'White', and 'Gender' are categorical.

Second, a multinomial logistic regression model creates a model with the same specification for each category in the dependent variable. Unlike having a uniform model for different categories, this appoach estimates the category-specific heterogeneous coefficients. This allows me to explore how each party affiliation has a differential impact on the probability of choosing each voting method.

Hypothesis

I examine the following hypothesis: Being a registered Republican (as opposed to being a Democrat) decreases the probability of voting by 'MAIL', even when controlling for voter age, race, sex, and distance to a designated polling location (controlled variables). I believe my hypothesis is true because many Republican politicians and members have continued their attacks on mail-in voting throughout the modern election history. During the 2020 presidential election, where many states expanded and encouraged mail-in voting due to the COVID-19 pandemic and the health risk of casting an in-person ballot, only 30% of Republicans voted by mail. In comparison, nearly 60% of Democrats cast a mail-in ballot.

I control for voter age and distance to a polling place because age and travel distance are two primary factors in deciding whether to vote by mail. The older the voter, and the longer the travel distance, the higher

the voter's incentive to VBM and save the trip. I also control for voter race and gender because white, and male voters are likely to have a higher socioeconomic status than non-white, and female voters, respectively, which is also one of the strong predictors for a high probability of voting by mail.

Data Description

Here, I use the dataset I created by (1) linking various election datasets - voter registration, voter history, and polling places - from North Carolina; (2) randomly sampling a subset of the entire state voter population; and (3) geocoding and calculating the distance between residence and a designated polling place for each sampled voter.

```
final_dist_no_outliers <- read_csv("final_dist_no_outliers.csv")
```

```
## Rows: 48666 Columns: 42

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (31): voting_method, voted_party_cd, pct_label, pct_description, ncid, v...
## dbl (11): voter_reg_num, county_id, zip_code, birth_year, latitude.x, longit...

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
data_for_mlogit <- final_dist_no_outliers %>%
  # Dependent variable - categorical
  mutate(VotingMethod =
           case_when(
             voting_method == "ABSENTEE BY MAIL" ~ "MAIL",
             voting_method ==  "IN-PERSON" ~ "IN-PERSON",
             voting_method %in%
               c("ABSENTEE CURBSIDE", "ABSENTEE ONESTOP") ~ "EARLY IN-PERSON",
             TRUE ~ "NA")) %>%
  # Remove methods outside these three
  filter(VotingMethod %in%
           c("MAIL", "IN-PERSON", "EARLY IN-PERSON")) %>%
  # Independent variable #1: Age
  mutate(Age = 2021 - birth_year) %>%
  # Independent variable #2: Party
  filter(voted_party_cd %in%
           c("UNA", "REP", "DEM")) %>%
  # Independent variable #3: White
  mutate(White = if_else(race_code == "W" & ethnic_code == "NL",
                         1, 0)) %>%
  # Independent variable #4: Gender
  filter(gender_code %in% c("M", "F")) %>%
  # Rename column names for more intuitive interpretation
  rename(Party = voted_party_cd,
         Gender = gender_code,
         TravelDistance = distance) %>%
  mutate(id = row_number())
```

```r
data_for_mlogit_final <- mlogit.data(data = data_for_mlogit, varying=NULL,
                                     shape = "wide", choice = "VotingMethod", id.var = "id")

ml_1 = mlogit(VotingMethod ~ 0 |
                as.factor(Party) + Age + White + as.factor(Gender) + TravelDistance,
              data=data_for_mlogit_final, reflevel = "EARLY IN-PERSON")
summary(ml_1)
```

```
##
## Call:
## mlogit(formula = VotingMethod ~ 0 | as.factor(Party) + Age +
##     White + as.factor(Gender) + TravelDistance, data = data_for_mlogit_final,
##     reflevel = "EARLY IN-PERSON", method = "nr")
##
## Frequencies of alternatives:choice
## EARLY IN-PERSON       IN-PERSON            MAIL
##         0.64694         0.18914         0.16391
##
## nr method
## 5 iterations, 0h:0m:5s
## g'(-H)^-1g = 0.000102
## successive function values within tolerance limits
##
## Coefficients :
##                                  Estimate  Std. Error  z-value  Pr(>|z|)
## (Intercept):IN-PERSON         -1.18436589  0.04232157 -27.9849 < 2.2e-16 ***
## (Intercept):MAIL              -1.85143328  0.04648306 -39.8303 < 2.2e-16 ***
## as.factor(Party)REP:IN-PERSON  0.31077237  0.03257989   9.5388 < 2.2e-16 ***
## as.factor(Party)REP:MAIL      -0.82313778  0.03676547 -22.3889 < 2.2e-16 ***
## as.factor(Party)UNA:IN-PERSON  0.25515078  0.03276697   7.7868 6.883e-15 ***
## as.factor(Party)UNA:MAIL      -0.05656222  0.03224555  -1.7541   0.07941 .
## Age:IN-PERSON                 -0.00975199  0.00068852 -14.1636 < 2.2e-16 ***
## Age:MAIL                       0.01325936  0.00072855  18.1997 < 2.2e-16 ***
## White:IN-PERSON                0.26310914  0.02805806   9.3773 < 2.2e-16 ***
## White:MAIL                     0.29925182  0.02928110  10.2200 < 2.2e-16 ***
## as.factor(Gender)M:IN-PERSON   0.14261649  0.02500609   5.7033 1.175e-08 ***
## as.factor(Gender)M:MAIL       -0.01776389  0.02679362  -0.6630   0.50734
## TravelDistance:IN-PERSON       0.01625965  0.00941495   1.7270   0.08417 .
## TravelDistance:MAIL           -0.10221785  0.01087130  -9.4025 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Log-Likelihood: -38965
## McFadden R^2:  0.023351
## Likelihood ratio test : chisq = 1863.3 (p.value = < 2.22e-16)
```

