Causal Inference

Zehao Qian

2024-03-05

```
# ------------------------------------------------------------
## clear the environment var area
# rm(list = ls())
## clear all plots
# graphics.off()
## clear the console area
#cat("\014")
```

Messages Designed to Increase Perceived Electoral Closeness Increase Turnout

Stata code explanation and R code implementation.

Read the data

- read DTA file
- Transfer it from Stata data file (*.dta) to csv format

```
# library(haven)
# data = read_dta(
#   "./data/APRCloseElections_Final_Publication_Replication_Dataset.dta"
# )
```

```
# write.csv(data, file = "./data/data.csv")
```

```
data = read.csv('./data/data.csv')
```

This Stata code is part of a replication file for a study on the effects of perceived electoral closeness on voter turnout. The code performs various statistical analyses and outputs the results. Here's a breakdown of the code, part by part:

Appendix Table A1: Balance Tests

1. Multinomial logistic regression (`mlogit`) is performed to test the balance of covariates across different treatment groups. The `robust` option is used for robust standard errors, and `baseoutcome(1)` sets the reference category for the dependent variable.

2. Local macros (`local`) are used to store statistics like p-values (`mlogitp`), degrees of freedom (`mlogitdf`), and chi-squared values (`mlogitchi`) from the model.
3. A note (`tablenotes2`) is prepared, summarizing the balance test results, and displayed using the `display` command.
4. `putexcel` commands are used to write the balance test results to an Excel file, creating a table with variable names, treatment groups, means, standard deviations, and the note prepared earlier.

```
library(nnet)
mlogit_model <-
  multinom(
    a_phone_treat_relplacebo_passed ~ d_yearssincereg + d_yearssincereg_miss
    + d_electiondayage + d_gender_male + d_gender_unknown + d_race_black
    + d_race_latino + d_race_miss + d_race_other + d_genvotes + d_primvotes
    + d_specvotes,
    data = data,
    weights = data$weight_allstatestreats
  )
```

```
## # weights:  56 (39 variable)
## initial  value 215843.257546
## iter  10 value 215830.151671
## iter  20 value 215807.949291
## iter  30 value 215800.365621
## iter  40 value 215795.213077
## final  value 215794.860735
## converged
```

```
summary(mlogit_model)
```

```
## Call:
## multinom(formula = a_phone_treat_relplacebo_passed ~ d_yearssincereg +
##     d_yearssincereg_miss + d_electiondayage + d_gender_male +
##     d_gender_unknown + d_race_black + d_race_latino + d_race_miss +
##     d_race_other + d_genvotes + d_primvotes + d_specvotes, data = data,
##     weights = data$weight_allstatestreats)
##
## Coefficients:
##     (Intercept) d_yearssincereg d_yearssincereg_miss d_electiondayage
## 1  -0.032964640    0.0007858479           0.08750893    -0.0008311004
## 10  0.000211515   -0.0007359491           0.10922055    -0.0003283361
## 11 -0.011546051   -0.0002820043           0.06450216    -0.0004212829
##    d_gender_male d_gender_unknown d_race_black d_race_latino d_race_miss
## 1     0.05736842       0.40488331  0.065524077    0.05439013   0.3158005
## 10    0.01014688       0.26755984  0.044889226   -0.14959959   0.2213804
## 11    0.03562879       0.09353999  0.009041932    0.09651350   0.5975183
##    d_race_other  d_genvotes   d_primvotes d_specvotes
## 1   -0.02982739 0.013687649 -0.0024667696 0.007241021
## 10  -0.12768551 0.006773196 -0.0002958713 0.008427966
## 11  -0.07600287 0.006020268  0.0022778769 0.007495242
##
## Std. Errors:
##    (Intercept) d_yearssincereg d_yearssincereg_miss d_electiondayage
## 1   0.03378481    0.0006259320           0.03460970     0.0005038792
```

```
## 10  0.03375436    0.0006285391           0.03442111     0.0005031458
## 11  0.03376136    0.0006276340           0.03475349     0.0005033578
##    d_gender_male d_gender_unknown d_race_black d_race_latino d_race_miss
## 1     0.01468116        0.1196274   0.03637734    0.06419893   0.1803780
## 10    0.01471265        0.1228912   0.03652982    0.06742543   0.1842329
## 11    0.01469106        0.1277490   0.03686184    0.06346600   0.1715563
##    d_race_other  d_genvotes d_primvotes d_specvotes
## 1    0.05888878 0.007074513 0.005794728 0.007762424
## 10   0.06025916 0.007072251 0.005796647 0.007761972
## 11   0.05961205 0.007070964 0.005791490 0.007758913
##
## Residual Deviance: 431589.7
## AIC: 431667.7
```

```r
library(dplyr)
```

```
##
##    'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
# Assuming 'data' is your DataFrame
# Define the treatment groups and variables
treatments <-
  c("t_placebo", "t_info_only", "t_closeness_1", "t_closeness_2")
variables <-
  c(
    "d_yearssincereg",
    "d_yearssincereg_miss",
    "d_electiondayage",
    "d_gender_male",
    "d_gender_unknown",
    "d_race_black",
    "d_race_latino",
    "d_race_miss",
    "d_race_other",
    "d_genvotes",
    "d_primvotes",
    "d_specvotes"
  )

# Initialize an empty data frame for storing the summary statistics
summary_df <-
  data.frame(
    Variable = character(),
    Treatment = character(),
```

```
    Mean = numeric(),
    SD = numeric(),
    stringsAsFactors = FALSE
  )

# Loop through variables and treatments
for (var in variables) {
  for (treat in treatments) {
    treat_data <- data %>% filter(!!sym(treat) == 1)
    mean_val <-
      weighted.mean(treat_data[[var]], treat_data$weight_allstatestreats,
                    na.rm = TRUE)
    sd_val <- sd(treat_data[[var]], na.rm = TRUE)

    # Add the summary statistics to the data frame
    summary_df <-
      rbind(
        summary_df,
        data.frame(
          Variable = var,
          Treatment = treat,
          Mean = mean_val,
          SD = sd_val
        )
      )
  }
}

# Add observations count for each treatment
obs_counts <-
  sapply(treatments, function(treat)
    sum(data[[treat]] == 1, na.rm = TRUE))
obs_df <-
  data.frame(
    Variable = "Observations",
    Treatment = treatments,
    Mean = obs_counts,
    SD = NA
  )

# Combine summary statistics and observations count
final_df <- rbind(summary_df, obs_df)
TableA1_BalanceTests = final_df

# Save the data frame to a CSV file
write.csv(TableA1_BalanceTests,
          "./result/TableA1-BalanceTests.csv",
          row.names = FALSE)
```

```
# clear variables
remove(obs_df,
       summary_df,
       mean_val,
```

```
       obs_counts,
       sd_val,
       treat,
       treatments,
       var,
       variables,
       final_df)
```