Results in Logit

First, I look at how the estimates related to Republicans change. The estimate of 0.31 on 'REP:IN-PERSON' can be interpreted as below: being a registered Republican, on average, increases the logit of voting "IN-PERSON" (compared to voting "EARLY IN-PERSON") by 0.31, controlling for voter age, race, gender, and distance to a designated polling location. The estimate of -0.82 on 'REP:MAIL' shows that: Being Repub-

lican, on average, decreases the logit of voting by "MAIL" (compared to voting "EARLY IN-PERSON") by 0.82, controlling for the same variables. Both estimates are statistically significant with near-zero p-values.

Then, I look at how the estimates related to the independent change. The estimate of 0.25 on 'UNA:IN-PERSON' can be interpreted as below: being an independent, on average, increases the logit of voting "IN-PERSON" (compared to voting "EARLY IN-PERSON") by 0.25, controlling for voter age, race, gender, and distance to a designated polling location. The estimate of -0.05 on 'UNA:MAIL' shows that: Being Republican, on average, decreases the logit of voting by "MAIL" (compared to voting "EARLY IN-PERSON") by 0.05, controlling for the same variables. Although the second estimate is not statistically significant with a p-value of 0.08, comparing the estimate (-0.05) with its equivalent for Republican (-0.82) implies that the preferance gap between "EARLY IN-PERSON" AND "MAIL" is much bigger for Republicans than the independent.

Because the estimates are in the unit of logit and have different standard errors, it is hard to compare them without computation. Therefore, I run the Z-test to see if the two estimates are indeed different, given their standard errors.

Z-test

```
test = ((0.82313778 - 0.05656222)^2)/(0.03676547^2 + 0.03224555^2)
format(pchisq(test, df = 1, lower.tail = FALSE), scientific = FALSE)
```

```
## [1] "0.00000000000000000000000000000000000000000000000000000002223893"
```

The result shows that the slope of 'REP' for going from "EARLY IN-PERSON" -> "MAIL" (0.82) is not equal to the slope of 'UNA' for going from "EARLY IN-PERSON" -> "MAIL" (0.05). The below 0.01 p-value from the Chi-square test above shows that we can reject the null hypothesis that the two slopes are the same at a 99% confidence level.

Results in Relative Risk Ratios (RRRs)

```
exp(coef(ml_1))
```

```
##         (Intercept):IN-PERSON              (Intercept):MAIL
##                     0.3059401                     0.1570120
## as.factor(Party)REP:IN-PERSON      as.factor(Party)REP:MAIL
##                     1.3644786                     0.4390518
## as.factor(Party)UNA:IN-PERSON      as.factor(Party)UNA:MAIL
##                     1.2906562                     0.9450077
##                  Age:IN-PERSON                      Age:MAIL
##                     0.9902954                     1.0133477
##                White:IN-PERSON                    White:MAIL
##                     1.3009687                     1.3488493
##  as.factor(Gender)M:IN-PERSON       as.factor(Gender)M:MAIL
##                     1.1532874                     0.9823930
##      TravelDistance:IN-PERSON           TravelDistance:MAIL
##                     1.0163926                     0.9028328
```

Exponentiating the coefficients produces relative risk ratios (RRRs). RRRs in a multinomial logistic model can be interpreted in the same way odds-ratios (ORs) can be in a binary logistic regression.

An RRR of 1 means that given a one-unit change in that independent variable, there is an equal chance of the event happening as not happening. An RRR above 1 indicates that given a one-unit change in that independent variable, there is a lower chance of the event happening than not happening. An RRR below 1, on the other hand, indicates that given a one-unit change in the independent variable, there is a higher chance of the event happening than not happening.