Table 2: Differences in Election and Turnout Context Across States

1. `sum` commands calculate the turnout rates for placebo subjects in various states. Local macros store these proportions.
2. Another `putexcel` set of commands creates an Excel table summarizing these turnout rates, along with information about the primary date, number of congressional districts, and the number of contested and uncontested primaries.

```
# Assuming 'data' is your DataFrame
# Calculate turnout rates for placebo subjects by state
turnout_rates <- data %>%
  filter(t_placebo == 1) %>%
  group_by(vf_state) %>%
  summarise(Turnout_Rate = mean(voted_2014_primary, na.rm = TRUE) * 100) %>%
  filter(vf_state %in% c("MA", "MI", "MN", "MO", "NH", "TN", "WI"))

# Define the state context information
state_context <- data.frame(
  State = c("Massachusetts", "Michigan", "Minnesota",
            "Missouri", "New Hampshire", "Tennessee", "Wisconsin"),
  Primary_Date = c("September 9", "August 5", "August 12",
                   "August 5", "September 9", "August 7", "August 12"),
  Number_of_Congressional_Districts = c(9, 14, 8, 8, 2, 9, 8),
  Democratic_Contested = c(2, 5, 1, 4, 0, 3, 3),
  Republican_Contested = c(1, 8, 2, 6, 2, 8, 5),
  Democratic_Uncontested = c(7, 9, 2, 3, 0, 5, 0),
  Republican_Uncontested = c(2, 6, 1, 1, 0, 1, 0)
)

# Merge turnout rates with the state context information
final_table <- cbind(state_context, turnout_rates)

# Rename and reorder columns to match the desired output
final_table <- final_table %>%
  select(State, Turnout_Rate,Primary_Date,
         Number_of_Congressional_Districts,
         Democratic_Contested, Republican_Contested,
         Democratic_Uncontested,
         Republican_Uncontested)
Table2_StateContext = final_table
# Save the table as a CSV file
write.csv(Table2_StateContext, "./result/Table2-StateContext.csv",
          row.names = FALSE)
```

```r
remove(turnout_rates, state_context, final_table)
```

Tables 3, A3, A5, and A7: Various Regressions and Proportions

- Variables `close350not2500` and `ageunder50` are generated to represent treatment conditions and age groups, respectively.
- The dataset is structured for panel data analysis using `xtset strata`.
- A series of regression analyses are conducted to assess the impact of closeness treatments on voter turnout, interactions with age, and other covariates. Results are outputted to Excel files.
- The `include` command suggests that another Stata script (`Closeness_SubProgramPRTestRegression.do`) is called multiple times to run specific regression models.

Figure 1 and Table A4: Comparative Effectiveness of Different Treatments

- Regression analyses compare the effectiveness of different treatments on voter turnout.
- `lincom` commands are used to compare the treatment effects.
- Variables for plotting (`tvar`, `beta`, `beta_lowci`, `beta_hici`) are prepared, and a figure is generated using the `twoway` command, showing the estimated treatment effects with confidence intervals.
- The figure is exported as a PDF, and unnecessary variables are dropped.

Table A6: Proportion Voting by Experimental Conditions, State, and Strata

- Proportions of voting by experimental condition, state, and voter strata are calculated and outputted to an Excel file.
- The `foreach` and `forvalues` loops iterate over treatment conditions, states, and voter history categories to summarize the data.

Table A8: Relationship Between Intention to Vote and Actual Turnout

- A regression model explores the relationship between the intention to vote and actual turnout, considering the experimental condition and other covariates.
- Results are outputted to an Excel file.