The RRR on 'REP:IN-PERSON' can be interpreted as below: For being a registered Republican (as opposed to Democrat), on average, the odds of voting "IN-PERSON", compared to voting "EARLY IN-PERSON", increases by 36% (1.36-1 = 0.36), controlling for voter age, race, sex, and distance to a designated polling location. The RRR on 'REP:MAIL' can be interpreted as below: For being Republican (as opposed to Democrat), the odds of voting by "MAIL" (compared to voting "EARLY IN-PERSON") decreases by 56% (0.44-1 = -0.56), controlling for the same variables. The RRR on 'UNA:MAIL' can be interpreted as below: For being an independent (as oppsed to Democrat), the odds of voting by "MAIL" (compared to voting "EARLY IN-PERSON") decreases by 5% (0.95-1 = -0.5), controlling for the same variables.

To sum up, holding voter age, race, gender, and distance to a polling place constant, being a registered Republican makes the voter to prefer voting in person on Election Day most over voting early in person or by mail. The disparity in the level of preference among Republicans between voting early in person and by mail is much bigger that that between voting early in person and voting in person on Election Day.

Also, net of voter age, race, gender, and distance to a polling place constant, Republicans have a bigger preference gap between 'EARLY IN-PERSON' and 'MAIL' than the independent. It implies that Republicans have a stronger distate for voting by mail than voting early in person compared to the independent.

Although more interpretive than logits, RRRs are still not as intuitive as probabilities. To yield specific probabilites of choosing each voting method for different populations, I create a predictions below.

Predictions

```
# Create a dataframe of values to use for predictions
data_party <- expand.grid(
  Age = mean(data_for_mlogit_final$Age, na.rm = TRUE), # let age as the mean
  Party = c("REP", "DEM", "UNA"),  # list all three voting method
  Gender = "M", # fix gender as male
  White = 1, # fix race at white
  TravelDistance = 1)

# Reformat the dependent variable to put in multinom() function
data_for_mlogit$VM <- relevel(as.factor(data_for_mlogit$VotingMethod),
                              ref = "EARLY IN-PERSON")

# Fit the model
ml_2 = multinom(VM ~ Age + as.factor(Party) + White + as.factor(Gender) + TravelDistance,
                data=data_for_mlogit)
```

```
## # weights:  24 (14 variable)
## initial  value 49075.010935
## iter  10 value 41043.884904
## iter  20 value 38964.728466
## final  value 38964.713817
## converged
```

```r
# Create a prediction for each voter
pred <- predict(ml_2, type = "probs")

predict(ml_2, newdata = data_party, type = "probs", se = TRUE)
```

```
##   EARLY IN-PERSON IN-PERSON      MAIL
## 1       0.6452717 0.2504338 0.1042945
## 2       0.6051203 0.1721196 0.2227601
## 3       0.5830934 0.2140591 0.2028475
```

```r
# Feed the dataframe created above to get predictions for specific populations
preds_party <- data.frame(
  Party = data_party$Party,
  predict(ml_2, newdata = data_party, type = "probs", se = TRUE))

print(preds_party)
```

```
##   Party EARLY.IN.PERSON IN.PERSON      MAIL
## 1   REP       0.6452717 0.2504338 0.1042945
## 2   DEM       0.6051203 0.1721196 0.2227601
## 3   UNA       0.5830934 0.2140591 0.2028475
```

The table above shows the probability of a typical 51-year-old white male living a mile away from his designated polling location, choosing each voting across the political spectrum. I set the 'age' as 51 because it is the mean of all voters in the dataset. I set 'race', 'gender', and 'travel distance' as white, male and 1 mile for simplicity's sake.

The table shows:

- A **Republican** 51-year-old white male living a mile away from his designated polling location typically has:
- 64% of probability of voting early in person;
- 25% of probability of voting in person on Election Day;
- 10% of probability of voting by mail.
- A **Democrat** 51-year-old white male living a mile away from his designated polling location typically has:
- 60% of probability of voting early in person;
- 17% of probability of voting in person on Election Day;
- 22% of probability of voting by mail.
- An **Independent** 51-year-old white male living a mile away from his designated polling location typically has:
- 58% of probability of voting early in person;
- 21% of probability of voting in person on Election Day;
- 20% of probability of voting by mail.

Such predictions, although without statistics to prove significance, confirm my hypothesis that being a registered Republican (as opposed to being a Democrat) decreases the probability of voting by mail, even when controlling for voter age, race, sex, and distance to a designated polling location. In fact, the numbers above show that a typical 51-year-old white male Democrat living a mile away from his designated polling location has more than twice the probability to vote by mail than his Republican counterpart.